Scene-Text Aware Image and Text Retrieval with Dual-Encoder

Shumpei Miyawaki[1], **Taku Hasegawa**[2], **Kyosuke Nishida**[2], **Takuma Kato**[1], **Jun Suzuki**[1]
[1]Tohoku University, [2]NTT Human Informatics Laboratories

Abstract

We tackle the tasks of image and text retrieval using a dual-encoder model in which images and text are encoded independently. This model has attracted attention as an approach that enables efficient offline inferences by connecting both vision and language in the same semantic space. However, whether an image encoder as part of a dual-encoder model can interpret scene-text, i.e., the textual information in images, is unclear. We propose pre-training methods that encourage a joint understanding of the scene-text and surrounding visual information. The experimental results demonstrate that our methods improve the retrieval performances of the dual-encoder models.

1 Introduction

When pre-trained on a large-scale corpus of image and text pairs, vision and language models can obtain effective multi-modal representations that bridge the semantic gap between visual and textual information. In general, two approaches are used: 1) the cross-encoder approach, in which textual and visual information are jointly fed into a single Transformer-based model (Vaswani et al., 2017), and 2) the dual-encoder approach, in which the textual and visual information are independently fed into two modality-specific encoders. Cross-encoder models use cross-modal attention, which facilitates the interpretation of the different modalities. However, such models are not suitable for image retrieval and other tasks requiring fast and large-scale inferences (Miech et al., 2021; Luan et al., 2021). In contrast, dual-encoder models can make quick inferences, but their interpretation of concomitant modalities is insufficient; in particular, such models have difficulty jointly interpreting scene-text and the surrounding visual information.

Given the above background, this paper investigates the effectiveness of incorporating scene-text into a dual-encoder. The contributions of this study are as follows. 1) We introduce pre-training methods for a dual-encoder to facilitate a joint interpretation of the textual information in the images and surrounding visual information (Figure 1). The performance of the model is then evaluated for image and text retrieval tasks. 2) We experimentally show that, similar to cross-encoder approaches, the joint scene-text and semantic representations improve the retrieval performance of the dual-encoder.

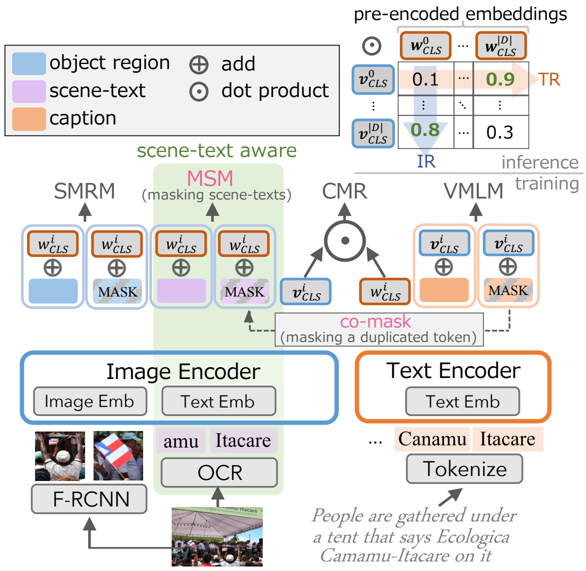


Figure 1: **Overview of the proposed architecture.** We propose pre-training methods to enable the image encoder to jointly interpret the scene-text and surrounding visual information.

2 Related Work

To make sense of visual and textual semantics, recent studies concerning vision and language pre-training, such as image captioning and text-aware VQA (Singh et al., 2019; Biten et al., 2019; Mishra et al., 2019; Mathew et al., 2021), incorporate concomitant textual information, such as scene-text and object tags, in terms of regions-of-interest to

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, pages 422 - 433
May 22-27, 2022 ©2022 Association for Computational Linguistics

enable cross-modal interactions using self-attention in a Transformer-based model (cross-encoder) (Hu et al., 2020; Li et al., 2020; Yang et al., 2021; Tanaka et al., 2021; Biten et al., 2021). However, cross-encoders are not suitable for image retrieval or other tasks requiring fast and large-scale inferences. Although cross-encoder models typically allow expressive token-wise interactions for an input pair of a query and retrieval target, the similarity score cannot be decomposed and is not indexable (Miech et al., 2021; Luan et al., 2021). Therefore, such models are impractical for application in tasks with many queries requiring quick responses, such as retrieval tasks.

In contrast, dual-encoder approaches (Sun et al., 2021; Alec et al., 2021; Jia et al., 2021; Yao et al., 2021) can successfully perform downstream tasks, enabling efficient offline inferences of all pre-encoded image and text embeddings. However, the effectiveness of incorporating concomitant modalities, such as scene-text, in dual-encoder models has not been thoroughly investigated or demonstrated in the community.

3 Scene-Text Aware Dual-Encoder

This paper proposes the incorporation of textual information in images into the dual-encoder architecture. We build our method based on the LightningDOT (Sun et al., 2021) framework, a cutting-edge dual-encoder that encodes both object-wise and token-wise representations. We first briefly introduce LightningDOT in its current use. We then describe the proposed method, including the learning objectives, to incorporate the textual information in the images into the image encoder.

3.1 LightningDOT

LightningDOT outputs a visual feature $\boldsymbol{V}$ and a textual feature $\boldsymbol{W}$[1]. To obtain a visual feature, LightningDOT first extracts multiple objects from an input image using a pre-trained object detector based on Faster R-CNN (Anderson et al., 2018). The obtained visual feature $\boldsymbol{V}$ is a list of vectors, namely, $\boldsymbol{V} = (\boldsymbol{v}_{\texttt{CLS}}, \boldsymbol{v}_1, \dots, \boldsymbol{v}_I)$, where I is the number of extracted objects and $\boldsymbol{v}_{\texttt{CLS}}$ is the vector for a special object "CLS." Similarly, the textual feature is a list of vectors $\boldsymbol{W} = (\boldsymbol{w}_{\texttt{CLS}}, \boldsymbol{w}_1, \dots, \boldsymbol{w}_J)$, where J is the number of tokens in a given caption and $\boldsymbol{w}_{\texttt{CLS}}$ is the vector for a special token "CLS." LightningDOT attempts three pre-training objectives: (1) visual-embedding fused masked language modeling (VMLM), (2) semantic-embedding fused masked region modeling (SMRM), and (3) cross-modal retrieval (CMR). Both VMLM and SMRM predict masked tokens from their surrounding context. Let $\mathcal{M}$ represent a set of mask indices. $\boldsymbol{W}_{\setminus \mathcal{M}}$ denotes $\boldsymbol{W}$ after substituting all m-th vectors of $m \in \mathcal{M}$ in $\boldsymbol{W}$ with the special vector assigned to the `[MASK]` token. Similarly, $\boldsymbol{V}_{\setminus \mathcal{M}}$ is $\boldsymbol{V}$ after substituting the m-th indices of all $m \in \mathcal{M}$ with the `[MASK]` vector[2]. The training objectives of VMLM and SMRM are formulated as follows:

$$\mathcal{L}_\theta^{(*)}(\mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathcal{L}_\theta^{(*)}(m, \mathcal{M}). \quad (1)$$

Here, the mask index for the caption feature $\mathcal{M}_w$ lies in the range of $2, \dots, I+1$ because an index of 1 corresponds to $\boldsymbol{w}_{\texttt{CLS}}$, which is not masked. The VMLM objective $\mathcal{L}_\theta^{(\texttt{VMLM})}(\mathcal{M}_w)$ can then be written by substituting $\mathcal{L}_\theta^{(*)}(m, \mathcal{M})$ into Eq. 1 with

$$\mathcal{L}_\theta^{(\texttt{VMLM})}(m, \mathcal{M}_w) = \ell_\theta(\boldsymbol{w}_m | \boldsymbol{W}_{\setminus \mathcal{M}_w}, \boldsymbol{v}_{\texttt{CLS}}), \quad (2)$$

where $\ell_\theta(\cdot) = -\log(P_\theta(\cdot))$. Similarly, the SMRM objective $\mathcal{L}_\theta^{(\texttt{VMLM})}(\mathcal{M}_v)$ can be obtained with

$$\mathcal{L}_\theta^{(\texttt{SMRM})}(m, \mathcal{M}_v) = \mathcal{D}_\theta(\boldsymbol{v}_m | \boldsymbol{V}_{\setminus \mathcal{M}_v}, \boldsymbol{w}_{\texttt{CLS}}) \quad (3)$$

where $\mathcal{M}_v = \{2, \dots, J+1\}$ and D_θ is any differentiable distance function[3].

The CMR task leverages the paired semantics between the visual and textual representations. Specifically, the similarity (obtained by calculating the inner product $\mathrm{sim}(\boldsymbol{w}_{\texttt{CLS}}, \boldsymbol{v}_{\texttt{CLS}}) = \boldsymbol{w}_{\texttt{CLS}} \cdot \boldsymbol{v}_{\texttt{CLS}}$ is optimized to promote pair matching with in-batch negative sampling. The details of CMR are omitted here because this objective is not related to the presented extensions of the proposed method.[4]

3.2 LightningDOT with scene-text

To obtain scene-text features from images, we apply an optical character recognition (OCR) system to each input image. Each token in the scene-text obtained by OCR is then converted to a d_v-dimensional token embedding ("Text Emb" in Figure 1). Let $\boldsymbol{s}_k$ be the embeddings corresponding to

[1] Appendix A provides additional details of LightningDOT.

[2] The `[MASK]` for the visual feature is the zero vector.

[3] The goal of the model prediction is to reconstruct the masked features themselves (masked region feature regression) or their object class (masked region classification with the Kullback–Leibler divergence)

[4] See Appendix A.2 for additional details concerning CMR.

the k-th token in the scene-text, and let K denote the number of tokens in the scene-text. We then modify and redefine the visual feature $\boldsymbol{V}$ as the concatenation of the visual features explained in Section 3.1 and the textual features $\boldsymbol{s}_k$ in the images, that is, $\boldsymbol{V} = (\boldsymbol{v}_{\texttt{CLS}}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_I, \boldsymbol{v}_{\texttt{SEP}}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_K)$, where $\boldsymbol{v}_{\texttt{SEP}}$ is a vector of separators.

3.3 Masked scene-text modeling (MSM)

This section proposes masked scene-text modeling (MSM) for training the scene-text features. We extended VMLM such that the mask prediction is applied directly to the scene-text. By masking only the textual information in the scene-text, the model can read the scene-text from the surrounding visual information. Let $\mathcal{M}_s = \{I+2, \ldots, I+K+2\}$. $\mathcal{M}_s$ is the mask for the scene-text.[5] Similar to the SMRM objective, the MSM objective $\mathcal{L}_\theta^{(\texttt{MSM})}(\mathcal{M}_s)$ can be obtained via Eq. 1 by substituting $\mathcal{L}_\theta^{(*)}(m, \mathcal{M})$ with

$$\mathcal{L}_\theta^{(\texttt{MSM})}(m, \mathcal{M}_s) = \ell_\theta(\boldsymbol{s}_m | \boldsymbol{V}_{\setminus \mathcal{M}_s}, \boldsymbol{w}_{\texttt{CLS}}). \quad (4)$$

3.4 Cross-modal VMLM (co-mask)

Inspired by Alexis and Guillaume (2019); Zhou et al. (2021), we also propose a cross-modal co-masking strategy (co-mask) that leverages the cross-modal correspondence. Following the same strategy as VMLM, we randomly replace a token from a caption and then simultaneously replace the duplicated token from the scene-text in `[MASK]` to promote cross-modal relationships. When at least one paired token exists between a caption and a scene-text and is outside the targets for masking, we randomly select one masked token and switch the masking target to the paired token. While both VMLM and MSM promote multi-modal relationships between the textual information in the images and a caption describing the scene image, the "co-mask" promotes textual semantic alignment to leverage cross-modal relationships.

4 Experiments

We designed experiments to investigate the effectiveness of incorporating the scene-text as an additional feature for visual features in image and text retrieval tasks.

4.1 Experimental setup

Dataset As the training and evaluation dataset, we selected TextCaps (Sidorov et al., 2020) because it provides "caption," "image," and "scene-text"[6] triples. TextCaps includes $22,953$ images and $109,764$ captions on training set, and $3,166$ images and $15,830$ captions on development set. Each image is described by five human-annotated captions. Textual information in an image context can be correctly extracted from the TextCaps data because 96.9% of the images and 81.3% of the captions contain scene-text.

Base model Following Sun et al. (2021), we used BERT (Devlin et al., 2019) as the text encoder and UNITER (Yen-Chun et al., 2020) as the image encoder. Note that we used UNITER as the image encoder only, not as the cross-encoder, although it can also simultaneously model text. This is because the inference speed of UNITER, as reported by Sun et al. (2021), is too slow for practical use in retrieval tasks[7]. In our setting, we employed the dual-encoder to model captions and images. However, the scene-text was concatenated with the visual features and input to the image encoder because this text is part of the visual information. The scene-text vocabulary of the image encoder was initialized with that of the text encoder.

Pre-training setting To pre-train LightningDOT with four tasks, MSM, CMR, VMLM (with co-mask), and SMRM, we randomly sampled one task for each mini-batch with $1:2:1:1$ weightings[8] for $300,000$ optimization steps.[9]

Conventional models To reveal the effectiveness of the proposed method, we compared its retrieval performance with those of the SCAN (Lee et al., 2018), VSRN (Li et al., 2019), and STARNet (Mafla et al., 2021) models, which were tested by Mafla et al. (2021). All models were trained on TextCaps and evaluated on its development set. We compared STARNet as a baseline for modeling the interaction among scene text, visual objects, and captions. The difference from the proposed method

[5]The index for the scene-text starts at $I+2$ because we redefine $\boldsymbol{V} = (\boldsymbol{v}_{\texttt{CLS}}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_I, \boldsymbol{v}_{\texttt{SEP}}, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_K)$.

[6]To obtaining the scene-text using OCR, Sidorov et al. (2020) employed Rosetta-en (Borisyuk et al., 2018).

[7]In an identical setting, the inference speed of LightningDOT is 639× faster than that of UNITER on the Flickr30K (Plummer et al., 2015) test set, in which the retrieval target includes 1K images

[8]SMRM was divided into MRFR and MRC-kl tasks. These weights were allocated with a ratio of $1:1$.

[9]Appendix A.3 describes the implementation details.

	IR@k			TR@k		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
VSRN	9.5	26.2	37.2	14.3	34.9	46.2
SCAN	14.1	37.6	52.1	23.2	50.5	63.5
STARNet	19.8	40.1	51.6	28.7	53.7	65.1
LightningDOT	16.6	36.0	46.2	21.3	43.6	54.5
w/ ST	38.7	60.4	68.4	50.6	73.7	81.3
w/ ST+co-mask	39.4	61.6	70.2	52.3	74.8	82.2
w/ ST+MSM	**40.5**	**63.0**	**71.1**	**52.9**	**76.4**	**83.2**

Table 1: **Results of the image (IR) and text retrieval (TR) performances with recall@k on the TextCaps development set.** We extended LightningDOT to input scene-text (w/ ST). In addition, we evaluated our proposed method with the co-mask and MSM.

modality	model	IR@k			TR@k		
		$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
IMG+ST	w/ ST	38.7	60.4	68.4	50.6	73.7	81.3
	+co-mask	39.4	61.6	70.2	52.3	74.8	82.2
	+MSM	**40.5**	**63.0**	**71.1**	**52.9**	**76.4**	**83.2**
IMG	w/ ST	11.6	28.2	37.9	14.1	31.3	41.6
	+co-mask	**13.3**	**31.5**	**42.1**	**16.0**	**34.1**	**45.3**
	+MSM	11.7	29.1	39.3	13.8	32.0	41.6
ST	w/ ST	0.0	0.1	0.3	5.0	15.4	24.7
	+co-mask	0.0	0.2	0.4	12.9	28.8	37.9
	+MSM	**16.7**	**31.4**	**37.8**	**16.1**	**33.3**	**42.0**

Table 2: **Ablation study on selecting visual modalities.** The "modality" indicates the input for the image encoder, which is used as the retrieval target in image retrieval (IR) and as the query in text retrieval (TR).

is that STARNet is trained by using the triplet ranking loss. Moreover, the final visual representations are obtained via a dot product following a graph convolutional network (Kipf and Welling, 2017)[10] on scene-text and visual objects.

Inference The visual and textual embeddings ($\boldsymbol{v}_{\texttt{CLS}}, \boldsymbol{w}_{\texttt{CLS}}$) from the development set were independently indexed using FAISS (Johnson et al., 2021). We then conducted an exact maximum inner product search (IndexFlatIP) for each query embedding, that is, for each $\boldsymbol{w}_{\texttt{CLS}}$ in the image retrieval (IR) and each $\boldsymbol{v}_{\texttt{CLS}}$ in the text retrieval (TR). The image retrieval (IR@k) and text retrieval (TR@k) tasks were evaluated in terms of the recall at k.

4.2 Retrieval results

Table 1 shows the retrieval performances of the tested methods on the TextCaps development set. In our experimental setting, the baseline Lighting-DOT model consistently delivered an inferior performance compared with that of STARNet. After considering scene-text (w/ ST), the performances in both the IR and TR settings were significantly improved and surpassed that of STARNet. Our proposal, which incorporates the co-mask (w/ ST+co-mask) and the MSM objective (w/ ST+MSM), further improved the retrieval performance. These observations indicate that modeling the scene-text directly is effective for modeling visual information that enhances semantic affinities with captions.

4.3 Ablation study on visual modalities

To investigate whether the image encoder can interpret the joint visual information in scene-text and object regions, we evaluated the retrieval performance by excluding one of the modalities. When the object regions or the scene-text alone was input into the image encoder, the retrieval performance was significantly reduced in the TR and IR settings (see Table 2). The cross-modal masking strategy (w/ ST+co-mask) improved the modeling compared with that of the scene-text strategy (w/ ST) on both modalities but was especially effective in the object regions. MSM (w/ ST+MSM) for multi-modal optimization improved the modeling of the scene-text but had a small effect on the images. These results suggest the necessity of modeling not only joint representations of visual and textual semantics in images but also fine-grained cross-modal relationships in future work.

4.4 Benefit of duplicated tokens

Here, we define the term **duplicated token** as a token that appears both in the caption and in the scene-text. To investigate whether the retrieval model leverages cross-modal relationships, we focus on the duplicated tokens because we will obtain a higher performance if such tokens share an adequate amount of information. For example, given a query that includes "Coca-Cola," the model was able to leverage the modality of the scene-text when retrieving an image of a can or a bottle that was labeled not as "Pepsi" but as "Coca-Cola." We evaluated the retrieval performance via accuracy@k on the development set in TextCaps (Sidorov et al., 2020) versus the number of duplicated tokens. We used spaCy[11] to narrow down the content tokens [12]

[10] The output of the scene-text and visual objects are fed into the average pooling layer and gated recurrent unit (Cho et al., 2014), respectively.

[11] https://spacy.io/

[12] Their part of speech tags are in "ADJ," "ADV," "NOUN," "PROPN," and "VERB".

task (retrieval targets)		IR (image and scene-text)				TR (caption)			
# of duplicated tokens		0	1	2	3	0	1	2	3
total # of tokens for retrieval targets		2,302	512	212	94	11,785	2,484	1,004	342
w/ ST	acc@1	**51.13**	47.85	50.94	52.13	36.25	41.14	51.10	**55.56**
	acc@5	**74.28**	71.48	73.58	68.09	57.92	62.80	71.31	**82.46**
	acc@10	81.75	80.08	**82.08**	76.60	66.26	70.57	78.39	**86.55**
w/ ST+co-mask	acc@1	52.48	**52.54**	51.89	50.00	37.07	41.14	50.10	**61.70**
	acc@5	**75.33**	74.61	70.75	70.21	59.30	62.76	72.11	**84.80**
	acc@10	82.41	81.64	79.72	**85.11**	68.25	71.30	78.69	**89.47**
w/ ST+MSM	acc@1	**53.52**	52.54	50.47	50.00	38.22	41.67	52.09	**62.28**
	acc@5	**77.15**	75.20	72.64	74.47	60.76	63.93	75.50	**80.12**
	acc@10	**84.06**	81.64	79.72	78.72	69.29	71.70	81.18	**86.55**

Table 3: **Retrieval accuracy versus the number of duplicated tokens between the caption and the scene-text.**

	IR@k			TR@k		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
LightningDOT	**17.2**	**37.7**	**48.9**	22.6	45.4	55.5
(mul - en)	$+0.6$	$+1.7$	$+2.6$	$+1.3$	$+1.8$	$+1.0$
w/ ST+MSM	0.0	44.5	57.8	**35.0**	**61.3**	**71.2**
(mul - en)	-40.5	-18.5	-13.3	-17.9	-15.1	-12.0

Table 4: **Retrieval performance on the development set in a multilingual setting.** We employed multilingual BERT and show differences obtained by subtracting the recall@k of the monolingual BERT (en) from that of the multilingual BERT (mul).

because the scene-text detected by an OCR system contains a large number of false positive tokens.

From Table 3, we can see that the retrieval performance in TR is proportional to the number of duplicated tokens. This indicates that duplicated tokens are one of the factors that enhance the semantic affinity between a caption and the scene-text[13]. In the IR setting, conversely, the retrieval performance does not depend on the number of duplicated tokens when the objectives are "w/ ST" and "w/ ST+co-mask." However, when using the MSM objective, the retrieval performance in IR is degraded depending on the number of duplicated tokens. According to these results, the performance gap is the result of differences in the modality of the retrieval target (textual or visual semantics) and in the inclusion of informative tokens between the scene-text and caption.

4.5 Effectiveness of multilingual text encoder

Modeling scene-text is not so easy; we have to essentially deal with various languages since they depend on where the picture was taken and where the product was made in scene-text (Chen et al., 2021). Recently, Biten et al. (2021) pre-trained a model on a large text corpus and reported the robustness of their model with respect to the OCR errors. We also investigated the model performance with multilingual (mul) BERT (Devlin et al., 2019) as the text encoder in the baseline LightningDOT and LightningDOT with MSM settings. Note that the vocabulary size $(119,547)$ of the multilingual BERT is approximately four times as large as that of its monolingual counterpart $(28,996)$.

Compared with the monolingual encoder, the multilingual encoder increased the retrieval performance in the baseline method (LightningDOT) but degraded the performance when using the scene-text (w/ ST+MSM). In the multilingual setting, the LightningDOT baseline could model the joint representations well because the pre-training corpus size and token fertility between the multilingual and monolingual BERT were nearly the same (Rust et al., 2021). In contrast, the degradation resulting from using scene-text in the multilingual setting indicates that scene-text may still be underrepresented or that false positive tokens due to OCR errors may harm the model. A better usage of multilingual BERT in scene-text needs to be explored in future work.

5 Conclusion

We proposed a framework that incorporates the textual information in images into the dual-encoder architecture. An evaluation on the TextCaps dataset confirmed that modeling the scene-text-aware cross-modal relationships benefited the dual-encoder architecture. In future research, we will attempt a more robust exploration of scene-text modeling (Singh et al., 2021; Wang et al., 2021b,a).

[13]Note that it may be possible to make the prediction easier because captions and images in TextCaps contain scene-text.

Acknowledgments

We thank the three reviewers for their valuable comments and suggestions to improve our paper.

References

Radford Alec, Wook Kim Jong, Hallacy Chris, Ramesh Aditya, Goh Gabriel, Agarwal Sandhini, Sastry Girish, Askell Amanda, Mishkin Pamela, Clark Jack, Krueger Gretchen, and Sutskever Ilya. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763.

Conneau Alexis and Lample Guillaume. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. 2021. LaTr: Layout-Aware Transformer for Scene-Text VQA. *CoRR*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. 2019. Scene Text Visual Question Answering. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*.

Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. 2018. 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, KDD 2018, London, UK, August 19-23, 2018*, pages 71–79.

Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. 2021. Text Recognition in the Wild: A Survey. *ACM Comput. Surv.*, pages 42:1–42:35.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9989–9999.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 4904–4916.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data*, pages 535–547.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations ICLR*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, pages 212–228.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4653–4661.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, pages 121–137.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations ICLR*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Trans. Assoc. Comput. Linguistics*, pages 329–345.

Andres Mafla, Rafael S. Rezende, Lluis Gomez, Diane Larlus, and Dimosthenis Karatzas. 2021. StacMR: Scene-Text Aware Cross-Modal Retrieval. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2220–2230.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A Dataset for VQA on Document Images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 9826–9836.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, pages 742–758.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA Models That Can Read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. TextOCR: Towards Large-Scale End-to-End Reasoning for Arbitrary-Shaped Scene Text. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8802–8812.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. LightningDOT: Pre-training Visual-Semantic Embeddings for Real-Time Image-Text Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. VisualMRC: Machine Reading Comprehension on Document Images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13878–13888.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. 2021a. Scene Text Retrieval via Joint Text Detection and Similarity Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4558–4567.

Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. 2021b. Improving OCR-Based Image Captioning by Incorporating Geometrical Relationship. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1306–1315.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. TAP: Text-Aware Pre-Training for Text-VQA and Text-Caption. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8751–8761.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. *CoRR*.

Chen Yen-Chun, Li Linjie, Yu Licheng, El Kholy Ahmed, Ahmed Faisal, Gan Zhe, Cheng Yu, and Liu Jingjing. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, pages 104–120.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal Cross-Lingual Cross-Modal Vision-and-Language Pre-Training. In *IEEE Conference* *on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4155–4165.

A Detailed Explanation of LightningDOT

A.1 Input tokens for the image encoder

As mentioned in Section 3.1, LightningDOT (Sun et al., 2021) first extracts multiple object regions from an input image using a pre-trained object detector based on Faster R-CNN (Anderson et al., 2018). Let I represent the number of extracted objects. In fact, the object detector provides two features: object regions and their locational features[14]. From these features, "Image Emb" (Figure 1) regenerates the input features to the image encoder. Specifically, an object feature and its locational feature are projected into the same d_v-dimensional space using an independent fully connected layer and then their embeddings are summed and finally fed into the normalization layer. By this means, input features $\boldsymbol{O}$ for object regions can be obtained, that is, $\boldsymbol{O} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_I)$.

The proposed method described in Section 3.2 also extracts multiple tokens of scene-text from an input image using an OCR system (Rosetta-en (Borisyuk et al., 2018)). Let K represent the number of tokenized tokens for the scene-text. In addition, we apply positional indices to each token instead of the locational features. Similar to "Image Emb," the input feature of the scene-text is obtained by "Text Emb" (Figure 1). Specifically, a scene-text token and its positional index are looked up in their d_v-dimensional embeddings and then their embeddings are summed and finally fed into the normalization layer. By this means, the input features $\boldsymbol{T}$ for the scene-text tokens can be obtained, that is, $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_K)$.

We denote an image encoder as f_{θ_v}. In the baseline setting, the image encoder encodes $\boldsymbol{V} = f_{\theta_v}(\tilde{\boldsymbol{v}}_{\texttt{CLS}}, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_I)$, where $\tilde{\boldsymbol{v}}_{\texttt{CLS}}$ is a special object "CLS." In our setting of a scene-text aware framework, the image encoder encodes $\boldsymbol{V} = f_{\theta_v}(\tilde{\boldsymbol{v}}_{\texttt{CLS}}, \boldsymbol{o}_1, \ldots, \boldsymbol{o}_I, \tilde{\boldsymbol{v}}_{\texttt{SEP}}, \boldsymbol{t}_1, \ldots, \boldsymbol{t}_K)$, where $\tilde{\boldsymbol{v}}_{\texttt{SEP}}$ is a special object "SEP."

A.2 Cross modal retrieval

Cross modal retrieval (CMR) is a task leveraging the paired semantics between the visual and textual representations. Specifically, the similarity according to the inner product $\mathrm{sim}(\boldsymbol{w}_{\texttt{CLS}}, \boldsymbol{v}_{\texttt{CLS}}) = \boldsymbol{w}_{\texttt{CLS}} \cdot \boldsymbol{v}_{\texttt{CLS}}$ is optimized to promote a matched pair and vice versa with in-batch negative sampling[15]:

$$\mathcal{L}^{(\texttt{CMR})}(B) = \frac{1}{2B} \sum_{b=1}^{B} \mathcal{L}^{(\texttt{TR})}(b) + \mathcal{L}^{(\texttt{IR})}(b) \quad (5)$$

$$\mathcal{L}^{(\texttt{TR})}(b) = -\log \left(\frac{e^{\mathrm{sim}(\boldsymbol{v}^b_{\texttt{CLS}}, \boldsymbol{w}^b_{\texttt{CLS}})}}{\sum_{j=1}^{B} e^{\mathrm{sim}(\boldsymbol{v}^b_{\texttt{CLS}}, \boldsymbol{w}^j_{\texttt{CLS}})}} \right) \quad (6)$$

$$\mathcal{L}^{(\texttt{IR})}(b) = -\log \left(\frac{e^{\mathrm{sim}(\boldsymbol{w}^b_{\texttt{CLS}}, \boldsymbol{v}^b_{\texttt{CLS}})}}{\sum_{i=1}^{B} e^{\mathrm{sim}(\boldsymbol{w}^b_{\texttt{CLS}}, \boldsymbol{v}^i_{\texttt{CLS}})}} \right), \quad (7)$$

where B is the number of instances in a single (mini-)batch during the training process.

A.3 Implementation details

The model dimensions of both encoders are set to 12 Transformer layers, 768 hidden dimensions, and 12 attention heads. In our masking strategy, following Devlin et al. (2019), we decomposed 15% of the total input tokens into 80% `[MASK]`, 10% random tokens, and 10% unchanged. We used AdamW (Loshchilov and Hutter, 2019) as the optimizer for pre-training with $\beta_1 = 0.9 and \beta_2 = 0.98$ and set the learning rate to $5e-5$. We adopted a learning rate warmup strategy, where the learning rate was linearly increased during the first $10,000$ training steps, followed by a linear decay to 0. We set the L2 weight decay to 0.01. We set the batch size to 4096 per GPU with six accumulation steps.

A.4 Qualitative examples

In this section, we show several qualitative results of the top-5 image retrievals using the TextCaps development set (Sidorov et al., 2020). We compare two models, "LightningDOT" and "LightningDOT w/ST+MSM," which showed the best scores in Table 1. Figure 2 and 3 show true positive examples when employing the MSM objective with the scene-text. The results indicate that both models can retrieve similar images given the entity level information and that the model using the MSM objective retrieved appropriate images, including the scene-text of *"Voll-Damm"* (Figure 2b) and *"Sibelius Symphonies from Minnesota Orchestra"* (Figure 3b). Figure 4 shows true negative examples. In the case when it is necessary to achieve reading comprehension, our proposed method does not work well. For a more robust and fine-grained comprehension, we need to consider the geometrical relationships between multiple scene-texts (Wang

[14] Each locational feature consists of seven-dimensional vectors: normalized top, left, bottom, and right coordinates, width, height, and area.

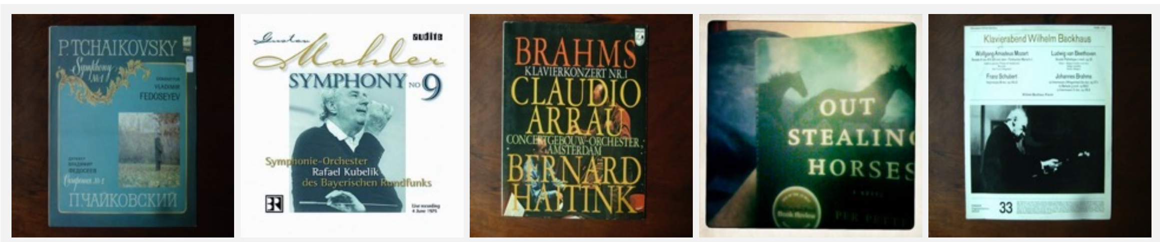
[15] Other images and captions in the mini-batch are selected as negative instances

et al., 2021b), as well as a pre-training framework with a large-scale text corpus (Biten et al., 2021), in future work.

(a) LightningDOT (out of top-100 range)

(b) LightningDOT w/ ST+MSM (1)

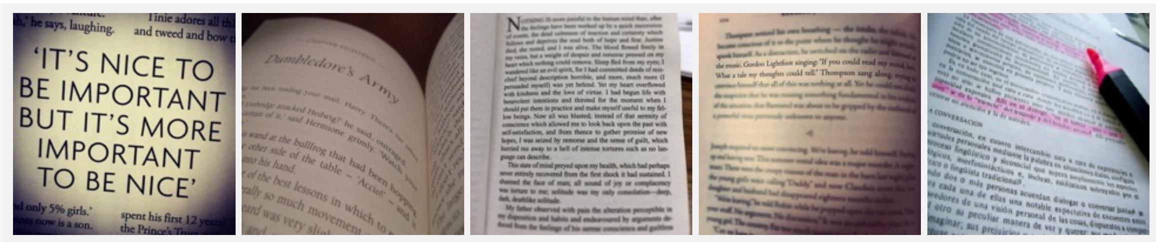
Figure 2: Top-5 retrieval images from the query *"A glass bottle and glass of Voll-Damm beer."* The ground truth is indicated by the green rectangle. The number in parentheses indicates the ranking index of the retrieval result for the positive image.



(a) LightningDOT (33)

(b) LightningDOT w/ ST+MSM (1)

Figure 3: Top-5 retrieval images from the query *"The music book cover with Sibelius Symphonies from Minnesota Orchestra."* The ground truth is indicated by the green rectangle. The number in parentheses indicates the ranking index of the retrieval result for the positive image.



(a) LightningDOT (86)

(b) LightningDOT w/ ST+MSM (20)

Figure 4: Top-5 retrieval images from the query *"Open book on a page that says the young man dried up his tears."* The number in parentheses indicates the ranking index of the retrieval result for the positive image.