Supplementary Material:
StacMR: Scene-Text Aware Cross-Modal Retrieval

1. Introduction

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

2. Additions to Baselines and Re-Ranking

2.1. Full Table of Results on CTC

Table 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 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).

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 [2] 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 [7] 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 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 [3] 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 highlights the important 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 are shown in rows (12-18) using ground-truth scene-text annotations and rows (19-21) using OCR predictions from [3]. In comparison to the models trained on Flickr30K+TextCaps, models trained on Flickr30K obtain similar improvement 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).

2.2. Performance on TextCaps

In order to describe why TextCaps is not fit as an evaluation dataset for StacMR, we performed similar experiments

	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++	✗	✓	✗	-	-	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	✗	✓	✗	-	-	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	-	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')	✗	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)	✗	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')	✗	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)	✗	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)	✗	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)	✗	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)	✗	Fasttext+FV uncleaned	✗	✗	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)	✗	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
(12)	VSE++	VSE++ GRU	✓	✗	GT	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)						LF	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)	VSRN	VSE++ GRU	✓	✗	GT	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)						PSC	30.6	59.3	69.5	16.2	43.2	58.2	14.8	38.8	50.2	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)	VSRN	VSE++ GRU	✓	✗	OCR	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)						PSC	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)	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
(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)						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
(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)	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
(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)						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
(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)	VSRN	VSE++ GRU			GT	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)			✓	✓		PSC	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)	VSRN	VSE++ GRU	✓	✓	OCR	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)			✓	✗		PSC	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)	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
(38)						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
(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)	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
(42)						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
(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 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 2.3)

to those described in Sections 5.1 of the main paper. The main results are shown in Table 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.

2.3. 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 consists 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 $s^{\star}_{LF}$, 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 re-

	Visual Model	Scene-Text Model	Trained on		Combination	TextCaps					
						Image to Text			Text to Image		
			F30K	TC		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 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 2.3)

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 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).

trieved positive item:

$$s^{\star}_{LF}(q,d) = \alpha^{\star}(q)s_v(q,d) + (1-\alpha^{\star}(q))s_t(q,d), \quad (1)$$

$$\alpha^{\star}(q) = \underset{\alpha\in[0,1]}{\arg\min}\left(\operatorname{Rank} s_{LF}(q,d)\right), \quad (2)$$

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 analyse 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, $\alpha^{\star}(q) \sim 1$ indicates that, for this query, the visual model is enough and the scene text should be ignored, $\alpha^{\star}(q) \sim 0$ means that the scene text is enough, and $\alpha^{\star}(q)$ 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 1 rows (33) and (40), and Table 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. We also provide, for the oracle late fusion of row (40), the histogram of optimal values of $\alpha^{\star}$ in 1. We observe that $\alpha^{\star}(q) \sim 1$ more common for text queries than image queries and more common for CTC-5k than CTC-1k. Indeed, text queries and CTC-5k queries have a higher probability to have a zero-word intersection between groundtruth scene text and positive captions, respectively, then image queries and CTC-1k queries, which favors $\alpha^{\star} = 1$.

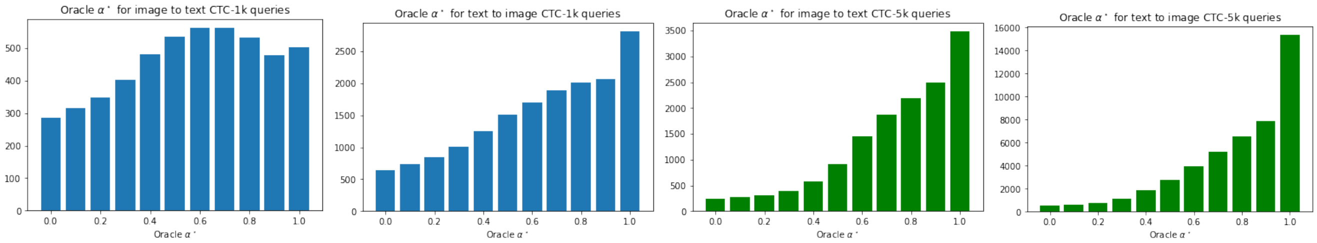


Figure 1: Histogram of α values for oracle late fusion, row (36) of Table 1. Blue histograms show oracle α for CTC-1k, green histograms for CTC-5k.

3. The STARNet Model

3.1. Implementation Details

In the baselines of supervised models, SCAN [5] and VSRN [6] use the same hyper parameters 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 [2] 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 [9] dataset is employed, we use the same training, validation and testing split as in [4], which contain $28,000$, $1,000$ and $1,000$ images respectively. When using only the TextCaps [8] 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.

3.2. Performance on Flickr30K and TextCaps

In Table 3 we show the performance of our proposed model with SCAN [5] and VSRN [6]. In order to obtain comparable results, we have obtained features from our implementation to extract visual regions as [1]. Publicly available code for SCAN [5] and VSRN [6] was used to train those models.

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.

4. Dataset Samples

Figure 2 showcases a few samples of image-caption pairs that belong to the full CTC dataset. On the other hand, in Figure 3 we depict image-caption pairs that belong to the explicit set of the CTC dataset, the bold words in captions reference to appearing scene text. We can note that scene text provides strong cues to better discriminate each image. Leveraging scene-text can provide with important complementary information for language and vision oriented tasks, such as in the case of cross-modal retrieval.

5. Qualitative Results

In Figure 4 we illustrate qualitative results when performing Image to Text cross-modal retrieval. Text contained within an image usually serve 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 5 the results obtained when performing Text to Image cross-modal retrieval. In the queries performed, scene-text work 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 modelling scene-text along with visual features does improve retrieval granularity thus yielding higher performing cross-modal retrieval pipelines.

References

[1] 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.

[2] 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.

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

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

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

[6] 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.

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

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

[9] 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.

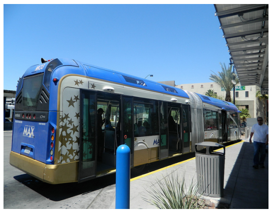
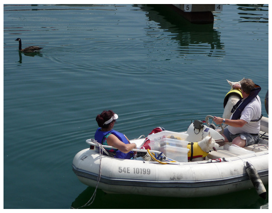
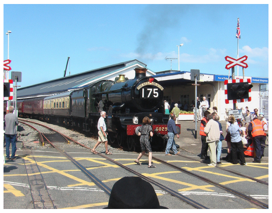
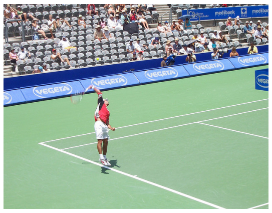
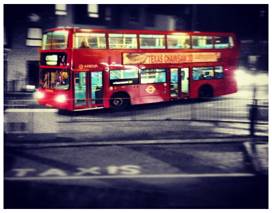
Image	Captions
	A blue bus at a bus stop with its doors open. A bus with its doors open is waiting at a bus stop. A bus sits parked on the side of a street. A picture of a bus on the side of the street. The blue and white trolley is waiting on passengers.
	A woman, man and two dogs in an inflatable raft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

Figure 2: 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.

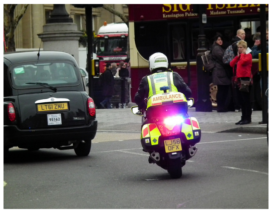
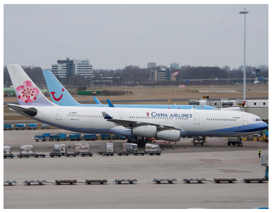
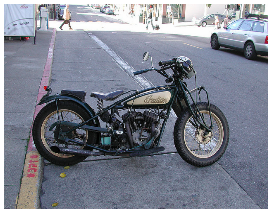
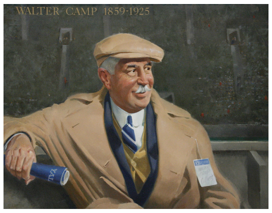
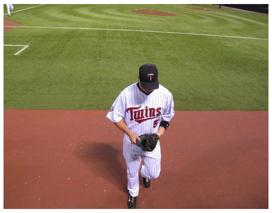
Image	**Captions**
	An emergency response person is on a motorcycle. A medical person riding a motorcycle with **ambulance** on back. A police officer on a motorcycle pulling over a black car. A police motorcycle gets down to business when someone speeds. A motorcycle with a sign on the back that says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

Figure 3: 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**).

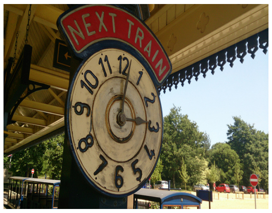
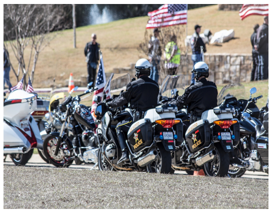
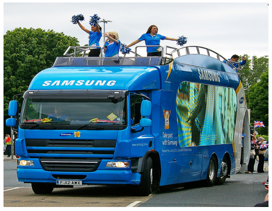
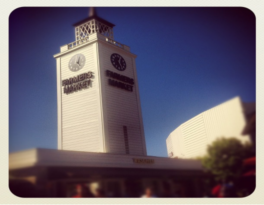
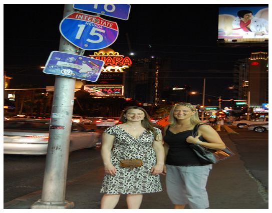
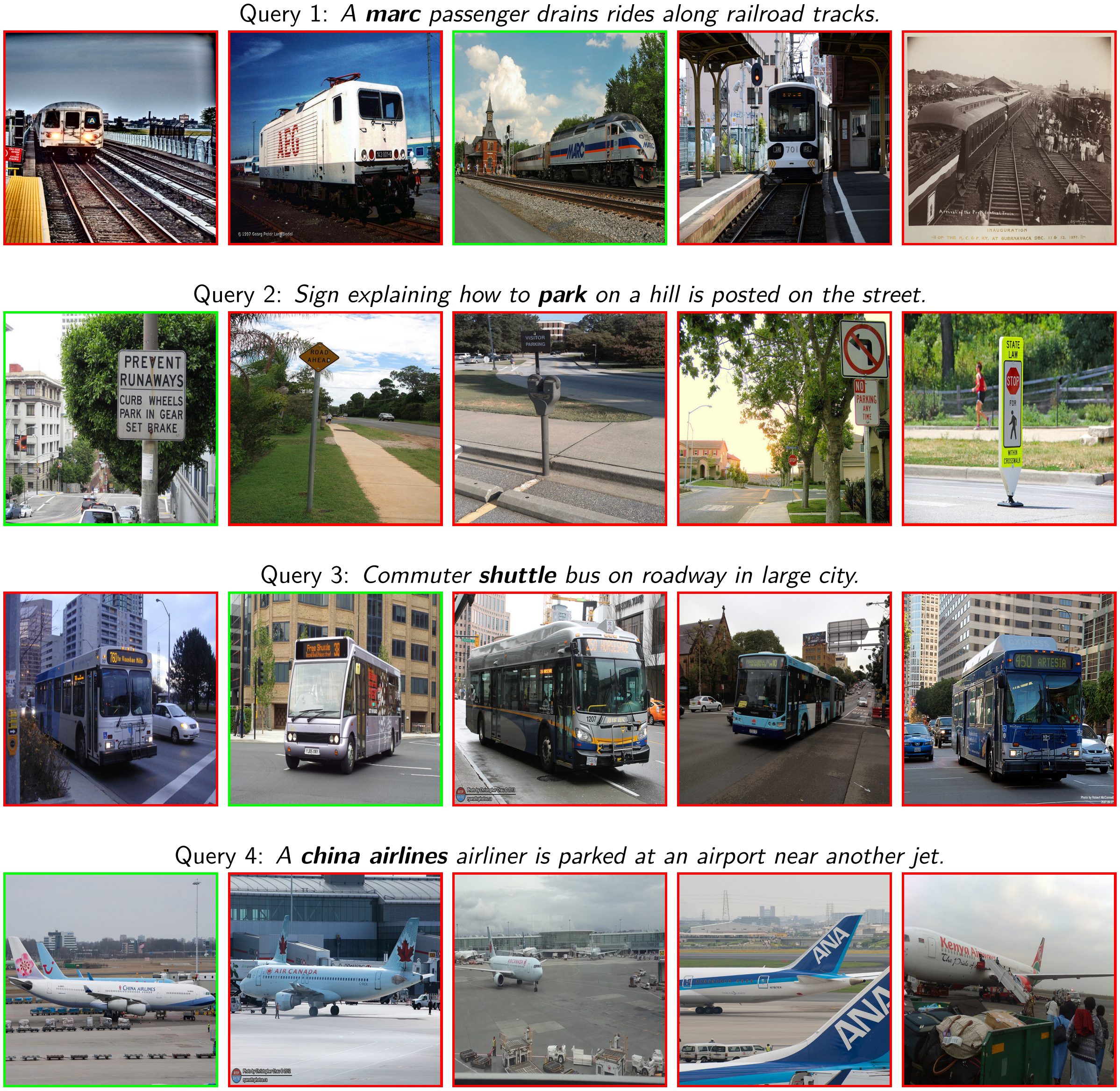
Queried Image	Retrieved Captions
	Clock at a **train** station showing the time of the next trains arrival. ✓
	A clock with the words **next train** written about it. ✓
	A clock on a **train** platform during day time. ✗
	A clock attached to a pole at a **train** station. † ✗
	A clock that is sitting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

Figure 4: 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 5: 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”.