



Universitat Autònoma de Barcelona

ADVERTIMENT. L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  http://cat.creativecommons.org/?page_id=184

ADVERTENCIA. El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <http://es.creativecommons.org/blog/licencias/>

WARNING. The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>



**Universitat Autònoma
de Barcelona**

**Leveraging Scene Text Information for Image
Interpretation**

A dissertation submitted by **Andrés Mafla Delgado**
at Universitat Autònoma de Barcelona to fulfil the
degree of **Doctor of Philosophy**.

Bellaterra, September 19, 2022

Director	<p>Dr. Dimosthenis Karatzas Universitat Autònoma de Barcelona Dept. Ciències de la Computació Centre de Visió per Computador</p>
Co-Director	<p>Dr. Lluís Gómez-Bigorda Universitat Autònoma de Barcelona Dept. Ciències de la Computació Centre de Visió per Computador</p>
Thesis committee	<p>Dr. Anand Mishra Indian Institute of Technology Jodhpur (IITJ) Jodhpur, India</p> <p>Dr. Ramon Baldrich Universitat Autònoma de Barcelona Dept. Ciències de la Computació Centre de Visió per Computador</p> <p>Dr. Jon Almazán Naver Labs Europe Grenoble, France</p>
International evaluators	<p>Dr. Andrew Bagdanov University of Florence Florence, Italy</p> <p>Dr. Rafael Sampaio de Rezende Naver Labs Europe Grenoble, France</p>



This document was typeset by the author using \LaTeX 2 ϵ .

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

This work is licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) © ⓘ 2022 by **Andrés Mafla Delgado**. You are free to copy and redistribute the material in any medium or format as long as you attribute its author. If you alter, transform or build upon this work, you may distribute the resulting work only under the same, similar or compatible license.

ISBN XYZ-xy-000000-x-x

Printed by Ediciones Gráficas Rey, S.L.

To my parents, brother and to the universe becoming aware of itself...

Acknowledgments

Without a fully supportive environment, in which free exchange of ideas and camaraderie emerges, research or any creative endeavor is not possible. Therefore, I would like to acknowledge and express my gratitude to all the people that I have come across, who in one way or another led me to finish this research work.

Firstly, I would like to thank my supervisor Dr. Dimosthenis Karatzas for giving me support during this long road. Initially, he was my former master's teacher, and his teachings transmitted interest, motivation, and hunger for knowledge on machine learning to me. He always stood up for me on every occasion, listened to my ideas, provide different perspectives, and provided rich criticism.

I also would like to thank my co-supervisor, Dr. Lluís Gomez, who had the initial patience required for me to start exploring computer vision and deep learning. I really admire his extensive coding skills as well as his analytical perspective founded on a love of research and creative thinking. I will always appreciate his friendship and cooperation in research endeavors.

Ali, I was lucky enough to share several years together since our days as master's students. You have always been there during good and bad times, constantly supporting me and offering different and fresh perspectives. You are one of the brightest people I have got the chance to interact with, and I learned a lot from your honest curiosity for knowledge, while being open to new ideas and perspectives. To that end, I consider that I didn't have only two supervisors, but three during my Ph.D. I consider you my closest friend, older brother (despite that I am older than you), workmate, and lawyer.

I would also like to thank Sounak, whom I met at the beginning of my Ph.D. He sat next to me and seeing his constant dedication and hard work was an inspiration to me. You had immense patience with me and taught me how to develop an idea into a research paper, and I was lucky to have collaborated with you. Additionally, you have been a supportive friend that taught me many life lessons.

I want to also acknowledge Sanket "Nicho, TLO, Snaket, etc.." Biswas, he has been like a brother to me during these years. His big heart was always keen to provide support and pure friendship. Our talks about friendship, music, and viewpoints of life have left me unvaluable experiences and memories.

I would like to thank Ruben as well, he taught me the richness of Catalan culture and showed me friendship from the master's until the finish of my Ph.D.

To Moha, Sergi, and Raul. We shared trips, laughs, research ideas and great moments. We have gotten closer during these years and I have learned to listen better

from Moha, laugh at sarcastic words from Sergi and try to enjoy every moment that life offers from Raul.

Also I would like to thanks all my friends from CVC and UAB, Esmitt, Pau, Armin, Maria, Gianmarco, Francesco, Marco, Emanuelle, David, Manu, Pietro, Enrico, Andrea, Giuseppe, Angelos, Suman, Khanh, Kai, and Fei. Without your lessons, talks, music shared and ping-pong matches this path would have be the same. I keep all of you in my memories.

Thanks to all the people that I have collaborated and worked with during my Ph.D, Marçal, Ernest, Diane, and Rafael, I have learned a lot from you. All your teachings have taught me how to deeply analyze ideas and how to build a compelling research paper.

I also would like to thank all the Centre de Visió per Computador (CVC) staff, specially Montse, Gigi, Marc, Kevin, Andrea, and Laura. You provide a valuable service to foreign students, and your friendship plays a big role to make CVC feel closer to home.

To my childhood friends, Xavier, Pocho, Daniel, Jimmy, Christian, Ramiro, Andres, Luis, and Alvaro. I owe you inspiration, friendship, and love. We were joined by life through music and art, and through creativity and friendship this link remains unchanged through time.

Special thanks to my family and people close to my heart. Estefania, thanks for all your friendship and love during my early adulthood and first years out of Ecuador, your best remains within me. Thanks to my brother Daniel, he is my friend and an example of hard work, professionalism, and most importantly an amazing human being. Rocio, thank you mother, you are a being of light, that taught me the love to read, and learn and provided me with infinite support and love. Thanks to my father Patricio, that despite long working hours, showed his love in his unique way. He is a role model of hard work and constant effort to study and perfect a professional environment. To the rest of my family and friends, I love you all.

Finally, I would like to thank the constant struggle of the universe against entropy so that it can become aware of itself, providing us with the unique experience of life.

Abstract

Until recently, most computer vision models remained illiterate, largely ignoring the semantically rich and explicit information contained as scene text. Recent progress in scene text detection and recognition has recently allowed exploring its role in a diverse set of open computer vision problems, e.g. image classification, image-text retrieval, image captioning, and visual question answering to name a few. The explicit semantic of scene text closely requires specific modeling similar to language. However, scene text is a particular signal that has to be interpreted according to a comprehensive perspective that encapsulates all the visual cues in an image. Incorporating this information is a straightforward task for humans, but if we are unfamiliar with a language or scripture, achieving a complete world understanding is impossible (e.a. visiting a foreign country with a different alphabet). Despite the importance of scene text, modeling it requires considering the several ways in which scene text interacts with an image, processing and fusing an additional modality. In this thesis, we mainly focus on two tasks, scene text-based fine-grained image classification, and cross-modal retrieval. In both studied tasks we identify existing limitations in current approaches and propose plausible solutions. Concretely, in each chapter: i) We define a compact way to embed scene text that generalizes to unseen words at training time while performing in real-time. ii) We incorporate the previously learned scene text embedding to create an image-level descriptor that overcomes optical character recognition (OCR) errors which is well-suited to the fine-grained image classification task. iii) We design a region-level reasoning network that learns the interaction through semantics among salient visual regions and scene text instances. iv) We employ scene text information in image-text matching and introduce the Scene Text Aware Cross-Modal retrieval StacMR task. We gather a dataset that incorporates scene text and design a model suited for the newly studied modality. v) We identify the drawbacks of current retrieval metrics in cross-modal retrieval. An image captioning metric is proposed as a way of better evaluating semantics in retrieved results. Ample experimentation shows that incorporating such semantics into a model yields better semantic results while requiring significantly fewer data to converge.

Keywords – Computer Vision, Pattern Recognition, Deep Learning, Scene Text Image Retrieval, Fine-grained image retrieval, Cross-modal retrieval, Image-Text matching, Vision and Language, Scene Text Aware Cross-modal retrieval, Semantic Adaptive Margin, COCO-Text Captioned (CTC) Dataset

Resum

Fins fa poc, la majoria dels models de visió per computador seguien sent analfabets, ignorant en gran mesura la informació explícita i semànticament rica continguda com a text d'escena. El progrés recent en la detecció i reconeixement de text d'escena ha permès recentment explorar el seu paper en un conjunt divers de problemes oberts de visió per computador, p. classificació d'imatges, recuperació de text d'imatges, subtítols d'imatges i resposta visual a preguntes, per nomenar-ne alguns. La semàntica explícita del text de l'escena requereix un modelatge específic similar al llenguatge. Tot i això, el text de l'escena és un senyal particular que s'ha d'interpretar d'acord amb una perspectiva integral que encapsuli tots els senyals visuals en una imatge. Incorporar aquesta informació és una tasca senzilla per als humans, però si no estem familiaritzats amb un idioma o escriptura, és impossible assolir una comprensió completa del món (per exemple, visitar un país estranger amb un alfabet diferent). Tot i la importància del text de l'escena, modelar-lo requereix considerar les diverses formes en què el text de l'escena interactua amb una imatge, processant i fusionant una modalitat addicional. En aquesta tesi ens centrem principalment en dues tasques, la classificació d'imatges de granularitat fina basada en text d'escena i la recuperació multimodal. En totes dues tasques estudiades identifiquem les limitacions existents als enfocaments actuals i proposem solucions plausibles. Concretament, a cada capítol:

- i) Definim una forma compacta de representar text d'escena que es generalitza a paraules invisibles en temps d'entrenament mentre es realitza en temps real.
- ii) Incorporarem la representació de text d'escena prèviament apresada per crear un descriptor de nivell d'imatge que supera els errors de reconeixement òptic de caràcters (OCR) que s'adapten bé a la tasca de classificació d'imatges de gra fi.
- iii) Dissenyem una xarxa de raonament a nivell de regió que aprèn la interacció a través de la semàntica entre regions visuals excel·lents i instàncies de text d'escena.
- iv) Fem servir informació de text d'escena a la coincidència d'imatge i text i introduïm la tasca StacMR de recuperació Cross-Modal conscient de text d'escena. Recopilem un conjunt de dades que incorpora el text de l'escena i dissenyem un model adequat per a la modalitat estudiada recentment.
- v) Identifiquem els inconvenients de les mètriques de recuperació actuals a la recuperació multimodal. Es proposa una mètrica de subtítols d'imatges com una forma d'avaluar millor la semàntica en els resultats recuperats. Una àmplia experimentació mostra que la incorporació de la dita semàntica en un model produeix millors resultats semàntics i requereix una quantitat significativament menor de dades per convergir.

Paraules Clau – Computer Vision, Pattern Recognition, Deep Learning, Scene Text Image Retrieval, Fine-grained image retrieval, Cross-modal retrieval, Image-Text matching, Vision and Language, Scene Text Aware Cross-modal retrieval, Semantic Adaptive Margin, COCO-Text Captioned (CTC) Dataset

Resumen

Hasta hace poco, la mayoría de los modelos de visión por computadora seguían siendo analfabetos, ignorando en gran medida la información explícita y semánticamente rica contenida como texto de escena. El progreso reciente en la detección y el reconocimiento de texto de escena ha permitido recientemente explorar su papel en un conjunto diverso de problemas abiertos de visión por computadora, ej. clasificación de imágenes, recuperación de texto de imágenes, subtítulos de imágenes y respuesta visual a preguntas, por nombrar algunos. La semántica explícita del texto de la escena requiere un modelado específico similar al lenguaje. Sin embargo, el texto de la escena es una señal particular que debe interpretarse de acuerdo con una perspectiva integral que encapsule todas las señales visuales en una imagen. Incorporar esta información es una tarea sencilla para los humanos, pero si no estamos familiarizados con un idioma o escritura, es imposible lograr una comprensión completa del mundo (por ejemplo, visitar un país extranjero con un alfabeto diferente). A pesar de la importancia del texto de la escena, modelarlo requiere considerar las diversas formas en que el texto de la escena interactúa con una imagen, procesando y fusionando una modalidad adicional. En esta tesis, nos centramos principalmente en dos tareas, la clasificación de imágenes de granularidad fina basada en texto de escena y la recuperación multimodal. En ambas tareas estudiadas identificamos las limitaciones existentes en los enfoques actuales y proponemos soluciones plausibles. Concretamente, en cada capítulo: i) Definimos una forma compacta de representar texto de escena que se generaliza a palabras no vistas en el entrenamiento, mientras su inferencia es realizada a tiempo real. ii) Incorporamos la representación de texto de escena previamente aprendida para crear un descriptor de nivel de imagen que supera los errores de reconocimiento óptico de caracteres (OCR) que se adapta bien a la tarea de clasificación de imágenes de grano fino. iii) Diseñamos una red de razonamiento a nivel de región que aprende la interacción a través de la semántica entre regiones visuales sobresalientes e instancias de texto de escena. iv) Empleamos información de texto de escena en la coincidencia de imagen y texto e introducimos la tarea de recuperación Cross-Modal consciente de texto de escena StacMR. Recopilamos un conjunto de datos que incorpora el texto de la escena y diseñamos un modelo adecuado para la modalidad recién estudiada. v) Identificamos los inconvenientes de las métricas de recuperación actuales en la recuperación multimodal. Se propone una métrica de subtítulos de imágenes como una forma de evaluar mejor la semántica en los resultados recuperados. Una amplia experimentación muestra que la incorporación de

dicha semántica en un modelo produce mejores resultados semánticos y requiere una cantidad significativamente menor de datos para converger.

Palabras Clave – Computer Vision, Pattern Recognition, Deep Learning, Scene Text Image Retrieval, Fine-grained image retrieval, Cross-modal retrieval, Image-Text matching, Vision and Language, Scene Text Aware Cross-modal retrieval, Semantic Adaptive Margin, COCO-Text Captioned (CTC) Dataset

Contents

1	Introduction	1
1.1	Vision	2
1.2	Language	3
1.3	Scene Text	4
1.3.1	Particularities of Scene Text	5
1.4	Outline, Research Questions and Contributions	6
2	Real-time Lexicon-free Scene Text Retrieval	9
2.1	Introduction	9
2.2	Related Work	11
2.2.1	Scene Text Detection	12
2.2.2	Scene Text Recognition	13
2.2.3	End-to-End Text Recognition	13
2.2.4	Scene Text Retrieval	14
2.3	Single Shot Scene Text Retrieval	15
2.4	Datasets	19
2.4.1	IIIT Scene Text Retrieval (STR)	19
2.4.2	IIIT Sports-10k Dataset	19
2.4.3	Street View Text (SVT) Dataset	19
2.4.4	Multi-lingual Scene Text (MLT) Datasets	19
2.4.5	Text in Videos (TiV) Dataset	20
2.5	Experimental Results	20
2.5.1	Training Details	20
2.5.2	Backbone Comparison	21
2.5.3	Effect of PHOC Dimension	23
2.5.4	Comparison with State-of-the-Art	24
2.5.5	Multi-Lingual Scene Text Retrieval	26
2.5.6	Real-time Text Spotting in Videos	28
2.6	Conclusions and Future Work	31

I	Leveraging Scene Text for Fine-grained Image Classification	33
3	Scene Text for Fine-grained Image Classification and Retrieval	35
3.1	Introduction	35
3.2	Related Work	36
3.2.1	Fine-Grained Classification	36
3.2.2	Multimodal Fusion	36
3.2.3	Attention and Reasoning	37
4	Combining Visual and Locally Pooled Textual Features	39
4.1	Introduction	39
4.2	Fine-grained Classification Model	41
4.2.1	Visual Features	42
4.2.2	Textual Features	42
4.2.3	Attention on features	44
4.3	Experiments and Results	45
4.3.1	Datasets	45
	Con-Text Dataset	45
	Drink Bottle Dataset	45
4.3.2	Implementation Details	45
4.3.3	Classification Results	46
4.3.4	Importance of Textual Features	46
4.3.5	Comparison of Models	49
4.3.6	Qualitative Results	50
4.3.7	Fine-grained Image Retrieval	51
4.4	Conclusions and Future Work	51
5	Multi-Modal Reasoning Graph for Fine-Grained Image Classification	53
5.1	Introduction	53
5.2	Multimodal Reasoning Model	55
5.2.1	Global Image Encoder	55
5.2.2	Local Feature Encoder	56
5.2.3	Text Encoder	56
5.2.4	Positional Encoder	57
5.2.5	Multi-modal Reasoning Graph	57
5.2.6	Classification	58
5.3	Experiments and Results	59
5.3.1	Datasets	59
5.3.2	Implementation Details	59
5.3.3	Comparison with the State-of-the-Art	60
5.3.4	Importance of Textual Features	61
5.3.5	Ablation studies	63
5.3.6	Qualitative Results	64
5.4	Fine-Grained Image Retrieval	65
5.4.1	Qualitative Retrieval Results	67
5.4.2	Relevance of Textual Features	67

5.4.3	Visualizing Reasoning	69
5.5	Conclusions and Future Work	71
II Image-Text Cross-Modal Retrieval		73
6	Image-Text Cross-Modal Retrieval	75
6.1	Introduction	75
6.2	Related Work	76
6.2.1	Cross-modal retrieval	76
6.2.2	Scene Text in Vision and Language	76
6.2.3	Image Captioning evaluation metrics	77
6.2.4	Semantics and Metric Learning.	78
7	Scene-Text Aware Cross-Modal Retrieval	79
7.1	Introduction	79
7.2	The CTC Dataset	81
7.2.1	Data Collection and Statistics	81
7.2.2	Dataset Samples	84
7.2.3	Comparison with other Datasets	84
7.3	Methods	87
7.3.1	Re-Ranking Strategies	88
7.3.2	STARNet: A Dedicated Trimodal Architecture	89
7.4	Experiments and Results	91
7.4.1	Implementation Details	91
7.4.2	Baselines and Re-Ranking Results	91
7.4.3	Supervised Results	93
7.4.4	Qualitative Results	95
7.5	Conclusion	98
8	Is An Image Worth Five Sentences? A New Look into Semantics for Image-Text Matching	99
8.1	Introduction	100
8.2	Metrics	101
8.2.1	Is an image worth 5 sentences?	101
8.2.2	Semantic Recall (SR)	102
8.2.3	Normalized Cumulative Semantic (NCS) Score	103
8.2.4	Correlation with Human Judgements	104
8.3	Semantic Adaptive Margin	104
8.3.1	SAM Formulation	105
8.4	Experiments	106
8.4.1	Implementation Details	106
8.4.2	Insights on State-of-the-Art Retrieval	108
8.4.3	Reduced Data Scenario	109
8.4.4	Comparison with State-of-the-Art	110
8.4.5	Effect of Temperature and Sampling	110

8.4.6 Dwelving into Temperature and Sampling	113
8.4.7 Qualitative Samples for the Reduced Data Scenario	113
8.5 Conclusion	122
9 Conclusions and Future Directions	123
9.1 Conclusions	123
9.2 Future Directions	125
A Appendix	127
A.1 Scene Text Aware Cross-Modal Retrieval	127
A.1.1 Introduction	127
A.1.2 Additions to Baselines and Re-Ranking	127
Full Table of Results on CTC	127
Performance on TextCaps	130
Oracle Late Fusion	130
Performance on Flickr30K and TextCaps	131
List of Contributions	133
Bibliography	135

List of Tables

2.1	Comparison of different frameworks and backbones in the scene text retrieval task. Each model predicts PHOCs of textual instances found and the most confident ones are employed as textual features for retrieval. The metric employed is the mean average precision (mAP).	23
2.2	Comparison of the effect of different PHOC sizes by employing different unigram and bigram levels. The backbone employed on all the experiments is a customized version of YOLOv2 [159]. The metric employed is the mean average precision (mAP) across all datasets.	24
2.3	Comparison to previous state of the art for text based image retrieval: mean average precision (mAP) for IIIT-STR, and Sports-10K, and SVT datasets. (*) Results reported by Mishra et al. in [142], not by the original authors. (†) Results computed with publicly available code from the original authors.	25
2.4	Comparison to the previous state of the art for text-based image retrieval: precision at n (P@n) for the Sports-10K dataset.	26
2.5	Comparison to previous state of the art method for text based image retrieval methods when queries are words already seen during the training process (IV) or not (OOV). The metric employed is the mean average precision (mAP).	28
2.6	Comparison to the previous state of the art method for text-based image retrieval methods when queries are words already seen during the training process (IV) or not (OOV): precision at n (P@n)	28
2.7	Top 15 most frequent words with their number of occurrences and the reached F-score.	31
4.1	Classification performance for two state-of-the-art methods and our proposed model on the Con-Text and Bottles dataset. The results presented by [12] depicted with * are based on an ensemble model.	46
4.3	Results obtained by employing different fusion strategies on both the Con-Text and Drink Bottle dataset. For presentation purposes, acronyms are used to represent each combination of text recognizers (Textspotter (T), E2E_MLT (E), PHOC (P)) and word embeddings (Word2Vec (W), GloVe (G), FastText (F), Fisher Vector (FV)). The † refers to the proposed model.	48

4.2	Visual only and Textual only results. The textual-only results were performed on the subset of images that contained spotted text. The metric depicted is the mean Average Precision (mAP in %).	48
4.4	Retrieval results on the evaluated datasets. The results on Con-Text are based on our implementation of the method by [12] since there is no publicly available code. The retrieval scores are depicted in terms of the mAP(%).	51
5.1	Classification performance of state-of-the art methods on the Con-Text and Drink-Bottle datasets. The results depicted with [†] are based on an ensemble model. The embeddings labeled as ¹ refer to a Bag of Bigrams, and ² is a probability vector along a dictionary. The acronym FV stands for Fisher Vector. The metric depicted is the mean Average Precision (mAP in %).	60
5.2	Classification performance of the proposed method on the subset of images from the test set of the Con-Text and Drink-Bottle datasets such that the images: contain scene-text (I + T) and do not contain scene-text (I - T). The metric depicted is the mean Average Precision (mAP in %).	61
5.3	Visual only and Textual only results. The textual-only results were performed on the subset of images that contained spotted text. The results with [†] were reported by [131]. The metric depicted is the mean Average Precision (mAP in %).	62
5.4	Quantitative results of the different components that form the proposed model. G_f : Global features, G_{fa} : G_f + Self-Attention, V_f : Local Features, T_f : Text Features, $bboxes$: Bounding Box information used by the Positional Encoder, MMR: Multi-modal Reasoning. Results are shown in terms of the mAP(%).	63
5.5	Results obtained by employing different Projection and Fusion strategies on all the modules of our pipeline. Results are shown in terms of the mAP(%).	64
5.6	Retrieval results on the evaluated datasets. The retrieval scores are depicted in terms of the mAP(%).	67
7.1	Datasets' statistics for standard benchmarks and the proposed CTC. [†] refers to COCO-Text filtered selecting machine printed, English and legible scene text only. [★] numbers obtained with method from [134]. [‡] numbers obtained with method from [22].	87
7.2	Results on CTC for visual and scene-text baselines, and their re-ranking combinations. Visual model and Scene-text model indicate image-caption and scene-text-caption retrieval, respectively. GT stands for ground-truth scene-text annotations and OCR for scene-text prediction obtained from [64]. Bold numbers denote the best performances of visual, scene-text, and re-ranking methods for each ensemble of models.	92

7.3	Retrieval results on the CTC-1K and CTC-5K test set of supervised models. The second-to-last row shows the result from the unsupervised re-ranking baseline described in Table 7.2, line 21. <i>OCR</i> stands for the textual features obtained from [64], whereas <i>GT</i> refers to the Ground-truth annotated scene text. Results depicted in terms of Recall@K (R@K).	94
8.1	Pearson-R correlation coefficient results between human judgments and image text matching metrics on the CrissCrossed [150] dataset.	104
8.2	Quantitative results on reduced training data samples. The acronyms used in the first column stand for Flickr30K (<i>F</i>), MSCOCO 1K (<i>C</i>). The (%) denotes the proportion of the training data used in relation to the original dataset size. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K). The † depicts that models are trained with the publicly available code released by the original authors.	109
8.3	Comparison of retrieval results of the original VSRN, CVSE and SGR models with and without the proposed SAM. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K). The column Rsum and Nsum is the summation of the overall retrieval scores in image-to-text and text-to-image for Recall and NCS respectively. The † depicts that models are trained with the publicly available code released by the original authors.	111
8.4	Experiments of the effect of (τ), soft negative (SN) sampling, and the whether the original triplet is kept (✓) or only our formulation is employed (✗). The acronym Nsum(N) refers that GT elements have been removed.	112
8.5	Experiments of the effect of (τ), random (RS), and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). The acronym Nsum(N) refers that GT elements have been removed.	112
8.6	Experiments on Flickr30K regarding the effect of (τ), soft (SN), random (RS), and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[203] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).	113
8.7	Experiments on Flickr30K (Non-GT) regarding the effect of (τ), soft (SN), random (RS) and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[203] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).	114

8.8	Experiments on MSCOCO 1K regarding the effect of (τ) , employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (\checkmark) or only our formulation is employed (\times). For all our experiments, we employ CVSE[203]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).	114
8.9	Experiments on MSCOCO 1K (Non-GT) regarding the effect of (τ) , employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (\checkmark) or only our formulation is employed (\times). For all our experiments, we employ CVSE[203]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).	115
8.10	Image-to-Text qualitative results in MSCOCO 1K. The initial row depicts the queried image and the associated ground truth captions. Each retrieved caption shows the Recall (R_i) and SPICE (S_i) score when compared with the GT captions. Each sample showcases the final per sample Recall ($R_i@5$) and NCS ($N_i@5$) score obtained. Bolded captions represent the correctly retrieved ground truth items.	116
8.11	Image-to-Text in Flickr30K. The first row shows the queried image and the GT captions. Metrics are Recall (R_i) and SPICE (S_i) score. Each row showcases the final per sample Recall ($R_i@5$) and NCS ($N_i@5$) score obtained.	117
A.1	Results on CTC-1k and CTC-5k for visual-only baselines, scene-text-only baselines and re-ranking combinations of these baselines. Bold results denote the best performance at each of visual model, scene-text model and re-ranking methods. † denotes theoretical upper-bounds to the linear combination re-rankings. (see Section A.1.2)	128
A.2	Results on TextCaps (validation set) for visual-only baselines, scene-text-only baselines and re-ranking combinations of these baselines. † denotes theoretical upper-bounds to the linear combination re-rankings. (see Section A.1.2)	129
A.3	Quantitative comparison of experimental results of image-to-text and text-to-image retrieval on the Flickr30K (test) and TextCaps (val) sets of supervised models. Metric depicted in terms of Recall@K (R@K).	131

List of Figures

1.1	Steps of creation a digital image. From left to right, Analog Image, Digital Sampling, and Pixel Quantization. The pixels determine the resolution of an image and the intensity represent the color on a specific location. . . .	2
1.2	There is a great variability and types of scene text. Natural images contain text in handwritten and machine-printed form, but it is prone to different styles, glare and occlusions, orientation, lightening conditions and context dependant, etc.	4
1.3	Textual and visual information in a given image may (from left to right) correlate, complement each other, reinterpret, and/or be totally orthogonal.	5
2.1	The visual appearance of different business places in images can be extremely variable. It seems impossible to correctly label them without reading the text in the images. Our scene text retrieval method returns all the images shown here within the top-10 ranked results among more than 10,000 distractors for the text query “hotel”.	10
2.2	Pyramidal histogram of characters (PHOC) [5] of the word “beyond” at levels 1, 2, and 3. The final PHOC representation is the concatenation of these partial histograms.	15
2.3	Our Convolutional Neural Network predicts at the same time bounding box coordinates x, y, w, h , an objectness score c , and a pyramidal histogram of characters (PHOC) of the word in each bounding box.	16
2.4	Anchor boxes used in the original YOLOv2 model for object detection in COCO (a) and PASCAL (b) datasets. (c) Our set of anchor boxes for text detection.	18
2.5	Synthetic training data generated with a modified version of the method of Gupta <i>et al.</i> [69]. We make use of a custom dictionary with the 90K most frequent English words, and restrict the range of random rotation to 15 degrees.	20
2.6	Top-5 retrieved images for the query “adidas”. Our model successfully retrieves partially occluded and blurred words.	26
2.7	Top 10 ranked images for the query “castrol”. Our model has not seen this word at training time.	27

2.8	Bounding box heat-maps for queried words “honda”, “police”, “tea” and “sony” respectively.	27
2.9	From top to bottom, top-5 ranked images for the queries “vodafone”, “uscita”, “werden” and “parkausweis”. Although our model has not seen these words at training time it is able to achieve a 100% P@5 for all of them. Best viewed in color.	29
2.10	From top to bottom, top-5 ranked images for the queries “apollo”, “bata”, “bawarchi”, “maruti” and “newsagency”. Although our model has not seen these words at training time it is able to achieve a 100% P@5 for all of them.	30
2.11	Error analysis: most of the errors made by our model come from text instances with a particular style, font type, size, etc. that is not well represented in our training data.	30
4.1	T-SNE Visualization [130] of the 300 dimensional PCAed PHOCs in a two dimensional space. Words with similar morphology are clustered together by a Gaussian Mixture Model, thus making such a descriptor suitable and powerful enough to discriminate text for a fine-grained classification task.	40
4.2	Proposed model pipeline. The PHOCs obtained from [135] are used to compute a Fisher Vector that yields a compact morphology-based descriptor suitable to discriminate features from visually similar objects.	41
4.3	Heat maps obtained according to the confidence detection score of the predicted PHOCs.	47
4.4	Classification results. The top-3 probabilities of a given image assigned by the output of our model are shown along the Ground Truth. Notice that without reading, the classification task is impossible to perform even for humans. Blue and red are used to display correct and incorrect predictions respectively.	50
5.1	The proposed model uses a Graph-based Multi-Modal Reasoning (MMR) module to enrich location-based visual and textual features in a combined semantic representation. The network learns at the output of the MMR to map strong complementary regions of visual (blue) and text (green) instances to obtain discriminative features to perform fine-grained image classification and retrieval.	54
5.2	Detailed model architecture. The proposed model combines features of regions of scene text and visual salient objects by employing a graph-based Multi-Modal Reasoning (MMR) module. The MMR module enhances semantic relations between the visual regions and uses the enriched nodes along with features from the Global Encoder to obtain a set of discriminatory signals for fine-grained classification and retrieval.	55

5.3	Classification predictions. The top-3 probabilities of a class are shown as well as the Ground Truth label performed on the test set. Without recognizing textual instances some images are extremely hard to classify even for humans. Text in blue and red is used to show correct and incorrect predictions respectively. Best viewed in color.	65
5.4	Qualitative results in Con-Text Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. The red border represents a mistaken retrieved image that does not correspond to the queried class. (Best viewed in color).	66
5.5	Qualitative results in Drink-Bottle Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. The red border represents a mistaken retrieved image that does not correspond to the queried class. (Best viewed in color).	66
5.6	Qualitative results in Con-Text Dataset when the text in the queried image is blurred. (Best viewed in color).	68
5.7	Qualitative results in Con-Text Dataset. Results are obtained when everything but the text is blurred in a queried image. (Best viewed in color).	69
5.8	Qualitative results in the Drink-Bottle Dataset when the text in the queried image is blurred. (Best viewed in color).	69
5.9	Qualitative results in the Drink-Bottle Dataset. Results are obtained when everything but the text is blurred in a queried image. (Best viewed in color).	70
5.10	Visualization of the learned attention and enriched nodes of the model in the Con-Text dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).	70
5.11	Visualization of the learned attention and enriched nodes of the model in the Drink-Bottle dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).	71
7.1	This Chapter introduces the scene-text aware cross-modal retrieval (StacMR) task and studies scene text as a third modality for cross-modal retrieval. For the example query above, the restaurant name provides crucial information to disambiguate two otherwise equally relevant results.	80
7.2	Proposed CTC dataset, which is designed to allow a proper evaluation of the task StacMR, as all entries contain three modalities: image, scene text and caption.	82
7.3	CTC full statistics. Cumulative histograms (as thresholds over similarity vary) of the semantic similarity between instances of scene-text tokens and a) all captions for an image (Images), b) individual captions (Captions), and c) individual words in captions (Words).	83
7.4	Image-caption pairs from the CTC dataset. These images belong to CTC explicit, i.e. their scene text and captions share at least one word (marked in bold).	84

7.5	Image-caption pairs taken from the full proposed CTC dataset, in which appearing scene-text does not have a semantic relation with the annotated captions, i.e. there are no scene-text and captions common words.	85
7.6	Image-caption pairs from the proposed CTC explicit dataset, i.e. the scene-text and captions have at least one word in common (marked in bold).	86
7.7	Histograms of the number of OCR tokens found in images (seen as sets of captions, left) and in individual captions (right) for the CTC and TextCaps datasets.	87
7.8	Our proposed STARNet model. Visual regions and scene-text instances are used as input to a GCN. The final learned representations are later combined to leverage complementary semantic information.	90
7.9	Qualitative samples obtained when an image is used as a query (Image to Text) in the proposed CTC explicit dataset. Correct results are marked with ✓. Incorrect results are marked with ✗. Reasonable mismatches are depicted with † but still marked by a ✗.	96
7.10	Qualitative samples when a caption is used as a query (Text to Image) in the proposed CTC explicit dataset. Correct results are marked in a green box. Incorrect results are marked in a red box. Words in bold in queried captions depict the scene-text that helps to discriminate retrieved images, which otherwise are ambiguous. Query 1 contains an annotator typo "drains".	97
8.1	According to the Recall@5 metric, defined for Image Text Matching, both methods A and B are equally good: it considers only one image as relevant for a given sentence query. We propose two metrics and an adaptive margin loss that takes into account that there might be other relevant images in the dataset. In this Figure, we represent the semantic similarity of images to the query by their colored border (the greener the more similar).	101
8.2	Comparison of a fixed margin loss function (left) and our adaptive margin (right). We consider an image anchor I , their positive sentences according to the ground truth (c^1, c^2, c^3), and four other sentences ($\hat{c}^1, \hat{c}^2, \hat{c}^3, \hat{c}^4$) that are negative according to the ground-truth but have some degree of semantic similarity with I . In our method, we dynamically adapt the margin of each possible triplet (anchor, positive, and negative items) to the value given by a similarity function ϕ that measures the semantic similarity of positive and negative items. In this Figure, we represent the similarity of sentences with the anchor by its color (the bluish the more semantically similar they are, the reddish the less similar).	105
8.3	Text-to-Image Top-5 retrieved results evaluated with Recall and the presented Semantic Recall for Non-GT items.	108
8.4	MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).	118

8.5 MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$). 119

8.6 Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$). 120

8.7 Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$). 121

Chapter 1

Introduction

Vision and language, two very different sources of information, have caught momentum lately in machine learning. Each discipline evolved independently until the recent convergence of common visio-linguistic representations. However, modeling the interaction between scene text as a particular visio-linguistic modality has remained fairly unexplored. Cues coming from different modalities, vision, and language are used to form semantic concepts, while at the same time different perspectives, ideas, and applications are explored.

The idea of enabling machines to think, reason, perceive and perform tasks as humans do has been a relentless goal of AI research. We can pinpoint this exceptional capability that humans have due to the evolution of brains, a centralized information-processing tool that has endowed humankind to be the rulers of our current world. As a result of emulating nature, common algorithmic design is brain-inspired. However, biological brains are capable of processing different input signals coming from the senses to allow us to process information coming from the world around us to create a mental representation of it. In order to have centralized information processing in computers, the incorporation and fusion of different modalities is a must, such that computers can reason, obtain information about their surroundings and act accordingly. In this thesis, we will focus on the incorporation and interaction of two different types of signals, vision, and language in order to construct similar semantic representations that allow computers to perform different multimodal reasoning tasks.

1.1 Vision

Human vision is one of the most well developed senses, which allows us to interpret subtle differences in our environments. According to Marr [138], vision is the process of discovering from images what is present in the world, and where it is. Humans acquire visual information through the eyes, which is later processed by the brain to build an internal world model. Visible light, is a range of the electromagnetic spectrum that our eyes are sensitive to. Our eyes process incoming light from objects that reflect or produce it, thus constructing depictions of color, shape, edges and locations of objects and their interactions with the environment. As with the brain, humans have taken inspiration from eyes and designed cameras that are sensible to the visible electromagnetic spectrum while some of them are even able to detect and process higher or lower frequencies. Current cameras often divide the visible light into three main components, red, green and blue (RGB), which are employed later to represent them digitally as color channels. The combination of such channels can recreate the human visual spectrum and given different of each channel create different colors and tonalities. However, a single color channel may exist and if that is the case, a grey-scale image is produced. On the other hand, the resolution of an image is given by pixels, which describe the intensity of a color in a specific position of an image.

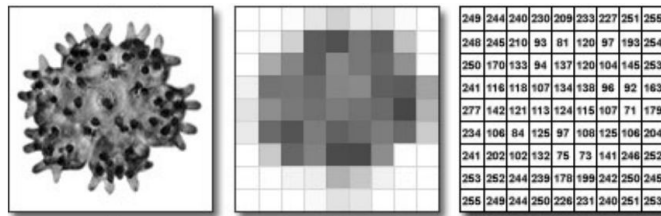


Figure 1.1: Steps of creation a digital image. From left to right, Analog Image, Digital Sampling, and Pixel Quantization. The pixels determine the resolution of an image and the intensity represent the color on a specific location.

Such pixels are involved in the digital sampling step, that involves manipulating an analog signal (real world) into a digital representation of the visual input, see Figure 1.1¹. Images in computers encode different light intensities of a pixel are as a 2-dimensional matrix for each color channel. Usually the pixel quantization step is used to represent the intensity of each channel with values ranging from 0 to 255 or are normalized between 0 and 1. Dealing with this way of representing an image provide several benefits to visualize, compress and retrieve large scale imaging data. However, such abstract representation is hardly useful to obtain semantically rich information (objects present, interactions, spatial relations, text, etc) from images by standard programming approaches. Therefore, computer vision approaches have emerged to develop algorithms that are able to see and interpret digital imagery. A compilation of hand-crafted methods were employed initially, however, due to the success of neu-

¹Image source: <http://hamamatsu.magnet.fsu.edu/articles/digitalimagebasics.html>

ral networks (NN), concretely convolutional neural networks (CNN) [107, 104] have been the to-go approach in order to learn representations from images. More recently, transformer-based [194] approaches [47] have been employed with success to obtain compact image embeddings for specific tasks.

1.2 Language

The ability of speakers to comprehend and construct grammatical sentences is what Chomsky [35] defines as language. Given that a language is a collection of (finite or infinite) sentences, it can be used to transmit complex and unique ideas. Due to the aforementioned reasons, one of the major evolutionary advantages of humans is the innate capability of understanding and building representations of the world, all shaped by language [70, 36]. Despite the importance of language, it has specific characteristics that make it challenging when it comes to the design of computer algorithms that emulate understanding human languages. By the usage of language, we can compose infinite amount of ideas, in a discrete, symbolic, and categorical manner by the usage of a signaling system. Thus as it is the case with images, there is not standard programming approach that can solve this task.

Natural language processing (NLP) emerged as the field that studies methods to design algorithms capable of understanding human languages. Numerically, we could represent words with a one-hot vector encoding the position of each word given a dictionary, however such approach will yield a very sparse and fixed dictionary incapable of incorporating new words. Recent approaches rely on neural networks and usually represent words as compact low-dimensional vector embeddings. Inherently, the main idea is to teach a network to predict the most probable words that fit a common context given by words. Such embeddings are designed to work with tokens that represent the whole word [139, 151] or a given word can be sub-divided into the most common n-grams [170] to obtain a semantic as well as a morphological embedding [21, 43]. The resulting pre-trained embeddings yield rich representations that cluster similar words together. Additionally, interesting mathematical properties appear in the learned embeddings as in the case of the vectors that represent the words:

$$king - man + woman \approx queen$$

$$Paris - France + Spain \approx Madrid$$

thus allowing the operation of algorithms in semantic spaces that allow models to better work with high-level concepts.

In order to encode longer sequences, current approaches directly employ single word pre-trained embeddings to be later fed to a recurrent neural network (RNN) [164] or variants as LSTM[75], GRU[34] or transformers[194].

1.3 Scene Text

Scene text refers to the symbols that represent text which appear on natural scene imagery. The great data diversity of text in the wild makes the scene text detection and recognition task a challenging one, encapsulating handwritten and machine-printed text happening in the wild, refer to Figure 1.2.



Figure 1.2: There is a great variability and types of scene text. Natural images contain text in handwritten and machine-printed form, but it is prone to different styles, glare and occlusions, orientation, lightening conditions and context dependant, etc.

Natural scene imagery contain a significant amount of variability in text patterns. The diversity of text, complexity of background and inference factors such as noise, blur, distortion, occlusion and variance among samples adds up to the problem of text detection and recognition in a non-trivial manner. According to [235], the three main difficulties while recognizing text in natural scene images come from:

- **Variability of Text in Scene Imagery:** Scriptures in documents usually contain a very homogeneous shape and style, in contrast, text in natural scenes have a wide range of variability and diversity of colors, shapes, fonts, scales and orientation.
- **Complex Background:** In natural scene images, the possible backgrounds that may appear are unpredictable and the textures can vary from simple to complex textures. Patterns found in stripes, nature, bricks, fences contain shapes that can be easily mistaken as text.
- **Interference and Imperfect Conditions:** Most of the images and video that contain text are produced in uncontrolled environments. This environments can arise problems such as low resolution, distortion, blur, partial occlusion, inappropriate angles among other factors, which may give rise to errors in detection and recognition.

The difficulties present in the text spotting task have been tackled by an ample use of methodologies, while the most successful ones can be linked to the rise of deep learning techniques. [125, 106, 90]. Despite the immense variability of scene text in the wild, recent advances in scene text recognition have made it feasible to explore new

computer vision challenges that previously appeared impossible. Some of these tasks include fine-grained image classification and retrieval, where it may be necessary to read and comprehend the text on the scene in order to distinguish between various visual categories, or tasks like visual question answering and image captioning, where it may be necessary to read and comprehend the text on an image in order to correctly respond to a question or create a descriptive caption for a specific image.

1.3.1 Particularities of Scene Text

Due to the nature of scene text, it plays a very particular role when used in computer vision applications. Interestingly and in order to exploit all the information that scene text entails, we have to treat such distinct cue as a combination of two modalities. Firstly, in order to detect and recognize textual symbols, the visual cues emerge with a predominant role. However, and in order to fully incorporate semantics, scene text has to be boarded with a language processing viewpoint. Therefore, to interpret scene text we require a careful usage of both modalities for a posterior information exploitation to solve a specific task. Similarly, the interaction of the overall visual cues and scene text have to be considered in order to achieve an holistic image understanding and interpretation. The ways the visual cues interact with scene text is very diverse and complex, nonetheless we can broadly categorize it in 4 different types:



Figure 1.3: Textual and visual information in a given image may (from left to right) correlate, complement each other, reinterpret, and/or be totally orthogonal.

Correlated: Similar information may be learned using either modality and when used together they can reinforce the idea learned using only a single input.

Complementary: Visual and textual cues convey distinct knowledge that, when combined, the message can be enhanced by the combination of both modalities.

Interpretative: When seen in the context of the other modality, the original message is altered; when combined, a new semantic interpretation results.

Orthogonal: The two modalities communicate separate information and are, in some ways, unrelated to one another; attempting to merge them can be harmful since they compete with one another.

However, modeling these categories is not straight-forward since various text in-

stances can co-exist continuously in an image and can be interpreted specifically according to the task. Due to the different nature of each of these modalities, diverse approaches have been explored to fuse them. While initial work proposed a diverse set of fusion mechanisms [16, 97, 71, 226, 17], current approaches rely on transformers, which more often than not require large annotated datasets [32, 202].

1.4 Outline, Research Questions and Contributions

In this section, we enumerate the research questions and provide the reader with a summary of each chapter that composes this thesis.

Chapter 2

Research Question 1: How can we embed scene text into a morphological representation that allows the generalization of unseen words during training time?

Research Question 2: What design choices can we propose to allow a detection-specific model to represent words rather than a classification vector while allowing real-time retrieval?

Previous state-of-the-art scene text retrieval pipelines comprise a multi-stage approach of detection, recognition, and search is performed. However, this approach has several drawbacks, including slow processing of images, not being well suited for out-of-vocabulary words, and not being efficient at a searching time to name a few. Therefore, we propose a fully convolutional neural network that represents scene text with a Pyramidal Histogram Of Characters (PHOC), therefore allowing out-of-vocabulary queries unseen at training time. This approach yields a compact representation of the scene text found in an image while performing in real-time. To perform retrieval, we simply cast the nearest neighbor search of the PHOC representation of the query and the predicted scene text.

Chapter 4

Research Question 1: How can we overcome OCR errors and rely on word morphology to obtain richer scene text embeddings?

Research Question 2: Can we obtain a single scene text descriptor that incorporates all the text instances found within an image?

We found out that models are prone to misclassify a given image if the OCR wrongly recognized a scene text instance. Therefore, we design a syntactically-based embedding by the incorporation of PHOCs, which encodes all the scene text instances within an image. The resulting embedding space clusters similar words according to the morphology, while at the same time representing all the scene text instances in an image,

thus being suitable to overcome OCR errors and yielding state-of-the-art performance in fine-grained image classification and retrieval.

Chapter 5

Research Question 1: Is it possible to design a model capable of reasoning of each scene text instance along with the salient objects of an image to perform fine-grained classification and retrieval?

Research Question 2: What design choices can we undertake to achieve richer region-level features that model the interaction of visual and textual cues?

The reasoning at the image level requires modeling specific pipelines that enrich features according to the interaction among visual regions with the occurring scene text. In this work, we construct a model capable of learning the semantic interaction between salient visual regions and text instances. We define a Graph Neural Network (GNN) that considers the semantics found in each modality to yield richer node features used to further classify a given image. We overcome the problem of employing high-dimensional scene text descriptors found in the model of the previous chapter, while at the same time obtaining a semantic space resulting in a significant boost in text-based fine-grained image classification and retrieval.

Chapter 7

Research Question 1: How can we train a retrieval network considering the scarcity of annotated datasets that model the interaction of scene text as a third modality aside from images and captions?

Research Question 2: Since scene text can have different interactions with the visual and language features, we pose the question of: how can we leverage scene text to perform selective image retrieval?

Given the explicit information that scene text provides in an image, we hypothesize whether text instances can be used to obtain better retrieval results in cross-modal pipelines. To the best of our knowledge, we are the first to study the role of this modality, therefore we propose the task of Scene Text Aware Cross-Modal Retrieval (StacMR). To this end, we gather a dataset and provide different approaches to try to model the interaction between language, images, and scene text. Moreover, we found an approach that improves the retrieved samples when considering scene text as a fine-grained element that needs to be incorporated selectively.

Chapter 8

Research Question 1: How much can we learn from the scarcity of many-to-many annotations regarding the relevance of images and captions in image retrieval pipelines to obtain richer semantics in the retrieved proposals?

Research Question 2: Is there a way to measure the degree of semantic relevance between each image and the retrieved captions?

We discover the problem of lacking a proper metric to assess semantic relevance in cross-modal retrieval literature. We employ image captioning metrics to define the degree of semantic relevance among queries and retrieved results. By utilizing such metrics, we define a semantic adaptive margin incorporated in a triplet loss that constructs a smoother retrieval space. When training models with such semantic margin we obtain an improved sorted set of results according to a semantic viewpoint while also improving on the commonly used recall metric. Importantly, the evaluated models that incorporate our margin require *significantly* fewer data to converge and outperform other approaches.

Chapter 9

Conclusion and Future Directions

We highlight this thesis's major contributions, as well as its important discoveries, and we provide open-ended research directions for further investigation.

Chapter 2

Real-time Lexicon-free Scene Text Retrieval

In this Chapter we address the problem of scene text retrieval: given a text query, the system must return all images containing the queried text. The proposed model uses a single-shot Convolutional Neural Network(CNN) architecture that predicts bounding boxes and builds a compact representation of spotted words. In this way, this problem can be modeled as a nearest neighbor search of the textual representation of a query over the outputs of the CNN collected from the totality of an image database. Our experiments demonstrate that the proposed model offers a significant increase in processing speed and unmatched expressiveness with samples never seen at training time. Several experiments to assess the generalization capability of the model are conducted in a multilingual dataset, as well as an application of real-time text spotting in videos.

2.1 Introduction

The development of language is one of the most influential inventions of humankind that allows the communication of abstract and complex ideas. Similarly, written text permits this set of complex ideas to be captured, stored, and communicated in an explicit manner. As is shown by several authors [196, 106], the text is present in a large percentage of real-life imagery, especially in urban scenarios and documents. Adding this to the fact that there is ample availability of visual data and the importance of text, it becomes essential to develop algorithms that allow efficient information retrieval

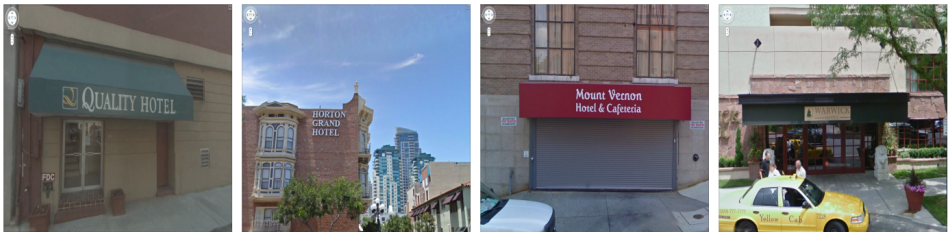


Figure 2.1: The visual appearance of different business places in images can be extremely variable. It seems impossible to correctly label them without reading the text in the images. Our scene text retrieval method returns all the images shown here within the top-10 ranked results among more than 10,000 distractors for the text query “hotel”.

by exploiting the richness of the textual content found in images and video. Leveraging text in scene imagery provides significant boosts to tasks such as image retrieval, scene understanding, instant translation, human-computer interaction, robot navigation, assisted reading for the visually impaired, and industrial automation.

In previous years significant advances have been accomplished, particularly since the introduction of AlexNet [187], architecture that won the ILSVRC2012 [165] contest by using deep learning techniques. Text spotting has been diverging from older approaches that used hand-crafted features to current ones that employ automatic feature learning by exploiting deep learning methodologies [235, 125]. Nonetheless, text spotting is not a trivial task and remains an open problem in the research community. Putting aside the complexity of spotting text in the wild, the importance that text encompasses is given by the high-level semantic and explicit information, which can not be leveraged by using visual cues alone. For example, there is a high degree of complexity involved in labeling images without considering the text found in them, even for humans. This effect is evident in Figure 2.1, in which the storefronts alone can belong to a wide plethora of businesses, but the exact label can be inferred if and only if the text contained is read and leveraged appropriately. Research conducted by Movshovitz *et al.* [144] showed that while training a shop classifier, the proposed model ended up associating specific visual representations to textual information as the only way of differentiating between diverse businesses. The described effect is evident and addressed explicitly in later works conducted by [87, 11, 131], which focus on the fine-grained classification of storefronts and bottles respectively. Additional tasks that require integration of scene text and visual information to generate a common domain knowledge have been proposed such as in [179, 20], which opens up new research paths.

Closely related to our work, Mishra *et al.* [142] proposed the task of scene text retrieval. The input to the system is a text query, which the system must employ to return all the images that contain the queried text. This task requires systems that are robust enough to perform fast word spotting while at the same time holding the capacity of generalizing to out-of-dictionary queries never seen before. An intuitive approach

to tackle such a problem is to make use of state-of-the-art reading systems, and use their output predictions to find the closest match with the given query. However, as it has been shown by [142], such attempts commonly have low performance caused by limitations in end-to-end reading systems. On one hand, end-to-end reading systems are evaluated on recognition, a different task that focuses on achieving high precision scores, often using a specific language dictionary [82] or as it is proposed by [91, 90] a short dictionary per image. On the other hand, a retrieval system requires a large number of proposals (high recall) which can be beneficial at the moment of finding close matching detections when compared to a query. It is also worth noting that end-to-end reading systems usually consist of at least two clearly defined stages that employ the encoder-decoder paradigm. The pipelines comprised by these two stages, more often than not are slow at the moment of generating predictions of text contained in an image. These time constraints hinder the use of such algorithms in real-time scenarios or at the stage of indexing large-scale collections of images and documents.

In order to exploit the particular requirements that need to be addressed by a retrieval system, we propose in this Chapter a real-time, high-performance word spotting method that detects and efficiently encodes text in a single calculation of a Fully Convolutional Neural Network (FCNN). Inspired by backbones used for object detection, we have modified them to predict a PHOC (Pyramidal Histogram Of Characters) [5, 184] descriptor of detected scene text instances. Our experiments show that by using a custom YOLOv2 model [159] we obtain the best architecture in terms of the trade-off between performance and efficiency. By employing this methodology, our model is able to perform text detection and encoding proposals in a single end-to-end calculation. This approach makes it suitable for real-time applications or for indexing large-scale image collections at an unmatched speed while achieving a generalization capability of unseen text instances at training time.

The main contributions of this Chapter are: (1) The usage of the PHOC as a word representation instead of a direct word classification over a closed dictionary, thus providing an elegant mechanism to generalize to any text string, allowing the method to tackle efficiently out-of-dictionary queries. (2) Differently from common scene text detection and recognition pipelines, we present a model comprised of an end-to-end trainable FCNN, capable to learn the morphology of scene text instances. (3) Due to its design, the adoption of this method achieves an unmatched inference time of 42 images per second when processing images while at the same time achieving an unmatched performance. Code has been released and can be found in the public repository¹.

2.2 Related Work

Initial approaches focused on hand-crafted features, thus the attention of the community was centered on the design of relevant features that allowed generalization. Early

¹<https://github.com/AndresPMD/Pytorch-yolo-phoc>

detection methods adopted Connected Components Analysis (CCA) [50, 78, 222] or classification by the use of Sliding Window (SW) such as in [108, 205, 207]. Methods that used CCA usually extract candidate regions by clustering them by color or according to textures that resemble text. The regions proposed are later classified by algorithms that are trained by the usage of hand-crafted features. In the Sliding Window method, windows of different sizes and scales slide over the image and output a binary classification of text or not text. The positive text regions were grouped into final detections by morphological operations or graph-based methods.

Early text recognition was mostly focused on the usage of feature-based methods. Label embeddings were used by [5, 65, 162] to perform a direct matching between input images and strings. Character segment methods were proposed by [174, 220] to recognize words. The recognition task has also been divided in a series of sub-tasks such as text line segmentation [221], character segmentation [175, 163], text binarization [141, 199] and single character recognition [171, 30].

In the past years, several advances in Deep Learning have been accomplished due to data availability and computing power [106], allowing deep learning models to surpass several benchmarks in a wide range of tasks. The main advantage of using deep learning methodologies is the possibility of automatic feature learning, rather than hand-crafted ones. Most literature [125, 235, 29] divide the existing methods into text detection, text recognition, and end-to-end systems.

2.2.1 Scene Text Detection

Deep learning methodologies usually follow a two-step pipeline that comprises an end-to-end trainable detection network and a post-processing step. A branch of research in text detection focuses on the pixel level of an image. In the work presented by [71], a CNN is used to predict if a given pixel belongs to a character, forms part of a text region, and its orientation. Analogously, in [15] an FCN is trained to classify whether a pixel belonged to a specific character. Yao *et al.* [219] propose a CNN that outputs text proposals, which are filtered by separating different text instances by employing a semantic segmentation model. Later works focus on simplifying the pipeline and thus improving the speed and training of models.

The work presented by [115, 114] named Textboxes, adopts a modified version of a popular object recognition model named Single Shot Detector [123]. It employs modified anchor boxes to regress the ground truth boxes followed by a non-maximum suppression step (NMS). A performance-focused approach is given by EAST [234], which up-samples feature maps gradually and uses [98] as the network backbone, and outputs a per pixel word or text line prediction followed by an NMS step.

Based on object detection frameworks proposed by R-CNN [59, 161, 72], alternative text detection pipelines have been explored. The common approach consists of a Region Proposal Network (RPN) that produces candidate text regions, which later are passed through a pooling layer that classifies the region as text or not text.

In the model presented by [129], rotated region proposals are presented, mostly to

handle arbitrary oriented text. Analogously, R2CNN [85] the Region of Interest(ROI) pooling stage uses different fixed sizes which are concatenated for regression and classification. The work conducted by [232] mainly focuses on adaptive weighted pooling in different scales to further predict and regress region proposals.

2.2.2 Scene Text Recognition

Initial approaches explored by Jaderberg *et al.* [81] tackle text recognition as a classification problem. After training a CNN on synthetically generated samples, the obtained features are used to predict a vector that classifies the input word over approximately 90,000 classes. After the introduction of the Connectionist Temporal Classification (CTC) by Graves *et al.* [67] in handwriting recognition, the same methodology has been widely used in scene text as well. The work proposed by [172] employs the CTC layer after passing the input image through a CNN that acts as the encoder and a Recurrent Neural Network (RNN) that act as the decoder. By employing a slightly different approach, [55] adopts stacked convolutional layers to capture contextual dependencies of the input sequence. The introduction of an attention mechanism was initially proposed by [9] in the task of machine translation. This mechanism was briefly adopted in several vision tasks, including text recognition. The work proposed by [33], namely Focus Attention Network, employs attention to supervise relevant locations for word recognition. Bai *et al.* [10] introduce an edit probability to handle the misalignment between the ground truth string and the attention output string. Jaderberg *et al.* [83] proposed the Spatial Transformer Network, which is used by [173] to align detected text horizontally to further employ an attention based recognizer.

2.2.3 End-to-End Text Recognition

Text detection and recognition tasks are highly correlated from an end-to-end perspective, in the sense that learned features can be used to solve both problems. An introductory approach proposed by Jaderberg *et al.* [82] employs a sliding window to extract proposals, which are filtered and a CNN is used to regress the bounding boxes. Later the filtered regions that surpass a threshold are classified. In another work, Gupta *et al.* [69] defined a Fully Convolutional Regression Network for text detection and bounding box regression and the same classification network proposed by [82] for text recognition, being one of the first models that were fully trainable based on deep learning methodologies solely. In [24] a YOLO[159] based CNN is adopted to detect text instances, which later are passed through a Connectionist Temporal Classification module for recognition. These two stages are trained separately and later connected together to form an end-to-end architecture. The research presented by [110] introduces a CNN that is used as an encoder and a Long Short-Term Memory (LSTM) along with an attention mechanism module as a decoder, both employed for detection and recognition. He *et al.* [74] use a CNN to extract proposals, which are fed into an LSTM to refine the bounding boxes that are later employed as input to yet another LSTM to per-

form recognition that fixes misalignment between attention maps and ground truth character labels. In parallel, additional work has been conducted on the development of multilingual scene text recognizers, such as the work of [25] which consists of two CNNs. The first one is optimized to detect text and a second one employs a Connectionist Temporal Classification (CTC) [67] module for recognition while training both in an end-to-end manner. In concurrent work, [124] uses EAST [234] to obtain text regions and employs a CTC recognition module to obtain an end-to-end reading system. Lyu *et al.* [128] use a variation of Mask R-CNN[72] to detect text in arbitrary shapes and segment an image in different instances to recognize similar text regions. Similarly, in order to detect and recognize oriented text, the model FOTS was proposed by [124], which comprises a CNN followed by a ROI Rotate operation, which applies a transformation on oriented text to obtain axis-aligned features maps that are further recognized by a bi-directional LSTM and a CTC decoder.

2.2.4 Scene Text Retrieval

In the scene text retrieval task, the goal is to retrieve all images that contain instances of the queried words in a dataset partition. Given a query, the database elements are sorted with respect to the probability of containing the queried word. We use the mean average precision as the accuracy measure, which is the standard measure of performance for retrieval tasks and is essentially equivalent to the area below the precision-recall curve. Notice that, since the system always returns a ranked list with all the images in the dataset, the recall is always 100%. An alternative performance measure consists of considering only the top- n ranked images and calculating the precision at this specific cut-off point ($P@n$).

It is important to note that the scene text retrieval problem slightly differs from classical scene text recognition methodologies. In a retrieval scenario, the user defines a textual query that he wants to retrieve, whereas most recognition approaches are based on employing a predefined vocabulary of the words one might come along within scene images. For instance, both Mishra *et al.*[142], who introduced the scene text retrieval task, and Jaderberg *et al.* [82], use a fixed vocabulary to create an inverted index that contains the presence of a word in the image. These approaches limit the freedom of queries to a set of predefined vocabulary words.

To address such a problem, text string descriptors based on n -gram frequencies, like the PHOC descriptor (Figure 2.2), have been successfully used for word spotting applications [5, 58]. By using a vectorial codification of text strings, users can query any string at inference time without being limited to a specific set of predefined vocabulary words. After the publication of this Chapter's work, an additional method [201] has achieved state-of-the-art. The pipeline from [201] directly optimizes scene text detection along with cross-modal similarity learning. Two branches, one specialized in text recognition and another that focuses on embedding text transcriptions are optimized for similarity in a common space. Queries are embedded through the second branch and the results are ranked according to the similarity of the text detected within

the image.

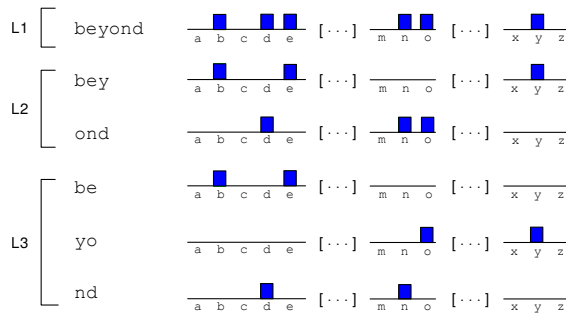


Figure 2.2: Pyramidal histogram of characters (PHOC) [5] of the word “beyond” at levels 1, 2, and 3. The final PHOC representation is the concatenation of these partial histograms.

In the proposed method, we make use of the PHOC descriptor along with an object detection framework based on YOLOv2 [159] that encodes found text instances in an image. We suggest that this approach brings many benefits, mostly due to the high recall and single shot calculation required to locate and recognize text contained within an image, accompanied by an unmatched processing speed.

2.3 Single Shot Scene Text Retrieval

The proposed architecture, illustrated in Figure 2.3, consists in a single shot CNN model that predicts at the same time bounding boxes and a compact text representation of the words within them. To accomplish this we adapt the YOLOv2 object detection model [157, 158] and recast it as a PHOC [5] predictor.

The YOLOv2 architecture is composed of 21 convolutional layers with a leaky ReLU activation and batch normalization [80] and 5 max pooling layers. It uses 3×3 filters and doubles the number of channels after every pooling step as in VGG models [177], but also uses 1×1 filters interspersed between 3×3 convolutions to compress the feature maps as in [9]. The backbone includes a pass-through layer from the second convolution layer and is followed by a final 1×1 convolutional layer with a linear activation with the number of filters matching the desired output tensor size for object detection. For example, in the PASCAL VOC challenge dataset (20 object classes) it needs 125 filters to predict 5 boxes with 4 coordinates each, 1 objectness value, and 20 classes per box $((4 + 1 + 20) \times 5 = 125)$. The resulting model achieves state-of-the-art object detection, has a smaller number of parameters than other single-shot models, and runs in real time.

A straightforward application of the YOLOv2 architecture to the word spotting task would be to treat each possible word as an object class. This way the one hot classifica-

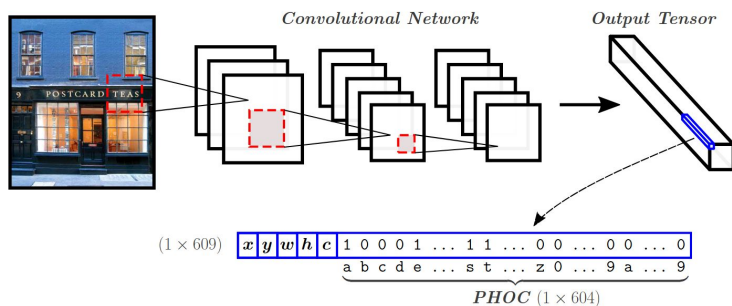


Figure 2.3: Our Convolutional Neural Network predicts at the same time bounding box coordinates x, y, w, h , an objectness score c , and a pyramidal histogram of characters (PHOC) of the word in each bounding box.

tion vectors in the output tensor would encode the word class probability distribution among a predefined list of possible words (the dictionary) for each bounding box prediction. The downside of such an approach is that we are limited in the number of words the model can detect. For a dictionary of 20 words the model would theoretically perform as well as for the 20 object classes of the PASCAL dataset, but training for a larger dictionary (e.g. the list of 100,000 most frequent words from the English vocabulary [82]) would require a final layer with 500,000 filters, and a tremendous amount of training data if we want to have enough samples for each of the 100,000 classes. Even if we could manage to train such a model, it would still be limited to the dictionary size and not able to detect any word not present on it.

Instead of the fixed vocabulary approach, we would like to have a model that is able to generalize to words that were not seen at training time. This is the rationale behind casting the network as a PHOC predictor. PHOC [5] is a compact representation of text strings that encodes if a specific character appears in a particular spatial region of the string (see Figure 2.2). Intuitively a model that effectively learns to predict PHOC representations will implicitly learn to identify the presence of a particular character in a particular region of the bounding box by learning character attributes independently. This way the knowledge acquired from training data can be transferred at test time for words never observed during training, because the presence of a character at a particular location of the word translates to the same information in the PHOC representation independently of the other characters in the word. Moreover, the PHOC representation offers unlimited expressiveness (it can represent any word) with a fixed length low-dimensional binary vector.

The PHOC version we propose in this model employs a higher dimensional binary vector of a length of 820 dimensions. This vector is formed by concatenating the L2 to the L6 unigram levels along with 2 levels of the 50 most common English language bigrams. In order to adapt the YOLOv2 object detection network for single-shot detection and PHOC prediction, it is necessary to define the nature of the proposed descriptor. Firstly, the PHOC descriptor does not resemble a one-hot vector as in a classification

scheme. To treat the PHOC as a multi-hot binary vector, the last layer does not employ a softmax function. Secondly, the predicted PHOC vector comprises a set of numbers that satisfy the condition given by:

$$S = \{x|x \in \mathbb{R}, 0 \leq x \leq 1\} \quad (2.1)$$

Where S represents the set of possible PHOC values. In order to have such a representation, a sigmoid activation function after the last convolutional layer is used to predict the PHOC vectors rather than the original softmax function.

To accomplish this, a sigmoid activation function was used in the last layer. Third, we propose to modify the original YOLOv2 loss function in order to help the model through the learning process. The original YOLOv2 model optimizes the following multi-part loss function:

$$L(b, C, c, \hat{b}, \hat{C}, \hat{c}) = \lambda_{box} L_{box}(b, \hat{b}) + L_{obj}(C, \hat{C}, \lambda_{obj}, \lambda_{noobj}) + \lambda_{cls} L_{cls}(c, \hat{c}) \quad (2.2)$$

where b is a vector with coordinates' offsets to an anchor bounding box, C is the probability of that bounding box containing an object, c is the one hot classification vector, and the three terms L_{box} , L_{obj} , and L_{cls} are respectively independent losses for bounding box regression, objectness estimation, and classification. All the aforementioned losses are essentially the sum-squared errors of ground truth (b, C, c) and predicted ($\hat{b}, \hat{C}, \hat{c}$) values. In the case of PHOC prediction, with c and \hat{c} being binary vectors but with an unrestricted number of 1 values we opt for using a cross-entropy loss function in L_{cls} as in a multi-label classification task:

$$L_{cls}(c, \hat{c}) = c \log \hat{c} + (1 - c) \log(1 - \hat{c}) \quad (2.3)$$

Similarly as in [157] the combination of the sum-squared errors L_{box} and L_{obj} with the cross-entropy loss L_{cls} is controlled by the scaling parameters λ_{box} , λ_{obj} , λ_{noobj} , and λ_{cls} .

Aside from the modifications made so far on top of the original YOLOv2 architecture, we also changed the number, the scales, and the aspect ratios of the pre-defined anchor boxes used by the network to predict bounding boxes. As an enhancement strategy when compared to our previous model, we ran a K-Means model with 13 centroids to obtain the initial priors for the anchor boxes that better capture the real distribution of text shapes. The centroids were obtained by gathering the width and height of all the text instances found in the datasets used at training time. Finally, the proposed model predicts 13 anchor boxes per grid cell as output features. Figure 2.4 illustrates the 13 bounding boxes found to be better suited for our training data and their

difference from the ones used in object detection models.

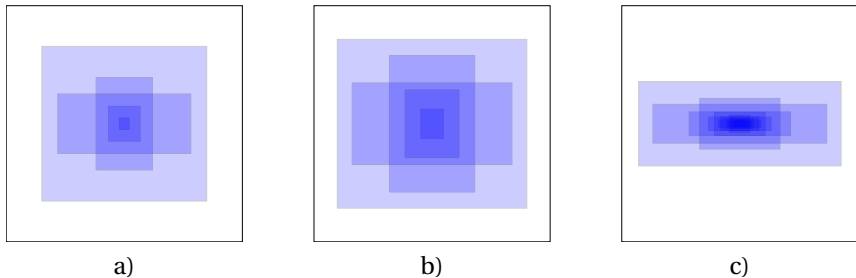


Figure 2.4: Anchor boxes used in the original YOLOv2 model for object detection in COCO (a) and PASCAL (b) datasets. (c) Our set of anchor boxes for text detection.

At test time, our model provides a total of $W/32 \times H/32 \times 13$ bounding box proposals, with W and H being the image input size, each one of them with an objectness score (\hat{C}) and a PHOC prediction (\hat{c}). The original YOLOv2 model filters the bounding box candidates with a detection threshold τ considering that a bounding box is a valid detection if $\hat{C} \max(\hat{c}) \geq \tau$. If the threshold condition is met, a non-maximal suppression (NMS) strategy is applied in order to get rid of overlapping detections of the same object. In our case the threshold is applied only on the objectness score (\hat{C}) but with a much smaller value ($\tau = 0.0025$) than in the original model ($\tau \approx 0.2$), and we do not apply NMS. The reason is that any evidence of the presence of a word, even if it is small, may be beneficial in terms of retrieval if its PHOC representation has a small distance to the PHOC of the queried word. With this threshold, we generate an average of 500 descriptors for every image in the dataset and all of them conform to our retrieval database.

In this way, the scene text retrieval of a given query word is performed with a simple nearest neighbor search of the query PHOC representation over the outputs of the CNN in the entire image database. While the distance between PHOCs is usually computed using the cosine similarity, we did not find any noticeable downside when using a Euclidean distance for the nearest neighbor search.

In this section, we explore different frameworks and backbones inspired by the most common object detection pipelines. Additionally, we present exhaustive experiments on the effect of different PHOC sizes, which yields an incremental improvement when compared to our previous model. We also analyze deeply the capacity of our model in dealing with out-of-vocabulary queries, by conducting several experiments in two multi-lingual datasets. These experiments prove that the proposed method is able to transfer knowledge acquired at training time to construct word representations of previously unseen text samples at inference time.

Lastly, we propose an application of real-time text spotting on video, in which the model corroborates its robustness to noise, blur, and distortions while at the same time maintaining its characteristic high processing speed.

2.4 Datasets

2.4.1 IIIT Scene Text Retrieval (STR)

The STR dataset [142] is a scene text image retrieval dataset composed of 10,000 images collected from the Google image search engine and Flickr. The dataset has 50 predefined query words and for each of them, a list of 10 – 50 relevant images (that contain the query word) is provided. It is a challenging dataset where relevant text appears in a wide range of different fonts and styles, and from different viewpoints, among many distractors (images without any text).

2.4.2 IIIT Sports-10k Dataset

The Sports-10k dataset [142] is another scene text retrieval dataset composed of 10,000 images extracted from sports video clips. It has 10 predefined query words with their corresponding relevant images' lists. Scene text retrieval in this dataset is especially challenging because images are low resolution and often noisy, occluded, or blurred, with small text generally located on advertisements signboards.

2.4.3 Street View Text (SVT) Dataset

The SVT dataset [205] is comprised of images harvested from Google Street View where text instances are present. The scene text found exhibits great variability and is usually related to business names and descriptions. It contains more than 900 words annotated in 350 different images. In our experiments, we use the official partition that splits the images into a train set of 100 images and a test set of 249 images. This dataset also provides a lexicon of 50 words per image for recognition purposes, but we do not make use of it. For the image retrieval task, we consider as queries the 427 unique words annotated on the test set.

2.4.4 Multi-lingual Scene Text (MLT) Datasets

These two datasets MLT2017 [1] and MLT2019 [2] are scene text detection and recognition datasets that contain 7,200 and 10,000 images respectively in 10 different languages (Chinese, Japanese, Korean, English, French, Arabic, Italian, German, Bangla, and Hindi) in equal proportions, representing 7 different scripts. These datasets mostly comprise focused text in natural images, and even though the main task is text detection and recognition, we adapted it to conduct text retrieval experiments. We employ this dataset to assess the generalization power of the PHOC representation of unseen words at training time.

2.4.5 Text in Videos (TiV) Dataset

The TiV dataset [3] contains 25 videos (13450 frames in total) and a test set of 24 videos (14374 frames in total) recorded from 4 different cameras. We use this dataset to assess the performance in real-time of our model at the moment of retrieving a specific text query. The challenge in this dataset remains in the fact that usually video frames contain a lower quality when compared to static images. The problems of text spotting usually relate to rotation, blur, and occlusion of text found on each frame due to movement and focusing issues while including loss of information at the moment of video compression.

2.5 Experimental Results

2.5.1 Training Details

The model was trained with a combination of several datasets. The first one is a modified version of the synthetic dataset of Gupta *et al.* [69]. The dataset generator has been evenly modified to use a custom dictionary with the 90K most frequent English words, as proposed by Jaderberg *et al.* [82], instead of the Newsgroup20 [105] dataset dictionary originally used by Gupta *et al.*. The rationale was that in the original dataset there was no control over the word occurrences, and the distribution of word instances had a large bias towards stop-words found in newsgroups' emails. Moreover, the text corpus of the Newsgroup20 dataset contains words with special characters and non-ASCII strings that we do not include in our PHOC representations. Finally, since the PHOC representation of a word with a strong rotation does not make sense under the pyramidal scheme employed, the dataset generator was modified to allow rotated text up to 15 degrees. This way we generated a dataset of 1 million images for training purposes. Figure 2.5 shows a set of samples of our training data. Additional 10k synthetic images proposed by [228] were added to the training data. This dataset has been shown to improve scene text detectors and recognizers. Real images taken from the 1,233 training images found in the ICDAR2013 [91] and ICDAR2015 [90] datasets were also added to the training set.

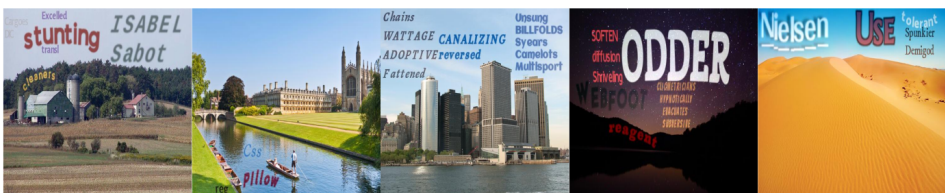


Figure 2.5: Synthetic training data generated with a modified version of the method of Gupta *et al.* [69]. We make use of a custom dictionary with the 90K most frequent English words, and restrict the range of random rotation to 15 degrees.

We trained the proposed model with an initial learning rate of 0.001 and a batch size of 16, however, a gradient accumulation strategy was employed every 4 iterations, thus resulting in a total batch size of 64. The training process involved the combination of two recent optimizers, RAdam [121] and the Lookahead [230] which when working in combination yielded the best results. We initialized the weights of our model with the YOLOv2 backbone pre-trained on Imagenet. At this stage, we employed a fixed input image size of 448×448 . After the initial 20 epochs, we decreased the learning rate to 0.0001 and trained the model for 5 additional epochs. At this point, we also adopted multi-scale training by randomly resizing input images among 13 possible sizes in the range from 352×352 to 768×768 to increase the generalization ability of the model towards different-sized images. As a way to fine-tune the model to predict text instances in real images, a final epoch was employed, which involved the usage of images only from the ICDAR2013 and ICDAR2015 datasets. Through the training process, we kept the parameter λ_{cls} set to 2.0 to increase the weight placed to learn accurate PHOCs, the parameter λ_{box} was set to 5.0 and the rest of the parameters were fixed to 1.0. During the whole training stage, we used a data augmentation strategy that involved randomly cropping the center of an image without affecting text instances and adding saturation and hue.

2.5.2 Backbone Comparison

Several works analyzing a wide range of backbones have been proposed to tackle the object detection task [117, 161, 123, 157]. More often than not, these frameworks are used as a starting point to construct text detectors and recognizers which rather than focusing on objects, are designed to detect and recognize textual instances. The model presented in our previous work [61] employs a customized backbone inspired by YOLOv2 [159].

Additionally, we include the results of the following frameworks: YOLOv3 [160], Faster R-CNN [161] and RetinaNet [117] modified to predict a PHOC embedding given that a text instance was detected. Adapting each of the mentioned frameworks for scene text retrieval required specific design choices and training methodologies that are explained as follows.

YOLOv3 [160] is an incremental improvement on the YOLOv2 model. A significantly bigger number of convolutional layers were added and if considering additional layers such as routing, upsampling, and shortcut layers, the total number increases to 106 layers. The custom YOLOv3 used in our experiments makes predictions at each grid cell through 3 different scales and employs 4 different anchors boxes per scale. The initialization of the model is given by the original YOLOv3 weights pre-trained on ImageNet [41]. Rather than predicting a class label, a PHOC prediction is performed on the detected text, and the most confident PHOCs are employed as scene text retrieval features. Since YOLOv3 already uses binary cross entropy to make predictions, we did not make any changes to the original loss function when working with a PHOC descriptor.

The second framework is taken from the work by [117], which presents a model

named RetinaNet. This model leverages a Feature Pyramid Network (FPN) [116] to extract features across different layers of a network. The resulting features are further used to perform bounding box regression and classification by two separate branches. A novel loss function that focuses on hard, misclassified samples while reducing the relative loss on well-classified ones is adopted. In our experiments we used as the FPN backbone a ResNet-18 [73] pre-trained on ImageNet. Since the RetinaNet model does not output a confidence parameter, we modified the bounding box regression branch. The network was trained to predict a confidence value based on the IoU between the predicted anchor boxes and the ground truth text annotations. We added an MSE Loss term to predict this new confidence output which was added to the total function to be optimized. The evaluated model outputs 9 anchor predictions per each FPN level, across 5 different levels (L3-L7) readjusted to capture text shape. Similar to the idea from [61], at inference time we use the confidence values to sort the most confident PHOCs predicted from the scene text contained in an image.

The third evaluated framework is a Faster R-CNN, a multi-stage object detector [161]. The Faster R-CNN comprises of two stages, a Region Proposal Network (RPN) and a Region Of Interest (ROI) Pooling stage used for later classification and regression. The backbone of this model is a VGG-16 [177] which extracts common features for the RPN and ROI layers. In the RPN layer, 9 anchors are proposed at each sliding window, yielding coordinates and scores. The coordinates refer to the location of regions from the convolutional features to be used as input into the ROI stage to be later regressed and used to predict PHOCs by two different branches. The scores indicate the presence of text and background according to the IoU between predicted anchors and ground truth labels. We use this score to sort the PHOC predictions at inference time accordingly.

It is paramount to indicate that training and finding a global optimum among all these models is not a trivial task. The additionally explored frameworks contain a significantly larger amount of parameters compared to the original model based on YOLOv2, which in turn produces a slower training procedure. Such models use several hyper-parameters that can be further tuned but are specific to each one. Due to the computational resources involved in fine-tuning and optimizing each hyper-parameter of the explored models individually, we trained all models for 10 epochs. The PHOC dimension was kept constant across all of them by using the most common size in literature 604-d [5][61][184]. A gradient accumulation policy was employed to train the additional frameworks in a way that the resulting gradient at back-propagation was applied to a larger batch of samples, in a similar manner as the original YOLOv2 model. This methodology was adopted because the explored backbones contained a big number of parameters that could not be optimized by using a single GPU otherwise.

The results obtained by using different models are summarized in Table 2.1. We note that by employing the YOLOv3 backbone better results are obtained in the Sports10K dataset. However, the outcome of increased performance on one dataset comes at the cost of a longer convergence time for the model at training stage and also a significant

decrease at inference (50%) measured by frames per second.

Framework	Backbone	Training Hours	IIIT	Sports10K	SVT	FPS
RetinaNet	ResNet 18	888	63.29	74.64	79.31	11
Faster R-CNN	VGG 16	960	66.21	76.08	80.53	5
YOLOv3	Darknet 53	288	67.99	79.33	83.59	20
YOLOv2	Darknet 19	144	69.92	78.01	85.12	43

Table 2.1: Comparison of different frameworks and backbones in the scene text retrieval task. Each model predicts PHOCs of textual instances found and the most confident ones are employed as textual features for retrieval. The metric employed is the mean average precision (mAP).

We should remark that Faster R-CNN and RetinaNet due to their slow processing speed are not suitable for the purpose of obtaining a real-time and efficient text spotter. The complexity of these two models is significantly higher than YOLOv2 or YOLOv3, bringing in a substantial time required to achieve convergence, a key asset, especially when working with limited computational resources. In order to preserve real-time processing capabilities, efficient large-scale indexing, and faster convergence time at training we have decided to employ the YOLOv2 backbone for the rest of our experiments.

2.5.3 Effect of PHOC Dimension

In order to assess the effect of the PHOC vector size, various retrieval experiments were conducted by incrementally adding up the pyramidal unigram levels and the redundancy of the bigrams that form the descriptor. Due to computational constraints and in order to provide a grounded comparison, all the evaluated models were trained for 10 epochs. The backbone employed was the custom variation of YOLOv2 [159] and the different loss function hyper-parameters were kept similar in all scenarios.

Table 2.2 shows the results obtained by employing different PHOC sizes. It is evident that using single unigram levels do not yield good results because spatial information of the characters found within a word is lost. It is important to point out that higher unigram levels contain richer spatial information that is discriminative when differentiating similar words, thus achieving a higher score than employing lower unigrams. Intuitively, concatenating several unigram levels provide a more informative vector of the positioning of characters in a word, thus resulting in higher precision. It is worth mentioning that combinations of PHOCs of higher levels usually yield better results than combinations of lower unigram levels since higher levels contain more discriminative information, particularly for longer words. Also, we should indicate that unigram levels contain information that is more relevant for the retrieval task than the one contained in bigrams, nonetheless adding bigram levels contribute to the improvement of the model. Lastly, by employing bigrams combined with several unigram levels, a more informative vector is obtained at the expense of a higher dimensional

PHOC Evaluated			Dataset		
Unigram Levels	Bigram Levels	Dimension	IIIT	Sports10K	SVT
L1	0	36	47.84	67.31	70.17
L2	0	72	54.83	69.54	73.07
L3	0	108	53.94	69.92	73.67
L4	0	144	54.69	69.33	73.21
L5	0	180	56.53	68.44	72.39
L6	0	216	57.14	69.47	73.56
L2+L3	0	180	58.77	68.56	74.68
L2+L3+L4	0	324	62.35	70.73	75.84
L3+L4+L5	0	432	61.46	68.87	76.51
L2+L3+L4+L5	0	504	62.19	72.52	76.98
L2+L3+L4+L5+L6	0	720	64.57	75.29	78.16
L1+L2+L3+L4+L5	0	540	61.29	72.76	76.19
L3+L4+L5	1	482	62.53	70.79	75.15
L3+L4+L5	2	532	63.35	73.28	76.72
L2+L3+L4+L5	1	554	62.62	73.41	77.55
L2+L3+L4+L5	2	604	62.89	73.88	77.76
L2+L3+L4+L5+L6	1	770	63.64	73.31	78.63
L2+L3+L4+L5+L6	2	820	66.60	74.35	79.19

Table 2.2: Comparison of the effect of different PHOC sizes by employing different unigram and bigram levels. The backbone employed on all the experiments is a customized version of YOLOv2 [159]. The metric employed is the mean average precision (mAP) across all datasets.

representation. As such, the best performing PHOC in our experiments is the one that combines 5 levels (L2-L6) jointly with the concatenation of 2 bigrams levels, resulting in a 820 dimensional vector.

2.5.4 Comparison with State-of-the-Art

The performance of the presented model as well as the previous state-of-the-art methods is shown in Table 2.3. The models are evaluated for scene text image retrieval on the IIIT-STR, Sports-10K, and SVT datasets. At inference time our model employs an input image size of (608 × 608), and the processing time has been calculated using a Titan X (Pascal) GPU with a batch size of 1.

Our method outperforms the previous best retrieval model [61] by more than 6 points on mAP on the Sports dataset. The improvement is smaller on the other two datasets, nonetheless, it is still relevant. However it is important to highlight that our model does not employ a multi-resolution strategy at inference time to predict PHOCs at different scales as in our previous work [61], and by using just a fixed input size of 608 × 608 is able to surpass previously obtained results. We should mention that our method achieves a slightly slower measurement in frames per second (fps) when

Method	IIIT (mAP)	Sports (mAP)	SVT (mAP)	fps
SWT [50]+ Mishra et al. [143]	-	-	19.25	
Wang <i>et al.</i> [205]	-	-	21.25*	
TextSpotter [148]	-	-	23.32*	1.0
Mishra <i>et al.</i> [142]	42.7	-	56.24	0.1
Ghosh <i>et al.</i> [57]	-	-	60.91	
Mishra [140]	44.5	-	62.15	0.1
Almazán <i>et al.</i> [5]	-	-	79.65	
TextProposals [60] + DictNet [81]	64.9 [†]	67.5 [†]	85.90 [†]	0.4
Jaderberg <i>et al.</i> [82]	66.5	66.1	86.30	0.3
Bušta <i>et al.</i> [26] ICCV 2017	62.94 [†]	59.62 [†]	69.37 [†]	44.21
Liu <i>et al.</i> [124] CVPR 2018	52.82 [†]	65.62 [†]	68.98 [†]	20.26
Bušta <i>et al.</i> [25] ACCV 2018	64.39 [†]	68.29 [†]	76.44 [†]	12.2
He <i>et al.</i> [74] CVPR 2018	50.16 [†]	50.74 [†]	72.82 [†]	1.25
He <i>et al.</i> [74] (With dictionary)	66.95 [†]	74.27 [†]	80.54 [†]	2.35
He <i>et al.</i> [74] (PHOC)	46.34 [†]	52.04 [†]	57.61 [†]	2.35
YOLO-PHOC 604-d (576 × 576) [61]	68.13	72.99	82.02	53.0
YOLO-PHOC 604-d (608 × 608) [61]	69.83	73.75	83.74	43.5
YOLO-PHOC 604-d (multi-res.) [61]	71.37	74.67	85.18	16.1
YOLO-PHOC 820-d (608 × 608)	71.67	80.96	85.74	42.2

Table 2.3: Comparison to previous state of the art for text based image retrieval: mean average precision (mAP) for IIIT-STR, and Sports-10K, and SVT datasets. (*) Results reported by Mishra et al. in [142], not by the original authors. (†) Results computed with publicly available code from the original authors.

compared to the model that uses a PHOC of 604 dimensions at the 608 × 608 input resolution, due to the usage of a bigger PHOC descriptor. Nevertheless, the previously mentioned effect is almost negligible when compared to single resolution models, especially if we consider the significant boost in mAP. The improvement comes into place due to larger preservation of spatial information of words by employing a larger descriptor with redundant information, resulting in greater performance achieved on all three evaluated datasets. Lastly, we compare the performance to additional state-of-the-art text recognizers, such as [24, 25, 124, 74]. We also include results from a variant of He *et al.* [74], but rather than direct string matching, the model’s results are first transformed to PHOC descriptors, and the look-up is based on similarity in the PHOC space. It can be seen that the PHOC space does not offer any advantage compared to end-to-end recognition methods.

Table 2.4 further compares the improved method to the previous state-of-the-art by showcasing the retrieval precision at 10 (P@10) and 20 (P@20) cut-off points on the Sports-10K dataset. Due to the significant improvement in the evaluated dataset, our YOLOv2-820-d model is able to achieve better precision and obtain state-of-the-art performance when compared to other methods.

Method	Sports-10K (P@10)	Sports-10K (P@20)
Mishra <i>et al.</i> [142]	44.82	43.42
Mishra [140]	47.20	46.25
Jaderberg <i>et al.</i> [82]	91.00	92.50
YOLO-PHOC 604-d	92.00	90.00
YOLO-PHOC 820-d	93.50	93.00

Table 2.4: Comparison to the previous state of the art for text-based image retrieval: precision at n (P@n) for the Sports-10K dataset.

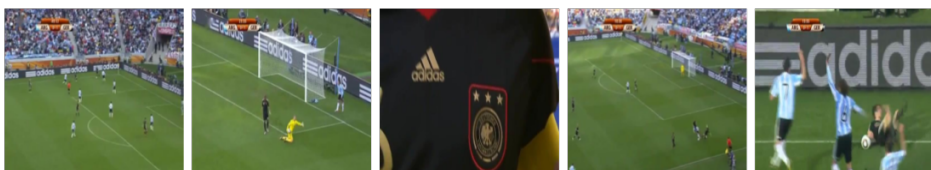


Figure 2.6: Top-5 retrieved images for the query “adidas”. Our model successfully retrieves partially occluded and blurred words.

In Figure 2.6 we provide qualitative results by showing the top-5 retrieved images for the query “adidas” and in Figure 2.7 for the query “castrol” on the Sports10k dataset. We achieve a 100% precision despite the fact that the images contain blurred, partially occluded, and rotated text instances.

To offer additional insights, Figure 2.8 depicts the heat maps of our model by calculating the closest matching PHOCs and the corresponding bounding boxes in relation to a given query. As can be seen in the showcased figure, several predicted PHOCs closely match the queried word. By revisiting the implementation details defined in the previous section, we confirm the fact that avoiding the usage of an NMS post-processing strategy is indeed a safe way to preserve high-matching PHOC proposals that could be discarded otherwise.

2.5.5 Multi-Lingual Scene Text Retrieval

As an extension to our previous work [61], we focus on analyzing the generalization capability of the proposed models. It becomes essential to note that designing an algorithm that learns to construct a compact representation of a string, such as the PHOC, paves the road to further development of models that are not constrained to a fixed dictionary or training data samples. In order to assess the expressiveness of our architecture, we make use of two Multi-lingual datasets 2017 [1] and 2019 [2] in which we can easily find out-of-vocabulary words (text not seen at training time) with different distributions and characteristics. These datasets are used by the research community



Figure 2.7: Top 10 ranked images for the query “castrol”. Our model has not seen this word at training time.

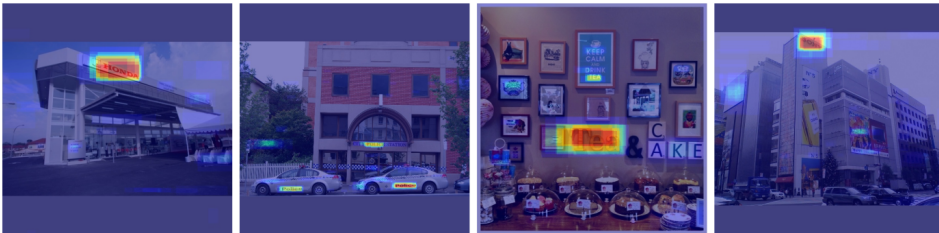


Figure 2.8: Bounding box heat-maps for queried words “honda”, “police”, “tea” and “sony” respectively.

to perform text detection and recognition tasks, but not text-based image retrieval. Therefore, we have selected a set of 100 queries for in-vocabulary experiments and another set of 100 queries for out-of-vocabulary experiments for each dataset taken from the training split. Out-of-vocabulary queries are selected by choosing the Latin words with the most occurrences after removing stop-words and words that contain non-alphanumeric characters. For in-vocabulary queries, we also remove stop-words and words with non-alphanumeric characters before searching for Latin words with similar frequencies as the ones obtained in the out-of-vocabulary queries. Therefore, we end up with queries that have comparable occurrences in both datasets with an average of 13.32 and 13.99 for IV and OOV queries respectively.

Tables 2.5 and 2.6 show the ability for our model to perform retrieval with the same accuracy for in-vocabulary queries and out-of-vocabulary queries in both datasets. As we stated previously, this is because our model is learning how to build a PHOC from text rather than performing a classification along a fixed dictionary. It is important to note that our model performs significantly better than a state-of-the-art reading system presented by [74] at the text retrieval task. Additionally, the method from [74] was trained using the dictionary from [105] which contains English words, thus performing

Method	MLT 2017		MLT 2019	
	IV	OOV	IV	OOV
He et al. [74]	24.79	19.47	27.6	24.99
YOLO-PHOC 604-d	46.52	46.87	46.41	46.03
YOLO-PHOC 820-d	47.61	47.75	47.55	47.19

Table 2.5: Comparison to previous state of the art method for text based image retrieval methods when queries are words already seen during the training process (IV) or not (OOV). The metric employed is the mean average precision (mAP).

Method	MLT 2017						MLT 2019					
	IV			OOV			IV			OOV		
	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20	P@5	P@10	P@20
He et al.[74]	0.51	0.37	0.22	0.46	0.33	0.20	0.62	0.44	0.27	0.60	0.40	0.23
YOLO-PHOC 604-d	0.77	0.57	0.34	0.78	0.59	0.34	0.80	0.64	0.41	0.80	0.64	0.40
YOLO-PHOC 820-d	0.79	0.59	0.35	0.80	0.61	0.36	0.82	0.66	0.43	0.83	0.65	0.41

Table 2.6: Comparison to the previous state of the art method for text-based image retrieval methods when queries are words already seen during the training process (IV) or not (OOV): precision at n (P@n)

poorly when dealing with out-of-vocabulary words mostly belonging to different languages. We can also observe that a slight increase in the performance of the proposed model is obtained when dealing with OOV queries rather than IV queries in the MLT 2017 dataset. Due to the similar occurrences of positive samples of queries both for IV and OOV words, this difference is negligible and rather proves the generalization and knowledge transfer capability of our model to construct PHOCs of unseen samples at training time.

Figure 2.9 shows the top-5 ranked images for the queries “vodafone” in IIIT-STR dataset, “uscita” (italian) and “parkausweis” (german) in MLT 2017 and “werden” (german) in MLT 2019, all of them being unseen samples at training time. In all of them our model reaches a 100% precision at 5.

2.5.6 Real-time Text Spotting in Videos

Given the high processing frame rates that we achieve (c.f. Table 2.3), we can use the proposed method for spotting text in video streams in real-time.

This application might be interesting in scenarios like assistance to driving systems, in order to spot certain words in the open world or to track advertisement exposure in sports broadcasting. In such cases, the user casts a textual query that has to be sought within videos. We shall take into account that video recorded in natural scenes contains text instances that are extremely susceptible to imperfect conditions. Low-quality recording devices and rapid camera movement tend to produce blurred



Figure 2.9: From top to bottom, top-5 ranked images for the queries “vodafone”, “uscita”, “werden” and “parkausweis”. Although our model has not seen these words at training time it is able to achieve a 100% P@5 for all of them. Best viewed in color.

and rotated content. Text found in video is also vulnerable to unintended occlusions that affect several consecutive frames. In order to test the performance of the proposed method in such a scenario, we have used the Text in Videos challenge dataset [91], in which the train partition consists of 25 videos, 13,450 frames in total, with their corresponding ground-truth annotation. We decided to use as queries the top 20 most occurring words in the dataset that have more than three characters in length. Having set a threshold on the distance between the query PHOC representation and the closest word hypothesis in each frame, we decide whether the queried word appears or not in that frame. We evaluate the text spotting in the videos task by using the F-score so that we penalize both missing frames where the query word appears and false positive frames. Overall we achieved an F-score of 76.70, and we provide some results for the topmost 15 queries in Table 2.7. Video demos are available in our public repository².

²<https://github.com/AndresPMD/Pytorch-yolo-phoc>

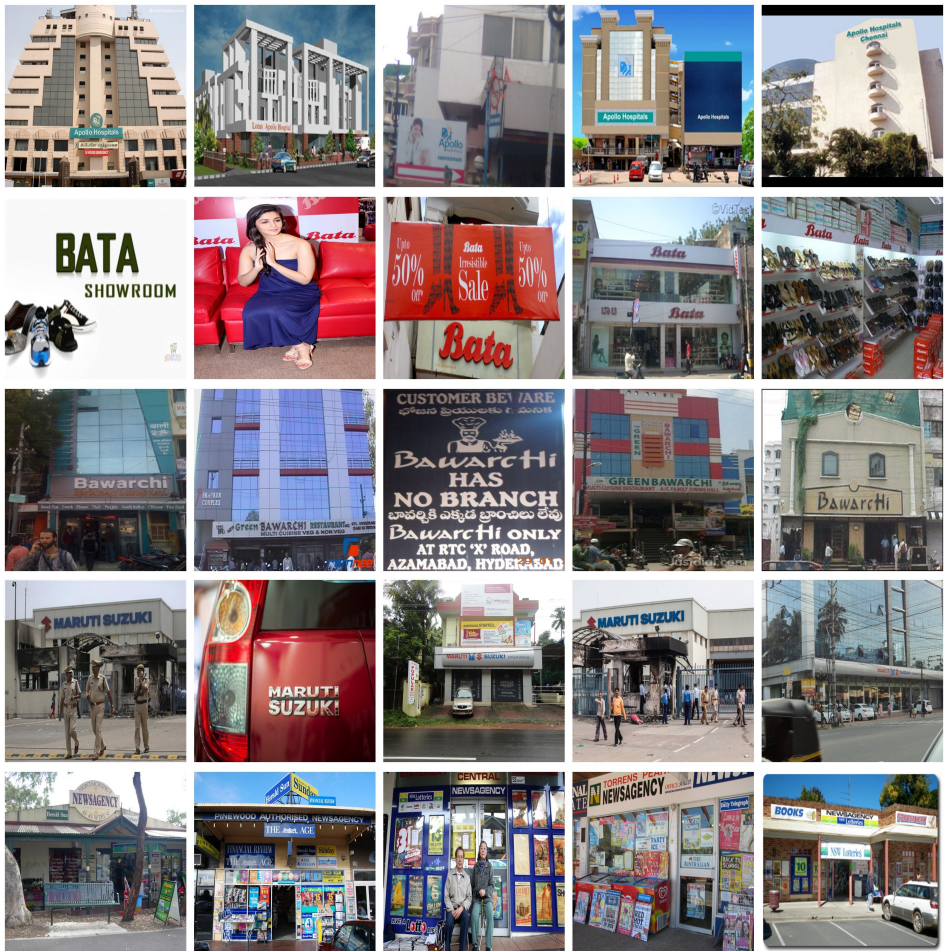


Figure 2.10: From top to bottom, top-5 ranked images for the queries “apollo”, “bata”, “bawarchi”, “maruti” and “newsagency”. Although our model has not seen these words at training time it is able to achieve a 100% P@5 for all of them.



Figure 2.11: Error analysis: most of the errors made by our model come from text instances with a particular style, font type, size, etc. that is not well represented in our training data.

Query	Occurrences	F-score
<i>flor</i>	539	94.05
<i>Marie</i>	426	83.89
<i>Renfe</i>	314	78.26
<i>createurs</i>	303	72.40
<i>Dixan</i>	278	87.54
<i>FONTANEDA</i>	261	84.44
<i>VOTRE</i>	257	91.01
<i>Digestive</i>	254	90.00
<i>USHIP</i>	245	75.35
<i>ACCASTILLEUR</i>	241	66.26
<i>Applus</i>	237	91.96
<i>Rectorat</i>	237	88.96
<i>CONSEIL</i>	230	83.18
<i>mundi</i>	230	85.24
<i>Accastillage</i>	199	61.41
<i>MISTOL</i>	186	57.51
Average	277.31	76.70

Table 2.7: Top 15 most frequent words with their number of occurrences and the reached F-score.

2.6 Conclusions and Future Work

In this Chapter, we presented a real-time performing word spotting method, based on a fully convolutional neural network that allows to detect and recognize text in a single pass which yields real-time processing capability.

A comparison of different backbones was presented and the results were analyzed. The proposed model achieves state-of-the-art performance while using fewer parameters, fewer computational resources, and converging in a considerably shorter training time. The effect of different PHOC sizes was undertaken as well. It was shown that a bigger PHOC, with redundant unigram and bigram levels, tends to better preserve and encode spatial information of a word yielding superior retrieval results at a minor cost of performance at inference time. The newly proposed model significantly improves the previous state-of-the-art results on the scene text retrieval task on the IIIT-STR and Sports-10K dataset while obtaining comparable results in the SVT Dataset. Moreover, it can do so by achieving speeds $50\times$ to $150\times$ faster when compared to other methods while using a single input resolution, which opens up the possibility of employing this model for real-time scenarios, such as video, and indexing large-scale databases.

Importantly, it has been shown that the proposed method is able to construct a compact vectorial representation of out-of-dictionary queries at inference time while keeping a similar performance as of words seen at training time. Achieving this re-

sult is only possible by employing the PHOC as a word representation instead of tackling the task as a direct word classification. The method showcased is able to generalize unseen samples in a powerful and efficient way, as the evidence strongly points out in experiments performed in two multilingual datasets. Future work can be conducted to investigate the use of word embeddings that exploit the morphology of a word other than PHOC or diverse tasks in which morphology can be leveraged. The code, pre-trained models, data, and demo videos used in this work are publicly available at <https://github.com/AndresPMD/Pytorch-yolo-phoc>.

Part I

Leveraging Scene Text for Fine-grained Image Classification

Chapter 3

Scene Text for Fine-grained Image Classification and Retrieval

3.1 Introduction

Written communication is arguably one of the most important human inventions that allows the transmission of information in an explicit manner. Given the fact that text is omnipresent in man-made environments [196, 90], as well as the implicit relation between visual information and scene text instances, the design of holistic computer vision models for scene interpretation is fundamental. With the purpose of designing such a holistic model, in this Chapter, we leverage textual information in the scene to address the problem of scene text-based fine-grained image classification and image retrieval.

Fine-grained image classification (FGIC) tackles the problem of classifying different object instances that are visually similar and difficult to discriminate. The complexity of this task lies in finding discriminative features which often require domain-specific knowledge [137, 212]. A lot of research on this problem has been oriented to differentiate visually similar objects such as birds[56], aircrafts [137], and dog breeds[94] among others, which more often than not require domain-specific knowledge. The importance often lies in extracting discriminative features that even though subtle, provide enough information to make a prediction. However, differentiating objects by leveraging available textual instances in the scene is an omnipresent practice in daily life. In this work, we focus on exploiting scene-text as the main discriminatory feature to perform FGIC. An early work that demonstrated the importance of text (domain-specific knowledge) for fine-grained storefront classification was put forward by Movshovitz *et al.* [144], in which the trained classifier learned automatically to attend to the text found in an image as the sole way of solving the task. In the case of blurred or occluded text instances, the classification task is extremely challenging for humans as well. Consequently, scene text found in an image serves as an addi-

tional discriminative signal that a model should incorporate into its design. Since then, there has been additional research that explicitly combines textual and visual cues. The works presented by Karaoglu *et al.* [89, 87] and Bai *et al.*[12] are the most related ones to our work.

3.2 Related Work

3.2.1 Fine-Grained Classification

Fine-grained classification is the task of classifying visually similar objects in which subtle differences are key to finding discriminative features between classes. Finding these subtle features is a challenging task that keeps this problem an active topic in computer vision. Lately ample research has been done, specially in task-specific domains such as animal[94], objects[136] and plants[186] classification. Recent works on fine-grained classification base their approach on localizing salient parts of an image [42, 215], and use the saliency maps to classify the objects. Later approaches such as the one of Tang *et al.* [190], use a weakly supervised method to find discriminative features and leverage them to perform the classification between similar instances. Other methods use existing prior knowledge from unstructured text to propose a semantic embedding that differentiates similar classes [213]. A self-supervision method is introduced in [217] that learns to propose significant image regions to find inter-class discriminative features.

Different from solely visual-based FGIC methods, there has been growing interest to use textual cues to achieve this task by incorporating two modalities. More related to our work, [87] tackles this task by extracting visual features with a pre-trained GoogleNet [188] and a Bag of Words feature to represent the text instances found in an image and further classify them. More recently, Bai *et al.* [12] use a similar approach and extract visual features using a GoogleNet and a combination of two models: [115] to detect and [172] to recognize text. The text found is represented as GloVe features [151], a word embedding that is further used with attention to the visual features to find a semantic relation between the two modalities to classify the image.

3.2.2 Multimodal Fusion

In computer vision tasks, the interaction of contents coming from different modalities usually can be described in multiple kinds of ways. The combination of different modalities provides a richer content description rather than one modality alone, therefore the contained knowledge should be leveraged to further exploit explicit information according to the task [182]. Traditional methods divide the fusion pipeline into early and late fusion. Early fusion integrates the features extracted from domain-specific models immediately after obtaining them. Late fusion integrates the features once each modality has made a decision such as regression or classification [13]. Ap-

proaches such as the work from [95] experiment with the early fusion of multimodal features by concatenating them before classifying them. Regarding similar work as ours, [87] and [12] both employ concatenation as a fusion method of combining textual and visual features. In this Chapter we explore other fusion methods used in multimodal learning, that show a performance increase, especially in tasks that require exploiting two modalities such as Visual Question Answering (VQA) and Visual Relationship Detection (VRD).

One of the initial works presented by [16], modeled a Tucker decomposition of the bilinear interaction of two distinct modalities. Later, a Multimodal Low-rank Bilinear Attention Network (MLB) was proposed by [97], in which the result of the fusion of two modalities was based on a low-rank bilinear pooling operation using the Hadamard product along with an attention mechanism. A factorized bilinear pooling (MFB) is proposed by [225], where each third mode section of the tensor is constrained by a rank. Later methods, such as a Multimodal Factorized High-order pooling (MFH) fusion presented by [226], use a high-order fusion formed by cascaded MFB modules. In the work conducted by [16], a bilinear pooling is performed where the tensor is represented as a Tucker decomposition. The obtained main tensor has the same rank constraint as the MFB technique. Lately, a Multimodal Bilinear Superdiagonal Block (Block) fusion strategy based on the work presented by [17], has achieved state-of-the-art results in VQA and Visual Reasoning.

3.2.3 Attention and Reasoning

To model the interaction among different modalities, attention-based [9] approaches also have been proposed [8, 214, 93]. With the aim of designing models capable of reasoning, the intrinsic synergy between visual and textual features has been explored. Work such as [227, 111] employ variations of an LSTM and a Gated Recurrent Unit (GRU) to perform reasoning in a sequential manner. However, significant advances have been made by the usage of Graph Convolutional Networks (GCN) [100], due to the proven capability of modeling relationships [167] between nodes in a given graph. Along this road, GCNs have been successfully used in tasks that require reasoning such as VQA [147, 178, 54], image captioning [112, 216] and image-sentence retrieval [111, 120].

Chapter 4

Combining Visual and Locally Pooled Textual Features

Text contained in an image carries high-level semantics that can be exploited to achieve richer image understanding. In particular, the mere presence of text provides strong guiding content that should be employed to tackle a diversity of computer vision tasks such as image retrieval, fine-grained classification, and visual question answering. In this Chapter, we address the problem of fine-grained classification and image retrieval by leveraging textual information along with visual cues to comprehend the existing intrinsic relation between the two modalities. The novelty of the proposed model consists of the usage of a PHOC descriptor to construct a bag of textual words along with a Fisher Vector Encoding that captures the morphology of text. This approach provides a stronger, multimodal representation for this task and as our experiments demonstrate, it achieves state-of-the-art results on two different tasks, fine-grained classification, and image retrieval.

4.1 Introduction

In this Chapter, we propose the use of a state-of-the-art text scene-text retrieval model introduced in the previous Chapter and published by Mafla *et al.* [134] to detect and obtain the Pyramidal Histogram of Characters (PHOC) of scene text. We use the PHOC descriptors extracted from images and explore different fusion strategies to merge the visual and textual modalities. Additionally, we construct a Fisher Vector (FV) encoding from the obtained PHOCs to obtain a fixed-length text feature in our pipeline and

further improve the classifier results. Our model leverages the visual features combined with the morphology of a word (refer to Figure 4.1), that belong to specific fine-grained classes, without the need to understand them semantically. Contrary to previous methods, this approach is especially useful when dealing with text recognition errors and named entities which are often difficult to encode in a purely semantic space. The combination of these two modalities produces an output probability vector that addresses the classification task at hand. As an additional application, we evaluate the proposed model on fine-grained image retrieval in available datasets.

Overall, the main contributions of our work are:

- We propose a novel architecture that achieves state of the art on fine-grained classification by jointly considering the textual and visual features of an image.
- We show that by using Fisher Vectors obtained from PHOCs of scene text, we obtain a more robust representation in which words with similar structures get encoded on the same Gaussian component, thus creating a more powerful discriminative descriptor than PHOCs alone.
- We provide exhaustive experiments in which we compare the performance of different alternative modules in our model and previous state of the art.

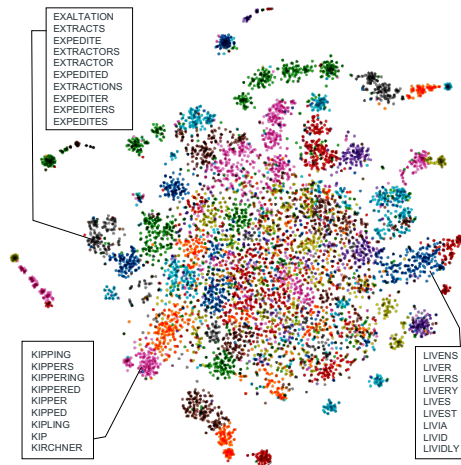


Figure 4.1: T-SNE Visualization [130] of the 300 dimensional PCAed PHOCs in a two dimensional space. Words with similar morphology are clustered together by a Gaussian Mixture Model, thus making such a descriptor suitable and powerful enough to discriminate text for a fine-grained classification task.

4.2 Fine-grained Classification Model

In this model, we leverage the Pyramidal Histogram Of Characters (PHOC) descriptor [5, 185] (see Figure 2.2) commonly used to query a given text instance in handwritten documents and natural scene images. The PHOC of a word encodes the position of a specific character in a particular spatial region of the detected text instance. Such a descriptor has proven to perform as the state of the art in scene text retrieval [62], and as our experiments show, encoding it with the Fisher Vector [152] provides an improved text descriptor for fine-grained classification.

The devised model consists mainly of four processing blocks: visual features extraction, textual features extraction, attention unit, and classification. The whole model pipeline is shown in Figure 4.2.

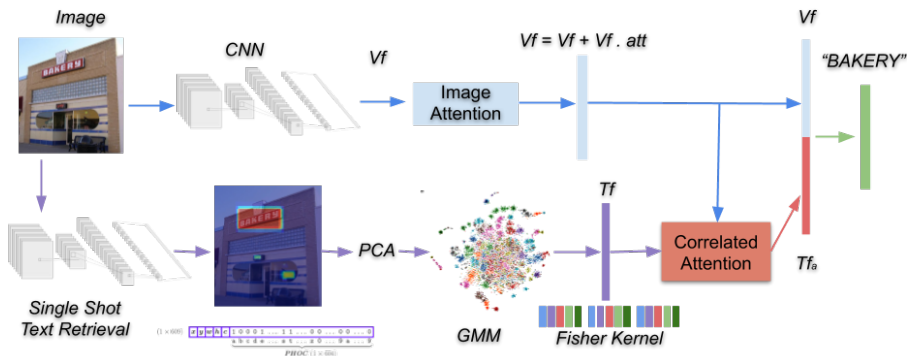


Figure 4.2: Proposed model pipeline. The PHOCs obtained from [135] are used to compute a Fisher Vector that yields a compact morphology-based descriptor suitable to discriminate features from visually similar objects.

The first block extracts the visual features from a given image and produces a fixed-size representation of it. The second block consists of extracting the PHOC representation of each text instance found in an image and using a pre-trained Gaussian Mixture Model (GMM) to obtain the corresponding FV descriptor. The third block consists of an attention unit that multiplies learned weights with the encoded FV depending on the visual features extracted previously. Finally, the last block consists of a concatenation of the two different modalities followed by a fully connected layer to obtain a probability output vector which is used for classification. For the rest of the Chapter, let \mathcal{C} be the set of all possible categories in a given dataset; $\mathcal{X} = \{x_i\}_i^N$ be the set of images; $l_x : \mathcal{X} \rightarrow \mathcal{C}$ be the labeling function.

4.2.1 Visual Features

In our model, we use a Convolutional Neural Network (CNN) [73] pre-trained on ImageNet [41] as a visual feature extractor, denoted as $\phi(\cdot)$. We use the output of the last convolutional block of $\phi(\cdot)$ before the last average pooling layer as the visual features, denoted as V_f . Attention to visual features has proven to yield improved performance on several tasks. As it is presented by [44], we compute a soft-attention mechanism due to its differentiable properties, thus allowing end-to-end learning. The proposed attention function learns an attention mask att which assigns weights to different regions of an image given a feature map V_f . The attention mask is learned by applying 1×1 convolution layers on the output features from the CNN. Lastly, to obtain the final output of the attention module along with the visual features, the operation is computed by: $V_{fa} = V_f + (V_f \times att)$.

4.2.2 Textual Features

Methods shown in previous works [87, 12] contain mainly three drawbacks. First, the employed text recognizers are bound to a fixed dictionary, which may or may not include the exact words that are present in the image. Second, some words that are contained in the fixed recognition dictionary may not exist in the proposed semantic embedding (GloVe, Word2Vec) such as license plates, brand names, acronyms, etc. Third, any mistake committed by the recognizer will yield a vector embedding that lies far from the semantic embedding of the correct word. Contrary, correct recognition of semantically similar words that might indicate different fine-grained classes will lead to embeddings close to each other, which are not discriminative enough to perform correct classification. This is the case of similar semantic words such as restaurant and steakhouse, cafe and bistro, Coke and Pepsi among some other sample classes from the datasets used.

In order to exploit the morphology of a word to obtain discerning features, we employ the PHOC representation. The PHOC representation employed in the proposed pipeline is composed of the concatenation of vectors from the levels 2 to 5 plus the 50 most common bi-grams in the English language. This yields a 604-dimensional discrete binary vector that represents the characters contained in a word (see Figure 2.2). A dictionary given by [82] is employed to obtain a PHOC per word, in this way, we populate a matrix of this compact representation. In order to reduce the dimensionality and to find linearly uncorrelated variables of this compact vector, a Principal Component Analysis (PCA) is performed. This procedure yields a more compact but at the same time informative vectorial representation of a given word.

The obtained data points were used to construct a Gaussian Mixture Model (GMM) [68] formed by K Gaussian components. We denote the parameters of the K -component GMM by $\lambda = \{w_k, \mu_k, \Sigma_k, k = 1, \dots, K\}$, where w_k , μ_k and Σ_k are respectively the mixture

weight, mean vector and covariance matrix of Gaussian k . We define:

$$u_\lambda(x) = \sum_{k=1}^K w_k u_k(x) \quad (4.1)$$

where u_k denotes Gaussian k :

$$u_k(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \right\} \quad (4.2)$$

and we require:

$$\forall_k : w_k \geq 0, \quad \sum_{k=1}^K w_k = 1 \quad (4.3)$$

Once the GMM model is trained, it will be used to extract a single Fisher Vector representation per image which encodes its contained textual information. The textual features per image are obtained by using the model from [62]. Given an input image, the model outputs a list of \mathcal{B} bounding boxes, each one containing a confidence score \mathbb{C} and a PHOC prediction. We get the top- m proposals ranked according to their conference score.

We get the top- m object proposals set $\mathcal{O}_m := \{o \in \mathbb{C}_i : o \geq c, \forall c \in \mathbb{C}_i\}$. The resulting PHOCs $\in [0, 1]^{d \times N}$, where d is the dimensionality of the PHOC embedding obtained and N the recognized words embedded in the PHOC space. It is essential to note that the model from [62] is able to generalize and construct PHOCs from previously unseen samples, out-of-vocabulary words, and different languages that employ a similar character set (e.g. Latin), making it suitable for the task at hand. Afterward, we project each embedded textual instance of the obtained descriptors into a reduced dimensional space by employing PCA. The resulting vectors are used to obtain the Fisher Vector [152] from the previously trained GMM. The GMM associates each PCAed vector o_i to a component k in the mixture model with a weight given by the posterior probability:

$$q_{ik} = \frac{\exp \left[-\frac{1}{2} (o_i - \mu_k)^T \Sigma_k^{-1} (o_i - \mu_k) \right]}{\sum_{t=1}^K \exp \left[-\frac{1}{2} (o_i - \mu_t)^T \Sigma_t^{-1} (o_i - \mu_t) \right]} \quad (4.4)$$

For each mode k , consider the mean and the co-variance deviation vectors

$$\begin{aligned} u_{jk} &= \frac{1}{N\sqrt{w_k}} \sum_{i=1}^N q_{ik} \frac{o_{ji} - \mu_{jk}}{\sigma_{jk}}, \\ v_{jk} &= \frac{1}{N\sqrt{2w_k}} \sum_{i=1}^N q_{ik} \left[\left(\frac{o_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \end{aligned} \quad (4.5)$$

where $j = 1, 2, \dots, D$ spans the vector dimensions. The FV of a given image I is simply the concatenation of the obtained vectors u_k and v_k for each of the K components in the Gaussian mixture model.

$$T_f = [\dots \quad \mathbf{u}_k \quad \dots \quad \mathbf{v}_k \quad \dots]^T \quad (4.6)$$

The FV and the GMM encode inherently similar information. This takes place because they both include statistics of order 0, 1 and 2 [166, 152]. However, the FV provides a vectorial representation that is more compact, faster to compute, and suitable for processing. The dimension of the FV obtained, noted as T_f , is given by $(2 \times d \times K)$, where d is the PHOC dimension after performing the PCA and K is the number of Gaussian clusters. The intuition captured by the FV is to compute the gradient of a PHOC sample (bag of textual features) that shows the probability of belonging to each of the Gaussian components, which can be understood as a probabilistic textual vocabulary based on its morphological structure (see Figure 4.1).

4.2.3 Attention on features

In the proposed fine-grained classification task we can intuitively state that there will be some recognized text that is more relevant than others at the moment of discriminating similar classes. Therefore, it is important to capture the inner correlation between the textual and visual features. To adhere this idea into our pipeline, we propose a modified attention mechanism inspired from [223]. The attention mechanism learns a tensor of weights W that is used between the visual features and the obtained FV. The implemented attention is defined by:

$$W_a = \text{Softmax}(\tanh(V_{fa}^T \cdot W \cdot T_f)) \quad (4.7)$$

$$T_{fa} = W_a \cdot T_f \quad (4.8)$$

The resulting tensor W_a , contains a normalized attention vector that is multiplied with the textual features T_f to obtain the final attended textual features T_{fa} .

The obtained attended textual features T_{fa} and the visual features V_{fa} are concatenated, such that the final features are formed by $F = [V_{fa}, T_{fa}]$. Finally, the resulting vector serves as input to a final classification layer that outputs the probability of a given class. The proposed network is trained to optimize the cross entropy loss function

$$J(\theta) = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{\mathcal{C}} l_i^n \log(\hat{l}_i^n) \quad (4.9)$$

4.3 Experiments and Results

4.3.1 Datasets

Con-Text Dataset

Originally presented by [89], is a dataset taken from the ImageNet [41] "building" and "place of business" sub-categories. It consists of 28 categories with 24,255 images in total. The classes from this dataset are visually similar (Pizzeria, Restaurant, Dinner, Cafe) and require text to successfully perform fine-grained classification. The dataset was not built for text recognition purposes, thus not all images contain text in them. High variability of text size, location, and font styles make text recognition on this dataset a challenging task.

Drink Bottle Dataset

Dataset presented by [12] comprises the sub-categories soft drink and alcoholic drink found on ImageNet[41]. There are 18,488 images divided into 20 categories. The dataset contains several not common, occluded, rotated, low quality, and blurred text instances which increase the difficulty of performing successful text recognition.

4.3.2 Implementation Details

The visual features of the proposed model are taken by attending to the features of the output of the last block layer of the Resnet152 before the last average pooling layer. These features are passed through a fully connected layer to down-sample them to a final dimension of 1×1024 . To construct the textual features, a maximum number of $N_{max} = 15$ PHOC proposals are obtained per image. If a lesser number of PHOC proposals are obtained, a zero padding scheme is employed to fix the size of the input features. The resulting PHOCs are reduced in size through PCA, to obtain features of a dimensionality of $N_{max} \times 300$.

The Fisher Vector is calculated from the PCA-ed PHOCs by employing a pre-trained Gaussian Mixture Model as it is described in Section 4.2.2. The trained GMM employs 64 Gaussian components thus yielding an FV of 1×38400 dimension. The obtained textual features are down-sampled by passing them through a fully connected layer to finally obtain a resulting size of 1×512 before the attention mechanism is computed. The attention between both modalities produces an output vector of 1×512 , that multiplies the learned weights to the textual features. As the last step, a concatenated vector of the visual and textual features ($dim = 1 \times 1536$) is used to produce the final classification probability vector.

The network is trained for 30 epochs with the combination of RAdam [122] and the Lookahead [230] optimizers. The batch size employed in all our experiments is 64, with

a learning rate of 0.001, the momentum of 0.9 that decays by 0.1 every 10 epochs.

4.3.3 Classification Results

When comparing our method to the current state of the art, it is evident that the proposed pipeline consistently outperforms previous approaches. The performance of our method is shown in Table 4.1. As it can be seen, our method surpasses [12] in the Drink Bottle dataset by a significant margin, however, this margin is smaller in the Con-Text dataset. Nonetheless, it is important to note that the method presented by [12] employs two additional classifiers to solve this task, thus relying on an ensemble model. Such kind of adopted approaches require longer training times, as well as more computation resources since several deep networks need to be trained. Therefore, when compared to the single classifier presented by [12], our model offers a significant improvement. In the upcoming sections, we provide explanations and exhaustive experimentation that shows the main strengths and advantages of our model.

Method	Con-Text	Bottles
Karaoglu[89]	39.0	–
Karaoglu[87]	77.3	–
Bai[12]	78.9	–
Bai*[12]	79.6	72.8
Ours	80.2	77.4

Table 4.1: Classification performance for two state-of-the-art methods and our proposed model on the Con-Text and Bottles dataset. The results presented by [12] depicted with * are based on an ensemble model.

4.3.4 Importance of Textual Features

Several baselines of growing complexity were defined in order to: assess the effectiveness of the proposed model, discern the added performance of employing textual features along with visual ones, and verify the improvement obtained from using a fusion mechanism.

Visual Only: This baseline assesses the performance of the CNN encoder based on visual features solely. To this end, the 2048 dimensional output features V_f , serve as the input to a fully connected layer according to the number of classes of the evaluated dataset.

Textual Only: We evaluate the performance of two current state of the art text recognizers: Textspotter [74] and E2E_MLT [25] along with the most confident PHOCs obtained from the model presented by [62].

We evaluate the performance of two current state-of-the-art text recognizers: Textspotter [74] and E2E_MLT [25] along with the most confident PHOCs obtained from the model presented by [62]. For illustration purposes, Figure 4.3 shows heat maps ob-



Figure 4.3: Heat maps obtained according to the confidence detection score of the predicted PHOCs.

tained by employing the model from [62] according to the confidence scores obtained when a text instance is detected. It is important to note that Textspotter [74] is bound to a dictionary to output the final recognized word, whereas the multilingual model E2E_MLT from [25] is not. The recognized text is embedded with pretrained versions of GloVe [151], FastText [21] and Word2Vec [139], finally outputting tensors of size $N_{max} \times 300$, which in our experiments $N_{max} = 15$. When working with PHOCs, the output vector has a size $N_{max} \times 604$. As we can observe in Table 4.2, in the visual-only baseline, the ResNet152 CNN [73] performed better in this task, due to the major expressiveness of the model and the residual block architecture that it is based on.

	Fusion	T+W	T+G	T+F	E+W	E+G	E+F	PHOC	FV(F)	FV(P)
Con-Text	Concat	73.84	74.11	74.33	77.04	77.58	77.77	77.45	77.31	80.21 [†]
	Block [17]	73.12	73.86	73.18	76.97	78.34	78.34	77.96	77.87	79.27
	Mutan [16]	72.46	72.08	73.47	77.67	77.26	78.05	76.97	76.01	78.51
	MLB [97]	73.17	72.18	74.09	77.45	76.28	78.81	76.96	76.46	78.49
	MFH [225]	73.62	73.23	74.42	77.68	76.79	78.55	77.56	76.27	78.03
	MFH [226]	72.95	72.43	74.48	77.3	76.64	78.23	77.42	76.39	77.58
Drink Bottle	Concat	75.05	75.12	75.25	74.62	74.91	75.4	75.93	75.15	77.38 [†]
	Block [17]	75.18	75.31	75.39	74.17	74.87	74.94	75.91	75.11	76.23
	Mutan [16]	74.48	73.91	74.72	73.62	75.12	76.05	75.95	74.48	75.97
	MLB [97]	74.34	73.02	75.54	73.55	75.42	75.19	76.37	75.07	76.18
	MFH [225]	74.25	74.25	75.21	74.23	74.88	75.84	76.21	74.78	76.01
	MFH [226]	73.99	73.61	75.36	74.77	75.26	75.72	75.98	74.56	75.85

Table 4.3: Results obtained by employing different fusion strategies on both the Con-Text and Drink Bottle dataset. For presentation purposes, acronyms are used to represent each combination of text recognizers (Textspotter (T), E2E_MLT (E), PHOC (P)) and word embeddings (Word2Vec (W), GloVe (G), FastText (F), Fisher Vector (FV)). The [†] refers to the proposed model.

	Model	Con-Text	Bottles
Visual	GoogLenet	61.21	64.93
	Resnet-152	63.70	66.56
Textual	Textspotter+w2v	35.09	50.68
	Textspotter+glove	34.52	50.26
	Textspotter+fasttext	36.71	51.93
	E2E_MLT+w2v	44.36	43.98
	E2E_MLT+glove	44.25	42.64
	E2E_MLT+fasttext	45.07	44.31
	PHOC	49.18	52.39
	Fisher Vector (PHOC)	63.93	62.41

Table 4.2: Visual only and Textual only results. The textual-only results were performed on the subset of images that contained spotted text. The metric depicted is the mean Average Precision (mAP in %).

In the text-only baseline, by using standard text recognizers we can observe that the E2E_MLT performs better in the Con-Text dataset, whereas the Textspotter model surpasses E2E_MLT in the Drink Bottle dataset. Nonetheless, both of them are outperformed by employing the PHOCs obtained from [62] as the word embedding. This effect is due to the inherent morphological nature of the PHOC embedding.

Overall, the best results in the textual-only baseline are obtained by the Fisher Vector obtained from the PHOCs. Qualitatively shown in Figure 4.1, the Gaussian Mixture gracefully captures the morphology of words obtained from PHOCs. Therefore, words with similar syntax are clustered together in the GMM, thus allowing the Fisher Vector

to be a powerful descriptor relevant for this task that yields even more discriminative features than other embeddings. It is important to note as well that in our experiments, FastText performs better than Word2Vec or GloVe because it can produce embeddings of out-of-vocabulary words while considering word n-grams which strengthens our conjecture on the importance of morphology of text to solve this task.

4.3.5 Comparison of Models

Extensive experiments were conducted regarding the different combinations of text recognizers, word embeddings, and fusion techniques. Table 4.3 show the results obtained in both the Con-Text and Drink Bottle datasets.

When introducing fusion techniques to the models, traditional text recognizers such as E2E_MLT performs better in Con-Text compared to Textspotter, thus achieving a higher mAP. The opposite effect is found in the Drink Bottle dataset, in which Textspotter behaves better than its E2E_MLT. It is interesting to note that the PHOCs obtained perform consistently in both datasets, yielding comparable results to the traditional recognizers employed. Regarding the embedding mechanism utilized, morphological embeddings (FastText, PHOC) work better than purely semantic embeddings due to the discriminative space learned.

We can observe that the usage of fusion techniques usually improves the mAP performance obtained on each method aside from the cases when the models employ Fisher Vector features. Nonetheless, in our experiments, we have not found a specific fusion technique that can be generalized for every tested method. Each fusion technique increases the performance for a specific model, being MFH and Block slightly more consistent than others. It is necessary to indicate that employing Fisher Vector features obtained from PHOCs consistently achieves the best performance in a general and consistent manner across both datasets.

In order to assess the efficacy of using the Fisher Vector along with another embedding that captures out-of-vocabulary words while at the same time considering the character morphology, we employ the Fisher Vector obtained from FastText. To this end, FastText employs character n-grams to construct a relevant vectorial representation of a word, thus it also uses the syntax of a detected word. The results of the conducted experiments using Fisher Vector features from FastText and PHOC are shown in the last two columns of Tables 4.3. There are two results to highlight obtained from this experiment. Firstly, working with PHOCs along FVs always yield better performance compared to Fasttext. The cause might be the information captured by Fasttext encapsulates morphology in the form of character n-grams, as well as semantics. Whereas the PHOC is a compact representation based solely on word morphology.

Secondly, by combining the explored fusion methods along with Fisher Vectors did not provide a significant advantage. A straightforward concatenation operation between the FV and the visual features reinforces the notion that both modalities contain discriminative and orthogonal features well suited for this task. As an additional ad-

vantage, by employing concatenation the model convergences faster while at the same time providing better performance.

4.3.6 Qualitative Results



Figure 4.4: Classification results. The top-3 probabilities of a given image assigned by the output of our model are shown along the Ground Truth. Notice that without reading, the classification task is impossible to perform even for humans. Blue and red are used to display correct and incorrect predictions respectively.

Fine-grained classification probabilities obtained from our model output are depicted in Figure 4.4. The textual features employed are able to generalize to unseen textual instances or named entities such as the case of bottle brands or business places. We can observe that our model has a hard time reading handwritten text or vertical textual occurrences, thus wrongly predicting a class, such as an example shown in the first row, seventh column. Nonetheless, the model seems to be capturing text morphology, as can be seen in the prediction of the class 'pawn shop'. Finally, on the last two samples on each row, there are not enough guiding textual features and the model relies only on similar visual features. Nonetheless, classifying these samples correctly is a hard task even for humans.

4.3.7 Fine-grained Image Retrieval

In the same manner, as in the work presented in [87] and [12], we conduct a retrieval experiment by utilizing the computed vector of the last output layer of the proposed model as retrieval features.

Method	Con-Text	Drink Bottle
Bai*[12]	62.87	60.80
Ours	64.52	62.91

Table 4.4: Retrieval results on the evaluated datasets. The results on Con-Text are based on our implementation of the method by [12] since there is no publicly available code. The retrieval scores are depicted in terms of the mAP(%).

We take the approach of query by example, that is, given a sample image that belongs to a specific class, the system must return a ranked list of similar classes as the query. The metric employed to conduct this experiment is the cosine similarity. The proposed method is more robust at the moment of employing a combination of visual and textual features which are discriminative enough to conduct a different task successfully as is the case in fine-grained image retrieval.

4.4 Conclusions and Future Work

In this Chapter, we have presented a deep neural network framework suitable for a fine-grained classification task. Through extensive experiments conducted, we have presented that leveraging textual information is a key approach to extracting information from images. Exploiting these textual cues can pave the road towards more holistic computer vision models of scene understanding. We have shown that current text recognizers that are limited by a dictionary are not the best alternative for this task, because it requires a recognizer able to generalize out of vocabulary words from unseen samples. Additionally, we have analyzed the fact that using semantic embeddings in a fine-grained classification task does not produce the best results due to the related semantic space shared across similar classes. By integrating state-of-the-art techniques and constructing a powerful morphological descriptor from the text contained in images, we show that a better-suited feature for this task can be learned. Such a feature proves to be useful for a fine-grained classification task as well as for query-by-example image retrieval. Leveraging this robust textual feature yields state-of-the-art results in both tasks across the assessed datasets. Successful classification and retrieval are possible due to the discriminative features learned by the model. In future work, we plan to develop a descriptor that captures the same discriminative features using a smaller feature dimension. A continuous valued embedding can replace the binary PHOC while preserving the generalization ability of unseen samples. We want to ex-

Explore the usefulness of this embedding in other computer vision tasks such as visual question answering [18, 180] and text-based image retrieval.

Chapter 5

Multi-Modal Reasoning Graph for Fine-Grained Image Classification

Scene text instances found in natural images carry explicit semantic information that can provide important cues to solve a wide array of computer vision problems. In this Chapter, we focus on leveraging multi-modal content in the form of visual and textual cues to tackle the task of fine-grained image classification and retrieval. First, we obtain the text instances from images by employing a text reading system. Then, we combine textual features with salient image regions to exploit the complementary information carried by the two sources. Specifically, we employ a Graph Convolutional Network to perform multi-modal reasoning and obtain relationship-enhanced features by learning a common semantic space between salient objects and text found in an image. By obtaining an enhanced set of visual and textual features, the proposed model greatly outperforms previous state-of-the-art in two different tasks, fine-grained classification and image retrieval.

5.1 Introduction

In this Chapter we propose a method to learn a richer set of visual features and model a more discriminative semantic space by employing a Graph Convolutional Network (GCN). To the best of our knowledge, this is the first approach that integrates multi-modal sources that come in the form of visual along with textual features jointly with positional encoding into a GCN pipeline that performs reasoning. We explore the role of such multi-modal cues, specifically in the form of visual and textual features.

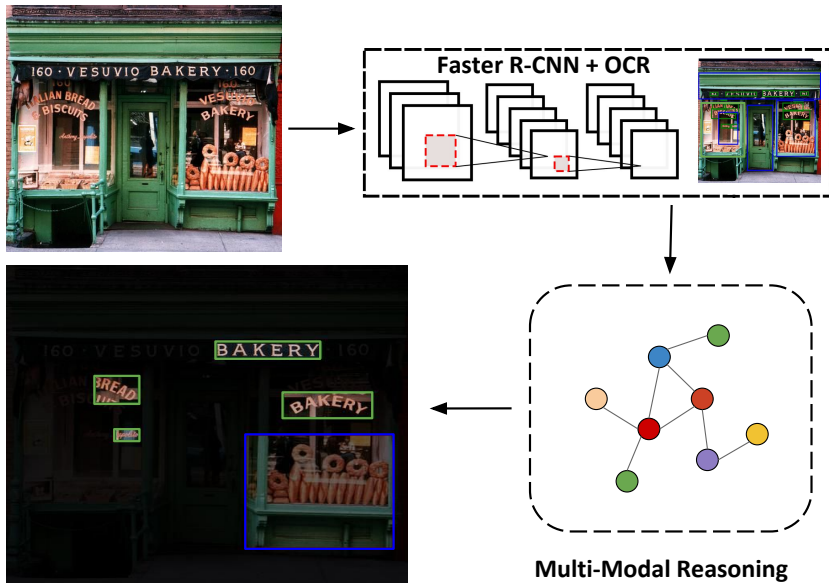


Figure 5.1: The proposed model uses a Graph-based Multi-Modal Reasoning (MMR) module to enrich location-based visual and textual features in a combined semantic representation. The network learns at the output of the MMR to map strong complementary regions of visual (blue) and text (green) instances to obtain discriminative features to perform fine-grained image classification and retrieval.

Departing from previous approaches, we exploit a structural representation between the studied modalities. Our work, summarized in Figure 5.1 with publicly available code at ¹, focuses on learning an enhanced visual representation that incorporates reasoning between salient regions of an image and scene text to construct a semantic space over which fine-grained classification is performed. In this example, we can observe that relevant regions such as the text "Bakery" and "Bread" are associated with a visual region that depicts pastry, both important cues to classify the given image. Additionally, we show experiments of fine-grained image retrieval, using the same multi-modal representation, in the two evaluated datasets. Overall, the main contributions can be summarized as follows:

- We propose a novel architecture that greatly surpasses previous state-of-the-art results in two datasets by more than 5% on fine-grained classification and 10% on image retrieval by considering text and visual features of an image.
- We design a fully end-to-end trainable pipeline that incorporates a Multi-Modal Reasoning module that combines textual and visual features that do not rely on ensemble models or pre-computed features.

¹https://github.com/AndresPMD/GCN_classification

- We provide exhaustive experiments in which we analyze the effectiveness of different modules in our model architecture and the importance of scene text towards comprehensive models of image understanding.

5.2 Multimodal Reasoning Model

In this section, we detail each of the components that comprise the proposed architecture. Figure 5.2 depicts the overall scheme of the proposed model, which is formed by 6 different modules: global image encoder, local feature encoder, text encoder, positional encoder, multi-modal reasoning graph and classification module. The local feature encoder employs features extracted based on the regions of interest obtained by a Faster R-CNN [161] in a similar manner as the bottom-up attention model [8]. The scene text encoder uses an OCR model to obtain scene text and further embed it into a common space. The goal is to obtain multi-modal node representations that leverage the semantic relationships found between salient objects and text instances within an image that are discriminative enough to perform fine-grained classification.

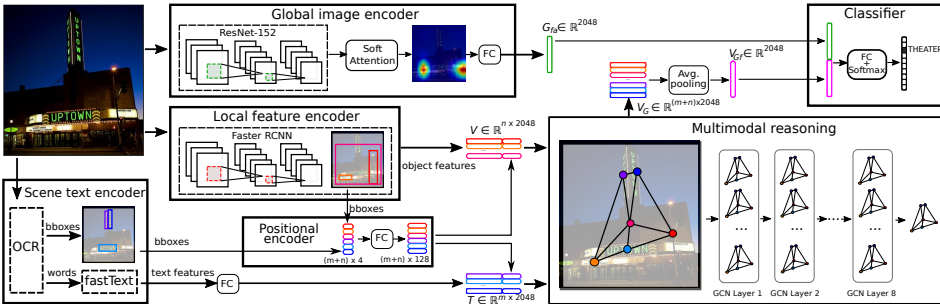


Figure 5.2: Detailed model architecture. The proposed model combines features of regions of scene text and visual salient objects by employing a graph-based Multi-Modal Reasoning (MMR) module. The MMR module enhances semantic relations between the visual regions and uses the enriched nodes along with features from the Global Encoder to obtain a set of discriminatory signals for fine-grained classification and retrieval.

5.2.1 Global Image Encoder

We employ a CNN as an encoder, which in our case is a ResNet-152 [73] pre-trained on ImageNet [41] to acquire global image features. Particularly, given an image I we take the output features before the last average pooling layer, which output is denoted as $G_f = \psi(I)$. In order to obtain a more descriptive set of global features and due to its differentiable properties, we compute a soft attention mechanism on top of the

global features. This self-attention mechanism yields an attention mask, $attn_{mask}$, that assigns weights on different regions of the input image. The attention weights are learned in an end-to-end manner by convolving 1×1 kernels projected into a single-dimensional filter and later followed by a Softmax function. In order to obtain the final attended global features, the attention mask is broadcasted and multiplied with the global features, which result is added to the global features G_f to later be used as input of a Fully-Connected layer, FC , in the form of:

$$G_{fa} = FC(G_f + (G_f \times attn_{mask})) \quad (5.1)$$

where $G_{fa} \in \mathbb{R}^{1 \times D}$, $G_f \in \mathbb{R}^{H \times W \times D}$, $attn_{mask} \in \mathbb{R}^{H \times W}$ stands for the final encoded global features, where $D = 2048$, $H = 7$ and $W = 7$.

5.2.2 Local Feature Encoder

Following [8], we employ a Faster R-CNN [161] pre-trained on Visual Genome [103] as the extractor of local visual features. This approach allows us to obtain salient image regions that are potentially discriminative for our task. We use an IoU threshold of 0.7 and a confidence threshold of 0.3, and sort the obtained predictions before the last average pooling layer to use the top n most confident regions of interest. Thus, we can represent the output of an image I with a set of region features $R_f = \{(r_1, bbox_{r_1}), \dots, (r_n, bbox_{r_n})\}$, $r_i \in \mathbb{R}^d$, where r_i is the i^{th} region of interest and $bbox_{r_i}$ is the r_i 's corresponding bounding box coordinates normalized with respect to the image. In our experiments, we set $n = 36$ and the obtained features have a dimension of $d = 2048$. In order to encode the local visual features, we project the features through a fully-connected layer.

In this manner, we obtain the final encoded local features that will serve as input to the multi-modal GCN in the form of $V_f = \{v_1, \dots, v_n\}$, $v_i \in \mathbb{R}^D$, where $D = 1920$ is the dimension of the final embedding space. We use $D = 1920$ to further add positional encoding information $D = 128$ to have a final feature representation of $D = 2048$. The bounding boxes obtained to represent these regions are later used as input into the positional encoder module. If there are less than $n = 36$ regions in an image, a zero padding scheme is adopted.

5.2.3 Text Encoder

To extract text contained in an image, we ran several public state of the art text recognizers as well as a commercial OCR model provided by Google². We extract the transcriptions of each word, denoted as w_i , as well as the corresponding bounding boxes, $bbox_{w_i}$. In particular, we extract the top m most confident textual instances found in an image. The transcriptions are embedded using fastText [21] and the bounding

²<https://cloud.google.com/vision/>

boxes will be used as input in the positional encoder branch. We employ the fastText embedding due to its capability of encoding word morphology in the form of n-grams as well as preserving a semantic space similar to Word2Vec [139] while at the same time dealing with out of vocabulary words. Analogously to the case of local features, we project the obtained embedded textual features by passing them through a fully-connected layer. The final textual features are represented by $T_f = \{t_1, \dots, t_m\}$, $t_i \in \mathbb{R}^D$, where $D = 1920$ is the dimension of the final embedding space and $m = 15$ is the number of text proposals extracted from an image. In the case that there is no text found in a given image, similarly to the local encoder module, zero padding is employed.

5.2.4 Positional Encoder

Encoding the position of objects and text instances within an image can provide important relational information about the scene. For example text found on top of a building often refers to its class in an explicit manner contrary to text found in any other location in the image. To meet this end, we design a positional encoding that takes as input a predicted bounding box of an object or text instance. The input to the positional encoder describes the top left (x_1, y_1) , and bottom right (x_2, y_2) coordinates normalized according to the image size, and is a concatenation of the bounding boxes of the local and text regions of interest. The bbox matrix is given by: $bboxes_{input} = \{bbox_{r_1}, \dots, bbox_{r_n}, bbox_{t_1}, \dots, bbox_{t_m}\}$ where $bbox_i = (x_1, y_1, x_2, y_2)$. In order to encode them, we pass the bounding boxes over a fully-connected in a similar way as the same as previous sections. The final encoded representation can be described as: $bboxes = \{bbox_{r_1}, \dots, bbox_{r_n}, bbox_{t_1}, \dots, bbox_{t_m}\}, bbox_i \in \mathbb{R}^b$, in which the dimension $b = 128$ represents the final encoded bounding boxes.

5.2.5 Multi-modal Reasoning Graph

Due to the showcased capability of graphs to describe reasoning between objects [178, 216, 54, 120], we construct a richer set of region-based visual descriptors that exploit the semantic correlation between visual and textual features. In order to do so, we initialize the node features as local visual features and textual features concatenated with their respective positional encoding of bounding boxes. We can describe the node features as:

$$V = \{(v_1, bbox_{r_1}), \dots, (v_n, bbox_{r_n}), (t_1, bbox_{t_1}), \dots, (t_m, bbox_{t_m})\}, V \in \mathbb{R}^{(n+m) \times D}$$

where n, m is the number of visual and textual features, respectively. In our case, $n + m = 51$ and $D = 1920 + 128 = 2048$. Furthermore, we construct the affinity matrix R which measures the degree of correlation of between two visual regions. The

construction of the affinity matrix is given by:

$$R_{ij} = \phi(k_i)^T \gamma(k_j) \quad (5.2)$$

where $k_i, k_j \in V$, $\phi(\cdot)$ and $\gamma(\cdot)$ are two fully connected layers that are learned end-to-end by back propagation at training time. If we define $k = n + m$, then the obtained affinity matrix consists of a shape $k \times k$. Once R is calculated, we can define our graph by $G = (V, R)$, in which the nodes are represented by the local and textual features V , and the edges are described by R . The obtained graph describes through the affinity matrix R the degree of semantic and spatial correlation between two nodes. We use the formulation of Graph Convolutional Networks given by [100] to obtain reasoning over the nodes and edges. Particularly, we use residual connections in the GCN formulation as it is presented by [111]. We can write the equation that describes a single Graph Convolution layer performed as:

$$V_g^l = W_r^l (R^l V^{l-1} W_g^l) + V^{l-1} \quad (5.3)$$

where $R \in \mathbb{R}^{k \times k}$ is the affinity matrix, $V \in \mathbb{R}^{k \times D}$ the local visual features, $W_g \in \mathbb{R}^{D \times D}$ is a learnable weights matrix of the GCN, $W_r \in \mathbb{R}^{k \times k}$ corresponds to the residual weights matrix and l is the number of GCN layer. Notice that passing V through the GCN layer, a richer set of multi-modal features is obtained. In order to find an enhanced representation of the visual features we apply $l = 8$ GCN layers in total, which finally yields a set of enriched nodes that represent the visual features V_G such that:

$$V_G = \{v_{g1}, \dots, v_{gk}\}, V_G \in \mathbb{R}^{k \times D}$$

5.2.6 Classification

In order to combine the global G_{fa} and the enriched local and textual V_G visual features, firstly we perform an average pooling of the V_G tensor. Specifically, we can rewrite the final local feature vector V_{Gf} as:

$$V_{Gf} = \frac{1}{k} \sum_{n=1}^k V_{gi} \quad (5.4)$$

Lastly, we simply concatenate the two obtained vectors V_{Gf} and G_{fa} , to obtain the final vector F that is used as input for the final fully-connected layer for classification denoted by: $F = [G_{fa}, V_{Gf}]$

By applying a softmax to the output of the final layer, we obtain a probability distribution of a class label given an input image. The model is trained in an end-to-end

fashion optimized with the cross entropy loss function described by:

$$J(\theta) = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^{\mathcal{C}} y_i^n \log(p_i^n) \quad (5.5)$$

Where, C is the number of classes, N the dataset samples such that each pair contains an annotation $\{x^{(n)}, y^{(n)}\}$ $n = 1, 2, \dots, N$, and p^n is the predicted output label.

5.3 Experiments and Results

This section presents an introduction to the datasets employed in this Chapter, as well as the implementation details, ablation studies performed, and a thorough analysis of the results obtained in the experiments conducted.

5.3.1 Datasets

The *Con-Text* dataset was introduced by Karaoglu *et al.* [89] and is a subset of ImageNet [41], constructed by selecting the sub-categories of "building" and "place of business". This dataset contains 24,255 images in total divided into three-folds to divide training and testing sets. This dataset introduces 28 visually similar categories of images such as Cafe, Pizzeria, and Pharmacy in which in order to perform fine-grained classification, text is a necessary cue to solve otherwise a very difficult task even for humans. This dataset closely resembles natural circumstances due to the fact that the images are taken without considering scene text instances, thus some images do not have text present in them.

The *Drink Bottle* dataset was presented by Bai *et al.* [12] and as the Con-Text dataset, it is a subset of images of ImageNet [41], specifically taken from the sub-categories of soft drink and alcoholic drink. The dataset is divided in three-folds as well and contains 18,488 images. There are 20 image categories which include visually similar instances such as Coca Cola, Pepsi Cola and Cream Soda. Akin to the Con-Text dataset, some images contain scene-text while others do not have it.

5.3.2 Implementation Details

In our experiments in order to extract visual regions of an image, we use the same settings as [8]. We take the top $n = 36$ ROIs and encode them along with their bounding boxes into a common space of 2048-d. The transcribed text is sorted by confidence score and we take the top $m = 15$ confident predictions. We embed the textual instances by using a pre-trained fastText model with 1 million 300_d word vectors, trained with sub-word information on Wikipedia2017, UMBC webbase corpus and

statmt.org news dataset. The obtained 300-d textual vectors are projected with the corresponding bounding boxes into a 2048-d space. The Faster R-CNN [161] from [8] and the OCR models, both employed as initial feature extractor modules use pre-trained weights and are not updated at training stage. The rest of the weights of each module in the model are learned in an end-to-end manner during training. The graph-based multimodal reasoning module employs 8 multi-modal GCN layers to obtain the final enriched visual features. In the last full-connected layer before classification, we employ a dropout rate of 0.3 to avoid over-fitting on the evaluated datasets. In general, we employ Leaky ReLU as an activation function in all layers except the last one, in which we use a Softmax to compute the class label probabilities. The proposed model is trained for 45 epochs, but an early stop condition is employed. We use a combination of optimizers comprised by RAdam [122] and Lookahead [230]. The batch size employed in all our experiments is 64, with a starting learning rate of 0.001 that decays by a factor of 0.1 on the epochs 15, 30 and 45. The momentum value used on the optimizers is 0.9 and the weight decay is 0.0005.

5.3.3 Comparison with the State-of-the-Art

We show the experimental results of our method compared to previous state-of-the-art on Table 5.1. We can note that the performance obtained in the Con-Text significantly surpasses the previous best performing method by 5.9%. The improvement in the Drink-Bottle dataset is more modest, of about 1.98%, however it is still significant.

Method	OCR	Emb.	Con-Text Bottles	
Karao.[89]	Custom	BoB ¹	39.0	–
Karao.[87]	Jaderberg	Probs ²	77.3	–
Bai[12]	Textboxes	GloVe	78.9	–
Bai[12] [†]	Textboxes	GloVe	79.6	72.8
Bai[12] [†]	Google OCR	GloVe	80.5	74.5
Mafla[131]	SSTR-PHOC	FV	80.2	77.4
Proposed	E2E-MLT	fastText	82.36	78.14
Proposed	SSTR-PHOC	PHOC	82.77	78.27
Proposed	SSTR-PHOC	FV	83.15	77.86
Ours	Google OCR	fastText	85.81	79.87

Table 5.1: Classification performance of state-of-the art methods on the Con-Text and Drink-Bottle datasets. The results depicted with [†] are based on an ensemble model. The embeddings labeled as ¹ refer to a Bag of Bigrams, and ² is a probability vector along a dictionary. The acronym FV stands for Fisher Vector. The metric depicted is the mean Average Precision (mAP in %).

We believe the improvement is greater in Con-Text due to the text instances found in it, which refer mostly to business places without much out of vocabulary words, therefore a semantic space for classification is more discriminative when compared to

Method	Con-Text		Drink Bottle	
	I + T	I - T	I + T	I - T
Bai [12]	78.92	71.63	71.61	62.25
Mafla [131]	80.94	72.59	78.57	68.97
Ours	86.76	74.31	82.75	69.19

Table 5.2: Classification performance of the proposed method on the subset of images from the test set of the Con-Text and Drink-Bottle datasets such that the images: contain scene-text (I + T) and do not contain scene-text (I - T). The metric depicted is the mean Average Precision (mAP in %).

the Drink-Bottle dataset. To provide further insights, we conducted experiments by employing the final model along with different OCRs and word embeddings in both datasets. It is essential to note that state-of-the-art results are achieved by the usage of other OCRs as well, showing that the proposed pipeline still outperforms previous methods.

When comparing to previous methods, it is worth revisiting previous approaches. The results reported by [12] used an ensemble of classifiers to reach the obtained performance. An additional experiment to showcase the effect of using the same OCR as our proposed model is included, and it shows that our model vastly outperforms the evaluated pipeline not because of the OCR system employed. On the other side, the work presented in the previous Chapter by [131] requires offline pre-computation of the Fisher Vector by training a Gaussian Mixture Model and tuning the hyper-parameters involved. In this manner, the method proposed in this model does not require an ensemble, and the features used are learned in an end-to-end manner at training time. We clearly show that the proposed pipeline surpasses other approaches even when employing a set of different scene-text OCRs.

With the aim of offering additional insights, we present in Table 5.2 the performance of the previous state-of-the-art methods compared with our proposed method in a subset of the test set such that the evaluated images either contain scene-text or not. The results show the average performance along with 3 different splits of each dataset. We can observe that our model is able to perform better than previous approaches in both scenarios while a more significant improvement is achieved in images that contain scene text, which we treat as the major discriminative feature to perform the task of fine-grained classification.

5.3.4 Importance of Textual Features

In order to assess the importance of the scene text found in images, we follow the previous works [87, 12, 131] by defining two different evaluation baselines, the visual features based and the textual features based. Moreover, due to the fact that the evaluated datasets do not contain text transcriptions as ground truth, we evaluated the effective-

	Model	Con-Text	Bottles
Visual	CNN	62.11	65.15
	CNN + Self Attention	63.78	66.62
Textual	Texspotter+w2v [†]	35.09	50.68
	Texspotter+glove [†]	34.52	50.26
	Texspotter+fastText [†]	36.71	51.93
	E2E_MLT+w2v [†]	44.36	43.98
	E2E_MLT+glove [†]	44.25	42.64
	E2E_MLT+fastText [†]	45.07	44.31
	FOTS+w2v	43.22	41.33
	FOTS+glove	43.71	41.85
	FOTS+fastText	44.19	42.69
	Google OCR+w2v	53.87	53.47
	Google OCR+glove	54.48	54.39
	Google OCR+fastText	55.61	55.16
	PHOC [†]	49.18	52.39
Fisher Vector (PHOC)[†]	63.93	62.41	

Table 5.3: Visual only and Textual only results. The textual-only results were performed on the subset of images that contained spotted text. The results with [†] were reported by [131]. The metric depicted is the mean Average Precision (mAP in %).

ness of the OCR employed in the fine-grained classification task.

The visual only evaluates all the test set images by only employing the global encoder features G_f in the first scenario and the global encoder along with the self-attention features G_{fa} in the second scenario. In both cases, the output of the global encoder, a 2048-d feature vector, is directly passed through a fully connected layer to obtain the final classification prediction. In the textual only, the baselines are evaluated only in the subset of images that contained spotted scene text. The results of each baseline by employing visual only, different OCRs, and word embeddings are shown in Table 5.3.

Following a previous approach [131], we employ $m = 15$ text instances and pre-trained word embeddings that yield 300-d vectors in the case of Word2Vec [139], GloVe [151] and fastText [21]. The textual tensor obtained is used as input to a fully connected layer, which output is used for classification purposes. In our experiments, we evaluate two additional state-of-the-art scene text recognizers, FOTS [124] and the commercially used Google OCR Cloud Vision based on an API. We note that the embedding that performs the best is fastText due to the capability of embedding out of vocabulary words by using character n-grams. Regarding the results, it was found that the best performing standard recognizer is the Google OCR, which employs a more compact (300-d) vector compared to a PHOC or a Fisher Vector. The PHOC embedding employs a 604-d feature vector along with $m = 15$ and the Fisher Vector is a single 38400-d vector in our experiments. Overall, by using only textual features, the Fisher Vector based on PHOCs remains the best-performing descriptor. However, besides the high dimensional vec-

Features	Con-Text	Drink Bottle
G_f	62.11	65.15
G_{fa}	63.78	66.62
without MMR		
$G_{fa} + V_f$	70.48	73.21
$G_{fa} + V_f + T_f$	78.72	76.43
$G_{fa} + V_f + T_f + bboxes$	80.12	77.51
with MMR		
$G_{fa} + V_f$	72.88	74.96
$G_{fa} + V_f + T_f$	82.51	77.46
$V_f + T_f + bboxes$	84.33	75.42
$G_{fa} + V_f + T_f + bboxes$	85.81	79.87

Table 5.4: Quantitative results of the different components that form the proposed model. G_f : Global features, G_{fa} : G_f + Self-Attention, V_f : Local Features, T_f : Text Features, $bboxes$: Bounding Box information used by the Positional Encoder, MMR: Multi-modal Reasoning. Results are shown in terms of the mAP(%).

tor employed, extensive offline pre-computation is required to obtain such a descriptor. Nonetheless, as it can be seen in Table 5.1, the FV descriptor does not achieve the best results in our final model.

5.3.5 Ablation studies

In this section, we show the incremental improvements and the effects obtained by the addition of each module that comprises the final architecture in the method proposed. Table 5.4 shows the quantitative results of adding components in the baseline model. Namely, we evaluate the effect of using self-attention and the multi-modal reasoning (MMR) module. We successively add to the attended global features (G_{fa}), local features (V_f), textual features (T_f), and the bounding boxes ($bboxes$) of both used in the Positional Encoder. In order to assess the effectiveness of the multi-modal reasoning graph module, we compare a model that uses the Faster R-CNN ROIs without the usage of the MMR. It is observed that solely by using the Faster R-CNN features, an important boost is achieved. One of the biggest improvements is reached by the usage of scene text, which enforces the idea that textual information is essential to successfully discriminate between visually similar classes. By the incorporation of scene text, an improvement of 9.7% is gained in Con-Text and 2.5% in the Drink-Bottle datasets. Nonetheless, the improvement is accentuated by the usage of the MMR module, which produces as output richer local and textual features coming from the graph nodes. Finally, by adding the positional encoder module into the MMR, another increase in the results is achieved. This encourages us to think that the MMR module learns relationships coming from semantic and spatial information. Insights into the attention masks learned and the reasoning coming from the MMR by using visual and textual regions

Projection	Fusion	Con-Text	Drink Bottle
Attention	MLB [97]	80.83	78.26
Attention	Block [17]	80.82	78.42
Attention	Concat	81.09	78.45
GRU	MLB [97]	83.12	78.21
GRU	Block [17]	83.8	78.74
GRU	Concat	83.93	78.89
Avg Pooling	MLB [97]	84.23	78.56
Avg Pooling	Block [17]	85.11	79.15
Avg Pooling	Concat	85.81	79.87

Table 5.5: Results obtained by employing different Projection and Fusion strategies on all the modules of our pipeline. Results are shown in terms of the mAP(%).

can be found in the Supplementary Material section.

Furthermore, we explore in our work several projections and fusion methods which are shown in Table 5.5. In our experiments, Projection refers to the strategy used to reduce the dimensionality of the output tensor coming from the MMR as V_G to obtain a single vector V_{Gf} . Late Fusion showcases the method employed to combine the features coming from V_{Gf} and G_{fa} . Due to several works showing performance gains by the usage of attention [223, 214] and Recurrent Neural Networks [111, 27] as reasoning modules, we explored those alternatives, however, no improvements were found. In the same manner, as it is presented by [131], we explored two additional fusion mechanisms, MLB [97] and Block [17] but no gains were obtained compared to feature concatenation.

5.3.6 Qualitative Results

Qualitative results of the fine-grained image classification task are shown in Figure 5.3. By reviewing the samples obtained, we can note that our model is capable of learning a semantic space that combines successfully visual and textual signals coming from a single image. Classified samples such as “Pizzeria”, “Tea House” and “Diner” often contain similar semantic classes ranked in second and third positions. Images belonging to the Drink Bottle dataset on the second row, are correctly classified even though text instances belong to specific brands, thus showing the generalization capability of our method. The seventh image on the first row is wrongly classified as “Theatre” due to OCR recognition errors and a lack of strong enough visual cues. The remaining wrongly classified images are very challenging and contain some degree of ambiguity even for humans.

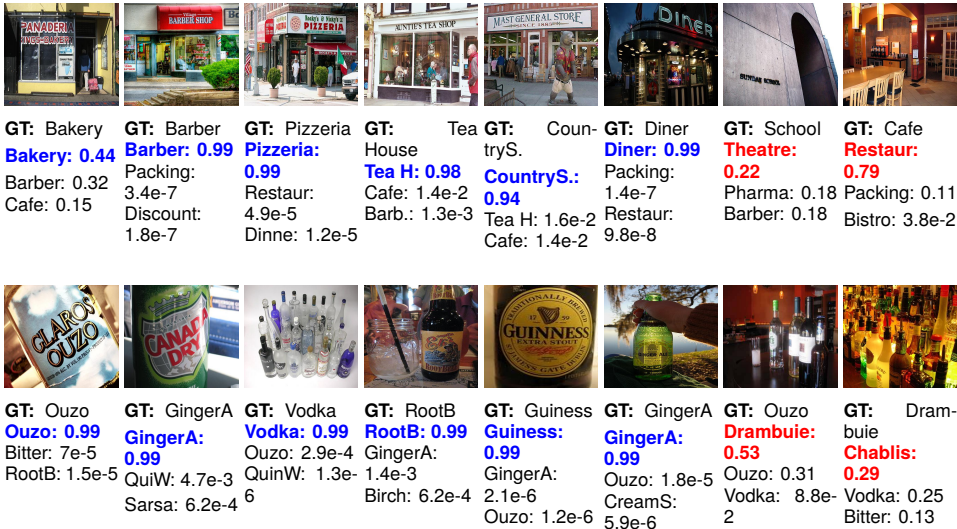


Figure 5.3: Classification predictions. The top-3 probabilities of a class are shown as well as the Ground Truth label performed on the test set. Without recognizing textual instances some images are extremely hard to classify even for humans. Text in blue and red is used to show correct and incorrect predictions respectively. Best viewed in color.

5.4 Fine-Grained Image Retrieval

As an additional experiment that highlights the capabilities of the proposed model, we show the results obtained in Table 5.6 by performing query-by-example (QbE) image retrieval. In QbE, a system must return images in the form of a ranked list that belongs to the same class as the image used as a query. To provide comparable results and follow the work from [12, 131], we use the vector of class probabilities as the image descriptor without using a specific metric-learning method. This vector is used to retrieve the nearest samples computed by the usage of the cosine similarity as a distance metric.

In our experiments, the query, as well as the database is formed by unseen samples at training time. The results demonstrate that a very significant boost of 10.98% and 2.48% in Con-Text and Drink-Bottle is achieved respectively. The lower gain in the Drink-Bottle dataset directly depends on the harder-to-recognize text instances, as well as the low image quality of several samples that directly affect the model performance.



Figure 5.4: Qualitative results in Con-Text Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. The red border represents a mistaken retrieved image that does not correspond to the queried class. (Best viewed in color).



Figure 5.5: Qualitative results in Drink-Bottle Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. The red border represents a mistaken retrieved image that does not correspond to the queried class. (Best viewed in color).

Method	Con-Text Drink Bottle	
Bai[12]	62.87	60.80
Mafla[131]	64.52	62.91
Proposed	75.50	65.39

Table 5.6: Retrieval results on the evaluated datasets. The retrieval scores are depicted in terms of the mAP(%).

5.4.1 Qualitative Retrieval Results

Qualitative results that show the robustness of the model, as well as experiments addressing the importance of text, are shown in this section. We present qualitative results based on the Query by Example (QbE) task on Figures 5.4 and 5.5 for the Con-Text and Drink-Bottle datasets respectively. In a QbE scenario, a system receives as an input a specific image, which belongs to a class seen in training time. The goal of the system is to retrieve a ranked list of the closest images that belong to the queried class. In order to measure the distance between the queried image and the retrieved samples, we employ the cosine similarity as it is described in the Fine-Grained Image Retrieval section. In all the Figures showcased, the first image (blue border) represents the image employed as a query. Images with a green border represent a correct retrieval, whereas images with a red border represent wrongly retrieved samples. We can observe successful retrieval results in both Figures except in the third row of Figure 5.4, in which the model fails to retrieve the first sample. This effect is due to an incorrect OCR recognition of the text from the queried image. Added this to the fact that the visual features closely resemble the wrongly retrieved sample, outcomes in a wrong prediction. The rest of the samples show that an appropriate space that clusters similar labels is learned by employing textual along with visual features.

5.4.2 Relevance of Textual Features

To further provide insights into the relevance of scene text as discriminative features, we perform qualitative experiments in two scenarios. In the first scenario, we blurry the text found in an area given by a text detector and preserve the remaining visual features. In the second scenario, we blurry all the non-textual regions in a queried image and preserve only the scene text. In both scenarios, we used the same queried images as in Figures 5.4 and 5.5. Depending on the images used as queries in the first scenario, blurring the text makes the retrieval task a very complicated problem to be solved even by humans. Figures 5.6 and 5.8 show the retrieved images for the Con-Text and Drink-Bottle datasets in the first scenario. It is worth noting the significant drop in the retrieved images in both Figures. Specifically, in the first row in Figure 5.6, we observe that the model learns to recognize pastry in the storefront, which results in some correct retrieved samples. On the third row, the effect is similar to the one based on visual cues alone. The remaining rows contain all wrongly retrieved images.



Figure 5.6: Qualitative results in Con-Text Dataset when the text in the queried image is blurred. (Best viewed in color).

A similar effect is found in Figure 5.8, which produces correct retrieved images in the first and fourth row due to the shared visual features between samples but incorrect retrievals at the second and third rows. Figures 5.7 and 5.9 depicts the outcomes of the second mentioned scenario. By using textual regions only, we can obtain better results than in the first scenario, strongly suggesting that there are several cases in images that contain scene text, in which textual features can be more discriminative than visual ones. Nonetheless, in the third row in Figure 5.7, wrong OCR recognition produces erroneous retrieval of samples. The effect is similar in Figure 5.9 on the fourth row in the case of the Drink-Bottle Dataset.



Figure 5.7: Qualitative results in Con-Text Dataset. Results are obtained when everything but the text is blurred in a queried image. (Best viewed in color).



Figure 5.8: Qualitative results in the Drink-Bottle Dataset when the text in the queried image is blurred. (Best viewed in color).

5.4.3 Visualizing Reasoning

To offer an understanding of the effect that the learned attention maps and the MMR module have on the predictions of the model, we show in Figure 5.10 and Figure 5.11 the original images, attention maps, and affinity visualizations of the Con-Text and Drink-Bottle dataset respectively. The attention map is simply a self-learned mask by the CNN over a 7×7 grid. The affinity visualizations are defined by selecting the high-



Figure 5.9: Qualitative results in the Drink-Bottle Dataset. Results are obtained when everything but the text is blurred in a queried image. (Best viewed in color).

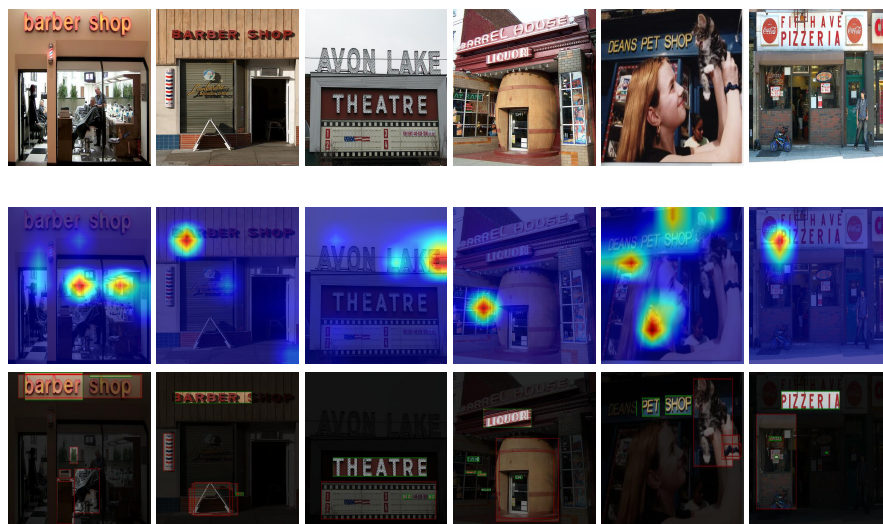


Figure 5.10: Visualization of the learned attention and enriched nodes of the model in the Con-Text dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).

est regions that present semantic correlation in the affinity matrix R in the last layer of the Graph-based MMR module. It is interesting to note that in Figure 5.10, third row, a strong semantic correlation is learned by attending to text regions and visual regions.



Figure 5.11: Visualization of the learned attention and enriched nodes of the model in the Drink-Bottle dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).

For example, in the first two images, the text "Barber" is highly correlated to people cutting hair and to the red, white, and blue barber pole. This effect is also evident in the text region that contains the text "Liquor" and the barrel located at the storefront. In the fifth sample, the text "pet shop" contains a strong semantic relation with a cat shown in the image. In the Drink-Bottle dataset, due to its noisy and low-quality nature, the local visual regions extracted do not contain as rich features as the ones in the Con-Text dataset. The model learns to attend to textual regions as well as the more salient visual regions that generalize to a specific class of image.

5.5 Conclusions and Future Work

In this paper, we have presented a simple end-to-end model that employs a Multi-Modal Reasoning graph to encounter semantic and positional relationships between text and salient visual regions. The learned space is composed of enriched features obtained from nodes in a graph, a module that acts as an appropriate reasoning scheme. Exhaustive experiments in two datasets and two different tasks validate the robustness of the presented model which achieves state-of-the-art results by a significant margin over previous methods. Moreover, our end-to-end pipeline does not require pre-computed handcrafted features or a collection of ensemble models as in earlier works. In the future, we expect to explore the effectiveness of this approach in other vision

and language-related tasks.

Part II

Image-Text Cross-Modal Retrieval

Chapter 6

Image-Text Cross-Modal Retrieval

6.1 Introduction

Language provides a medium to explain our perceptual experience while being discretely infinite. “Discrete infinity” is referred to as a property in which language is constructed by using a few discrete elements albeit giving an infinite variety of interpretations [35, 183]. In other words, the language’s discrete infinity property dictates that a potentially infinite number of semantically correct sentences can be used to express the same idea, for example, in describing an image. Framing the previous notion into consideration, we explore the task of Image-Text Matching (ITM) in a cross-modal retrieval scenario. Image-text matching refers to the problem of retrieving a ranked list of relevant representations of the query portrayed by a different modality. However, we identify mainly two unexplored areas in the ITM task. First, the incorporation of scene text information into a retrieval pipeline, and, secondly, the lack of semantic annotations to evaluate and train models in commonly used datasets.

As a starting point, the textual content is omnipresent in most man-made environments and plays a crucial role as it conveys key information to understand a visual scene. Scene text commonly appears in natural images, especially in urban scenarios, for which about half of the images habitually contain text [196]. This is especially relevant when considering vision and language tasks, and in particular, related to cross-modal retrieval. Scene text is rich, explicit, and semantic source of information that can be used to disambiguate the fine-grained semantics of a visual scene and can help to provide a suitable ranking for otherwise equally probable. Thus explicitly taking advantage of this third modality should be a natural step toward more efficient retrieval models. Nonetheless, and to the best of our knowledge, scene text has never been used for the task of cross-modal retrieval, and the community lacks a benchmark to properly address this research question. Chapter 7 addresses these two open directions. Chapter 8, tackles the lack of annotations in the ITM task. Current datasets assume

that only 5 sentences correctly describe a single image, labeling it in a binary manner as relevant or irrelevant. Consequently, the lack of many-to-many annotations causes a direct effect on the way the ITM task is evaluated. Sentences that are not relevant according to the ground truth can describe an image in various degrees of correctness and coverage, thus making the way we evaluate current models on ITM *incomplete*.

6.2 Related Work

6.2.1 Cross-modal retrieval

The task of ITM is a subset of cross-modal retrieval which aims to map the images and sentences in such a way that a suitable space for retrieval is learned, where the query and the search data come from distinct modalities. Initial approaches [51, 53] learned to align the global visual and textual features by applying a learned non-linear transformation to project both modalities into a common space. A similar pipeline is proposed by [146] with the incorporation of an attention mechanism. However, the main drawback of such approaches is that semantic concepts fall short of capturing fine-grained semantic interactions between visual regions and sentences. In the work presented by [8], several visual regions that describe an image in a more detailed manner are used for the task of Visual Question Answering (VQA) and Image Captioning. Initial works [149] incorporated visual regions along with a hierarchical Long Short-Term Memory (LSTM) [75] module. Following up, [109] proposed a stacked cross-attention network to model the latent alignments between image regions and words. Additional models have explored the role of attention mechanisms [119, 181, 209, 211, 231], and Graph Convolutional Neural Networks (GCN) [45, 100, 111, 120]. External modules have been explored to improve retrieval results such as the usage of an iterative recurrent attention module [28] and an external consensus knowledge base [203]. In order to use language as a visual altering feature, Vo *et al.* [198] proposes to use text modifiers along with images to retrieve relevant images.

More recently, Transformers [194] have been used to learn intra and inter-modality attention maps for a wide range of visual and language tasks [113, 84, 127, 126], often achieving state-of-the-art. However, these approaches require an additional order of magnitude of training samples, giving rise to a large increase in computational costs.

6.2.2 Scene Text in Vision and Language

Methods for vision and language tasks typically align both modalities and often perform visual reasoning. Several tasks that combine vision and natural language require developing models that are capable of scene understanding, visual reasoning, language semantics, and cross-modal alignment. Considerable efforts and improvements have been accomplished in a wide arrange of vision and language tasks. However, it is essential to take advantage of scene text due to the explicit semantic information it

contains in order to get closer to holistic computer vision systems capable of performing full scene reasoning. Works such as Text-VQA [180] and Scene-Text VQA [18] focus on models capable of reading text in the wild as well as reasoning about the inherent relations with visual features to properly answer a question given in natural language. Scene text has also been used to perform fine-grained image classification: [12, 88, 132] learn a shared semantic space between visual features and text to perform classification while [131] uses the Pyramidal Histogram Of Characters (PHOC) [5, 62, 134] descriptor as a way of overcoming OCR limitations and learning a morphological space suitable for the task. Other works [61, 142] perform scene-text-based image search, where we query with a word and retrieve images containing such query. Additionally, the TextCaps dataset [176] includes scene text in the task of generating textual descriptions of an image.

6.2.3 Image Captioning evaluation metrics

Image captioning is the task of transcribing an image into natural language. There are metrics proposed specifically for image captioning models, specifically, CIDEr [195] and SPICE [6]. These recently proposed metrics not only have been widely accepted to evaluate the captioning models but also they have been shown to correlate better with human judgments across different datasets [6, 195] when compared to machine translation metrics. Machine translation metrics use n-gram statistics to calculate precision or recall to evaluate the language models. For example, BLEU-n is a precision-based metric calculated according to n-gram match between the ground truth set (reference) and the generated samples (hypothesis). However, the limitation of BLEU is that it treats each n-gram as having the same weight. This results in stop words and nouns having the same importance on the resulting score. On the other hand, CIDEr solves this limitation by using tf-idf [86] to weigh the importance of each n-gram. Moreover, CIDEr employs cosine distance between tf-idf values of reference and hypothesis instead of using a direct match, which accounts for both precision and recall. SPICE is the first captioning metric that does not follow the classic n-gram matching. Instead, they run a rule-based version of the Stanford Scene Graph Parser [168] to obtain triplets consisting of object classes, relation types, and attribute types. These triplets are later used to build tuples of (c) , (c, a) , (c, a, r) . Finally, the SPICE score is an F-score calculated by direct matching between the tuples of reference and hypothesis. Despite their strengths, CIDEr and SPICE still have certain limitations regarding synonyms, word order, style, and sentence grammar among others [48, 96]. Aside from the limitations, they remain good automatic metrics to measure semantic similarity, especially when data comes from a similar distribution. Hence we employ the aforementioned metrics in a classical retrieval scenario.

6.2.4 Semantics and Metric Learning.

Initial work [145] highlights the main flaws of current metric learning approaches, which shows that metrics are not consistent for the task at hand. Also, it is shown that the gap between methods is less significant when evaluation is properly done. In this work, we refer to the problem of captions that can correctly describe an image that is not annotated in the GT, as semantic gap. Trying to overcome the existing semantic gap in current datasets, [79] employs a network to predict the semantic concepts from an image, however they rely to a binary annotation of relevance. Other works [63, 66] propose a model to learn a visual embedding where the image similarity is correlated with the paired text. Similarly [191] proposes a novel within-modality loss that leverages semantic coherency between text and images, which do not necessarily align with visually similar images. In order to address non-paired relevant images and captions, [229] proposes to build denotation graphs to link these two modalities. Trying to overcome the non-exhaustive annotation in datasets, [37] models the probability of an image belonging to a set of specific contexts. A newly introduced CrissCrossed [150] dataset, is an extension of MS-COCO that collects human judgments on the matching images and sentences. In parallel, [233] proposes the usage of a ladder loss among samples based on BERT [43] to define similarities. However, calculating BERT for every sample is very expensive computationally, thus they rely on a threshold given by a CBOW [200] to refine the comparison. Another drawback of this approach is that the similarity is computed among two captions alone, thus not all information available (5 captions) is leveraged.

Chapter 7

Scene-Text Aware Cross-Modal Retrieval

Recent models for cross-modal retrieval have benefited from an increasingly rich understanding of visual scenes, afforded by scene graphs and object interactions to mention a few. This has resulted in an improved matching between the visual representation of an image and the textual representation of its caption. Yet, current visual representations overlook a key aspect: the text appearing in images, which may contain crucial information for retrieval. In this Chapter, we first propose a new dataset that allows exploration of cross-modal retrieval where images contain scene-text instances. Then, armed with this dataset, we describe several approaches that leverage scene text, including a better scene-text aware cross-modal retrieval method that uses specialized representations for the text from the captions and text from the visual scene, and reconciles them in a common embedding space. Extensive experiments confirm that cross-modal retrieval approaches benefit from scene text and highlight interesting research questions worth exploring further.

7.1 Introduction

Scene text is a rich, explicit and semantic source of information which can be used to disambiguate the fine-grained semantics of a visual scene and can help to provide a suitable ranking for otherwise equally probable results (see example in Figure 7.1). Thus explicitly taking advantage of this third modality should be a natural step towards more efficient retrieval models. Nonetheless, and to the best of our knowledge, scene text has never been used for the task of cross-modal retrieval, and the community lacks

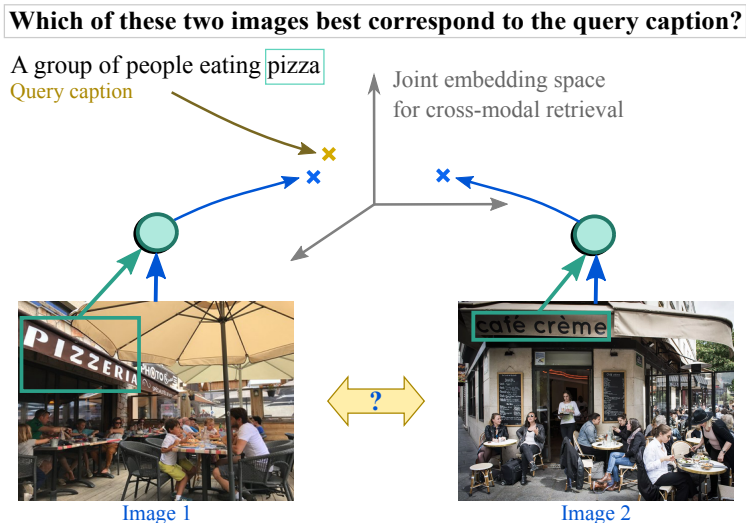


Figure 7.1: This Chapter introduces the scene-text aware cross-modal retrieval (StacMR) task and studies scene text as a third modality for cross-modal retrieval. For the example query above, the restaurant name provides crucial information to disambiguate two otherwise equally relevant results.

a benchmark to properly address this research question.

Scene text has been successfully leveraged to improve several semantics tasks in the past, such as fine-grained image classification [12, 88, 131, 144], visual question answering (VQA) [18, 180] or image captioning [176]. Current mainstream methods tackle cross-modal retrieval by either learning to project images and their captions into a joint embedding space [52, 101, 111, 203] or by directly comparing image regions and caption fragments to compute a similarity score [92, 109]. Although significant gaps have been overcome by previous methods, and the lack of integration between scene text and the other modalities still hinder fuller image comprehension. The intuition that serves as the foundation of this work stems from the notion that scene text, found in natural images, can be exploited to obtain stronger semantic relations between images and their captions. Obtaining such relations opens up the path toward improved retrieval systems in which scene text can serve as a guiding signal to provide more relevant and precise results.

This Chapter introduces the Scene-Text Aware Cross-Modal Retrieval (StacMR) a task that aims to capture the interplay between captions, scene text, and visual signals. To overcome the data scarcity of the proposed task, we have constructed a dataset based on COCO images [118] which we name COCO-Text Captioned (CTC). It exhibits unique characteristics compared to other datasets employed for multi-modal tasks and does not share their bias towards scene text as the main component present in

an image.

We also evaluate the performance of different state-of-the-art cross-modal retrieval models, and their limitations, and we propose distinctive baselines to solve this task. Concretely, the contribution of this Chapter is threefold. First, we introduce a new task called Scene-Text Aware Cross-Modal Retrieval (or StacMR in short), as an extension to cross-modal retrieval. In this task, leveraging the additional modality provided by scene text is crucial to further reduce the heterogeneity gap between images and captions. Second, we describe a new dataset, COCO-Text Captioned (CTC), as the first dataset properly equipped to evaluate the StacMR task. We highlight the importance of the role that incidental scene text plays when interpreting an image and its positive impact on retrieval results. We also compare the properties of our CTC dataset with similar existing datasets containing scene text and captions.

Finally, we provide an extensive analysis of CTC. In particular (1) we benchmark the combination of different cross-modal baselines to model the interaction between scene text, visual, and caption information, and (2) we propose and evaluate a new model, STARNet, which explicitly learns to combine visual and scene-text cues into a unified image-level representation.

7.2 The CTC Dataset

This section introduces the proposed COCO-Text Captioned (CTC) dataset. We first describe how it was gathered and tailored for the new StacMR task, which extends traditional cross-modal retrieval to leverage information from a third modality: *scene text*. Then we present CTC statistics and discuss the dataset in the light of other benchmarks and in particular the most related dataset: TextCaps [176] (Section 7.2.3).

7.2.1 Data Collection and Statistics

Building the Dataset. A suitable dataset for the proposed StacMR task requires the availability of these three modalities: *images*, *captions* and *scene text*. The most commonly used datasets for the cross-modal retrieval task [49, 52, 102, 109, 111, 189, 203, 206] are COCO Captions [31], commonly known as MS-COCO in the cross-modal literature, and Flickr30K [224]. Only very few images from Flickr30K contain scene text (see Table 7.1), so we decided to start from COCO Captions, a subset of the COCO dataset [118]. Additionally, the reading systems community commonly uses the COCO-Text dataset [196]. It contains a sample of 63,686 COCO images with fully annotated scene-text instances. Among the COCO-Text images, we selected the ones that contain machine-printed, legible text in English, leading to a total of 17,237 images. In order to gather only images with the three modalities, we finally select the intersection between the filtered COCO-Text and COCO Captions. This leads to a multimodal dataset of 10,683 items, each item consisting of an image with scene text and five captions, referred to as *COCO-Text Captioned (CTC)*. Note that the resulting CTC dataset shares

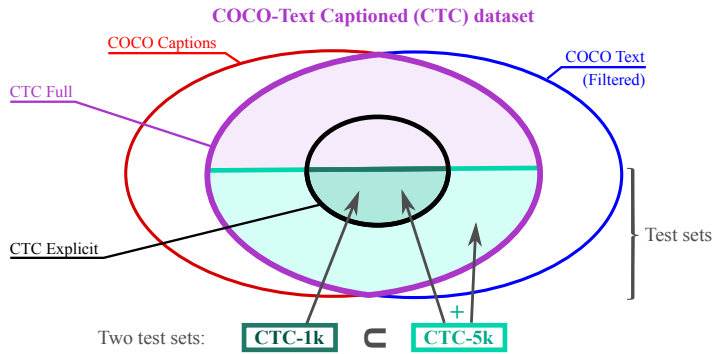


Figure 7.2: Proposed CTC dataset, which is designed to allow a proper evaluation of the task StacMR, as all entries contain three modalities: image, scene text and caption.

92.47% of its images with the original COCO caption training split. As a consequence, *we can not use any models trained on COCO caption in our experiments*, as their training set would inevitably share images with our test set. The dataset’s construction is illustrated in Figure 7.2.

Statistics. Our only driver for building the CTC dataset has been to identify samples where all three modalities are available, without explicitly requiring at any point that scene text had any semantic relation to the captions. This is the most important requirement for a dataset where scene text is truly incidental and captions are not biased towards this additional modality. Despite this, to be coherent with the StacMR task definition, it is paramount to show that the proposed CTC dataset contains some inherent semantic relations between scene text found in an image and the captions that describe it. To this end, we design three scenarios which illustrate this semantic relevance at the *image*, *caption* and *word*-level.

More precisely, we first remove stop-words from captions and scene-text annotations, and embed each remaining word with Word2Vec [139] vectors trained on the Google News dataset. The semantic relevance between two words is defined as the cosine similarity between their Word2Vec embeddings. We then consider three scenarios to showcase the relevance of scene text to image captions. The first scenario considers the highest semantic similarity between any scene-text word and any word from the set of 5 captions, for each image. This scenario visualizes the *semantic relation with images*, seen as sets of captions. The second scenario considers the highest semantic similarity between any scene-text word and any word from a corresponding caption. It highlights the *semantic relation with individual captions*. The third scenario considers how many caption words are related to scene-text words. This captures the *semantic relation with individual words* in captions.

The three histograms of Figure 7.3 correspond to the three previously described scenarios. The fact that many words have a strong similarity at all three levels confirms that scene text can be used to model the semantics between the three studied

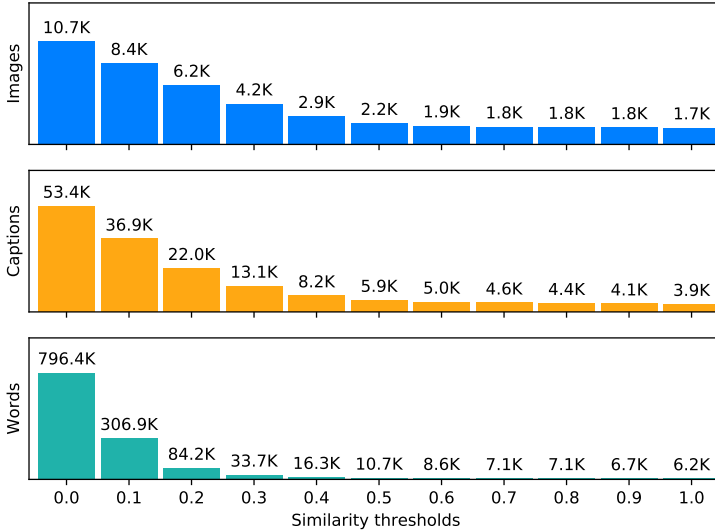


Figure 7.3: CTC full statistics. Cumulative histograms (as thresholds over similarity vary) of the semantic similarity between instances of scene-text tokens and a) all captions for an image (Images), b) individual captions (Captions), and c) individual words in captions (Words).

modalities to further leverage them in order to obtain a better performing cross-modal retrieval system.

As scene text provides fine-grained semantic information, its importance is query-dependant and it should be used selectively. An algorithm designed for the task should be able to decide, for each image, to which extent scene text is relevant for the cross-modal retrieval task. In order to better capture this, we define two partitions of the CTC dataset. CTC presents a natural semantic split that is evident in Figure 7.3 - a) that quantifies semantic similarity at the image level. The first quantization (threshold = 1) corresponds to images for which at least one word appears in both the scene text and one of the captions. We refer to this set of 1,738 images as *CTC explicit*. We expect scene text from this set to often be relevant to the retrieval task. We employ the full CTC dataset, here referenced as *CTC full* to avoid ambiguity, and to evaluate the more generic scenario where the role of scene text for retrieval is a priori unknown. This second set contains the previously mentioned explicit partition as well as images in which scene text is less relevant according to the annotated captions. Example image-caption pairs from *CTC explicit* are shown in Figure 7.4. This illustrates that scene text provides a strong cue and fine-grained information for cross-modal retrieval.

For evaluation purposes, we define two test splits. The first one, which we refer to as *CTC-1K*, is a subset of *CTC explicit*. The second test set, *CTC-5K*, contains the previous 1,000 explicit images of *CTC-1K* plus 4,000 non-explicit images. The remaining 738 explicit plus 4,945 non-explicit images are used for training and validation purposes.




Image	Captions
	<p>Sign warns against runaway vehicles along a hilly roadway. A white signing telling people how to park their cars on a steep hill. A sign explaining how to park on a hill is posted on the street. A warning sign is fastened to a post. Street sign with instructions on parking the hilly city roads.</p>
	<p>A person holding up a tasty looking treat. A person holding up a gummy hot dog in their hand. a closeup of a candy gummy hot dog in plastic packaging. A hotdog that appears to be a gummy hotdog. A gummy hot dog that is for sale.</p>
	<p>Parked school bus with a banner attached to it and people looking at it. A man and a woman outside a school bus. A school bus parked outside of a building. A school bus sits parked as people walk by. A school bus sitting on the side of the road near a pink car.</p>

Figure 7.4: Image-caption pairs from the CTC dataset. These images belong to CTC explicit, i.e. their scene text and captions share at least one word (marked in bold).

7.2.2 Dataset Samples

Figure 7.5 showcases more samples of image-caption pairs that belong to the full CTC dataset. On the other hand, in Figure 7.6 we depict image-caption pairs that belong to the explicit set of the CTC dataset, the bold words in captions reference appearing scene text. We can note that scene text provides strong cues to better discriminate each image. Leveraging scene-text can provide important complementary information for vision and language-oriented tasks, such as in the case of cross-modal retrieval.

7.2.3 Comparison with other Datasets

Table 7.1 provides a comparison with the previously mentioned datasets with statistics on the three modalities. Scene-text from COCO Captions [31] and Flickr30K [224] was acquired using a scene-text detector [134]. As mentioned earlier, none of the existing benchmarks contains samples where all three modalities are annotated.

Closely related to the proposed CTC dataset, TextCaps [176] is an image captioning dataset that contains scene-text instances in all of its 28,408 images. TextCaps is biased by design, as annotators were asked to describe an image in one sentence which would require reading the text in the image. From the statistics shown in Figure 7.7 it can be seen first, that TextCaps images were selected to contain more text tokens than should be naturally expected, and second, that many more of these tokens end up being used in the captions compared to the unbiased captions of CTC. The existing bias in TextCaps is also evident by analyzing the intersection of 6,653 images it has with the recently published Localized Narratives dataset [154].






Image	Captions
	<p>A blue bus at a bus stop with its doors open.</p> <p>A bus with its doors open is waiting at a bus stop.</p> <p>A bus sits parked on the side of a street.</p> <p>A picture of a bus on the side of the street.</p> <p>The blue and white trolley is waiting on passengers.</p>
	<p>A woman, man and two dogs in an inflatable raft on some water.</p> <p>The two ladies are in the row boat.</p> <p>Three people in a raft on the lake.</p> <p>A boat with people on it with a dog in water with a goose in it.</p> <p>Man and woman with two dogs on a power boat on a lake.</p>
	<p>A train on the tracks with people standing and walking by it</p> <p>A crowd of people are walking in front of a train</p> <p>A stopped train at a train crossing with people crossing the tracks.</p> <p>A black train parked at a train station as people walk across the train tracks.</p> <p>People at a train station, gathering around a black locomotive.</p>
	<p>A man holding a tennis racquet on a court.</p> <p>A man swinging a tennis racket during a tennis match.</p> <p>A tennis player in mid air action on the court.</p> <p>A tennis player about to serve the ball as a small crowd looks on.</p> <p>A tennis player is in the air making an overhead swing.</p>
	<p>A red double decker bus on street next to building.</p> <p>A bus that is driving in the street.</p> <p>A ride double-decker bus stands out against a black and white background.</p> <p>A double decker bus with few passengers turning at a corner.</p> <p>A red double decker bus driving down a city street.</p>

Figure 7.5: Image-caption pairs taken from the full proposed CTC dataset, in which appearing scene-text does not have a semantic relation with the annotated captions, i.e. there are no scene-text and captions common words.

From those 6,653 images, only 512 (10%) of them were annotated with captions that make use of any text tokens in the Localised Narratives dataset, where annotators

Image	Captions
	<p>An emergency response person is on a motorcycle.</p> <p>A medical person riding a motorcycle with ambulance on back.</p> <p>A police officer on a motorcycle pulling over a black car.</p> <p>A police motorcycle gets down to business when someone speeds.</p> <p>A motorcycle with a sign on the back that says ambulance.</p>
	<p>A China Airlines Airplane sitting on a waiting area of an airport.</p> <p>A big commuter plane sits parked in a air port.</p> <p>A China Airlines airliner is parked at an airport near another jet.</p> <p>Some white red and blue jets at an airport.</p> <p>China airplane airline is parked at a dock.</p>
	<p>A motorcycle parked in a parking lot next to a car.</p> <p>An antique Indian motorcycle is parked next to the sidewalk.</p> <p>Motorcycle parked on the edge of a street.</p> <p>An old Indian motorcycle parked at the curb of a street.</p> <p>A motorcycle parked on a sidewalk next to a street.</p>
	<p>Looks like a portrait of a distinguished gentleman.</p> <p>A painting of Walter Camp, siting on bench.</p> <p>A painting of a man in brown jacket and hat sitting at a bench.</p> <p>This a painting of Walter Camp in a trench coat.</p> <p>A painting of an older man on a city bench holding a rolled up magazine.</p>
	<p>A professional baseball player standing on the field while holding a mitt.</p> <p>A baseball player wearing a catchers mitt on top of a field.</p> <p>A Twins baseball player holding his glove walking on the field.</p> <p>The pitcher is resigned to losing the important game.</p> <p>A Twins baseball player walking to the dugout.</p>

Figure 7.6: Image-caption pairs from the proposed CTC explicit dataset, i.e. the scene-text and captions have at least one word in common (marked in bold).

were not instructed to always use the scene text.

According to our statistics, this is already higher than expected in the real world.

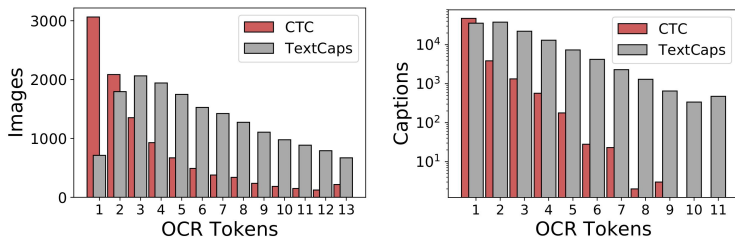


Figure 7.7: Histograms of the number of OCR tokens found in images (seen as sets of captions, left) and in individual captions (right) for the CTC and TextCaps datasets.

Dataset	Total Images		Annotations	
	Images	w/ Text	Scene Text	Captions
Flickr30K [224]	31,783	3,338 [★]	✗	✓
TextCaps [176]	28,408	28,408 [‡]	✗	✓
COCO Captions [31]	123,287	15,844 [★]	✗	✓
COCO-Text [196]	63,686	17,237 [†]	✓	✗
COCO-Text Caps	10,683	10,683[†]	✓	✓

Table 7.1: Datasets’ statistics for standard benchmarks and the proposed CTC. † refers to COCO-Text filtered selecting machine printed, English and legible scene text only. ★ numbers obtained with method from [134]. ‡ numbers obtained with method from [22].

This is because the Localised Narratives captions are long descriptions and tend to venture to fine-grained (localized) descriptions of image parts where text is more relevant. The proposed CTC is a much less biased dataset in terms of caption generation. The objective is to provide realistic data that permit algorithms to explore the complex, real-life interaction between captions, and visual and scene-text information, avoiding assuming or forcing any semantic relation between them.

7.3 Methods

This section describes approaches to tackle the StacMR task. First, we propose strategies to directly apply standard pretrained cross-modal retrieval models to our new task and its three modalities: images, captions and scene text (Section 7.3.1). Second, we propose an architecture to learn a joint embedding space for cross-modal retrieval in which the image embedding function learns to fuse both the visual and the scene-text information (Section 7.3.2).

7.3.1 Re-Ranking Strategies

This subsection considers the image-to-caption retrieval task. Note that everything can easily be rewritten to consider the caption-to-image case. For StacMR, images are multimodal objects: they contain visual information as well as textual information coming from scene text. On the other hand, captions contain textual information only. This asymmetry allows decomposing the StacMR task into two independent retrieval problems: visual-to-caption and scene-text-to-caption. The first *visual-to-caption* retrieval task performs comparisons between a purely visual descriptor of the query image and the textual descriptor of the captions. This corresponds to the standard cross-modal retrieval task as performed on Flickr30K or COCO Captions. The second, *scene-text-to-caption* retrieval task, performs comparisons between the textual descriptors of the scene text from the query image and the captions. Any textual descriptor could be used. In our experiments, we use the textual descriptor of a cross-modal retrieval model as it has been tailored for capturing concepts relevant to images.

A pretrained cross-modal retrieval model relies on a metric space equipped with a similarity function that can indistinguishably compare visual and textual descriptors and allows ranking all database elements according to a query.

Notations. Given a query image q and a caption from the gallery d , let $s_v(q, d)$ be the score between q and d using the image-to-caption similarity from a cross-modal retrieval model and $s_t(q, d)$ the score between q and d using the scene-text-to-caption similarity from that same model.

Re-Ranking Strategies. The most straightforward way to obtain StacMR results is simply to perform a *late fusion (LF)* of the ranking results obtained using both s_v and s_t . More formally, we compute the linear combination s_{LF} of the scores s_v and s_t , using a parameter α :

$$s_{LF}(q, d) = \alpha s_v(q, d) + (1 - \alpha) s_t(q, d). \quad (7.1)$$

One weakness of the late fusion strategy is that it combines all gallery items. Instead, we can limit the influence of the tails to avoid misranking by using different fusion strategies. Given $k > 0$, let I_k be the indicator function that a gallery item is in the top- k ranked items according to s_t , i.e. $I_k(q, d) = 1$ if d is in the top- k results when querying with q and similarity s_t , and $I_k(q, d) = 0$ otherwise. Following [4, 39, 40], we then define the *late semantic combination (LSC)* and *product semantic combination (PSC)* with Equations (7.2) and (7.3) respectively. Note that LSC is equivalent to the late fusion if k is equal to the gallery size.

$$s_{LSC}(q, d) = \alpha s_v(q, d) + (1 - \alpha) s_t(q, d) I_k(q, d) \quad (7.2)$$

$$s_{PSC}(q, d) = s_v(q, d) s_t(q, d) I_k(q, d) \quad (7.3)$$

These different re-ranking strategies do not require any training and rely on existing pretrained cross-modal retrieval models. We simply use the part of CTC disjoint from

the two test sets to choose the hyper-parameters α and k .

7.3.2 STARNet: A Dedicated Trimodal Architecture

All previously described approaches rely on a pretrained cross-modal retrieval model. Here, we introduce a new architecture able to handle the trimodality of the StacMR task. We start from the model presented in [111] and extend it to integrate scene text. First, we assume that scene text has been detected within an image. Then we adapt the model of [111] to be able to read scene-text instances. We include a positional information encoder along with a scene-text Graph Convolutional Network (GCN) and a customized fusion module into the original pipeline. Sharing intuition with [198], we assume that scene text acts as the modifier in the joint embedding space applied to the visual descriptor of an image.

We propose the STARNet (**Scene-Text Aware Retrieval Network**) model, illustrated in Figure 7.8. It is composed of the following modules: a joint encoder Φ for both an image and its scene text, a caption encoder Θ , and a caption generation module Ψ . Given an image I_i and its scene-text OCR_i , the global feature encoding for both modalities are $I_{fi} = \Phi(I_i, OCR_i)$. The image encoder follows [8] and uses a customized Faster R-CNN [161] to extract visual features for all regions of interest represented by V_i . Similarly, the employed OCR [64] extracts scene-text instances as well as their bounding boxes and is represented by T_i .

For both modalities, image and scene text, we use a GCN [100] to obtain richer representations. For notation purposes, we refer to the visual or textual features as F_i since the formulation of both visual and textual GCNs is similar. The inputs to each GCN are features $F_{fi} \in \mathbb{R}^{k \times D}$, where $D = 2048$ and, $k = 36$ in the case of V_i and $k = 15$ in the case of T_i . A zero padding scheme is employed for both modalities if the number of features is smaller than k . We define the affinity matrix R , which computes the correlation between two regions and is given by: $R_{ij} = \rho(k_i)^T \omega(k_j)$, where k_i, k_j represent the two features being compared and $\rho(\cdot)$ and $\omega(\cdot)$ are two fully connected layers that are learned in an end-to-end manner by backpropagation.

The obtained graph can be defined by $F_{fi} = (F_i, R)$, in which the nodes are represented by the features F_i and the edges are described by the affinity matrix R . The graph describes through R the degree of semantic relation between two nodes. In our method, we employ the definition of Graph Convolutional Networks given by [100] to obtain a richer set of features from the nodes and edges. The equation that describes a single Graph Convolution layer is:

$$F_g^l = W_r^l (R^l F_i^{l-1} W_g^l) + F_i^{l-1} \quad (7.4)$$

where $R \in \mathbb{R}^{k \times k}$ is the affinity matrix, $F_i \in \mathbb{R}^{k \times D}$ are the input features of a previous layer, $W_g \in \mathbb{R}^{D \times D}$ is a learnable weights matrix of the GCN, $W_r \in \mathbb{R}^{k \times k}$ is a residual weights matrix and l is the number of GCN layer. Particularly, we employ a total number of $l = 4$ for both GCNs used in the proposed pipeline.

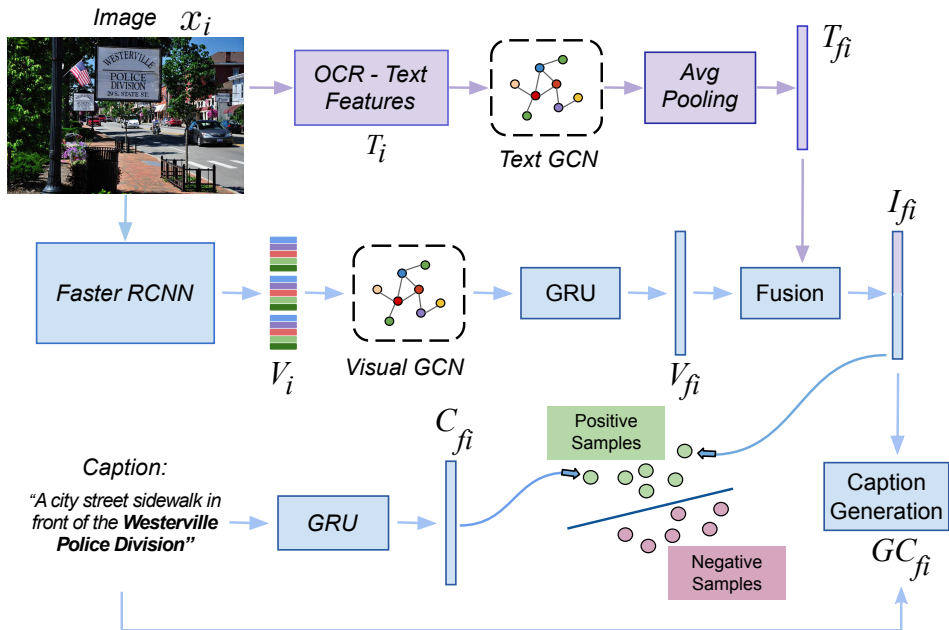


Figure 7.8: Our proposed STARNet model. Visual regions and scene-text instances are used as input to a GCN. The final learned representations are later combined to leverage complementary semantic information.

The output of the visual GCN goes through a GRU [38] to obtain the global image representation denoted by V_{fi} . Textual features from the output of the scene-text GCN are average-pooled to obtain a final textual representation denoted by T_{fi} .

The final image representation I_{fi} is the dot product between the visual and final scene-text features (which act as a modifier) added to the original visual features: $I_{fi} = V_{fi} \odot T_{fi} + V_{fi}$.

Caption C_i from the corresponding training image-caption pair is encoded with a GRU [38, 52], leading to $C_{fi} = \Theta(C_i)$. To align image features with their caption features in a joint embedding space, we train Φ and Θ using a triplet ranking loss [52, 109] by employing the hardest negative sample on each mini-batch.

In order to provide the model with a stronger supervision signal, the learned image representation I_{fi} is also used to generate a caption as an auxiliary task. We train the third encoder Ψ so that the generated caption equals to: $GC_{fi} = \Psi(I_{fi})$. This sequence-to-sequence model uses an attention mechanism similar to [197] and we optimize the log-likelihood of the predicted output caption given the final visual features and the previously generated word.

7.4 Experiments and Results

We present results on CTC. They are split into two parts: visual-only and scene-text-only baselines, as well as their unsupervised re-ranking (Section 7.4.2), and supervised trimodal fusion results from STARNet (Section 7.4.3). Following cross-modal retrieval (CMR) evaluation standards, we report performance with recall at K ($R@K$) for K in $\{1, 5, 10\}$ for both image-to-text and text-to-image retrieval.

7.4.1 Implementation Details

In the baselines of supervised models, SCAN [109] and VSRN [111] use the same hyperparameters as the correspondent work published and it is based on public code available. We introduce modifications to each of those models, in a way that scene-text instances are treated similarly to visual regions. We expanded the number of visual region inputs from the original 36 to add 15 scene-text instances that sum in total 51 combined visual and textual regions. Text instances are sorted according to the confidence value. If text is not present, or the instances are less than 15, we use a zero-padding scheme.

The proposed supervised model, STARNet was trained for 30 epochs along with a batch size of 128 samples per iteration on each experiment. The learning rate employed was 0.0002 and was decreased by a factor of 10 every 10 epochs. The visual features have a dimension of 2048-d. The FastText [21] textual vectors that serve as input to the model have a dimension of 300-d, which are linearly projected into a similar feature space of 2048-d as the visual features. We use 4 GCN-based reasoning layers on the visual and textual GCN to enrich and reason from the visual and scene-text features. The final semantic space learned contains 2048-d, which is used to project the final image representation and captions.

In our experiments, when the Flickr30K [224] dataset is employed, we use the same training, validation and testing split as in [92], which contain 28,000, 1,000 and 1,000 images respectively. When using only the TextCaps [176] dataset, the original training set is used and the validation set is employed as the evaluation set since the test set is currently publicly unavailable. At the moment of training the proposed STARNet model, we employ the validation set of TextCaps to achieve the best-performing weights.

7.4.2 Baselines and Re-Ranking Results

This section first introduces visual-only CMR models. These allow observing how standard CMR models tackle the StacMR task on CTC. Then, we propose scene-text-only metric spaces, where the only information extracted from the image is its scene text. These baselines should help judge the semantic relevance of the scene text with respect to the captions. The remaining results correspond to different combinations: a

	Visual Model	Scene-text Model	Trained on		Scene-text Source	Re-rank	CTC-1K						CTC-5K					
			F30K	TC			Image to Text			Text to Image			Image to Text			Text to Image		
							R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
(1)	VSE++ [52]	✗	✓	✗	-	-	20.5	42.8	54.5	15.4	35.2	48.4	13.3	30.2	40.2	8.4	21.5	30.1
(2)	VSE++	✗	✓	✓	-	-	23.9	50.6	63.2	16.5	39.6	53.3	12.6	30.1	40.2	7.9	21.0	29.7
(3)	VSRN [111]	✗	✓	✓	-	-	27.1	50.7	62.0	19.7	42.8	55.7	19.2	38.6	49.4	12.5	29.2	39.1
(4)	VSRN	✗	✓	✓	-	-	35.6	64.4	76.0	24.1	50.1	63.8	22.7	45.1	56.0	14.2	32.1	42.6
(5)	✗	VSE++ GRU	✓	✓	GT	-	26.3	40.4	47.3	10.0	20.3	25.6	4.4	7.1	8.2	1.6	3.5	4.7
(6)	✗	VSRN GRU	✓	✓	GT	-	12.3	25.1	30.1	6.8	15.3	20.0	1.9	4.0	5.2	1.1	2.8	3.8
(7)	✗	Fasttext+FV	✗	✗	GT	-	21.7	36.5	44.3	3.2	6.6	9.0	3.5	5.9	7.5	0.6	1.3	1.7
(8)						AVG	34.6	53.1	61.0	14.5	31.0	39.4	10.0	21.5	29.5	5.0	14.1	21.4
(9)					GT	LF	31.0	60.0	72.3	20.4	44.7	57.3	13.4	30.9	41.5	7.4	20.5	29.1
(10)	VSE++	VSE++ GRU	✓	✓		PSC	37.4	62.8	73.6	15.5	42.6	57.1	12.2	32.1	42.4	4.1	19.3	29.2
(11)						LSC	31.6	57.8	70.2	20.3	44.7	57.8	13.7	31.7	41.6	7.7	21.0	29.6
(12)						AVG	36.8	62.2	72.9	18.6	40.5	52.9	15.3	33.5	44.3	6.4	18.9	28.0
(13)	VSRN	VSRN GRU	✓	✓	GT	LF	40.3	68.5	79.9	23.9	49.9	63.4	22.6	45.0	56.3	11.8	29.5	40.0
(14)						PSC	33.5	65.9	78.2	15.8	48.1	64.3	18.5	44.5	56.0	5.3	28.7	41.0
(15)						LSC	38.6	67.5	78.5	24.3	50.4	64.0	23.4	45.6	56.5	12.1	30.6	41.1
(16)						LF	45.8	72.7	81.4	26.5	52.7	66.1	24.2	46.1	57.1	12.9	31.0	41.2
(17)	VSRN	VSE++ GRU	✓	✓	GT	PSC	42.2	71.5	82.8	18.9	51.1	66.4	20.1	46.4	57.5	6.7	29.5	41.6
(18)						LSC	45.3	71.5	80.7	26.7	53.0	66.2	24.4	46.9	57.4	13.2	31.8	42.3
(19)						LF	41.5	70.1	79.8	25.1	51.2	64.3	23.3	45.0	58.9	12.6	30.5	41.1
(20)	VSRN	VSE++ GRU	✓	✓	OCR	PSC	38.5	69.6	80.6	17.9	50.1	65.1	19.8	45.7	57.2	7.0	29.8	41.7
(21)						LSC	42.2	68.6	78.5	25.5	51.8	64.9	19.8	45.7	57.2	13.2	31.5	42.2

Table 7.2: Results on CTC for visual and scene-text baselines, and their re-ranking combinations. Visual model and Scene-text model indicate image-caption and scene-text-caption retrieval, respectively. *GT* stands for ground-truth scene-text annotations and *OCR* for scene-text prediction obtained from [64]. Bold numbers denote the best performances of visual, scene-text, and re-ranking methods for each ensemble of models.

naive average of visual and scene-text embeddings for metric spaces that allow it and the different re-ranking strategies introduced in Section 7.3.1.

Visual-only Baselines. We use two CMR models based on global features for both images and captions,

VSE++ [52] and VSRN [111]. Both works provide public training code, used for all models in this section, with the exception of the VSE++ model trained on Flickr30K, for which we use the model provided by [52]. We train these architectures either with Flickr30K or Flickr30K + TextCaps. As mentioned in Section 7.2.1, models pretrained on COCO Captions are not considered due to the overlap between the training set of COCO Captions and our test sets.

Results are presented in Table 7.2, rows (1-4). VSRN surpasses VSE++, mirroring their relative performance from CMR benchmarks. Furthermore, models trained on the additional data of TextCaps outperform models trained only on Flickr30k. This is interesting, as TextCaps image-captions pairs are more dependent on their scene text than those from Flickr30k. Enlarging the dataset size with the inclusion of TextCaps explains this improvement to an extent, as the training set of Flickr30k is relatively small. Moving forward, we only report models trained on F30K+TC.

Scene-Text only Baselines. We use the textual embedding part of our two previously used CMR models (denoted by VSE++ GRU and VSRN GRU respectively). We also consider FastText [21] word embeddings followed by a Fisher vector encoding [152]

(denoted by FastText+FV), which is able to deal with out-of-vocabulary words. For these experiments, we use ground-truth OCR annotations as scene text. Results are presented in Table 7.2, rows (5-7). We observe much weaker results than the purely visual baselines. For CTC-1K, this approach can rely on shared words between scene text and one of the captions. For the more realistic CTC-5K, we see that scene text brings very little in isolation. Note that the VSE++ GRU outperforms VSRN GRU for this task, while VSRN is better for the purely visual case. This motivates the hybrid strategies merging both models that we report later. Fasttext+FV yields strong results on image-to-caption retrieval on CTC-1K, but shows poor results on the other evaluated scenarios.

Average Embedding. If an image and scene text are represented using the same CMR model, all three modalities are represented in the same embedding space. This allows a naive combination that consists in averaging visual and scene-text embeddings to represent the image, reported as AVG on the Table 7.2, rows (8) and (12). This brings a non-negligible improvement on CTC-1K Image to Text compared to their respective visual-only baseline and it is the first proof that scene text, even naively used, improves on some StacMR queries.

However, we observe a decline in CTC-5K in the same comparison. This hints at the fact that scene text provides fine-grained information that should be used selectively, and giving equal weight to both modalities is too crude an approach.

Re-Ranking Results. Some re-ranking results are presented in Table 7.2, rows (9-21). We test the best pairing of visual-only and scene-text-only models with three combination strategies: late fusion (LF), product semantic combination (PSC), and late semantic combination (LSC). Hyper-parameters of each re-ranking strategy are chosen for VSRN with VSE++ GRU and applied to all other combinations as is. We use the part of CTC explicit which is not used for testing as validation. For LF, $\alpha = 0.8$. For PSC, $\alpha = 0.95$ and $k = 3$. For LSC, $\alpha = 0.8$ and $k = 100$.

When compared to the unimodal baselines, all combinations improve results on CTC-1K. Both LF and LSC match the results of their visual baselines on CTC-5K, showing that these methods are more robust to scene-text information unrelated to the captions.

For the three best-performing re-ranking variants, we repeat the experiment using OCR predictions instead of the ground-truth scene-text annotations. Results are shown in rows (19-21). When compared with their counterparts in rows (16-18), we observe an R@10 loss on average of 1.7% in CTC-1k and stable results for CTC-5k. This validates the stability of these re-ranking strategies to loss of information due to imperfect OCR predictions.

7.4.3 Supervised Results

The latest cross-modal retrieval models rely on region-based visual features [109, 111, 203] rather than a global image representation [52]. In this section, we include the

Model	Uses Scene Text	Scene-Text Source	Trained on			CTC-1K						CTC-5K					
						Image to Text			Text to Image			Image to Text			Text to Image		
			F30K	TextCaps	CTC	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [109]	✗	-	✓	✗	✗	26.4	48.6	61.1	15.2	36.8	49.3	17.5	36.7	47.1	7.6	21.2	30.4
	✓	OCR	✗	✓	✗	19.5	43.8	57.1	10.2	28.7	42.1	7.0	20.0	29.7	3.2	11.7	18.1
	✓	OCR	✓	✓	✗	35.0	62.9	74.4	19.3	44.0	58.3	21.1	43.0	54.6	9.6	25.4	35.6
	✓	OCR	✓	✗	✓	27.5	48.9	61.9	16.5	37.7	51.1	18.6	37.3	47.6	8.1	21.6	30.6
	✓	OCR	✓	✓	✓	36.3	63.7	75.2	26.6	53.6	65.3	22.8	45.6	54.3	12.3	28.6	39.9
VSRN [111]	✗	-	✓	✗	✗	27.1	50.7	62.0	19.7	42.8	55.7	19.2	38.6	49.4	12.5	29.2	39.1
	✓	OCR	✗	✓	✗	18.6	40.4	52.2	11.7	31.0	44.2	6.6	17.9	25.8	4.5	13.0	19.8
	✓	OCR	✓	✓	✗	35.6	64.3	76.0	24.0	50.1	63.1	22.6	45.0	55.9	14.2	32.1	42.5
	✓	OCR	✓	✗	✓	36.1	64.1	75.8	26.2	53.1	65.2	24.6	48.1	58.8	15.4	35.7	46.9
	✓	OCR	✓	✓	✓	38.2	67.4	79.1	26.6	54.2	66.2	23.7	47.6	59.1	14.9	34.7	45.5
STARNet	✗	OCR	✓	✗	✗	29.4	52.3	62.6	21.8	44.3	57.2	19.9	39.6	50.1	13.4	30.7	40.4
	✓	OCR	✗	✓	✗	23.4	48.0	61.0	14.2	34.9	47.3	5.1	15.1	22.3	3.9	11.9	25.1
	✓	OCR	✓	✓	✗	39.3	65.4	76.8	25.9	52.3	65.2	21.1	41.8	52.9	13.8	31.8	42.0
	✓	OCR	✓	✗	✓	36.5	64.6	74.3	26.4	53.8	65.6	25.5	48.4	59.8	15.7	35.3	46.6
	✓	OCR	✓	✓	✓	44.1	74.8	82.7	31.5	60.8	72.4	26.4	51.1	63.9	17.1	37.4	48.3
Re-rank. (21)	✓	OCR	✓	✓	✗	42.2	68.6	78.5	25.5	51.8	64.9	19.8	45.7	57.2	13.2	31.5	42.2
STARNet - GT	✓	GT	✓	✓	✓	45.4	74.9	83.9	32.0	61.2	73.3	26.8	51.4	64.1	17.4	37.8	48.7

Table 7.3: Retrieval results on the CTC-1K and CTC-5K test set of supervised models. The second-to-last row shows the result from the unsupervised re-ranking baseline described in Table 7.2, line 21. *OCR* stands for the textual features obtained from [64], whereas *GT* refers to the Ground-truth annotated scene text. Results depicted in terms of Recall@K (R@K).

results of two state-of-the-art models, SCAN [109] and VSRN [111] that employ such region-based visual features.

The original cross-modal retrieval models, SCAN and VSRN are used only when trained on Flickr30K. In order to leverage scene text, we have modified them to include OCR features. In both models, the OCR features are projected into the same space as the visual features, and the default hyper-parameters are employed. All the obtained results are reported in Table 7.3. The second column depicts the usage of scene-text instances by each model, and the third column depicts the source of the scene text. We make the following observations.

First, we see that using standard models trained on a common cross-modal retrieval dataset, such as Flickr30k, does not yield good performances on the StacMR task. Second, we note the different behaviors when each dataset is used for training and testing is done on the CTC test sets. In particular, it is worth noting that by training solely on TextCaps [176], the performance of any model decreases significantly, especially in the CTC-5K dataset. This effect is caused by the bias in Textcaps that places a big focus on scene-text instances to describe an image, rather than combining visual and textual features in an unbiased way.

However, all datasets provide complementary statistics to train the STARNet model. For instance, Flickr30k focuses on relevant visual regions, whereas the combination of TextCaps and CTC can be seen as a reciprocal set of datasets that aim toward modeling the relevance of scene-text from an image in a more natural manner.

It is worth pointing out that STARNet almost doubles the performance in the CTC-

1K subset when compared to common retrieval models. We believe this effect is due to the explicit scene-text instances that reinforce the notion of the relevance of this modality. A smaller improvement is achieved in the CTC-5K. This result is caused by the fact that even though scene text does not appear explicitly in the captions, a varying degree of semantics between image and scene text can still be found.

Finally, we also show an upper bound at test time assuming a perfect OCR (using ground truth scene-text annotations in CTC), which adds a slight boost to the proposed method. This effect shows and confirms the importance of accurate scene-text recognizers in the StacMR task.

7.4.4 Qualitative Results

In Figure 7.9 we illustrate qualitative results when performing Image to Text cross-modal retrieval. The text contained within an image usually serves as discriminatory signal, such as the word "*samsung*" in the third image and the number "*15*" in the fifth query. Scene text also provides a strong complementary cue to be used along with visual features as the rest of the queried samples suggest.

It is important to note, that even though some samples are not entirely correct, the model still preserves semantics between image and retrieved captions.

We illustrate in Figure 7.10 the results obtained when performing Text to Image cross-modal retrieval. In the queries performed, scene-text works as fine-grained and discriminative information to retrieve correctly an image. Similarly to the previous scenario, wrongly retrieved samples still preserve semantics.

By exploring the qualitative results obtained, added to the quantitative tables in previous sections, we can reinforce the notion that modeling scene-text along with visual features does improve retrieval granularity thus yielding higher performing cross-modal retrieval pipelines.






Queried Image	Retrieved Captions
	<p>Clock at a train station showing the time of the next trains arrival. ✓</p> <p>A clock with the words next train written about it. ✓</p> <p>A clock on a train platform during day time. ✗</p> <p>A clock attached to a pole at a train station. † ✗</p> <p>A clock that is sitting on the side of the pole. ✓</p>
	<p>A large number of police motorcycles are lined up. † ✗</p> <p>A bunch of police officers on motorcycles waiting for something. ✓</p> <p>A group of police officers that are riding on motorcycles. † ✗</p> <p>A police on motorcycles are parked beside a crowd. † ✗</p> <p>A line of police are riding motorcycles down the street. ✗</p>
	<p>People riding on the upper level of a samsung bus in a parade. ✓</p> <p>A blue tow truck carrying a boat. ✗</p> <p>A blue truck is pulling a white boat. ✗</p> <p>A police vehicle on a tow truck that is being taken away. ✗</p> <p>A group of police standing at the back of a moving truck. ✗</p>
	<p>A tall lighthouse sign with a clock on the tower of a plaza. ✓</p> <p>A tall church building with a massive clock on front of it. ✗</p> <p>A modern clock tower is embellishing a market which sits beneath a clear blue sky. ✓</p> <p>Tall tower with clock near well lit building at night. ✗</p> <p>A large tower that has a clock on the very top of it. † ✗</p>
	<p>Two woman near the interstate 15 sign in las vegas. ✓</p> <p>Two women standing on a sidewalk next to a street sign at night while cars drive on the street next to them and behind them. ✓</p> <p>Two young ladies standing on the sidewalk under a street sign. ✓</p> <p>Two people standing on a street with a street sign. ✓</p> <p>Two women on street next to cars and traffic signs. ✓</p>

Figure 7.9: Qualitative samples obtained when an image is used as a query (Image to Text) in the proposed CTC explicit dataset. Correct results are marked with ✓. Incorrect results are marked with ✗. Reasonable mismatches are depicted with † but still marked by a ✗.

Query 1: A *marc* passenger **drains** rides along railroad tracks.



Query 2: Sign explaining how to **park** on a hill is posted on the street.



Query 3: Commuter **shuttle** bus on roadway in large city.



Query 4: A *china airlines* airliner is parked at an airport near another jet.

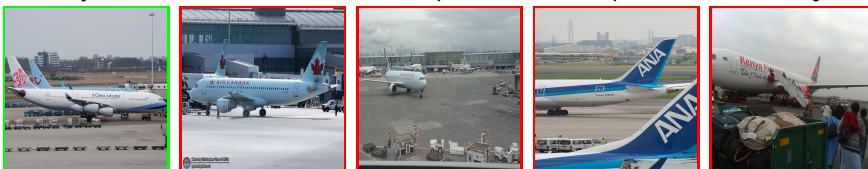


Figure 7.10: Qualitative samples when a caption is used as a query (Text to Image) in the proposed CTC explicit dataset. Correct results are marked in a green box. Incorrect results are marked in a red box. Words in bold in queried captions depict the scene-text that helps to discriminate retrieved images, which otherwise are ambiguous. Query 1 contains an annotator typo "drains".

7.5 Conclusion

In this Chapter, we highlight the challenges stemming from including scene-text information in the cross-modal retrieval task. Although of high semantic value, scene text proves to be a fine-grained element in the retrieval process that should be used selectively. We introduce a realistic dataset, *CTC*, where annotations for both scene text and captions are available. Contrary to datasets constructed with scene text in mind, *CTC* is unbiased in terms of scene-text content and of how it is employed in the captions. A comprehensive set of baseline methods showcase that combining modalities is beneficial, while a simple fusion cannot tackle the newly introduced task of scene-text aware cross-modal retrieval. Finally, we introduce *STARNet* a supervised model that successfully combines all three modalities. Public available code and collected dataset can be found at <https://github.com/AndresPMD/StacMR> and <https://europe.naverlabs.com/stacmr>.

Chapter 8

Is An Image Worth Five Sentences? A New Look into Semantics for Image-Text Matching

The task of image-text matching aims to map representations from different modalities into a common joint visual-textual embedding. However, the most widely used datasets for this task, MSCOCO, and Flickr30K, actually image captioning datasets that offer a very limited set of relationships between images and sentences in their ground-truth annotations. This limited ground truth information forces us to use evaluation metrics based on binary relevance: given a sentence query we consider only one image as relevant. However, many other relevant images or captions may be present in the dataset. Therefore, we propose two metrics that evaluate the degree of semantic relevance of retrieved items, independently of their annotated binary relevance. Additionally, we incorporate a novel strategy that uses an image captioning metric, CIDEr, to define a Semantic Adaptive Margin (SAM) to be optimized in a standard triplet loss. By incorporating our formulation into existing models, a large improvement is obtained in scenarios where available training data is limited. We also demonstrate that the performance on the annotated image-caption pairs is maintained while improving on other non-annotated relevant items when employing the full training set.

8.1 Introduction

According to [35, 183], language offers a discretely infinite number of opportunities to explain perceived information. However, this notion of “discrete infinity” is disregarded in current cross-modal retrieval metrics and models. We study the task of Image-Text Matching (ITM) in a cross-modal retrieval situation while keeping the discrete infinity idea in mind. However, widely used datasets for the image-text matching (ITM) task lack thorough annotations of many-to-many mappings between pictures and captions, which seems to go counter to the idea of discrete infinity.

Consequently, the lack of many-to-many annotations causes a direct effect on the way the ITM task is evaluated. Sentences that are not relevant according to the ground truth can describe an image in various degrees of correctness and coverage, thus making the way we evaluate current models on ITM *incomplete*. We can see an example of such a problem in Figure 8.1. The widely adopted metric employed to evaluate the performance of a model in the ITM task is Recall@K [51, 109, 111, 113, 120, 127]. The Recall@K as it is used in IMT is binary by definition: it returns 1 if at least one of the relevant items according to the ground truth is retrieved within the top-k positions for a given query, otherwise, it returns 0. Due to this binary definition, the metric can not fully assess the degree of accuracy and coverage of the retrieved sentences given an image or the other way around.

Additionally, the to-go strategy from standard approaches for ITM, firstly introduced by [51], relies on hard-negative mining at the moment of constructing samples to be used in a Triplet loss function to be optimized. Current state-of-the-art methods rely heavily on such formulation, which requires a carefully handcrafted fixed margin [51, 109, 111, 120, 133, 209]. In this Chapter, we propose solutions to the aforementioned problems by introducing the usage of image captioning metrics such as SPICE [6] and CIDEr [195] as a part of an additional metric formulation for the evaluation of the ITM task. Image captioning metrics have been widely studied and accepted as automatic tools to evaluate the similarity of sentence meanings that closely correlate with human judgment. We utilize such metrics that allow a transition from a traditional recall to a Normalized Cumulative Semantic (NCS) Recall by incorporating the continuum of language into the evaluation. Secondly, considering the continuous nature of language, we re-formulate a triplet loss by introducing a Semantic Adaptive Margin (SAM). We calculate a SAM according to image captioning metrics, which does not rely on a hard-negative mining approach (see Figure 8.2). Our formulation employed in scenarios with limited data achieves state-of-the-art by a *significant* retrieval improvement.

Our contributions are as follows: (1) We identify shortcomings from the commonly used Recall@K in the ITM task. By adopting image captioning metrics we model the many-to-many semantic relationships between images and captions. (2) We propose a novel Semantic Adaptive Margin (SAM) that takes into consideration image captioning metrics to define the similarity among samples. (3) We show that by relying on image captioning metrics and incorporating them into our proposed adaptive margin,

Text Query: A baseball player attempting to score a run before the catcher tags the player out.

Method A (Top-5 results)



Method B (Top-5 results)



Figure 8.1: According to the Recall@5 metric, defined for Image Text Matching, both methods A and B are equally good: it considers only one image as relevant for a given sentence query. We propose two metrics and an adaptive margin loss that takes into account that there might be other relevant images in the dataset. In this Figure, we represent the semantic similarity of images to the query by their colored border (the greener the more similar).

a substantial boost is achieved in scenarios with reduced training data. (4) We provide exhaustive experiments on two benchmark datasets, which show that by incorporating our adaptive margin formulation an increase in performance is achieved across a variety of state-of-the-art pipelines.

8.2 Metrics

Before we move with our formulation, we introduce the reader the nomenclature used in the rest of this work. First, the image and caption sets will be referred to as I and C , while the respective test set will be represented by I_T, C_T . We refer as G_l to all the ground truth captions corresponding to an image $l \in I$. We use ϕ to indicate an evaluation metric function such as CIDEr or SPICE. Finally, Q_{ik} represents the *retrieved* items for a given query i at a top- k cut-off threshold.

8.2.1 Is an image worth 5 sentences?

Both of the most commonly used datasets in ITM, namely Flickr30k [153] and

MSCOCO [118], contain 5 ground truth sentences per image. A direct outcome is that the current evaluation solely considers those 5 sentences as relevant to a single

image. However, it is a known fact that in MSCOCO or Flickr30k there are many sentences that can perfectly describe a non-paired image [150, 204, 208]. In other words, there are sentences (images) that are relevant to images (sentences) even though they are not defined as such in the retrieval ground truth. We refer to these samples as non-ground truth (non-GT) relevant items. Specifically, ITM models are tested on 5k images and 25k sentences in MSCOCO. In the case of image-to-text retrieval, recall completely ignores the retrieved order of the remaining 24995 sentences (99.98% of the test set). Yet, it is crucial to consider all semantically relevant items (including non-GT) to properly evaluate a model’s capability.

Aside from the prior problems, Recall@K ($R@K$) as used in the ITM task is a binary metric, i.e. it is a hard metric that does not take into account the semantic continuum of language. When it comes to language, even ground truth paired sentences do not explain a given image to the same degree as they are not paraphrases from each other.

Another identified drawback is that the recall formulation used in ITM is different than the original recall employed in information retrieval. The recall metric used in ITM, referred to as R^V , takes the definition from [76]. In the image-to-text scenario, the R^V only cares about the first GT annotated caption retrieved in the top- k relevant results. This formulation discards the remaining 4 annotated samples in the GT. On the other hand, recall defined by [169], referred as R , considers *all* other relevant items in the formulation. It is important to note that both formulations agree on the text-to-image scenario due to the existence of only 1 relevant image in the GT. Both recall formulations can be appreciated better in Equation 8.1 and 8.2.

$$R@k = \frac{1}{|I_T|} \sum_{i \in I_T} R_i@k, \text{ where } R_i@k = \frac{|G_i \cap Q_{ik}|}{|G_i|} \quad (8.1)$$

$$R^V@k = \frac{1}{|I_T|} \sum_{i \in I_T} R_i^V@k, \text{ where } R_i^V@k = \mathbb{1}_{\{G_i \cap Q_{ik} \neq \emptyset\}} \quad (8.2)$$

When formulating our metrics in the following sections, we use R instead of R^V , as it includes the remaining 4 items at evaluation. Nevertheless, it is important to note here that both R and R^V completely disregard the possible semantic relevance of non-GT samples. The existent limitations of employing solely recall as a metric lie in the fact that it misses evaluating those non-GT-relevant items.

8.2.2 Semantic Recall (SR)

Our metrics rely on the evaluations of captions with CIDEr and SPICE to decide which images are semantically similar to other sentences in the test set. Concretely, for a

given image i and sentence j such that $i \in I_T, c_j \in C_T$, we construct a matrix N where:

$$N_{ij} = \phi(G_i, c_j) \quad (8.3)$$

where $N_{ij} \in R$ and ϕ is one of the captioning metrics (CIDEr or SPICE). Once the similarity matrix N is defined, we can easily extend the ground truth relevant items for each possible query. Formally, we define \tilde{G}_i as the extension of ground truth relevant items for a query image i as the most similar m sentences from N_i . Now we define the Semantic Recall (SR) metric as follows:

$$R_i^{SR@k} = \frac{|\tilde{G}_i \cap Q_{ik}|}{|\tilde{G}_i|} \quad (8.4)$$

This metric allows a transition from the classic recall R^V to a metric that considers semantic relevance. However, the limitation on binary scoring associated with recall still persists. Another drawback is how to select a threshold m that captures how many non-GT images or sentences are relevant in the whole data corpus.

8.2.3 Normalized Cumulative Semantic (NCS) Score

The Normalized Cumulative Semantic Score (NCS) aims at addressing the limitations of the Semantic Recall (SR) described in the previous section. The NCS score is calculated as the division between the image captioning similarity ϕ of the retrieved samples and the maximum image captioning similarity score ϕ at a cut-off point K . Formally, we define our metric as:

$$N_i@k = \frac{\sum_j N_{ij}}{\sum_l N_{il}}, \text{ for } j \in \tilde{G}_i \cap Q_{ik} \text{ and } l \in \tilde{G}_i \quad (8.5)$$

For illustrative purposes, methods A and B from Figure 8.1, both equally good at recall (R^V), will score very differently at NCS. Method A will achieve a maximum score of 0.2. On the contrary, Method B will achieve a higher score since the retrieved samples contain a closer degree of semantics compared to the query.

With this formulation, we specify a solution to the binary nature of Recall@K ($R@K$) when it addresses the semantics of the language. Moreover, NCS can properly take into account the non-GT items when evaluating a model without the need of selecting a threshold m . In Section 8.4 we use these metrics to provide us with additional insights about the current model's performance.

8.2.4 Correlation with Human Judgements

Related to our work, the recently introduced CrissCrossed [150] dataset, is an extension of MS-COCO that comprises human judgments on the degree of similarity between captions and images. In this dataset, each annotator assesses how well an image and a sentence match on a 5-point Likert scale on MSCOCO. They collect these judgments not only for the predefined ground truth pairs but also for other pairs. Despite the *extensive* annotation process required, the test set contains 44k judgment pairs, of which 25k are ground truth pairs. We utilize these human judgments to calculate the Pearson-R correlation coefficient for Recall and NCS.

As it can be seen in Table 8.1, when all the pairs are considered, our metric has a better correlation with human judgments [6] with both SPICE and CIDEr. We observe that CIDEr has a better correlation when we take into account the 44k pairs, nonetheless, SPICE is better on Non-GT, Which is why we always evaluate our models with SPICE. Furthermore, this also extends to the case of non-ground truth relevant pairs. In non-GT relevant pairs the classic recall is uninformative due to the metric definition, while the NCS provides an acceptable estimation that correlates well with human judgment.

	All	Non-GT
Binary relevance (GT)	0.711	0.00
NCS with SPICE	0.729	0.536
NCS with CIDEr	0.734	0.453

Table 8.1: Pearson-R correlation coefficient results between human judgments and image text matching metrics on the CrissCrossed [150] dataset.

8.3 Semantic Adaptive Margin

In this section, we introduce our Semantic Adaptive Margin (SAM) formulation, which aims to alleviate common problems of the usage of a triplet loss on non-exhaustive many-to-many data mappings. Before we elaborate on the details, we present the reader with the original triplet formulation along with a formal definition of the ITM task. Let $D = \{(i_n, c_n)\}_{n=1}^N$ be the training set of image and caption pairs. These pairs are further divided into positive and negative samples where (i_p, c_p) are considered as positive samples while $(i_k, c_m)_{(k,m) \neq p}$ as negative samples. Then, the embedded images and captions are represented as $e_{c_p} = \sigma_c(c_p)$ and $e_{i_p} = \sigma_i(i_p)$ where σ_c, σ_i are embedding functions for captions and images respectively. Given a similarity function ψ , the classic formulation of the triplet loss in ITM [51], L_T , is defined as:

$$L_T = \max[\alpha + \psi(e_{i_p}, e_{c_m}) - \psi(e_{i_p}, e_{c_p}), 0] + \max[\alpha + \psi(e_{i_k}, e_{c_p}) - \psi(e_{i_p}, e_{c_p}), 0] \quad (8.6)$$

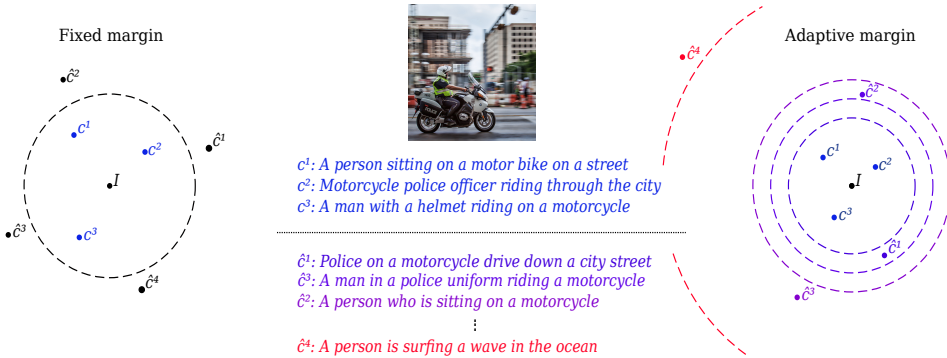


Figure 8.2: Comparison of a fixed margin loss function (left) and our adaptive margin (right). We consider an image anchor I , their positive sentences according to the ground truth (c^1, c^2, c^3), and four other sentences ($\hat{c}^1, \hat{c}^2, \hat{c}^3, \hat{c}^4$) that are negative according to the ground-truth but have some degree of semantic similarity with I . In our method, we dynamically adapt the margin of each possible triplet (anchor, positive, and negative items) to the value given by a similarity function ϕ that measures the semantic similarity of positive and negative items. In this Figure, we represent the similarity of sentences with the anchor by its color (the bluish the more semantically similar they are, the reddish the less similar).

where α is known as the margin. The intuition behind the triplet formulation is that given an n -sphere with radius α , positive samples should be projected inside and negative samples on the external region of the n -sphere. This can be observed in the left section of Figure 8.2. It is important to remark that the margin employed in the triplet loss is fixed despite the relatedness of hard-negative pairs.

8.3.1 SAM Formulation

Even though a fixed margin might be acceptable in image-to-image metric learning tasks, a fixed margin can not capture the continuum of language properly. Looking at the right on Figure 8.2, we can acknowledge that even the non-GT items can properly explain the provided image. Therefore, using a fixed margin and treating every negative as equal is unfeasible if the semantics is to be modeled properly. Due to this fact, creating an adaptive margin is imperative to teach our models the continuous nature of language.

Consequently, we formulate the Semantic Adaptive Margin (SAM) to dynamically calculate the similarity between images and sentences. More formally, given a positive pair (i_p^+, c_p^+) with negative samples $(i_l^-, c_l^-), (i_m^-, c_m^-)$, we use the ground truth caption

set G_p to calculate the triplet loss by incorporating a SAM (L_{SAM}):

$$\begin{aligned}
 L_{SAM} = & \max[\alpha_{i2t} + \psi(e_{i_p^+}, e_{c_m^-}) - \psi(e_{i_p^+}, e_{c_p^+}), 0] + \\
 & \max[\alpha_{i2i} + \psi(e_{i_l^-}, e_{c_p^+}) - \psi(e_{i_p^+}, e_{c_p^+}), 0] \\
 \alpha_{i2t} = & (\phi(G_p, c_p^+) - \phi(G_p, c_l^-)) / \tau \\
 \alpha_{i2i} = & (\phi(G_p, c_p^+) - \phi(G_p, c_m^-)) / \tau
 \end{aligned} \tag{8.7}$$

where ψ is a similarity function such as cosine similarity, ϕ stands for an aforementioned captioning metric (SPICE or CIDEr) and τ is a temperature parameter to be controlled on how wide or small the margin is desired. In other words, τ is used as a scaling factor. In essence, if c_l^- (a negative caption) is close to G_p then α_{i2t} will be lower and when it is farther away, the margin will be higher. As it can be appreciated by Equation 8.7, we incorporate a SAM into the original triplet formulation, which assigns a unique margin value specific to each sampled pair. SAM still can be optimized jointly with the original triplet formulation.

8.4 Experiments

In this section, we present the results obtained by evaluating state-of-the-art models with and without the adoption of the proposed SAM. Section 8.4.2 shows the performance of state-of-the-art methods evaluated on the introduced Semantic Recall metric. In Section 8.4.3 we present the significantly better performance achieved at retrieval when using considerably less training data compared to current state-of-the-art models. Section 8.4.4 showcases several state-of-the-art models with and without the adoption of our adaptive margin formulation. Finally, Section 8.4.5 presents the effects of employing the original triplet formulation, different values of a temperature parameter τ , and different sampling strategies.

In all our experiments, we employ publicly available code from the authors and train the models from scratch according to the original strategy and hyper-parameters. In order to perform a fair comparison, we do not use ensembles in our experiments. We employed CIDEr to assess the similarity between samples at training time (ϕ). With the purpose of avoiding training and evaluating in similar metrics, we employ SPICE when NSC is used as an evaluation metric.

8.4.1 Implementation Details

In this section, we describe the hyper-parameters and the training procedure used to obtain the models shown in the main paper for the reduced data scenario (Table 2 - main paper) and the state-of-the-art comparison (Table 3 - main paper). Specifically, for each model, we employ the training procedure described in the original paper. Each

of the models use the 36 most confident regions obtained by a Faster R-CNN [59] from the object detector proposed by [7]. The visual features used by each model are the same in all cases. The best model is selected according to the sum of NCS with CIDEr as a similarity metric on the validation set. It is important to note that NCS with SPICE is the one we employ for evaluation.

For the model VSRN [111] + SAM, we start with the public pretrained weights. We add our SAM loss function and re-train the model for 30 epochs. For Flickr30K we employ a soft negative sampling, with a temperature parameter $\tau = 3$ and a weight on the SAM triplet loss of 20. When training on MSCOCO, we employ a soft negative sampling, with a temperature parameter $\tau = 10$ and a weight on the SAM triplet loss of 5. In both datasets, the original triplet loss is kept alongside with our SAM formulation. As in the original paper, the word embedding size is 300-d and the dimension of the final joint embedding space is 2048-d. The mini-batch size employed is 128. At training, the Adam optimizer [99] is used. The original triplet loss margin is 0.2. The model is trained for 30 epochs, with a learning rate of 0.0002 for the initial 15 epochs, and is divided by 10 for the remaining 15 epochs.

In the CVSE [203] + SAM model, we train the model from scratch alongside the proposed SAM loss function as follows. In Flickr30K, we employ a random sampling strategy with a temperature parameter $\tau = 7$ and a weight on the SAM triplet loss of 5 alongside with the original triplet. In MSCOCO, we employ a soft negative sampling strategy, with a temperature parameter $\tau = 5$ and a weight on the SAM triplet loss set to 5, and only our SAM triplet loss is used. Following the original CVSE model, the joint space dimension is 1024-d. The consensus exploitation is performed with a 300-d GloVe [151] representation. The loss formulation contains the following weights for each term are kept, $\lambda_1 = 3, \lambda_2 = 5, \lambda_3 = 1, \lambda_4 = 2$. The mini-batch size employed is 128. At training, the Adam optimizer [99] is used. The original triplet loss margin is 0.2. The model is trained for 30 epochs, with a learning rate of 0.0002 for the initial 15 epochs and is divided by 10 for the remaining 15 epochs.

Finally, for the SGR [45] + SAM model, we train the model from scratch in both datasets. In Flickr30K, we employ a random sampling strategy, with a temperature parameter $\tau = 10$ and a weight on the SAM triplet loss of 10. In MSCOCO, we use a random sampling strategy, with a temperature parameter $\tau = 5$ and a weight on the SAM triplet loss of 5. In both datasets, the original triplet is kept alongside with our SAM triplet. Training of the original model is performed as described by the authors. The word embedding size is 300-d and the number of hidden states is 1024-d. The dimension of the similarity representation is 256. The original triplet loss margin is 0.2. The number of reasoning steps is 3. The initial learning rate is set to 0.0002 for 10 epochs and is decreased by a 10 on the final 10 epochs on MSCOCO. For Flickr30K, the initial learning rate is kept for 30 epochs and it decays by 0.1 for the next 10 epochs.

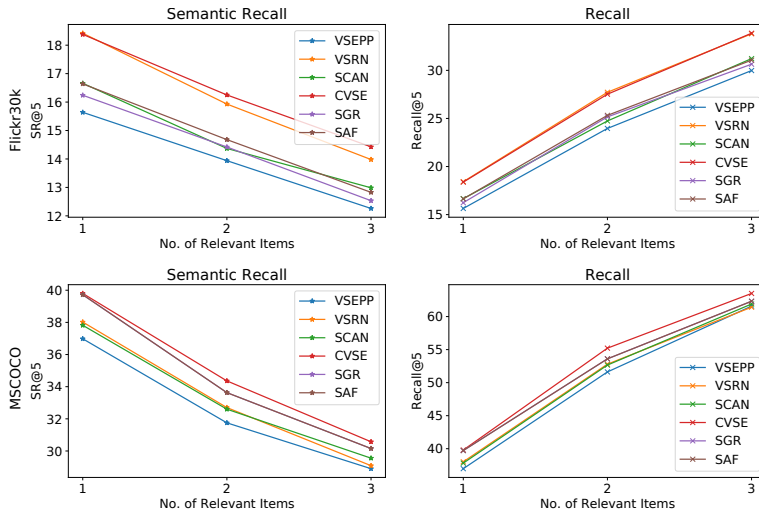


Figure 8.3: Text-to-Image Top-5 retrieved results evaluated with Recall and the presented Semantic Recall for Non-GT items.

8.4.2 Insights on State-of-the-Art Retrieval

In this section, we compare the behavior of existing systems by evaluating them on the newly proposed metrics. We evaluate the following state-of-the-art models: VSE++ [51], SCAN [109], VSRN [111], CVSE [203], SGR and SAF [45]. The experiment depicting the top-5 text-to-image retrieval scores for non-ground truth relevant items is shown in Figure 8.3. The scores shown are in terms of Recall and Semantic Recall at a cut-off point 5.

It is worth noting in Figure 8.3 that according to the recall ($R@5$), the models have a steady raise in recall scores as the number of relevant images m increases. However, the opposite effect is found when the models are evaluated with the previously introduced Semantic Recall (SR) formulation. The behavior of the models according to these two metrics seems to have an inversely proportional relation. The reason is due to the different definition between R and R^V . Merely evaluating the models on the first correctly retrieved item does not provide a complete landscape of its performance. Instead, our formulation shows that the models tend to have a decreasing score when more relevant items are considered. Furthermore, we observe that the big difference in numbers between models seems to diminish when we increase the relevant items for both metrics. Our conclusion is that the performance boost we obtain in the literature is not reflected well on the non-GT relevant items, suggesting that the generalization power of the models is overestimated.

Method		Recall						Normalized Cumulative Semantic Score							
		I2T			T2I			I2T			T2I			Nsum	
		R@1	R@5	R@10	R@1	R@5	R@10	Rsum	N@1	N@5	N@10	N@1	N@5		N@10
F-10%	CVSE [203]†	13.0	16.2	23.9	12.5	30.9	42.2	138.7	19.0	24.4	28.3	25.7	34.0	36.3	167.6
	CVSE+SAM	34.6	60.4	70.9	23.9	50.6	62.8	303.2	37.4	40.2	44.1	38.6	46.3	47.1	253.6
	SGR [45]†	0.3	0.7	1.3	0.2	0.6	1.1	4.2	2.8	4.2	5.7	5.7	9.7	11.7	39.8
	SGR+SAM	37.9	64.8	77.5	26.6	53.4	64.4	324.6	40.1	41.0	44.6	41.4	48.1	48.5	263.7
F-25%	CVSE [203]†	30.4	47.5	59.6	28.8	58.1	69.7	294.1	30.8	39.9	43.7	42.4	48.0	47.7	252.5
	CVSE+SAM	48.7	73.2	81.8	37.9	66.2	76.0	383.8	46.2	49.7	52.4	51.0	52.7	50.9	302.8
	SGR [45]†	11.0	29.3	40.0	7.3	21.4	31.6	140.6	20.3	24.1	29.7	22.7	33.4	37.3	167.6
	SGR+SAM	54.0	81.4	87.8	41.3	68.0	77.5	410.0	52.7	52.5	53.8	54.2	53.9	51.9	319.0
C-3%	CVSE [203]†	28.2	49.0	63.6	24.9	58.5	74.4	298.6	34.9	40.6	43.3	44.6	57.5	59.8	280.9
	CVSE+SAM	41.1	71.8	82.1	32.2	65.4	77.2	369.8	47.2	49.5	51.6	50.6	59.8	59.8	318.5
	SGR [45]†	0.2	1.1	2.1	0.1	0.6	1.2	5.3	4.1	5.1	5.9	4.2	6.7	8.1	34.2
	SGR+SAM	23.7	59.0	74.3	24.9	56.4	72.3	310.6	33.8	39.2	43.5	45.3	58.6	61.7	282.1
C-5%	CVSE [203]†	48.5	77.8	85.7	36.2	69.7	81.6	399.5	51.4	53.8	54.8	54.5	62.7	62.3	339.7
	CVSE+SAM	48.6	77.3	86.5	37.9	71.1	82.6	404.0	50.7	54.2	55.4	55.9	62.9	62.3	341.5
	SGR [45]†	1.0	3.0	5.3	0.2	0.5	1.3	11.3	6.9	7.9	9.0	2.9	4.2	5.7	36.6
	SGR+SAM	30.4	62.6	79.1	29.4	63.1	77.6	342.2	39.2	42.6	46.9	49.6	61.3	63.6	303.3

Table 8.2: Quantitative results on reduced training data samples. The acronyms used in the first column stand for Flickr30K (*F*), MSCOCO 1K (*C*). The (%) denotes the proportion of the training data used in relation to the original dataset size. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K). The † depicts that models are trained with the publicly available code released by the original authors.

8.4.3 Reduced Data Scenario

Further investigating our suggested formulation, we experiment on the low data regimen. We hypothesize that our adaptive margin formulation based on CIDEr is better equipped to deal with scarce training data scenarios, as it can better exploit the semantics over the whole data. More explicitly, we set aside a similar proportion of training samples from Flickr30k (29,000) and MSCOCO (113,287). In Flickr30K we employed 10% and 25% of the training set, resulting in 2,900 and 7,250 samples respectively. In the case of MSCOCO we employed 3% and 5% of the training set, thus yielding 3,398 and 5,664 data points. We evaluate all the models on the standard 1K test set split of each dataset. We employ two state-of-the-art methods, CVSE [203] and SGR[45] for experimentation. Similar to the previous section, all the experiments are performed with publicly available code as described by the authors disregarding the adoption or not of our formulation. The results of these experiments can be found in Table 8.2.

In the 10% data scenario of Flickr30k, CVSE with SAM achieves almost 3 times the performance when compared to the original model. It is paramount to note that by the adoption of our formulation SGR achieves an enormous improvement. On the other hand, the original SGR model is barely capable to learn useful information due to the bigger number of parameters compared to CVSE.

As more data is used on each scenario, the original models tend to improve in performance and the retrieval gap decreases. Results in Flickr30k tend to be stronger when adopting the proposed SAM. This is due to the significantly higher descriptive nature of

captions found in Flickr30K training dataset compared to the less granular ones found in MSCOCO.

The significant improvement in scarce training data also translates into an increased rate of convergence. By employing an adaptive margin with CIDEr, a model exploits a strong guiding cue about the semantic space to be learned according to weighted n-gram statistics.

8.4.4 Comparison with State-of-the-Art

The results obtained by comparing state-of-the-art methods with and without our formulated SAM are shown in Table 8.3. First, by incorporating SAM, calculated from an image captioning metric into a state-of-the-art pipeline, a boost in the recall is obtained. A similar effect is achieved in most of the models when they are also evaluated with the proposed NCS metric. Second, both depicted metrics have a strong degree of correlation, however, obtaining an improvement in recall does not necessarily translate into an increase in NCS. This effect can be observed in particular with the MSCOCO dataset with the SGR model.

This is due to the fact that Recall and NCS are inherently different metrics that provide complementary information. Recall shows how well a model ranks a single image or sentence labeled as relevant. Whereas, the NCS shows what is the degree of semantics captured by a model at a cut-off point k . Therefore, an increase in Recall or NCS should not necessarily be treated as equally significant.

Third, it is evident that a greater improvement is achieved on Flickr30k than in MSCOCO. All of our models on Flickr30k perform better than the baselines on every metric, while in MSCOCO, the boost attained is more conservative. The reason is that captions of Flickr30k are more detailed and longer compared to the ones in MSCOCO, which are shorter and less specific. This difference in the nature of the captions allows CIDEr to provide a more precise and discriminative margin per sample in intricate captions, due to the CIDEr formulation which relies on a weighted tf-idf n-gram matching.

Finally, it is important to note that while the Recall score increases as the cut-off point increases, in our proposed NCS metric this behaviour is not present. The NCS shows the normalized capability of a model to capture the greatest amount of semantic similarity on a specific cut-off point.

8.4.5 Effect of Temperature and Sampling

In this section, we study the effect of the temperature parameter τ , sampling techniques, and whether the original triplet is kept or only a SAM is employed. Several sampling techniques are explored to find the negative items in our SAM formulation, namely random (RS), hard negative (HN), and soft negative (SN). In HN, the negative item in each triplet is selected as the closest to the anchor in a batch [51]. We refer to

Flickr30K														
Method	Recall						Normalized Cumulative Semantic Score							
	I2T			T2I			Rsum	I2T			T2I			Nsum
	R@1	R@5	R@10	R@1	R@5	R@10		N@1	N@5	N@10	N@1	N@5	N@10	
VSRN[111]†	68.1	88.4	93.9	51.6	78.3	85.8	466.1	60.3	62.9	62.9	63.2	58.8	55.0	363.4
VSRN+SAM	68.4	89.7	94.8	52.4	78.7	86.6	470.6	60.2	62.7	63.1	64.1	59.4	55.7	365.2
CVSE [203]†	68.6	87.7	92.7	53.2	81.1	88.3	471.6	59.0	63.5	63.1	64.1	59.6	55.5	364.9
CVSE + SAM	70.0	89.2	93.1	55.0	82.6	89.0	478.9	59.6	64.6	64.2	65.5	59.8	55.3	387.9
SGR[45]†	74.4	92.9	96.3	55.8	81.1	87.9	488.4	68.1	65.6	63.8	66.0	58.6	54.5	376.7
SGR+SAM	75.9	92.4	96.6	57.6	83.1	89.7	495.3	69.4	66.2	64.0	67.5	59.2	55.0	381.4

MSCOCO 1K														
Method	Recall						Normalized Cumulative Semantic Score							
	I2T			T2I			Rsum	I2T			T2I			Nsum
	R@1	R@5	R@10	R@1	R@5	R@10		N@1	N@5	N@10	N@1	N@5	N@10	
VSRN[111]†	72.4	94.7	97.8	61.2	89.3	94.9	510.3	68.3	72.1	68.2	74.4	71.1	66.4	420.6
VSRN+SAM	74.6	93.6	97.5	61.5	89.6	94.9	511.7	69.3	72.2	68.1	74.5	70.9	66.5	421.5
CVSE[203]†	77.0	94.2	97.3	64.3	91.1	95.9	519.8	69.7	73.3	69.3	76.2	71.4	67.1	427.2
CVSE+SAM	79.8	95.1	97.7	67.0	93.0	97.3	529.9	71.8	76.3	71.0	78.6	72.9	69.1	439.6
SGR[45]†	79.9	97.4	98.3	63.2	90.5	95.4	524.7	74.7	73.1	67.9	76.1	70.7	67.2	429.9
SGR+SAM	80.7	97.2	98.6	63.8	90.5	95.9	526.7	73.2	72.9	67.8	76.2	70.9	67.4	428.5

MSCOCO 5K														
Method	Recall						Normalized Cumulative Semantic Score							
	I2T			T2I			Rsum	I2T			T2I			Nsum
	R@1	R@5	R@10	R@1	R@5	R@10		N@1	N@5	N@10	N@1	N@5	N@10	
VSRN[111]†	48.4	78.9	87.9	37.2	67.9	79.6	399.9	55.8	58.6	61.2	60.2	63.4	62.8	362.2
VSRN+SAM	49.1	79.0	87.4	37.5	68.1	79.5	400.6	56.4	58.8	61.7	60.6	63.5	62.9	363.9
CVSE[203]†	53.1	79.6	88.0	40.5	72.2	83.1	416.5	57.1	61.1	63.2	62.2	64.4	63.4	371.4
CVSE+SAM	56.4	82.4	90.1	42.3	73.9	84.5	429.6	59.2	63.0	64.5	63.8	65.3	64.4	380.2
SGR[45]†	56.0	83.3	90.7	40.1	69.3	80.2	419.6	60.4	59.1	60.4	61.7	62.5	61.6	366.0
SGR+SAM	55.7	83.2	91.2	40.5	69.7	80.5	420.8	59.5	59.3	60.4	62.0	62.4	61.4	365.0

Table 8.3: Comparison of retrieval results of the original VSRN, CVSE and SGR models with and without the proposed SAM. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K). The column Rsum and Nsum is the summation of the overall retrieval scores in image-to-text and text-to-image for Recall and NCS respectively. The † depicts that models are trained with the publicly available code released by the original authors.

random sampling when a negative item is randomly picked in a batch. SN refers to picking the furthest negative item to the anchor within the batch. We investigate the effect of these parameters by employing CVSE [203] model as a baseline. The majority of the best performing models obtained were by employing a Soft Negative (SN) sampling, thus we provide the results on both datasets in Table 8.4. The results of the effect of Random Sampling (RS) and Hard Negative (HN) sampling in Flickr30k are shown in Table 8.5. In both tables, we provide the sum of the Recall and NCS metrics at the top 1, 5, and 10 in image-to-text and text-to-image scenarios. When the NCS is employed, we show two variations. One, by preserving the GT images labeled as relevant, and the second one by removing only the GT images, denoted with the acronym N in Table 8.4 and 8.5.

Initially, it is important to notice that improvements on the recall score do not necessarily go in hand with better scores at NCS. This can be seen in the MSCOCO-1K results between the fifth row and the first row in Table 8.4. In these experiments, we obtain a score of 521 on Rsum and 429 on NCS sum in the fifth row. Comparing it to

τ	S	T	F30K			MSCOCO-1K		
			Nsum	Nsum(N)	Rsum	Nsum	Nsum(N)	Rsum
3	SN	✓	371.29	257.42	479.1	429.41	312.2	517.9
3	SN	✗	369.73	258.12	476.5	427.32	313.06	515
5	SN	✓	369.97	257.21	477.5	429.74	311.53	520.5
5	SN	✗	371.31	258.75	478.2	428.2	312.89	518
10	SN	✓	369.74	257.64	477.2	429.09	306.51	521.3
10	SN	✗	370.17	257.85	475.8	429.88	309.27	520.6

Table 8.4: Experiments of the effect of (τ), soft negative (SN) sampling, and the whether the original triplet is kept (✓) or only our formulation is employed (✗). The acronym Nsum(N) refers that GT elements have been removed.

τ	S	T	F30K		
			Nsum	Nsum(N)	Rsum
3	RS	✓	355.87	257.93	460.1
3	RS	✗	344.44	257.67	441.7
5	RS	✓	367.02	258.25	473.4
5	RS	✗	363.27	259.83	468.3
10	RS	✓	370.09	257.42	478.7
10	RS	✗	365.72	258.82	471.2
3	HN	✓	338.94	249.04	435.1
3	HN	✗	344.19	258.27	439.5
5	HN	✓	369.94	257.97	478.8
5	HN	✗	351.68	257.64	450.4
10	HN	✓	369.04	256.85	477.2
10	HN	✗	351.18	257.83	448.6

Table 8.5: Experiments of the effect of (τ), random (RS), and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). The acronym Nsum(N) refers that GT elements have been removed.

the first row, there is a 4% drop in recall, however, the score of 429 remains on NCS. Although NCS and Recall are correlated, they provide different information about our models.

In general, we obtain our best NCS scores when the temperature parameter τ is increased to 10. The smaller margin gives the model more freedom in shaping the space on which to project the data points. By increasing the margin, we restrict the models on where to project the positive and negative samples, resulting in a drop in NCS and Recall. However, a trade-off between NCS scores on GT and non-GT items exists. The increase in the margin (lower values in τ) seems to improve the results on non-GT items, this is especially evident on MSCOCO. We discover that on average, we obtain the best results with SN. Regarding the usage of the original triplet formulation, we notice that it is complementary to SAM since each one focuses on learning a different task. The hard negative focuses solely on GT samples, while SAM learns to measure the degree of similarity.

8.4.6 Dwelving into Temperature and Sampling

In this section, we extend the results of the effect of setting different temperature values τ and sampling strategies. In all the experiments in this section, CVSE [203] is used. We evaluate the impact of these parameters on Flickr30K on Table 8.6 and removing GT items (non-GT) on Table 8.7. When calculating the results on MSCOCO 1K, we keep the soft negative (SN) sampling strategy while the impact of different values of τ and the usage of the original triplet is measured. Table 8.8 shows the results with the inclusion of GT and Table 8.9 depicts the results obtained with non-GT relevant items.

τ	S	T	Recall					Normalized Cumulative Semantic Recall								
			I2T			T2I		Rsum	I2T			T2I			Nsum	
			R@1	R@5	R@10	R@1	R@5		R@10	N@1	N@5	N@10	N@1	N@5		N@10
3	SN	✓	69.8	87.8	93.1	55.5	83.2	89.7	479.1	59.6	65.4	64.6	65.9	60.1	55.8	371.3
3	SN	✗	70.0	88.1	92.6	54.1	82.4	89.3	476.5	60.0	64.7	64.5	64.8	60.0	55.8	369.7
5	SN	✓	69.7	88.4	93.1	54.5	82.4	89.4	477.5	59.9	65.2	64.4	65.2	59.7	55.5	370.0
5	SN	✗	70.4	88.2	92.5	55.1	82.6	89.4	478.2	60.3	65.6	64.3	65.5	59.8	55.9	371.3
10	SN	✓	69.6	89.1	92.3	54.9	82.2	89.1	477.2	60.0	65.4	64.2	65.3	59.6	55.3	369.7
10	SN	✗	70.6	86.5	92.9	54.1	82.3	89.4	475.8	60.9	65.1	64.3	64.7	59.6	55.7	370.2
3	RS	✓	65.4	85.7	91.6	50.9	79.1	87.4	460.1	56.3	61.1	61.5	62.1	59.1	55.9	355.9
3	RS	✗	60.9	84.3	89.4	46.9	76.0	84.2	441.7	53.8	58.9	59.7	59.0	58.0	55.1	344.4
5	RS	✓	68.3	88.1	92.7	54.0	81.5	88.8	473.4	59.1	63.9	63.8	64.7	59.6	56.0	367.0
5	RS	✗	67.5	87.4	92.1	52.9	80.3	88.1	468.3	58.1	63.0	63.0	63.8	59.4	56.0	363.3
10	RS	✓	70.7	87.8	93.7	54.7	82.4	89.4	478.7	60.9	64.8	64.0	65.2	59.8	55.4	370.1
10	RS	✗	68.0	87.9	93.0	52.8	81.2	88.3	471.2	59.5	63.9	63.5	63.8	59.5	55.5	365.7
3	HN	✓	59.6	81.4	89.3	44.9	75.6	84.3	435.1	54.1	57.4	58.9	56.9	57.2	54.4	338.9
3	HN	✗	60.5	85.8	91.5	44.6	73.9	83.2	439.5	55.5	58.7	59.6	57.8	57.5	55.0	344.2
5	HN	✓	71.1	88.7	92.7	54.4	82.6	89.3	478.8	60.6	64.6	64.1	65.1	59.8	55.7	369.9
5	HN	✗	62.9	86.3	92.7	47.4	76.7	84.4	450.4	57.6	60.1	60.9	59.6	58.2	55.3	351.7
10	HN	✓	70.0	88.3	92.6	54.7	82.4	89.2	477.2	60.2	64.5	64.0	65.1	59.8	55.5	369.0
10	HN	✗	63.5	85.1	92.0	47.2	76.3	84.5	448.6	57.2	60.6	60.8	59.5	58.1	55.0	351.2

Table 8.6: Experiments on Flickr30K regarding the effect of (τ), soft (SN), random (RS), and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[203] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

8.4.7 Qualitative Samples for the Reduced Data Scenario

In this section, we provide qualitative samples on image-to-text and text-to-image in Flickr30K and MSCOCO 1K coming from the reduced data scenario by only using 10% of the training set in Flickr30K and 5% in MSCOCO. To offer additional insights, we provide the Recall (R^v) and NCS per sample. It is evident from Tables 8.10, 8.11 and Figures 8.4, 8.5, 8.6 and 8.7, that the incorporation of the proposed SAM improves not only the standard Recall metric but also the semantics of retrieved non-GT items on both scenarios, image-to-text and text-to-image.

τ	S	T	Normalized Cumulative Semantic Recall (Non-GT)								
			I2T			T2I			Nsum		
			N@1	N@5	N@10	N@1	N@5	N@10			
3	SN	✓	40.9	42.4	43.1	42.7	44.1	44.4	257.5		
3	SN	✗	41.2	42.3	43.1	42.8	44.2	44.5	258.1		
5	SN	✓	41.4	42.3	42.8	42.8	43.9	44.1	257.2		
5	SN	✗	41.5	42.6	43.4	42.7	44.0	44.6	258.8		
10	SN	✓	42.3	42.3	42.8	42.6	43.7	43.9	257.6		
10	SN	✗	41.3	42.5	43.0	42.7	44.0	44.4	257.8		
3	RS	✓	40.4	42.1	43.1	42.6	44.6	45.1	257.9		
3	RS	✗	40.7	42.2	43.0	42.4	44.4	45.0	257.7		
5	RS	✓	40.9	42.6	43.1	42.5	44.2	45.0	258.3		
5	RS	✗	41.0	43.1	43.5	42.6	44.5	45.1	259.8		
10	RS	✓	41.7	42.6	42.8	42.6	43.8	43.9	257.4		
10	RS	✗	41.5	43.0	43.3	42.5	44.1	44.4	258.8		
3	HN	✓	38.4	40.5	41.4	41.1	43.5	44.3	249.0		
3	HN	✗	41.2	41.9	42.7	42.8	44.4	45.3	258.3		
5	HN	✓	41.2	42.4	43.0	42.9	44.0	44.4	258.0		
5	HN	✗	40.1	42.3	43.0	42.6	44.4	45.2	257.7		
10	HN	✓	41.0	42.4	43.0	42.3	43.9	44.1	256.8		
10	HN	✗	41.2	42.6	43.0	42.1	44.3	44.6	257.8		

Table 8.7: Experiments on Flickr30K (Non-GT) regarding the effect of (τ), soft (SN), random (RS) and hard negative (HN) sampling. The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all the experiments shown, CVSE[203] is employed. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

τ	S	T	Recall						Normalized Cumulative Semantic Recall							
			I2T			T2I			I2T			T2I				
			R@1	R@5	R@10	R@1	R@5	R@10	Rsum	N@1	N@5	N@10	N@1	N@5	N@10	Nsum
3	SN	✓	76.2	93.7	96.3	63.6	91.7	96.4	517.9	69.2	73.7	69.7	75.8	72.5	68.6	429.4
3	SN	✗	74.8	93.5	96.8	62.2	91.5	96.2	515.0	68.3	73.2	69.4	74.8	72.7	69.0	427.3
5	SN	✓	76.4	94.0	97.3	64.2	92.2	96.4	520.5	68.8	74.1	70.0	76.1	72.5	68.4	429.7
5	SN	✗	76.1	93.8	96.9	63.2	91.4	96.6	518.0	68.2	73.6	69.6	75.5	72.4	68.9	428.2
10	SN	✓	76.9	94.2	97.7	64.4	91.8	96.3	521.3	69.9	73.8	69.6	76.3	71.9	67.7	429.1
10	SN	✗	76.8	94.2	97.4	64.2	91.8	96.2	520.6	70.1	73.7	69.7	76.3	72.1	68.0	429.9

Table 8.8: Experiments on MSCOCO 1K regarding the effect of (τ), employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all our experiments, we employ CVSE[203]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).

τ	S	T	Normalized Cumulative Semantic Recall (Non-GT)						
			I2T			T2I			Nsum
			N@1	N@5	N@10	N@1	N@5	N@10	
3	SN	✓	45.4	46.0	46.0	56.9	58.9	59.0	312.2
3	SN	✗	45.2	46.2	46.2	56.9	59.1	59.5	313.1
5	SN	✓	45.4	46.2	46.2	56.6	58.6	58.6	311.5
5	SN	✗	45.3	46.2	46.2	56.9	59.1	59.3	312.9
10	SN	✓	44.1	45.2	45.1	56.6	57.8	57.7	306.5
10	SN	✗	44.4	45.7	45.5	57.1	58.3	58.2	309.3

Table 8.9: Experiments on MSCOCO 1K (Non-GT) regarding the effect of (τ), employing a soft negative sampling strategy (SN). The third column (T) shows whether the original triplet is kept (✓) or only our formulation is employed (✗). For all our experiments, we employ CVSE[203]. Results are depicted in terms of Recall@K (R@K) and Normalized Cumulative Semantic Score (N@K).


	<p>Ground Truth: Three zebras and other wild animals out in a semi-green field. Three zebras and two other animals grazing. Wildlife standing near water area in natural setting. Three zebras near the shore line of a body of water. A group of animals stand next to a watering hole.</p>
CVSE + SAM	<p>Three zebras near the shore line of a body of water. $R_i : 1, S_i : 0.46$ A heard of zebra on the plains at a watering hole. $R_i : 0, S_i : 0.21$ A group of animals stand next to a watering hole. $R_i : 1, S_i : 0.33$ A group of zebras and birds are gathered around water. $R_i : 0, S_i : 0.17$ There is a herd of zebras standing around. $R_i : 0, S_i : 0.06$ $R_i@5 = 1, N_i@5 = 0.61$</p>
CVSE	<p>Three zebras near the shore line of a body of water. $R_i : 1, S_i : 0.46$ There is a herd of zebras standing around. $R_i : 0, S_i : 0.06$ A heard of zebra on the plains at a watering hole. $R_i : 0, S_i : 0.21$ There are several zebras grazing near the water as a bird flies over them. $R_i : 0, S_i : 0.05$ A group of zebras and birds are gathered around water. $R_i : 0, S_i : 0.17$ $R_i@5 = 1, N_i@5 = 0.47$</p>
SGR + SAM	<p>Three zebras near the shore line of a body of water. $R_i : 1, S_i : 0.46$ Two zebras fighting in a cloud of dust. $R_i : 0, S_i : 0.05$ Three zebra in the middle of a field with a body of water in the distance. $R_i : 0, S_i : 0.24$ Three zebras standing in a sandy desert area. $R_i : 0, S_i : 0.17$ Three zebras and other wild animals out in a semi-green field. $R_i : 1, S_i : 0.42$ $R_i@5 = 1, N_i@5 = 0.67$</p>
SGR	<p>A zebra grazing on long dry grass in a field. $R_i : 0, S_i : 0.05$ Four zebras are grazing at a nature reserve. $R_i : 0, S_i : 0.06$ A group of animals stand next to a watering hole. $R_i : 0, S_i : 0.06$ Three zebras standing in a sandy desert area. $R_i : 0, S_i : 0.17$ The small herd of sheep are grazing on the grassy field. $R_i : 0, S_i : 0.05$ $R_i@5 = 0, N_i@5 = 0.19$</p>

Table 8.10: Image-to-Text qualitative results in MSCOCO 1K. The initial row depicts the queried image and the associated ground truth captions. Each retrieved caption shows the Recall (R_i) and SPICE (S_i) score when compared with the GT captions. Each sample showcases the final per sample Recall ($R_i@5$) and NCS ($N_i@5$) score obtained. Bolded captions represent the correctly retrieved ground truth items.


	<p>Ground Truth: Two children and a woman are sitting on a sofa, one of the children has a camera. Three Asian children sitting on a couch with tapestries hanging in the background. An Asian woman and her two children sit at a table doing crafts. A woman in a red shirt sitting with two young girls in dresses. Three young girls in arts and crafts room.</p>
<p>CVSE + SAM</p>	<p>Two children and a woman are sitting on a sofa, one of the children has a camera. $R_i : 1, S_i : 0.39$ An Asian woman and her two children sit at a table doing crafts. $R_i : 1, S_i : 0.43$ Three Asian children sitting on a couch with tapestries hanging in the background. $R_i : 1, S_i : 0.39$ A woman and three children are in a room full of toys. $R_i : 0, S_i : 0.18$ A group of children sitting on the floor, eating snacks at school $R_i : 0, S_i : 0.05$ $R_i@5 = 1, N_i@5 = 0.72$</p>
<p>CVSE</p>	<p>Three college-age women sit in upholstered chairs. $R_i : 0, S_i : 0.05$ Three young women face each other while sitting on red plush chairs. $R_i : 0, S_i : 0.04$ A plat is sitting on the floor next to a blond girl. $R_i : 0, S_i : 0.05$ Three girls talking in a lobby. $R_i : 0, S_i : 0.10$ Two kids sitting at a table eating. $R_i : 0, S_i : 0.17$ $R_i@5 = 0, N_i@5 = 0.16$</p>
<p>SGR + SAM</p>	<p>Three Asian children sitting on a couch with tapestries hanging in the background. $R_i : 1, S_i : 0.39$ An Asian woman and her two children sit at a table doing crafts. $R_i : 1, S_i : 0.43$ Six children are sitting around taking notes together. $R_i : 0, S_i : 0.05$ Woman on four way seesaw with 2 kids. $R_i : 0, S_i : 0.14$ A group of mostly asian children sitting at cubicles in blue chairs. $R_i : 0, S_i : 0.09$ $R_i@5 = 1, N_i@5 = 0.54$</p>
<p>SGR</p>	<p>A child playing in the ocean. $R_i : 0, S_i : 0.05$ Construction workers deal with removing railroad tracks. $R_i : 0, S_i : 0.00$ A mural of children on a brick wall. $R_i : 0, S_i : 0.05$ Four people in the subway are having fun. $R_i : 0, S_i : 0.00$ Several elderly men are grouped around a table. $R_i : 0, S_i : 0.05$ $R_i@5 = 0, N_i@5 = 0.07$</p>

Table 8.11: Image-to-Text in Flickr30K. The first row shows the queried image and the GT captions. Metrics are Recall (R_i) and SPICE (S_i) score. Each row showcases the final per sample Recall ($R_i@5$) and NCS ($N_i@5$) score obtained.



Figure 8.4: MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).



Figure 8.5: MSCOCO 1K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).



Figure 8.6: Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).



Figure 8.7: Flickr30K text-to-image qualitative samples. Each retrieved image shows the SPICE (S_i) score when compared with the GT. Recall (R_i) is shown as green (1) or red (0) border on retrieved images. The final score per sample is presented in terms of Recall ($R_i@5$) and NCS ($N_i@5$).

8.5 Conclusion

In this Chapter, we highlight the challenges stemming from the lack of annotations in the task of image-text matching. Inspired by image captioning metrics, we present a formulation that addresses the many-to-many mapping problem between images and captions. The introduced metric, namely Normalized Cumulative Semantic Score (NCS), shows a higher degree of semantic correlation to human judgment compared to the standard Recall. Additionally, we show a comprehensive set of experiments that considers the usage of IC metrics to learn an adaptive margin. The incorporation of such margin yields a big improvement in scenarios when training data is scarce (e.g. semi-supervised learning), as well as increasing the semantics of the retrieved non-GT items.

Chapter 9

Conclusions and Future Directions

This chapter provides a summary of this thesis' contributions to the machine learning and computer vision domain, with a focus on exploiting scene text for a more holistic image understanding. We also point out the primary successes and shortcomings of the suggested methods from previous chapters. We direct the reader toward potential future study directions and logical expansions of the suggested approaches.

9.1 Conclusions

Modeling the rich and diverse information contained in scene text found in natural imagery is still a challenging task that intersects between scene text recognition and vision and language. In this PhD thesis, our goal was to exploit textual information and incorporate it into the tasks of fine-grained image classification and scene text and cross-modal retrieval pipelines. The final goal is to integrate this relatively new modality to obtain holistic computer vision models capable of a full image understanding. Following, we present the conclusions of each chapter on the studied tasks.

In Chapter 2, we introduced a novel approach to perform lexicon-free single-shot scene text retrieval in real time. The proposed model incorporates a hierarchical and morphological way of representing scene text given by PHOCs, which enables the network to learn how to construct out of vocabulary words unseen at training time. More specifically, our design choices customize an object detection model and adapt it to embed text to perform retrieval. Conducted experiments show that the suggested technique yields state-of-the-art performance while performing *significantly* faster than

previous methods. Moreover, our approach is able to outperform other approaches due to the capability of generalizing unseen text instances.

In Chapter 4, we identify the limitation of current approaches while employing semantic embeddings due to OCR mistakes. We devise a pipeline that employs the model introduced in Chapter 2 to represent scene text with a morphological embedding given by PHOC. We construct an image-level scene text descriptor, that clusters similar words in a common space, which is later used as features to perform scene text-based image classification and retrieval. We evaluate the importance of scene text as well as different fusion methods when undertaking these tasks. Ample experimentation shows state-of-the-art results of both tasks, showing the versatility of the constructed scene text descriptor.

In Chapter 5, we hypothesize that a model capable of reasoning about the interaction of regions within an image can provide rich information for the fine-grained image classification task. To meet this end, we employ a GCN that learns to reason about salient objects and scene text instances within an image. Interestingly, this approach maps common scene text instances and objects in a semantic manner to better perform the aforementioned task. Relations of the GCN nodes, given by edges show the degree of the reasoning and correlation among visual and textual regions. Finally, this method improves previous state-of-the-art by a great margin, thus allowing us to model the interaction of scene text in a more specific manner.

Chapter 7 contemplates the possibility of improving image-text cross-modal retrieval by leveraging scene text information. Due to the lack of exploration in current research, we propose the StacMR task and gather a dataset Coco-Text Captioned (CTC) which contains all the 3 modalities studied, images, captions, and scene text. We further propose two subsets, explicit and non-explicit splits according to the scene text occurrences in the captions. We present several benchmarks alongside two possible approaches designed for this task. Finally, we discover that a nontrivial gain can be obtained by employing scene text in current cross-modal pipelines, however, it should be used selectively to obtain fine-grained retrieval results. We conclude that more annotated data is needed to capture the complex interaction among all three modalities, especially the human caption-centered viewpoint when employing scene text.

In Chapter 8, we explore current approaches and metrics employed in the image-text matching literature. We found out that Recall only captures whether a retrieved item was annotated as relevant while discarding the rest of the ranked items. In order to obtain a more descriptive metric, we employ image captioning metrics to define a way to assess the semantic relevance of ranked results, namely Normalized Cumulative Semantic (NCS) score. Furthermore, we incorporate the semantic relatedness given by captioning metrics among images and captions to define a margin. The Semantic Adaptive Margin (SAM) is plugged-in a triplet loss, which yields a model that learns a smoother semantic space well-suited for retrieval. Extensive experimentation show: firstly, that older and current state-of-the-art models under-perform if GT elements are not included. Secondly, the proposed metric captures well the semantic similarity among query and retrieved results. Thirdly, a model trained with SAM im-

prove the results measured by the standard Recall and NCS. Finally, by employing our formulation, models converge to a suitable solution while employing several magnitudes of less data.

We conclude this research with two final ideas. First, the incorporation of the information that scene text conveys into different computer vision tasks was largely unexplored. Until recently, most computer vision models were illiterate and thanks to recent progress in text detection and recognition, textual features have been incorporated into novel and complex tasks that require holistic image understanding. We have solidly shown through extensive experimentation that integrating scene text is key in order to improve visio-linguistic applications.

Secondly, the incorporation of scene text as an additional modality is not straightforward. Despite several improvements in late research, scene text modeling often requires a context-specific perspective that encloses the interplay of three modalities. Moreover, the relevance of specific scene text instances relies on a human perspective that may consider consistent or not to a specific query. Despite such challenges, designing proper pipelines and gathering data proves to be paramount in order to achieve models that are able to reason and understand images in an “intelligent” manner.

9.2 Future Directions

In this section, we will further discuss the future research directions as well as the open challenges that remain when employing scene text as an additional source of information. Currently, the incorporation of high computational models along with the capability of scrapping bigger datasets than ever have allowed significant advances in machine learning, especially with the introduction of the Transformer [194] in Natural Language Processing (NLP). High-performing models such as BERT [43], GPT-3 [23], T5 [156] and their variants have allowed the incorporation of such models into vision and language tasks, despite the huge computational requirements, billions of parameters, and Petabytes of web crawling as in C4 (Colossal Clean Crawled Corpus) data. These advances have translated into computer vision, with a transformer variant, the ViT [47] and inspired variants [193, 192, 14, 46], that yield state-of-the-art in several vision tasks. Additionally, visio-linguistic models have emerged, such as CLIP [155], in which vision pipelines can benefit from language as a supervisory signal.

First, we believe that the usage of pre-trained language models can benefit all the tasks studied in this thesis, therefore opening a path for future research. On one hand, the previously mentioned high-performing language models, contain a well-defined pre-training stage. It has been shown that incorporating pre-trained models helps to boost performance in transfer learning tasks. On another hand, computer vision models can benefit from current NLP models and the emergent abilities [210] that incorporate few-shot learning, mathematics, words in context, and world-knowledge extraction. This performance boost has been seen in the Scene Text Visual Question Answering (STVQA) and Image Captioning tasks [218, 19], and cross-modal retrieval [32].

Therefore, successfully incorporating scene text as an additional modality requires a specific viewpoint that combines a human-centered and language-based perspective as well as world knowledge that can be obtained from pre-trained models.

Secondly, we believe that fully understanding the way deep learning works remains a goal far from being reached in the short term. To this end, explainable research can be undergone. Concretely, due to the complex synergy between language and vision, explainability can offer new insights into how computer models fuse the information of totally different information sources. Additionally, the incorporation of scene text can pave new methods that study such interaction by an analysis of attention maps in transformer-based models or by GCNs given the existing relationship among nodes through edges by crafted design. Learning these interactions, can well be inspired and provide intuitions on how humans grasp and reason about the world giving all the modalities that interplay to form mental representations.

Thirdly, as we showed in Chapter 8, current metrics in image and text cross-modal retrieval are not well suited to assess the real performance of a retrieval model. Due to the complexity of measuring semantic relations among languages, as in the case of image captioning, coming up with metrics for two modalities (image and text) comes at a magnitude higher complexity. Despite the major consensus in the research community on employing a specific metric, we firmly believe that metrics such as the proposed NCS should be incorporated and explored into further research. The resulting advantage of employing such an idea i.e. into a semantic margin yields major improvements while requiring significantly less labeled data, thus probably producing better few-shot learning models.

Additionally, we would like to point out that current models are computationally expensive, thus increasing the gap between institutions that can afford such technological requirements and academy that usually lags behind. On top of that, environmental impact is a must in order to preserve current limited resources and the effect it has on global warming. Considering the previous notions, we can opt for smaller and more efficient models. It has been shown that large language models are somehow inefficient [77], therefore the need for algorithms that model the interaction of different modalities in more efficient manners, namely, pre-training, distillation, network pruning, or knowledge retrieval to name a few. As a final remark, we would like to point out that incorporating scene text and the visual/semantic information that it entails comes as a natural progression of truly intelligent systems capable of efficient exploitation of available cues that could enrich and simulate human-like mental representations.

Appendix A

Appendix

A.1 Scene Text Aware Cross-Modal Retrieval

A.1.1 Introduction

In this document, we provide additional details about the proposed CTC dataset as well as experiments that offer more insights into the different re-ranking strategies and the proposed supervised model that we describe in the main dissertation.

A.1.2 Additions to Baselines and Re-Ranking

Full Table of Results on CTC

Table A.1 presents a more extensive version of the results presented in Section 5.1 from the main paper. This section dives into some parts of these results.

Scene-Text-only Baselines. Here we discuss additional scene-text baselines we applied to our task. As described in the main paper, we first experimented with the GRU (textual embedding) of the cross-modal models to describe the scene text and compare it to the captions. Their results are shown in Table A.1, rows (5-8). In contrast to the visual model, where VSRN consistently outperformed VSE++, for scene text the later performs better than the former. Models trained on Flickr30K + TextCaps also perform better than their counterparts trained on Flickr30K only.

We also experimented with training a GRU for a caption-to-scene-text retrieval in Flickr30K. We directly applied the training code of VSE++ to these two modalities (scene text and captions) and simulated the scene text of an image as the intersection between two of its captions. The results of this method, called GRU++, are presented in row (9).

	Scene-text Model	Train		Scene-text Source	Re-rank	CTC-1K						CTC-5K						
		F30K	TC			I2T			T2I			I2T			T2I			
						R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
(1)	VSE++	X	✓	X	-	-	20.5	42.8	54.5	15.4	35.2	48.4	13.3	30.2	40.2	8.4	21.5	30.1
(2)	VSE++	X	✓	✓	-	-	23.9	50.6	63.2	16.5	39.6	53.3	12.6	30.1	40.2	7.9	21.0	29.7
(3)	VSRN	X	✓	X	-	-	27.1	50.7	62.0	19.7	42.8	55.7	19.2	38.6	49.4	12.5	29.2	39.1
(4)	VSRN	X	✓	X	-	-	35.6	64.4	76.0	24.1	50.1	63.8	22.7	45.1	56.0	14.2	32.1	42.6
(5)	X	VSE++ GRU	✓	✓	GT	-	17.4	29.9	37.1	8.3	17.5	23.2	2.4	4.8	5.8	1.3	3.0	4.2
(5')	X	VSE++ GRU	✓	✓	OCR	-	12.4	21.7	26.0	6.5	14.5	18.9	1.9	3.6	4.4	1.1	2.6	3.6
(6)	X	VSE++ GRU	✓	✓	GT	-	26.3	40.4	47.3	10.0	20.3	25.6	4.4	7.1	8.2	1.6	3.5	4.7
(6')	X	VSE++ GRU	✓	✓	OCR	-	19.9	30.8	36.4	8.8	16.1	20.8	3.4	5.4	6.3	1.5	3.0	4.0
(7)	X	VSRN GRU	✓	✓	GT	-	7.7	18.8	26.0	5.2	12.7	18.8	1.1	2.4	3.3	0.9	2.2	3.3
(8)	X	VSRN GRU	✓	✓	GT	-	12.3	25.1	30.1	6.8	15.3	20.0	1.9	4.0	5.2	1.1	2.8	3.8
(9)	X	GRU++	✓	✓	GT	-	16.0	29.9	35.1	8.7	17.7	22.4	1.4	2.5	3.5	0.8	2.0	2.9
(10)	X	Fasttext+FV(u)	X	X	GT	-	19.5	35.8	43.1	0.5	1.4	2.1	3.1	5.4	7.1	0.1	0.3	0.4
(11)	X	Fasttext+FV	X	X	GT	-	21.7	36.5	44.3	3.2	6.6	9.0	3.5	5.9	7.5	0.6	1.3	1.7
(12)					AVG		31.1	54.5	65.7	17.2	37.2	47.6	7.2	16.4	24.0	4.7	13.5	20.7
(13)	VSE++	VSE++ GRU	✓	X	GT		25.3	51.9	63.6	17.3	39.5	52.2	13.4	30.1	40.4	7.5	20.3	29.2
(14)					PSC		25.8	51.7	63.2	13.5	37.4	51.0	10.9	30.5	41.3	4.2	19.8	29.5
(15)					LSC		25.9	51.8	63.1	17.2	39.4	52.5	13.6	31.1	41.5	7.9	20.8	30.0
(16)					LF		35.6	61.2	71.3	21.8	45.4	58.0	19.2	39.2	50.2	10.7	26.7	36.9
(17)	VSRN	VSE++ GRU	✓	X	GT		30.6	59.3	69.5	16.2	43.2	58.2	14.8	38.8	50.6	6.0	26.4	38.1
(18)					LSC		38.0	60.3	70.3	21.9	45.8	58.2	20.3	40.0	50.6	11.1	27.8	38.2
(19)					LF		32.2	58.3	69.3	20.3	43.5	56.5	18.3	37.8	48.5	10.6	27.0	36.8
(20)	VSRN	VSE++ GRU	✓	X	OCR		26.7	56.0	66.7	15.0	44.2	57.4	14.5	38.1	49.5	6.2	26.4	38.0
(21)					LSC		32.8	57.0	68.5	20.7	44.0	57.1	19.7	39.6	50.3	11.3	27.9	38.3
(22)					AVG		34.6	53.1	61.0	14.5	31.0	39.4	10.0	21.5	29.5	5.0	14.1	21.4
(23)					LF		31.0	60.0	72.3	20.4	44.7	57.3	13.4	30.9	41.5	7.4	20.5	29.1
(24)	VSE++	VSE++ GRU	✓	✓	GT		37.4	62.8	73.6	15.5	42.6	57.1	12.2	32.1	42.4	4.1	19.3	29.2
(25)					LSC		31.6	57.8	70.2	20.3	44.7	57.8	13.7	31.7	41.6	7.7	21.0	29.6
(26)					AVG		36.8	62.2	72.9	18.6	40.5	52.9	15.3	33.5	44.3	6.4	18.9	28.0
(27)					LF		40.3	68.5	79.9	23.9	49.9	63.4	22.6	45.0	56.3	11.8	29.5	40.0
(28)	VSRN	VSRN GRU	✓	✓	GT		33.5	65.9	78.2	15.8	48.1	64.3	18.5	44.5	56.0	5.3	28.7	41.0
(29)					LSC		38.6	67.5	78.5	24.3	50.4	64.0	23.4	45.6	56.5	12.1	30.6	41.1
(30)					LF		41.7	68.6	78.9	25.1	52.0	65.5	22.5	44.4	55.7	12.8	31.0	41.3
(31)	VSRN	VSE++ GRU	✓	✓	GT		32.8	67.3	79.9	17.6	49.4	64.9	16.1	44.6	56.2	6.5	29.3	41.3
(32)					LSC		42.2	67.9	78.5	25.5	52.0	65.6	23.1	45.9	56.1	13.3	31.7	42.2
(33)					Oracle LF		†63.2	†82.9	†89.3	†37.9	†64.3	†75.5	†31.0	†53.9	†64.5	†19.7	†39.3	†49.6
(34)					LF		39.1	66.7	79.1	24.1	50.3	64.3	21.2	43.8	55.4	12.8	31.8	43.0
(35)	VSRN	VSE++ GRU	✓	✓	OCR		31.6	65.2	78.5	16.6	48.6	64.6	15.8	43.9	55.8	6.7	29.4	41.4
(36)					LSC		39.3	67.4	78.7	24.7	50.9	64.6	22.7	45.3	56.3	13.3	31.6	42.2
(37)					LF		45.8	72.7	81.4	26.5	52.7	66.1	24.2	46.1	57.1	12.9	31.0	41.2
(38)	VSRN	VSE++ GRU	✓	✓	GT		42.2	71.5	82.8	18.9	51.1	66.4	20.1	46.4	57.5	6.7	29.5	41.6
(39)					LSC		45.3	71.5	80.7	26.7	53.0	66.2	24.4	46.9	57.4	13.2	31.8	42.3
(40)					Oracle LF		†67.9	†84.8	†91.1	†39.2	†64.8	†76.2	†32.9	†55.3	†65.2	†20.1	†39.7	†50.3
(41)					LF		41.5	70.1	79.8	25.1	51.2	64.3	23.3	45.0	58.9	12.6	30.5	41.1
(42)	VSRN	VSE++ GRU	✓	✓	OCR		38.5	69.6	80.6	17.9	50.1	65.1	19.8	45.7	57.2	7.0	29.8	41.7
(43)					LSC		42.2	68.6	78.5	25.5	51.8	64.9	19.8	45.7	57.2	13.2	31.5	42.2

Table A.1: Results on CTC-1k and CTC-5k for visual-only baselines, scene-text-only baselines and re-ranking combinations of these baselines. Bold results denote the best performance at each of visual model, scene-text model and re-ranking methods. † denotes theoretical upper-bounds to the linear combination re-rankings. (see Section A.1.2)

Using GRU trained for cross-modal retrieval (CMR) as scene-text descriptors has its limitations. The scene text is described with a descriptor learned to represent captions, which is not optimal. For scene text, the order of the words is not as relevant as for a caption. However, since the CMR models use a GRU, the scene-text representation is dependent on the order their words are fed to the model. The Fasttext+FV baseline aims to address these limitations. FastText [21] uses a larger vocabulary than other Word2Vec based models, and uses word n-grams to embed words. In this manner, FastText is a more robust embedding that learns the syntax as well as the semantics of a given word. On top of FastText, a Fisher kernel [152] is employed to aggregate word embeddings. Additionally, an advantage of such an approach is that the scene-text instances are not order dependent and the only training required is at the moment of

	Visual Model	Scene-Text Model	Trained on		Combination	TextCaps					
			F30K	TC		Image to Text			Text to Image		
						R@1	R@5	R@10	R@1	R@5	R@10
(1)	VSE++	✗	✓	✗	-	5.6	15.1	21.5	4.1	11.1	16.6
(2)	VSRN	✗	✓	✗	-	6.2	14.5	20.2	4.5	11.7	16.6
(3)	VSE++	✗	✗	✓	-	14.7	30.9	40.4	10.0	24.3	32.9
(4)	✗	VSE++ GRU	✓	✗	-	11.5	18.7	22.0	10.3	17.5	20.1
(5)	✗	VSE++ GRU	✗	✓	-	34.6	45.7	49.7	25.1	35.0	37.9
(6)	VSE++	VSE++ GRU Rosetta OCR	✗	✓	AVG	42.8	56.6	62.8	30.8	46.2	52.7
(7)					LF	33.5	54.7	63.7	22.6	40.8	50.2
(8)					PSC	40.0	56.3	64.6	24.7	42.3	50.7
(9)					LSC	25.7	46.0	56.1	18.0	36.0	45.3
(10)					Oracle LF	†57.3	†72.3	†78.0	†39.6	†55.9	†63.0

Table A.2: Results on TextCaps (validation set) for visual-only baselines, scene-text-only baselines and re-ranking combinations of these baselines. † denotes theoretical upper-bounds to the linear combination re-rankings. (see Section A.1.2)

constructing a Gaussian Mixture Model (GMM) that models the FastText vocabulary distribution. The best performing implementation of Fasttext+FV approach is presented in row (11). On top of it, we show in row (10) a first implementation of this method before lemmatisation and removal of stop words.

Finally, we show results for the two best models (two different flavors of VSE++ GRU) when using OCR prediction from [64] in rows (5') and (6'). These models are also used in combination with visual-only baselines in rows (19-21), (34-36) and (41-43). We observe a considerable decline in performance between (5) and (5'), (6) and (6'). This can be attributed to errors in OCR prediction. Indeed, COCO-Text is a very challenging dataset for scene-text recognition due to its many small bounding boxes, and CTC inherits these annotations. These results highlight the importance of good scene-text recognition for StacMR. When comparing combinations to their equivalents with ground-truth annotations, the decline in performance is less pronounced.

Models trained on Flickr30K In the main paper, we highlighted how the best performance are obtained from cross-modal retrieval models trained on Flickr30K+TextCaps. We recommend models trained on this combination of datasets for benchmark on CTC. For completeness, we include here re-ranking results for combining models trained on Flickr30K only. Their performance is shown in rows (12-18) using ground-truth scene-text annotations and rows (19-21) using OCR predictions from [64]. In comparison to the models trained on Flickr30K+TextCaps, models trained on Flickr30K obtain similar improvements on CTC-1K and more significant gains on CTC-5K.

In addition to these, a few hybrid models (where visual-only models are trained on F30K+TC and scene-text-only models are trained on F30K) are shown in rows (30-36).

Performance on TextCaps

In order to describe why TextCaps is not fit as an evaluation dataset for StacMR, we performed similar experiments to those described in the Chapter Chapter 5.1 of the main paper. The main results are shown in Table A.2. Here we see how a model trained for cross-modal retrieval with no access to the scene-text information performs better as a scene-text model than a visual model. This highlights the bias of the dataset towards scene text as its main information and the fact that purely visual information comes second.

Oracle Late Fusion

In addition to providing strong multimodal baselines from separated visual and scene-text models, combination methods are very intuitive to understand. For example, late fusion scores of two models consist of a linear combination of the scores given by two different models. The hyper-parameter α corresponds to the best linear combination factor when averaging for all queries, both images, and captions.

A natural extension to the late fusion combination is to make α a parameter dependent on the values of the the image-to-caption similarity $s_v(q, d)$ and the scene-text-to-caption score $s_t(q, d)$. Based on this extension, we propose an oracle combination method s_{LF}^* , called *oracle late fusion*, where the parameter α is query dependent and hand-picked to optimize the ranking for the query. More precisely, this oracle optimizes the median rank of the first retrieved positive item:

$$s_{LF}^*(q, d) = \alpha^*(q)s_v(q, d) + (1 - \alpha^*(q))s_t(q, d), \quad (\text{A.1})$$

$$\alpha^*(q) = \underset{\alpha \in [0,1]}{\operatorname{argmin}} (\operatorname{Rank}_{s_{LF}}(q, d)), \quad (\text{A.2})$$

where Rank denotes the rank of the first retrieved positive item. Given a visual-only and a scene-text-only model, the oracle late fusion provides us with a theoretical upper-bound to the performance of any combination obtained by linearly combining these models. Moreover, we can analyze the values of α obtained for each query to understand how often does a combination prefers to use the visual model or the scene-text model. Indeed, $\alpha^*(q) \sim 1$ indicates that, for this query, the visual model is enough and the scene text should be ignored, $\alpha^*(q) \sim 0$ means that the scene text is enough, and $\alpha^*(q)$ in between implies a balanced optimal weighting of both modalities.

We present the performance for oracle late fusion, evaluated both for CTC and TextCaps, on Table 7.2 rows (33) and (40), and Table A.2 row (10). We observe a considerable improvement compared to combination methods. While for instance, looking at $R@10$ results, row (39) improved upon row (4) by 4.7%, 2.4%, 1.4% and -0.3%, row (40) beats row (39) by 10.4%, 10%, 7.8% and 8%. More importantly, these theoretical upper-bounds show the unexplored potential of combining visual and scene-text information to improve StacMR results.

Performance on Flickr30K and TextCaps

In Table A.3 we show the performance of our proposed model with SCAN [109] and VSRN [111]. In order to obtain comparable results, we have obtained features from our implementation to extract visual regions as [8]. Publicly available code for SCAN [109] and VSRN [111] was used to train those models.

Model	Trained on			Flickr30K						TextCaps					
				Image to Text			Text to Image			Image to Text			Text to Image		
	F30K	TextCaps	CTC	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN	✓	✗	✗	57.2	84.4	90.5	38.6	68.4	79.1	9.3	21.7	29.8	4.7	14.1	21.2
	✗	✓	✗	14.1	34.6	45.0	7.8	22.7	32.1	23.2	50.5	63.5	14.1	37.6	52.1
	✓	✓	✗	57.6	85.3	92.4	39.2	70.0	80.2	16.6	36.6	48.7	9.3	25.4	36.4
	✓	✗	✓	58.1	83.2	91.5	39.6	69.8	81.3	4.4	11.2	16.2	2.4	7.2	11.3
	✓	✓	✓	55.1	79.6	87.1	35.5	67.2	77.3	15.4	35.2	46.9	13.4	37.1	51.8
VSRN	✓	✗	✗	63.1	86.5	92.1	47.1	75.3	83.8	6.3	14.9	21.4	4.2	11.4	16.6
	✗	✓	✗	11.7	30.1	40.2	9.2	23.7	32.8	14.3	34.9	46.2	9.53	26.2	37.2
	✓	✓	✗	62.5	86.1	92.3	48.1	76.8	84.3	19.6	41.9	53.1	13.9	32.8	43.8
	✓	✗	✓	64.9	88.0	93.2	49.0	76.9	84.9	8.21	18.6	25.4	5.56	14.0	19.5
	✓	✓	✓	60.7	85.2	90.4	45.7	73.9	81.8	18.7	38.6	50.1	12.4	30.0	41.2
STARNet	✓	✗	✗	63.9	86.9	92.4	48.6	76.7	84.7	6.79	15.5	21.6	4.6	12.1	17.5
	✗	✓	✗	13.3	29.6	39.6	9.8	24.5	34.1	28.7	53.7	65.1	19.8	40.1	51.6
	✓	✓	✗	62.4	85.8	92.1	47.1	76.1	84.1	24.0	48.9	60.7	17.3	37.9	49.8
	✓	✗	✓	63.2	87.2	92.5	49.5	78.1	85.2	7.5	17.5	25.1	5.2	13.6	19.5
	✓	✓	✓	67.5	88.1	93.6	50.7	78.0	85.4	29.5	53.8	65.3	20.8	42.9	53.6

Table A.3: Quantitative comparison of experimental results of image-to-text and text-to-image retrieval on the Flickr30K (test) and TextCaps (val) sets of supervised models. Metric depicted in terms of Recall@K (R@K).

Results show that by leveraging scene-text retrieval improvements can be achieved. It is important to note the effect of employing different datasets in the training procedure. As it is expected, training on TextCaps and due to the dataset nature that focuses only on scene text instances, as well as their captions, it does not yield good results when used alone. Even adding samples from the CTC dataset at training time, can yield an improvement when evaluated on the TextCaps validation set.

It is worth noting as well that in standard cross-modal retrieval models, adding TextCaps training data achieve a minor improvement (SCAN) or lower the performance (VSRN) when compared in the Flickr30k dataset. However a slight improvement is achieved when adding the CTC training set.

However, the proposed model learns to model the interactions between scene-text and visual descriptors to combine them appropriately. STARNet achieves better a performance among both datasets even when scene-text is not widely available in Flickr30k.

List of Publications

Topics

The main topic of this dissertation, is the incorporation of scene text information into fine-grained image classification and cross-modal retrieval systems. Nonetheless, the novel contributions can also be applied into another different vision and language tasks. However, this thesis has also yielded additional contributions in other computer vision topics which we following enumerate.

International Journals

- **Andres Mafla**, Rubèn Tito, Sounak Dey, Lluís Gomez, Marçal Rossinyol and Dimosthenis Karatzas, "Real-Time Lexicon-Free Scene Text Retrieval", in *Pattern Recognition*, 2020.
- Lluís Gomez, Ali Furkan Biten, Rubèn Tito, **Andres Mafla**, Marçal Rusiñol, Ernest Valveny, Dimosthenis Karatzas, "Multimodal grid features and cell pointers for scene text visual question answering", in *Pattern Recognition Letters*, 2021.

International Conferences

- Lluís Gomez*, **Andres Mafla***, Marçal Rossinyol and Dimosthenis Karatzas, "Single shot scene text retrieval", in *European Conference on Computer Vision (ECCV)*, 2018.
- Ali Furkan Biten*, Ruben Tito*, **Andres Mafla***, Lluís Gomez, Marçal Rusiñol, Ernest Valveny, CV Jawahar, Dimosthenis Karatzas, "Scene Text Visual Question Answering", *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019
- Ali Furkan Biten*, Ruben Tito*, **Andres Mafla***, Lluís Gomez, Marçal Rusiñol, Minesh Mathew, CV Jawahar, Ernest Valveny, Dimosthenis Karatzas, "Icdar 2019

competition on scene text visual question answering", *International Conference on Document Analysis and Recognition (ICDAR)*, 2019

- **Andres Mafla**, Sounak Dey, Ali Furkan Biten, Lluís Gomez and Dimosthenis Karatzas, "Fine-grained Image Classification and Retrieval by Combining Visual and Locally Pooled Textual Features", in *Winter Application in Computer Vision (WACV)*, 2020.
- **Andres Mafla**, Sounak Dey, Ali Furkan Biten, Lluís Gomez and Dimosthenis Karatzas, "Multi-Modal Reasoning Graph for Scene-Text Based Fine-Grained Image Classification and Retrieval", in *Winter Application in Computer Vision (WACV)*, 2021.
- **Andres Mafla**, Rafael Sampaio de Rezende, Lluís Gomez, Diane Larlus and Dimosthenis Karatzas, "Stacmr: scene-text aware cross-modal retrieval", in *Winter Application in Computer Vision (WACV)*, 2021.
- Ali Furkan Biten*, **Andres Mafla***, Lluís Gomez, Dimosthenis Karatzas, "Is An Image Worth Five Sentences? A New Look into Semantics for Image-Text Matching", *Winter Application in Computer Vision (WACV)*, 2022

International Workshops

- Emanuele Vivoli, Ali Furkan Biten, **Andres Mafla**, Lluís Gomez and Dimosthenis Karatzas, "MUST-VQA: MULTilingual Scene-text VQA", *European Conference on Computer Vision (ECCV)*, 2022
- Sergi Garcia-Bordils*, **Andres Mafla***, Ali Furkan Biten*, Oren Nuriel, Aviad Aberdam, Shai Mazor, Ron Litman and Dimosthenis Karatzas, "Out-of-Vocabulary Challenge Report", *European Conference on Computer Vision (ECCV)*, 2022

arXiv

- Mohamed Ali Souibgui, Sanket Biswas, **Andres Mafla**, Ali Furkan Biten, Alicia Fornés, Yousri Kessentini, Josep Lladós, Lluís Gomez, Dimosthenis Karatzas, "Text-DIAE: Degradation Invariant Autoencoders for Text Recognition and Document Enhancement", in *arXiv*, 2022.

Under Review

- Van Khanh Nguyen, , Ali Furkan Biten, **Andres Mafla**, Lluís Gomez and Dimosthenis Karatzas, "Show, Interpret and Tell: Entity-aware Contextualised Image Captioning in Wikipedia", under review, 2022.

Bibliography

- [1] I. Bizid, J. Chazalon, H. Choi, Y. Feng, D. Karatzas, W. Khlif, Z. Luo, M. Luqman, N. Nayef, U. Pal, C. Rigaud, F. Yin J. Matas, N. Nayef, U. Pal, Y. Patel. ICDAR2017 Competition on Multi-lingual scene text detection and script identification. <http://http://rrc.cvc.uab.es/?ch=8>, 2017. [Online, accessed 22-April-2019].
- [2] M. Bušta, D. Karatzas, W. Khlif, J. Matas, N. Nayef, U. Pal, Y. Patel. ICDAR 2019 Robust Reading Challenge on Multi-lingual scene text detection and recognition. <http://http://rrc.cvc.uab.es/?ch=11>, 2019. [Online, accessed 30-April-2019].
- [3] M. Iwamura, L. Gomez, D. Karatzas. Robust Reading Challenge on Text in videos 2013-2015. <http://http://rrc.cvc.uab.es/?ch=11>, 2019. [Online, accessed 22-April-2019].
- [4] Julien Ah-Pine, Stephane Clinchant, Gabriela Csurka, Florent Perronnin, and Jean-Michel Renders. Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval. In *ImageCLEF*, pages 315–342. Springer, 2010.
- [5] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(12):2552–2566, 2014.
- [6] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, 2016.
- [7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *arXiv preprint arXiv:1707.07998*, 2017.
- [8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

-
- [9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [10] Fan Bai, Zhanzhan Cheng, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Edit probability for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1508–1516, 2018.
- [11] Xiang Bai, Mingkun Yang, Pengyuan Lyu, and Yongchao Xu. Integrating scene text and visual appearance for fine-grained image classification with convolutional neural networks. *arXiv preprint arXiv:1704.04613*, 2017.
- [12] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018.
- [13] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.
- [14] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.
- [15] Dena Bazazian, Raúl Gómez, Angelos Nicolaou, Lluís Gómez, Dimosthenis Karatzas, and Andrew D Bagdanov. Fast: Facilitated and accurate scene text proposals through fcn guided pruning. *Pattern Recognition Letters*, 119:112–120, 2019.
- [16] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.
- [17] Hedi Ben-Younes, Rémi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. *arXiv preprint arXiv:1902.00038*, 2019.
- [18] Ali Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. October 2019.
- [19] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16548–16558, 2022.
- [20] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

- [21] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- [22] Fedor Borisjuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79, 2018.
- [23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [24] Michal Busta, Lukas Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.
- [25] Michal Bušta, Yash Patel, and Jiri Matas. E2e-mlt-an unconstrained end-to-end method for multi-language scene text. In *Asian Conference on Computer Vision*, pages 127–143. Springer, 2018.
- [26] Michal Buvsta, Lukavs Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proc. of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.
- [27] Charles Chen, Ruiyi Zhang, Eunye Koh, Sungchul Kim, Scott Cohen, Tong Yu, Ryan Rossi, and Razvan Bunescu. Figure captioning with reasoning and sequence-level training. *arXiv preprint arXiv:1906.02850*, 2019.
- [28] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663, 2020.
- [29] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *arXiv preprint arXiv:2005.03492*, 2020.
- [30] Xilin Chen, Jie Yang, Jing Zhang, and Alex Waibel. Automatic detection and recognition of signs from natural scenes. *IEEE Transactions on image processing*, 13(1):87–99, 2004.
- [31] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [32] Mengjun Cheng, Yipeng Sun, Longchao Wang, Xiongwei Zhu, Kun Yao, Jie Chen, Guoli Song, Junyu Han, Jingtuo Liu, Errui Ding, et al. Vista: Vision and scene text aggregation for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2022.

- [33] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5076–5084, 2017.
- [34] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- [35] Noam Chomsky. *The architecture of language*. OUP India, 2006.
- [36] Noam Chomsky, Chomsky Noam, et al. *On nature and language*. Cambridge University Press, 2002.
- [37] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. *arXiv preprint arXiv:2101.05068*, 2021.
- [38] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [39] Stéphane Clinchant, Julien Ah-Pine, and Gabriela Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, pages 1–8, 2011.
- [40] Gabriela Csurka and Stéphane Clinchant. An empirical study of fusion operators for multimodal image retrieval. In *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2012.
- [41] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [42] Jia Deng, Jonathan Krause, and Li Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2013.
- [43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [44] Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019.
- [45] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368*, 2021.

- [46] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.
- [47] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiao-hua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [48] Desmond Elliott and Frank Keller. Comparing automatic evaluation measures for image description. In *Annual Meeting of the Association for Computational Linguistics*, 2014.
- [49] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proc. CVPR*, 2018.
- [50] Boris Epshtein, Eyal Ofek, and Yonatan Wexler. Detecting text in natural scenes with stroke width transform. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2963–2970, 2010.
- [51] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.
- [52] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2018.
- [53] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. 2013.
- [54] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12746–12756, 2020.
- [55] Yunze Gao, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Reading scene text with attention convolutional sequence modeling. *arXiv preprint arXiv:1709.04303*, 2017.
- [56] ZongYuan Ge, Chris McCool, Conrad Sanderson, Peng Wang, Lingqiao Liu, Ian Reid, and Peter Corke. Exploiting temporal information for DCNN-based fine-grained object classification. In *International Conference on Digital Image Computing: Techniques and Applications*, 2016.
- [57] Suman K Ghosh, Lluís Gomez, Dimosthenis Karatzas, and Ernest Valveny. Efficient indexing for query by string text retrieval. In *Proc. of the IEEE International Conference on Document Analysis and Recognition*, pages 1236–1240, 2015.

- [58] Suman K Ghosh and Ernest Valveny. Query by string word spotting based on character bi-gram indexing. In *Proc. of the IEEE International Conference on Document Analysis and Recognition*, pages 881–885, 2015.
- [59] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.
- [60] Lluís Gómez and Dimosthenis Karatzas. Textproposals: a text-specific selective search algorithm for word spotting in the wild. *Pattern Recognition*, 70:60–74, 2017.
- [61] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 700–715, 2018.
- [62] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [63] Lluís Gómez, Yash Patel, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017.
- [64] Google. *Cloud Vision API*, 2020 (accessed June 3, 2020).
- [65] Albert Gordo. Supervised mid-level features for word image representation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2956–2964, 2015.
- [66] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6589–6598, 2017.
- [67] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.
- [68] J Gregor. An algorithm for the decomposition of a distribution into gaussian components. *Biometrics*, pages 79–93, 1969.
- [69] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324, 2016.
- [70] Yuval Noah Harari. Sapiens: A brief history of humankind. *Asian Review of World Histories*, 3(2):265–267, 2015.

- [71] Dafang He, Xiao Yang, Chen Liang, Zihan Zhou, Alexander G Ororbia, Daniel Kifer, and C Lee Giles. Multi-scale fcn with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3519–3528, 2017.
- [72] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- [74] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5020–5029, 2018.
- [75] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [76] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
- [77] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [78] Weilin Huang, Yu Qiao, and Xiaoou Tang. Robust scene text detection with convolution neural network induced msr trees. In *Proc. of the European Conference on Computer Vision*, pages 497–511. Springer, 2014.
- [79] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.
- [80] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [81] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.
- [82] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

- [83] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.
- [84] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.
- [85] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: Rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579*, 2017.
- [86] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 1972.
- [87] Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold WM Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5):1063–1076, 2017.
- [88] Sezer Karaoglu, Ran Tao, Jan C van Gemert, and Theo Gevers. Con-text: Text detection for fine-grained object classification. *IEEE Transactions on Image Processing*, 26(8):3965–3980, 2017.
- [89] Sezer Karaoglu, Jan C van Gemert, and Theo Gevers. Con-text: text detection using background connectivity for fine-grained object classification. In *Proc. of the ACM International Conference on Multimedia*, pages 757–760, 2013.
- [90] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. ICDAR 2015 competition on robust reading. In *Proc. of the IEEE International Conference on Document Analysis and Recognition*, pages 1156–1160, 2015.
- [91] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluís Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluís Pere De Las Heras. ICDAR 2013 robust reading competition. In *Proc. of the IEEE International Conference on Document Analysis and Recognition*, pages 1484–1493, 2013.
- [92] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137, 2015.
- [93] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.
- [94] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, page 1, 2011.

- [95] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, 2014.
- [96] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating Automatic Metrics for Image Captioning. 2016.
- [97] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.
- [98] Kye-Hyeon Kim, Sanghoon Hong, Byungseok Roh, Yeongjae Cheon, and Minje Park. Pvanet: deep but lightweight neural networks for real-time object detection. *arXiv preprint arXiv:1608.08021*, 2016.
- [99] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [100] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [101] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.
- [102] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proc. CVPR*, 2015.
- [103] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.
- [104] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NeurIPS*, pages 1097–1105, 2012.
- [105] K. Lang and T. Mitchell. Newsgroup 20 dataset. 1999.
- [106] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [107] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

- [108] Jung-Jin Lee, Pyoung-Hean Lee, Seong-Whan Lee, Alan Yuille, and Christof Koch. Adaboost for text detection in natural scene. In *2011 International Conference on Document Analysis and Recognition*, pages 429–434. IEEE, 2011.
- [109] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. ECCV*, 2018.
- [110] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. *arXiv preprint arXiv:1707.03985*, 2017.
- [111] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4654–4662, 2019.
- [112] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, 21(8):2117–2130, 2019.
- [113] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.
- [114] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.
- [115] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggong Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 4161–4167, 2017.
- [116] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [117] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.
- [118] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [119] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 3–11, 2019.

- [120] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020.
- [121] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [122] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [123] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *Proc. of the European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- [124] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.
- [125] Shangbang Long, Xin He, and Cong Ya. Scene text detection and recognition: The deep learning era. *arXiv preprint arXiv:1811.04256*, 2018.
- [126] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.
- [127] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.
- [128] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.
- [129] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018.
- [130] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [131] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluís Gomez, and Dimosthenis Karatzas. Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2950–2959, 2020.

- [132] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluís Gomez, and Dimosthenis Karatzas. Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. *arXiv preprint arXiv:2009.09809*, 2020.
- [133] Andrés Mafla, Rafael S Rezende, Lluís Gomez, Diane Larlus, and Dimosthenis Karatzas. Stacmr: Scene-text aware cross-modal retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2220–2230, 2021.
- [134] Andrés Mafla, Rubèn Tito, Sounak Dey, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, and Dimosthenis Karatzas. Real-time lexicon-free scene text retrieval. *Pattern Recognition*, page 107656, 2020.
- [135] Andrés Mafla, Rubèn Tito, Sounak Dey, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, and Dimosthenis Karatzas. Real-time lexicon-free scene text retrieval. *Pattern Recognition*, 110:107656, 2021.
- [136] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.
- [137] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [138] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.
- [139] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [140] Anand Mishra. *Understanding Text in Scene Images*. PhD thesis, International Institute of Information Technology Hyderabad, 2016.
- [141] Anand Mishra, Karteek Alahari, and C.V. Jawahar. An mrf model for binarization of natural scene text. In *2011 International Conference on Document Analysis and Recognition*, pages 11–16. IEEE, 2011.
- [142] Anand Mishra, Karteek Alahari, and C.V. Jawahar. Image retrieval using textual cues. In *Proc. of the IEEE International Conference on Computer Vision*, pages 3040–3047, 2013.
- [143] Anand Mishra, Karteek Alahari, and C.V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2687–2694, 2012.
- [144] Yair Movshovitz-Attias, Qian Yu, Martin C Stumpe, Vinay Shet, Sacha Arnoud, and Liron Yatziv. Ontological supervision for fine grained classification of street view storefronts. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1693–1702, 2015.

- [145] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision*, pages 681–699. Springer, 2020.
- [146] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.
- [147] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in neural information processing systems*, pages 2654–2665, 2018.
- [148] Lukáš Neumann and Jiří Matas. Real-time scene text localization and recognition. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3538–3545, 2012.
- [149] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 1881–1889, 2017.
- [150] Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. *arXiv preprint arXiv:2004.15020*, 2020.
- [151] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [152] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007.
- [153] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *IEEE International Conference on Computer Vision*, 2015.
- [154] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. *arXiv preprint arXiv:1912.03098*, 2019.
- [155] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

- [156] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
- [157] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition(2016)*, pages 779–788, 2016.
- [158] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.
- [159] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.
- [160] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.
- [161] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. of the International Conference on Neural Information Processing Systems*, pages 91–99, 2015.
- [162] Jose A Rodriguez-Serrano, Albert Gordo, and Florent Perronnin. Label embedding: A frugal baseline for text recognition. *International Journal of Computer Vision*, 113(3):193–207, 2015.
- [163] Partha Pratim Roy, Umapada Pal, Josep Lladós, and Mathieu Delalandre. Multi-oriented and multi-sized touching character segmentation using dynamic programming. In *2009 10th International Conference on Document Analysis and Recognition*, pages 11–15. IEEE, 2009.
- [164] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
- [165] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [166] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.
- [167] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

- [168] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015.
- [169] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.
- [170] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- [171] Karthik Sheshadri and Santosh Kumar Divvala. Exemplar driven character recognition in the wild. In *BMVC*, pages 1–10, 2012.
- [172] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304, 2017.
- [173] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4168–4176, 2016.
- [174] Cunzhao Shi, Chunheng Wang, Baihua Xiao, Yang Zhang, Song Gao, and Zhong Zhang. Scene text recognition using part-based tree-structured character detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2961–2968, 2013.
- [175] Palaiahnakote Shivakumara, Souvik Bhowmick, Bolan Su, Chew Lim Tan, and Umapada Pal. A new gradient based character segmentation method for video text recognition. In *2011 International Conference on Document Analysis and Recognition*, pages 126–130. IEEE, 2011.
- [176] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020.
- [177] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [178] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4602–4612, 2019.
- [179] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)*, pages 8317–8326, 2019.

- [180] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.
- [181] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2019.
- [182] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.
- [183] Michael Studdert-Kennedy and Louis Goldstein. Launching language: The gestural origin of discrete infinity. *Studies in the Evolution of Language*, 3:235–254, 2003.
- [184] Sebastian Sudholt and Gernot A Fink. Phocnet: A deep convolutional neural network for word spotting in handwritten documents. In *Proc. of the IEEE International Conference on Frontiers in Handwriting Recognition*, pages 277–282, 2016.
- [185] Sebastian Sudholt, Neha Gurjar, and Gernot A Fink. Learning deep representations for word spotting under weak supervision. *arXiv preprint arXiv:1712.00250*, 2017.
- [186] Milan Šulc and Jiří Matas. Fine-grained recognition of plants from images. *Plant methods*, 13(1):115, 2017.
- [187] Ilya Sutskever, Geoffrey E Hinton, and A Krizhevsky. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [188] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
- [189] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [190] Peng Tang, Xinggang Wang, Bin Feng, and Wenyu Liu. Learning multi-instance deep discriminative patterns for image classification. *IEEE Transactions on Image Processing*, 26(7):3385–3396, 2017.
- [191] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *European Conference on Computer Vision*, pages 317–335. Springer, 2020.

- [192] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.
- [193] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.
- [194] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.
- [195] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [196] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. COCO-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.
- [197] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015.
- [198] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6439–6448, 2019.
- [199] Toru Wakahara and Kohei Kita. Binarization of color character strings in scene images using k-means clustering and support vector machines. In *2011 International Conference on Document Analysis and Recognition*, pages 274–278. IEEE, 2011.
- [200] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.
- [201] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4558–4567, 2021.
- [202] Hao Wang, Junchao Liao, Tianheng Cheng, Zewen Gao, Hao Liu, Bo Ren, Xiang Bai, and Wenyu Liu. Knowledge mining with scene text for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4624–4633, 2022.

- [203] Haoran Wang, Ying Zhang, Zhong Ji, Yanwei Pang, and Lin Ma. Consensus-aware visual-semantic embedding for image-text matching. *arXiv preprint arXiv:2007.08883*, 2020.
- [204] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *European Conference on Computer Vision*, pages 370–386. Springer, 2020.
- [205] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *Proc. of the IEEE International Conference on Computer Vision*, pages 1457–1464, 2011.
- [206] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020.
- [207] Tao Wang, David J Wu, Adam Coates, and Andrew Y Ng. End-to-end text recognition with convolutional neural networks. In *Proc. of the IEEE International Conference on Pattern Recognition*, pages 3304–3308, 2012.
- [208] Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards unique and informative captioning of images. *arXiv preprint arXiv:2009.03949*, 2020.
- [209] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proc. ICCV*, 2019.
- [210] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [211] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950, 2020.
- [212] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaying Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015.
- [213] Huapeng Xu, Guilin Qi, Jingjing Li, Meng Wang, Kang Xu, and Huan Gao. Fine-grained image classification by visual-semantic embedding. In *IJCAI*, pages 1043–1049, 2018.

- [214] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.
- [215] Shulin Yang, Liefeng Bo, Jue Wang, and Linda G Shapiro. Unsupervised template learning for fine-grained object recognition. In *Advances in neural information processing systems*, pages 3122–3130, 2012.
- [216] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.
- [217] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 420–435, 2018.
- [218] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761, 2021.
- [219] Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016.
- [220] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A learned multi-scale representation for scene text recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4042–4049, 2014.
- [221] Qixiang Ye, Wen Gao, Weiqiang Wang, and Wei Zeng. A robust text detection algorithm in images and video frames. In *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, volume 2, pages 802–806. IEEE, 2003.
- [222] Xu-Cheng Yin, Xuwang Yin, Kaizhu Huang, and Hong-Wei Hao. Robust text detection in natural scene images. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):970–983, 2014.
- [223] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.
- [224] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2:67–78, 2014.

- [225] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830, 2017.
- [226] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018.
- [227] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.
- [228] Fangneng Zhan, Shijian Lu, and Chuhui Xue. Verisimilar image synthesis for accurate detection and recognition of texts in scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 249–266, 2018.
- [229] Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. Learning to represent image and text with denotation graph. *arXiv preprint arXiv:2010.02949*, 2020.
- [230] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In *Advances in Neural Information Processing Systems*, pages 9593–9604, 2019.
- [231] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3536–3545, 2020.
- [232] Sheng Zhang, Yuliang Liu, Lianwen Jin, and Canjie Luo. Feature enhancement network: A refined scene text detector. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [233] Mo Zhou, Zhenxing Niu, Le Wang, Zhanning Gao, Qilin Zhang, and Gang Hua. Ladder loss for coherent visual-semantic embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13050–13057, 2020.
- [234] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jijun Liang. East: an efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651. IEEE, 2017.
- [235] Yingying Zhu, Cong Yao, and Xiang Bai. Scene text detection and recognition: Recent advances and future trends. *Frontiers of Computer Science*, 10(1):19–36, 2016.

