Improving Cross-Modal Retrieval with Set of Diverse Embeddings

— Supplementary Material —

Dongwon Kim[1] Namyup Kim[1] Suha Kwak[1,2]
[1]Dept. of CSE, POSTECH [2] Graduate School of AI, POSTECH
https://cvlab.postech.ac.kr/research/DivE/

This supplementary material provides details and additional results of our method that have not been presented in the main paper due to the page limit. In Section A, we first provides further implementation details for each feature extractor settings. We present additional experimental results including in-depth analysis of the model, results on the additional benchmarks, more ablation studies, and embedding space visualization in Section B. Finally, Section C presents more qualitative results for the set prediction module.

A. Implementation Details

Our model is implemented with PyTorch [9] v 1.8.1. Automatic mixed precision is used for faster and more efficient training. In addition to implementation details provided in the main paper, training settings vary based on feature extractors.

ResNet-152 + bi-GRU: In this setting, the model is trained for 120 epochs with the initial learning rate of 1e-3 and 2e-3 on COCO and Flickr30K, respectively. The learning rate for the set prediction module is scaled by 0.1 and 0.01 on COCO and Flickr30K, respectively. The learning rate decays by a multiplicative factor of 0.1 for every 10 epochs. Following [10], the CNN is not trained for the first 50 epochs. Training is performed using a single RTX 3090 GPU.

Faster-RCNN + bi-GRU: The learning rate for the set prediction module is scaled by 0.1 and 0.05 on COCO and Flickr30K, respectively. Following [4], we drop 20% of ROI features and words during training. Training is performed using a single RTX 3090 GPU.

ResNeXt-101 + BERT: In this setting, we construct a batch with 128 images and their entire matching captions. The model is trained for 50 epochs with the initial learning rate of 1e-4, where the learning rate decays by a multiplicative factor of 0.1 for every 20 epochs. Following [4], the learning rate for CNN is scaled by 0.1. The statistics of the batch normalization [3] layer are fixed during training. The CNN is not trained in the first epoch. In the first epoch, triplet loss without mining is used, whereas the hardest negative mining is used for later epochs. Training is performed using two A100 PCIe GPU.

	Image-to-text				Text-to-image			
	ECCV Caption			CxC	ECCV Caption			CxC
	mAP@R	R-P	R@1	R@1	mAP@R	R-P	R@1	R@1
VSRN	30.8	42.9	73.8	55.1	**53.8**	**60.8**	89.2	42.6
VSE_∞	34.8	45.4	81.1	67.9	50.0	57.5	**91.8**	53.7
Ours	**36.0**	**46.4**	**84.7**	**72.3**	51.0	58.5	91.6	**55.5**

Table A1. mAP@R, R-Precision, and Recall@1 are reported for both ECCV Caption and CrissCrossed Caption (CxC). Results on image-to-text retrieval and text-to-image retrieval are reported.

B. Additional Experiments

B.1. In-Depth Computation Cost Analysis

In addition to the computation cost analysis presented in the main paper, we compare our method with SCAN [5] and VSE_∞ [4], focusing on FLOP and latencies. FLOPs and latencies are measured during computing similarity score between data and then obtaining nearest top-10 retrieval result on Flickr30K validation. Given the cost of VSE_∞ as 1 in terms of FLOPs, our method has an approximate cost of 16, while SCAN has a cost of 1,280. VSE_∞, Ours, and SCAN have latencies of 159ms, 168ms, and 198,121ms, respectively. While our model is substantially more efficient than cross-attention based methods like SCAN, it demands more computation than single embedding methods such as VSE_∞.

Nevertheless, we observed that when increasing the embedding dimension of VSE_∞ to match the FLOPs of ours, it results in performance drop of 10.8%p on Flickr30K RSUM. This finding indicates that the improvement we achieved is not merely due to the additional computation, but rather stems from our set-based embedding approach.

B.2. Results on ECCV Caption and CrissCrossed Caption

Recently, benchmarks for the cross-modal retrieval, such as CrissCrossed Caption (CxC) [8] and ECCV Caption [2], have been proposed to address the missing correspondences issue in conventional benchmarks. In particular, within the COCO dataset, each caption associated with only one image, while each image is matched with five different captions. This missing correspondence leads to a large numbers

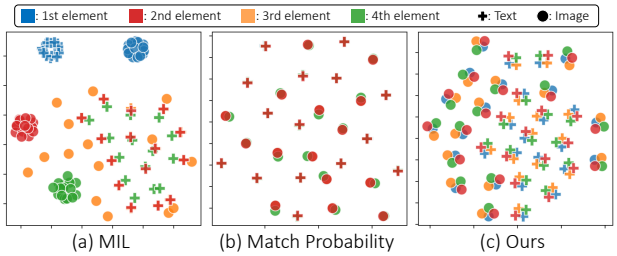

Figure A1. t-SNE visualization of the embedding spaces. Visualized marker represent elements of embedding sets. Color and shape of the markers denote corresponding slots and modality of the data, respectively.

of false-negatives, as captions that may accurately describe other images are overlooked during testing, thus obstructing the correct evaluation of the models. CxC and ECCV caption mitigate the false-negative issue by introducing re-established correspondences between images and captions in the COCO test split.

For the comprehensive evaluation of our model, we report the results of our best model on CxC and ECCV caption, in the Table A1. We compare our method with VSRN [6] and VSE$_\infty$ [4], which are reported to achieve previous best results on ECCV caption and CxC, respectively [2]. It is important to note, however, that VSRN is one of the machine annotators used to construct the ECCV Caption dataset itself, which may introduce potential machine bias into the dataset and result in inflated evaluations. Despite this, our work achieves the best or second-best performance in every metric, which is particularly noteworthy given that the results on ECCV caption are known to have a low correlation with those on conventional benchmarks.

B.3. Ablation Study of Architectural Modification to Slot Attention

As described in the Section 2 of the main paper, we made three modifications to the original slot attention [7]: (1) using learnable embeddings for initial element slots, (2) replacing GRU [1] with a residual sum, and (3) adding a global feature into the final element slots. Without these modifications, training failed, yielding a COCO 5K RSUM of 0.74. Ablations of (1), (2), and (3) result in COCO 5K RSUM of 427.2 (-3.5%p), 423.3 (-7.4%p), and 342.0 (-88.7%p), respectively. It is evident that (3) has the most substantial impact on retrieval performance. This is because a global feature effectively addresses samples with little ambiguity, particularly during the early stages of training when addressing semantic ambiguity is challenging for the network. However, this does not imply that global features dominate the embedding set, as verified by the high circular variance in Table 3 of the main paper.

B.4. t-SNE Visualization of Embedding Space

Figure A1 visualizes embedding spaces of our model trained with different similairty functions through t-SNE. Each marker represents an embedding set element, and its color and shape indicate the corresponding slot and modality, respectively. The visualization shows two side effects of MIL and MP: sparse supervision and set collapsing. MIL leaves some slots untrained, which is observed as clusters of elements produced from the same slots. Conversely, MP suffers from the set collapsing, making it difficult to distinguish set elements in t-SNE, as also verified with the small within-set circular variance in Table 3 of the main paper. Unlike MIL and MP, our smooth-Chamfer similarity enables learning of a model that takes account of every set element, maintaining sufficient within-set variance to encode semantic ambiguity.

C. Additional Qualitative Results

In Figure A2 and Figure A3, we present additional visualization of attention map from the visual set prediction module $f^{\mathcal{V}}$. Visualizations of attention maps, including ones presented in the main paper, are obtained from the model using ResNeXt + BERT feature extractors. Attention maps from each iteration are presented together, where $t = 4$ is the last iteration. For each attention map from the last iteration, its corresponding element of embedding set and nearest caption are provided together. Results show that the aggregation block produces heterogeneous attention maps capturing various semantics, such as different objects (1st row of Figure A2) and action (2nd row of Figure A3). Moreover, in every case, we can observe that element slots are progressively updated to capture distinctive semantics, starting from sparse and noisy attention maps.

Specifically, in Figure A3, we present the examples where the nearest captions of multiple elements are the same. For instance, in the 2nd row of Figure A3, each element attends to individual entities (sky, larger giraffe, grassy area, and baby giraffe), but their nearest captions, which describe the entire scene, are the same. Results imply that by fusing element slots with the global feature, elements of the embedding set can preserve the global context while focusing on distinctive semantics. These characteristics help model when samples with little ambiguity are given, such as a caption describing the entire scene (2nd row of Figure A3) or an image containing a single iconic entity (3rd row of Figure A3).

References

[1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In

Proc. Empirical Methods in Natural Language Processing (EMNLP), 2014. 2

[2] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[3] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning (ICML)*, 2015. 1

[4] Chen Jiacheng, Hu Hexiang, Wu Hao, Jiang Yuning, and Wang Changhu. Learning the best pooling strategy for visual semantic embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[5] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 1

[6] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[7] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020. 2

[8] Zarana Parekh, Jason Baldridge, Daniel Matthew Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for ms-coco. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020. 1

[9] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *AutoDiff, NIPS Workshop*, 2017. 1

[10] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

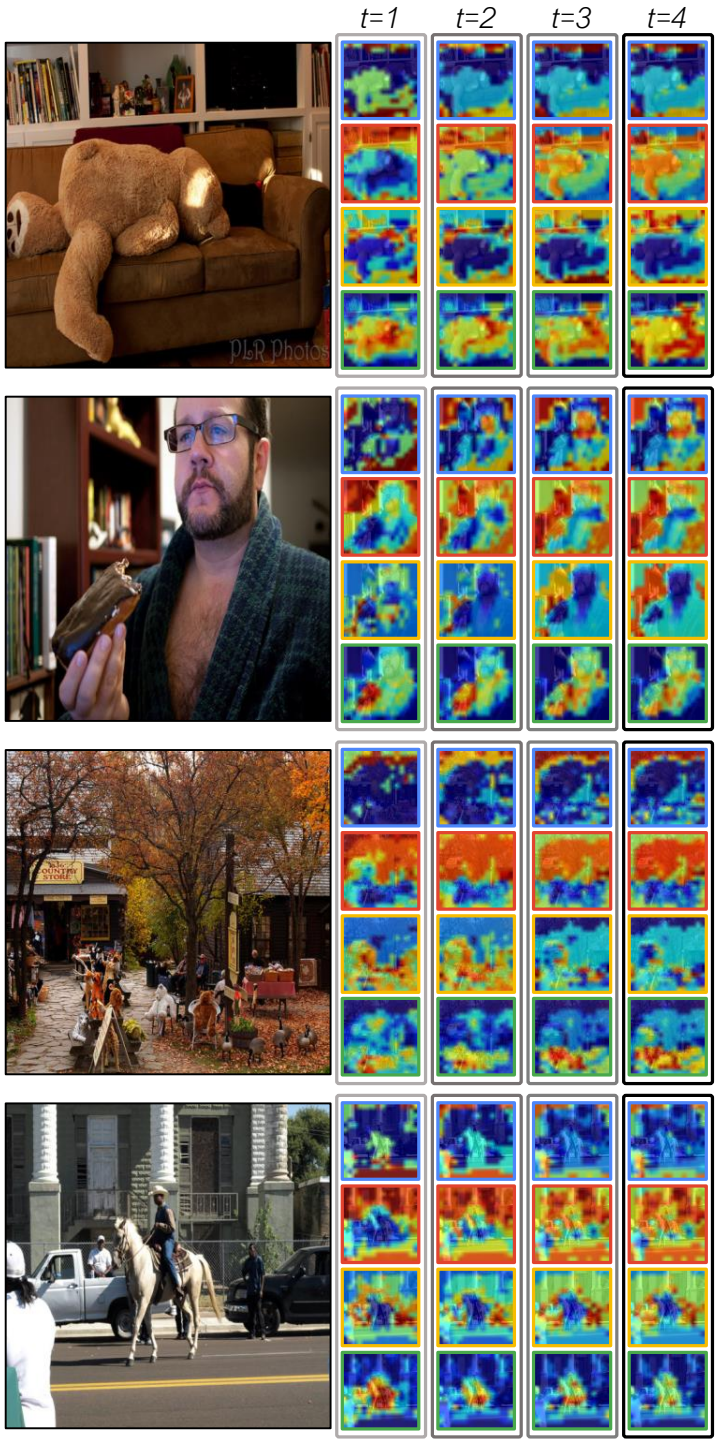

Figure A2. For each element of the image embedding set, we present its attention map and the caption nearest to the element in the embedding space. Matching captions are colored in green. Entities corresponding to the attention maps are underlined.

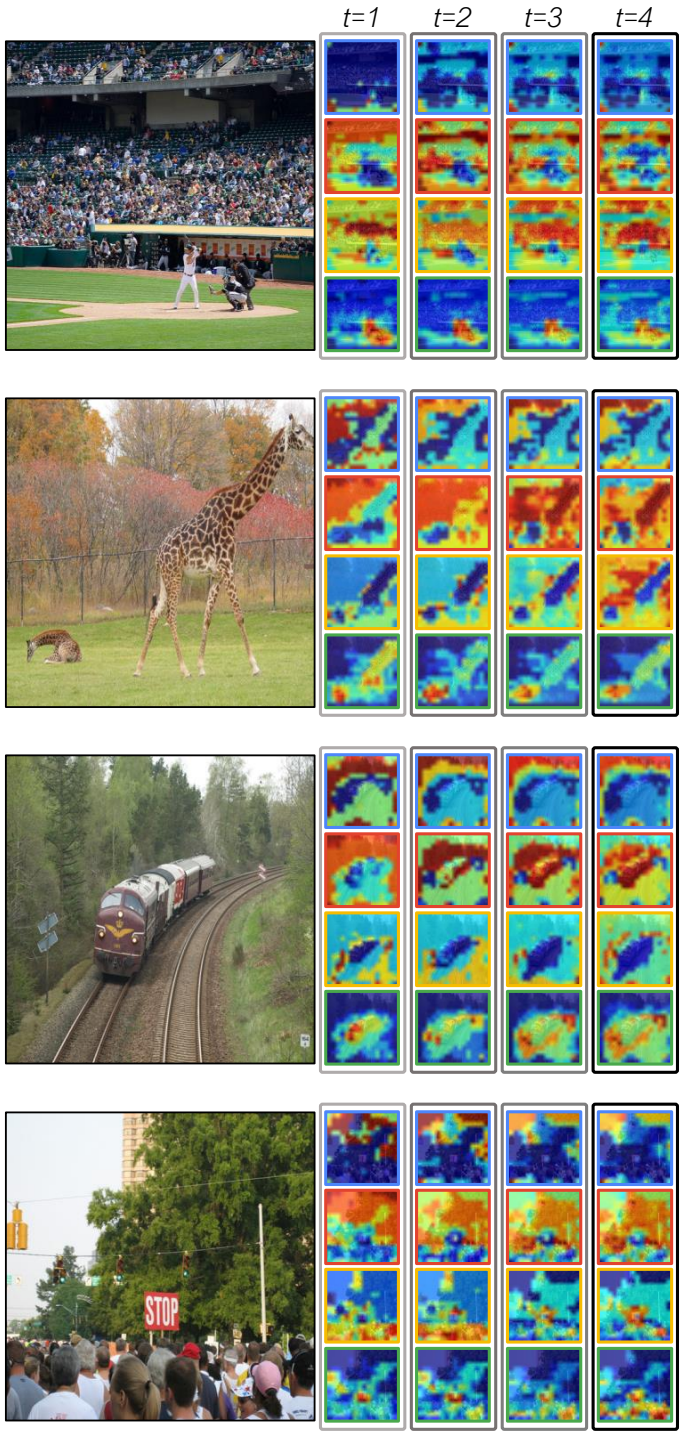

Figure A3. For each element of the image embedding set, we present its attention map and the caption nearest to the element in the embedding space. Matching captions are colored in green. Entities corresponding to the attention maps are underlined.