StacMR: Scene-Text Aware Cross-Modal Retrieval

Andrés Mafla† Rafael S. Rezende‡ Lluís Gómez† Diane Larlus‡ Dimosthenis Karatzas†
† Computer Vision Center, Universitat Autónoma de Barcelona ‡ NAVER LABS Europe

Abstract

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 paper, 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 which leverage scene text, including a better scene-text aware cross-modal retrieval method which uses specialized representations for text from the captions and text from the visual scene, and reconcile 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. Dataset and code are available at europe.naverlabs.com/stacmr.

1. Introduction

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 [51]. This is especially relevant when considering vision and language tasks, and in particular, related to our work, cross-modal retrieval. 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 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 a benchmark to properly address this research question. Our work tackles these two open directions.

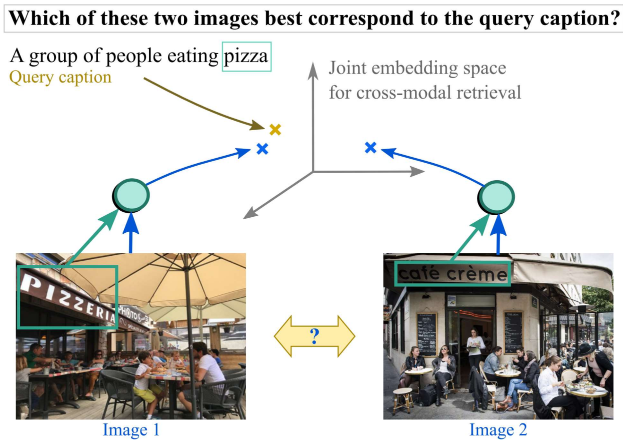

Figure 1: This paper introduces the **scene-text aware cross-modal retrieval** (StacMR) task and studies scene text as a third modality for cross-modal retrieval. For the example query above, the restaurant name provides crucial information to disambiguate two otherwise equally relevant results.

Scene text has been successfully leveraged to improve several semantics tasks in the past, such as fine-grained image classification [4, 21, 34, 40], visual question answering (VQA) [5, 47] or image captioning [46]. Current mainstream methods tackle cross-modal retrieval by either learning to project images and their captions into a joint embedding space [15, 25, 28, 54] or by directly comparing image regions and caption fragments to compute a similarity score [22, 27]. Although significant gaps have been overcome by previous methods, the lack of integration between scene text and the other modalities still hinder a 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 paper introduces the Scene-Text Aware Cross-Modal Retrieval (StacMR) task which 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 [30] 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. In this work, we also evaluate the performance of different state-of-the-art cross-modal retrieval models, their limitations, and we propose distinctive baselines to solve this task.

Concretely, the contribution of this paper 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 a 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.

2. Related Work

Scene-Text Detection and Recognition. Due to the large variance in text instances found in the wild [10, 64], scene text detection and recognition is still an active research field. Methods such as EAST [63], Textboxes++ [29] or LOMO [61] draw inspiration from general object detectors [19, 31, 44, 45] and typically localize text instances by regressing pre-defined anchor boxes or pixels.

Moreover, pipelines trained end-to-end often benefit from both tasks, detection and recognition. Mask Textspotter [32] is an end-to-end segmentation-based approach which detects and recognizes text in arbitrary shapes. Similarly, [20] extracts image features with a CNN that are later refined by two Long-Short Term Memories (LSTMs) along with a text-alignment layer to perform these two tasks jointly. In another approach, [60] employs a semantic reasoning network to mitigate transcriptions by projecting textual regions in a learned semantic space.

Scene Text in Vision and Language. Methods for vision and language tasks typically align both modalities and often perform visual reasoning. Only recently have they started including scene text as an additional modality. Works such as Text-VQA [47] and Scene-Text VQA [5] 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: [4, 21, 35] learn a shared semantic space between visual features and text to perform classification while [34] uses the Pyramidal Histogram Of Characters (PHOC) [2, 16, 36] descriptor as a way of overcoming OCR limitations and learn a morphological space suitable for the task. Other works [17, 39] perform scene-text based image search, where we query with a word and retrieve images containing such word. Closer to our work, the TextCaps dataset [46] includes scene text into textual descriptions. We discuss further the link with our work in Section 3.

Cross-Modal Retrieval. Most cross-modal retrieval (CMR) approaches learn a joint representation space together with visual and textual embedding functions which produce similar representations for semantically related input, *e.g*. an image and its captions. Often, the visual embedding function is a CNN and the textual one a recurrent neural network [15, 33, 37, 55]. Other approaches use regions of interest given by a detector [3]. These approaches align each visual region with a corresponding caption word to get a finer-grained image representation [8, 23, 27, 28, 54, 62]. Some methods also use attention mechanisms [27, 41, 48] that model detailed interactions between captions and image regions. More recently, transformers [50] have been combined [49, 57, 58] to perform multi-layered self-attention operations in order to better align visual and textual features. Other works [28, 56] perform visual reasoning by employing graph convolutional networks [24] which yield a rich set of features by defining a relational graph between paired images and sentences. Closer to our work, Vo *et al*. [53] propose to use text modifiers along with images to retrieve relevant images.

3. 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*. (Section 3.1). Then we present CTC statistics and discuss the dataset in the light of other benchmarks and in particular the most related dataset: TextCaps [46] (Section 3.2).

3.1. 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 [14, 15, 26, 27, 28, 49, 54, 56] are COCO Captions [9], commonly known as MS-COCO in the cross-modal literature,

Dataset	Total Images	Images w/ Text	Annotations	
			Scene Text	Captions
Flickr30K [59]	31,783	3,338⋆	✗	✓
TextCaps [46]	28,408	28,408‡	✗	✓
COCO Captions [9]	123,287	15,844⋆	✗	✓
COCO-Text [51]	63,686	17,237†	✓	✗
COCO-Text Caps	**10,683**	**10,683**†	✓	✓

Table 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 [36]. ‡ numbers obtained with method from [7].

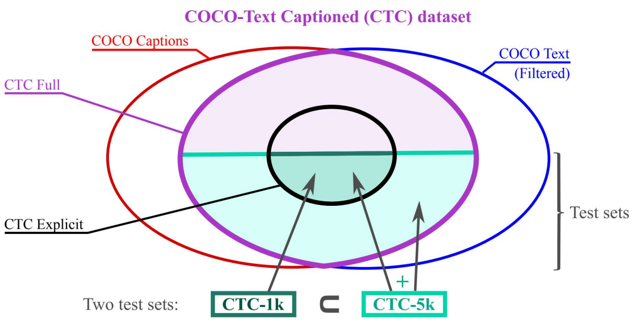

Figure 2: **Proposed CTC dataset**, which is designed to allow a proper evaluation of the STACMR task, as all entries contain three modalities: image, scene text and caption.

and Flickr30K [59]. Only very few images from Flickr30K contain scene text (see Table 1), so we decided to start from COCO Captions, a subset of the COCO dataset [30]. Additionally, the reading systems community commonly uses the COCO-Text dataset [51]. 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 as *COCO-Text Captioned (CTC)*.

Note that the resulting CTC dataset shares 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 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.

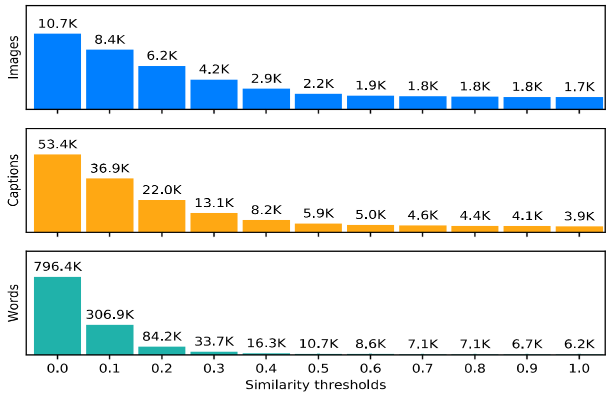

Figure 3: **CTC full statistics.** Cumulative histograms (as thresholds over similarity vary) of the semantic similarity between instances of scene-text tokens and a) all captions for an image (Images), b) individual captions (Captions), and c) individual words in captions (Words).

More precisely, we first remove stop-words from captions and scene-text annotations, and embed each remaining word with Word2Vec [38] 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 3 correspond to the three previously described scenarios. The fact that many words have a strong similarity at all three levels confirm that scene text can be used to model the semantics between the three studied 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 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, 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 5. 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.

3.2. Comparison with other Datasets

Table 1 provides a comparison with the previously mentioned datasets with statistics on the three modalities. Scene-text from COCO Captions [9] and Flickr30K [59] was acquired using a scene-text detector [36]. As mentioned earlier, none of the existing benchmarks contains samples where all three modalities are annotated.

Closely related to the proposed CTC dataset, TextCaps [46] 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 4 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 analysing the intersection of 6,653 images it has with the recently published Localized Narratives dataset [43]. 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 were not instructed to always use the scene text. According to our statistics, this is already higher than expected in the real world. This is because the Localised Narratives captions are long

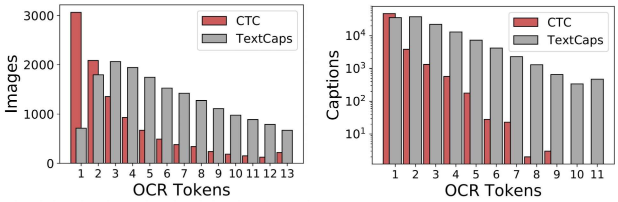

Figure 4: Histograms of the number of OCR tokens found in images (seen as sets of captions, left) and in individual captions (right) for the CTC and TextCaps datasets.

Image	Captions
	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.
	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.
	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.

Figure 5: **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**).

descriptions and tend to venture to fine-grained (localised) descriptions of images 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, visual and scene-text information, avoiding to assume or force any semantic relation between them. More experiments showing the bias between TextCaps' captions and scene-text are provided in Section 5 and in the supplementary material.

4. Method

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 4.1). 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 4.2).

4.1. 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 for images.

A pretrained cross-modal retrieval model relies on a metric space equipped with a similarity function which can indistinguishably compare visual and textual descriptors and allows to rank 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_{LT} of the scores s_v and s_t, using a parameter α:

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

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 [1, 12, 13], we then define the *late semantic combination (LSC)* and *product semantic combination (PSC)* with Equations (2) and (3) respectively. Note that LSC is equivalent to the late fusion if k is equal to the gallery size.

$$s_{LSC}(q, d) = \alpha s_v(q, d) + (1 - \alpha)s_t(q, d)I_k(q, d) \quad (2)$$

$$s_{PSC}(q, d) = s_v(q, d)s_t(q, d)I_k(q, d) \quad (3)$$

These different reranking 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 hyperparameters α and k.

4.2. 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 [28] 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 [28] 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 [53], we assume that scene text acts as a modifier in the joint embedding space, applied to the visual descriptor of an image.

We propose the STARNet (**S**cene-**T**ext **A**ware **R**etrieval **Net**work) model, illustrated in Figure 6. 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 is $I_{fi} = \Phi(I_i, OCR_i)$. The image encoder follows [3] and uses a customized Faster R-CNN [45] to extract visual features for all regions of interest represented by V_i. Similarly, the employed OCR [18] 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 [24] 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 are similar. The inputs to each GCN are features $F_{fi} \in R^{k \times D}$, 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 $\rho(.)$ and $\omega(.)$ are two fully connected layers that are learned in an end-to-end manner by back propagation.

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 [24] to obtain a

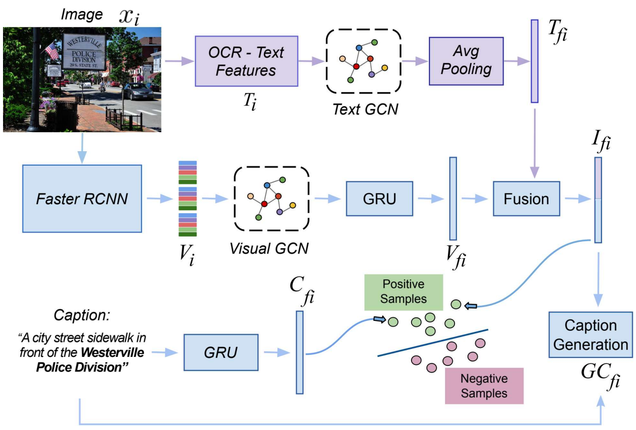

Figure 6: **Our proposed STARNet model.** Visual regions and scene-text instances are used as input to a GCN. The final learned representations are later combined to leverage complementary semantic information.

richer set of features from the nodes and edges. The equation that describes a single Graph Convolution layer is:

$$F_g^l = W_r^l(R^l F_i^{l-1} W_g^l) + F_i^{l-1} \tag{4}$$

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.

The output of the visual GCN goes through a Gated Recurrent Unit (GRU) [11] 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 [11, 15], 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 [15, 27] 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 similarly to [52] and we optimize the log-likelihood of the predicted output caption given the final visual features and the previous generated word.

5. Experiments

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 5.1), and supervised tri-modal fusion results from STARNet (Section 5.2). 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.

5.1. 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 naive average of visual and scene-text embeddings for metric spaces that allow it, and the different re-ranking strategies introduced in Section 4.1.

Visual-only Baselines. We use two CMR models based on global features for both images and captions, VSE++ [15] and VSRN [28]. 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 [15]. We train these architectures either with Flickr30K or Flickr30K + TextCaps. As mentioned in Section 3.1, 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 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 images-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 [6] word embeddings followed by a Fisher vector encoding [42] (denoted by FastText+FV), which is able to deal with out-of-vocabulary words. For these experiments, we use the ground-truth OCR annotations as scene text. Results are presented in Table 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 outper-

	Visual Model	Scene-text Model	Trained on		Scene-text Source	Re-rank	CTC-1K						CTC-5K					
							Image to Text			Text to Image			Image to Text			Text to Image		
			F30K	TC			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++ [15]	✗	✓	✗	-	-	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 [28]	✗	✓	✗	-	-	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)	VSE++	VSE++ GRU	✓	✓	GT	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)						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)						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)	VSRN	VSRN GRU	✓	✓	GT	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)						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)	VSRN	VSE++ GRU	✓	✓	GT	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)						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)	VSRN	VSE++ GRU	✓	✓	OCR	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)						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 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 [18]. **Bold** numbers denote the best performances of visual, scene-text, and re-ranking methods for each ensemble of models.

forms 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. A discussion of several scene-text only baselines is available in the supplementary material.

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 which consists in averaging visual and scene-text embeddings to represent the image, reported as AVG on the Table 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 a first proof that scene text, even naively used, improves on some StacMR queries. However, we observe a decline on 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 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 a 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.

5.2. Supervised Results

Latest cross-modal retrieval models rely on region-based visual features [27, 28, 54] rather than a global image representation [15]. In this section, we include results of two state-of-the-art models, SCAN [27] and VSRN [28] that employ such region-based visual features. The original cross-modal retrieval models, SCAN and VSRN are used

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 [27]	✗	-	✓	✗	✗	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 [28]	✗	-	✓	✗	✗	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 Comb. (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 3: Retrieval results on the CTC-1K and CTC-5K test set of **supervised** models. Second-to-last row shows the result from the unsupervised re-ranking baseline described in Table 2, line 21. *OCR* stands for the textual features obtained from [18], whereas *GT* refers to the Ground-truth annotated scene text. Results depicted in terms of Recall@K (R@K).

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 hyperparameters are employed, details are showed in the supplementary material. All the obtained results are reported on Table 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 [46], the performance of any model decreases significantly, specially 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 towards 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. Additional experiments regarding the performance of the baseline supervised models have been conducted in Flickr30K and TextCaps datasets along with qualitative results available on the supplementary material.

6. Conclusion

In this work, 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.

References

[1] 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.

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *PAMI*, 36(12):2552–2566, 2014.

[3] 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.

[4] 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.

[5] Ali Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proc. ICCV*, October 2019.

[6] 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.

[7] Fedor Borisyuk, 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.

[8] 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.

[9] 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.

[10] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *arXiv preprint arXiv:2005.03492*, 2020.

[11] 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.

[12] 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.

[13] 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.

[14] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proc. CVPR*, 2018.

[15] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *Proc. BMVC*, 2018.

[16] Lluis Gomez, Andres Mafla, Marcal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proc. ECCV*, 2018.

[17] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proc. ECCV*, 2018.

[18] Google. *Cloud Vision API*, 2020 (accessed June 3, 2020).

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proc. ICCV*, pages 2961–2969, 2017.

[20] 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.

[21] Sezer Karaoglu, Ran Tao, Jan C van Gemert, and Theo Gevers. Con-text: text detection for fine-grained object classification. *TIP*, 26(8):3965–3980, 2017.

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. CVPR*, pages 3128–3137, 2015.

[23] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proc. NeurIPS*, pages 1889–1897, 2014.

[24] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[25] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[26] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *Proc. CVPR*, 2015.

[27] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. ECCV*, 2018.

[28] 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.

[29] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *TIP*, 27(8):3676–3690, 2018.

[30] 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. ECCV*, 2014.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Proc. ECCV*, pages 21–37. Springer, 2016.

[32] 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.

[33] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[34] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis 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.

[35] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Multi-modal reasoning graph for scene-text based fine-grained image classification and retrieval. *arXiv preprint arXiv:2009.09809*, 2020.

[36] 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.

[37] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[38] 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.

[39] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3040–3047, 2013.

[40] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1693–1702, 2015.

[41] 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.

[42] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. CVPR*, 2007.

[43] 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.

[44] Joseph Redmon and Ali Farhadi. YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NeurIPS*, pages 91–99, 2015.

[46] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020.

[47] 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.

[48] 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.

[49] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[56] 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.

[57] 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.

[58] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2088–2096, 2019.

[59] 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.

[60] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In *Proc. CVPR*, pages 12113–12122, 2020.

[61] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. Look more than once: An accurate detector for text of arbitrary shapes. In *Proc. CVPR*, pages 10552–10561, 2019.

[62] 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.

[63] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proc. CVPR*, pages 2642–2651. IEEE, 2017.

[64] 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.