Supplementary Material: Multi-Scale Visual Semantics Aggregation with Self-Attention for End-to-End Image-Text Matching

Zhuobin Zheng[1,2] † ZHENGZB16@MAILS.TSINGHUA.EDU.CN
Youcheng Ben[1,2] † BYC16@MAILS.TSINGHUA.EDU.CN
Chun Yuan[2,3] * YUANC@SZ.TSINGHUA.EDU.CN
[1] *Department of Computer Science and Technologies, Tsinghua University, Beijing, China*
[2] *Graduate School at Shenzhen, Tsinghua University, Shenzhen, China*
[3] *Peng Cheng Laboratory, Shenzhen, China*

Editors: Wee Sun Lee and Taiji Suzuki

This is a supplementary material for the main paper. Section 1 shows our implementation details. Section 2 presents additional ablation studies on the multi-scale self-attention. Section 3 demonstrates qualitative examples of attended image instances, sentence retrieval for give image queries, and image retrieval for given sentence queries.

1. Implementation Details

For both datasets, input images are resized to 224×224 with a random crop, and we use the 152-layer ResNet He et al. (2016) pre-trained on ImageNet Jia et al. (2009) to extract visual features. Different scales of feature maps ($28 \times 28, 14 \times 14, 7 \times 7$) are captured by the last residual blocks in different stages (*conv3_x*, *conv4_x*, *conv5_x*) of ResNet. For the adaptive self-attention (ASA), we follow Wang et al. (2018) to initialize the weight layers. Note that BatchNorm Ioffe and Szegedy (2015) is crucial to the convergence of ASA. We set the dimensionality of instance candidates and instance-level features as $C = 2048$ and $D = 1024$ respectively. For sentences, the dimensionality of word embeddings is set as $d = 300$, and the pre-trained GloVe Jeffrey et al. (2014) is taken to initialize the word embedding matrix W_x. We set the number of hidden units in the bi-directional GRU and the dimensionality of joint embedding space to 1024.

We follow [Lee et al. (2018); Faghri et al. (2018)] to set the margin m of triplet loss to 0.2 and threshold of maximum gradient norm to 2.0 for gradient clipping. The Adam optimizer Kingma and Ba (2015) is leveraged to train all models with a mini-batch size of 128. Besides, two training stages are applied to each model, where we freeze ResNet in the first 20 epochs with an initial learning rate of $5e^{-4}$, and the whole network is fine-tuned in the next 15 epochs with an initial learning rate of $2e^{-5}$. We follow Lee et al. (2018) to set the hyper-parameters.

† The first two authors contribute equally to this work.
* Corresponding author: Chun Yuan.

R@*K* ($K = 1, 5, 10$) is commonly used to evaluate the retrieval performance, which is defined as the percentage of queries where at least one ground-truth is retrieved among the top K results. **Med r** is another metric, which denotes the median rank of the top-ranked ground-truth. We also compute **Sum** to evaluate the overall performance of cross-modal retrieval, where

$$\text{Sum} = \underbrace{R@1 + R@5 + R@10}_{\text{Sentence Retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{Image Retrieval}}.$$

2. Ablation Studies

For all experiments, we report results under four different settings of similarity measurement Lee et al. (2018), *i.e.* *i-t* AVG, *i-t* LSE, *t-i* AVG and *t-i* LSE. *i-t* denotes Image-Text. *t-i* denotes Text-Image. AVG and LSE refer to average and LogSumExp pooling respectively.

We perform extended ablation studies in terms of the following two aspects:

1) **Effect of different numbers of instance candidates.** As is introduced in the main paper, different scales of feature maps correspond to different numbers of instance candidates. Since we aggregate instance-level visual semantics for each instance candidate, the number of these candidates directly decides the number of image instances captured by our model (duplicates exist). To evaluate its effect on cross-modal retrieval, we apply self-attention (ASA) to 2 scales of feature maps (14×14, 7×7), which obtain 289 and 49 instance candidates respectively.

Table 1 presents results on Flickr30K. Compared to 7×7 feature maps, serious drops are observed in retrieval performance by all measures for 14×14 feature maps, which yield excessive instance candidates for the retrieval task.

2) **Effect of different down-sampling approaches.** To verify the effect of different down-sampling approaches, we compare max and average pooling in the fusion of 2 scales of instance-level features (14×14, 7×7) by down-sampling feature maps of scale 14×14 to 7×7 and then concatenating them together for cross-modal retrieval.

Table 2 shows retrieval results on Flickr30K. We observe that max pooling is better than average pooling in most measures under three different settings (*i-t* LSE, *i-t* AVG and *t-i* LSE). However, the overall performance of average pooling is similar to that of max pooling.

3. Qualitative Results

3.1. Retrieval Examples

Figure 1 and Figure 2 show the qualitative results of sentence retrieval given image queries on Flickr30K and MS-COCO respectively. Figure 3 and Figure 4 illustrate the qualitative results of image retrieval given sentence queries on Flickr30K and MS-COCO respectively. Each sentence corresponds to one ground-truth image. For each image or sentence query, we showcase the top-5 retrieved results.

Scale	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
i-t LSE:						
14×14	19.6	48.1	62.1	16	41.1	53.6
7×7	**59.1**	**86.5**	**91.3**	**45.1**	**74.9**	**83**
i-t AVG:						
14×14	19.7	44.6	56.8	15.2	39.8	52.4
7×7	**61.7**	**85.7**	**90.6**	**45.6**	**74.5**	**83.1**
t-i LSE:						
14×14	4.4	15.4	24.5	3.1	11.8	19.4
7×7	**60.9**	**84.4**	**90**	**44.1**	**73.9**	**82.2**
t-i AVG:						
14×14	13.2	36.1	49.1	7.6	24.1	34.4
7×7	**61.1**	**86.2**	**91.3**	**46.7**	**75.2**	**83.3**

Table 1: Comparison of self-attention applied to two scales of feature maps on Flickr30K. Results are reported in terms of Recall@K(R@K).

Method	Sentence Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
i-t LSE:						
max	**63.6**	**87.7**	**93.4**	46.2	**76.4**	**84.8**
avg	63.3	86.6	93.2	**46.6**	76.1	84.2
i-t AVG:						
max	**61.3**	**88.6**	**93.8**	**46.8**	**75.9**	**84.3**
avg	59.9	86.1	92.1	46.1	75.2	84
t-i LSE:						
max	61.1	**86.2**	**91.8**	**44.3**	73.5	82.2
avg	**61.4**	85	91.4	43.5	**74.6**	**83**
t-i AVG:						
max	62.6	**86.5**	92.2	46.3	74.8	83.9
avg	**64**	86.4	**92.6**	**46.9**	**75.9**	**84.2**

Table 2: Comparison of max and average pooling for down-sampling instance-level features on Flickr30K. Results are reported in terms of Recall@K(R@K).

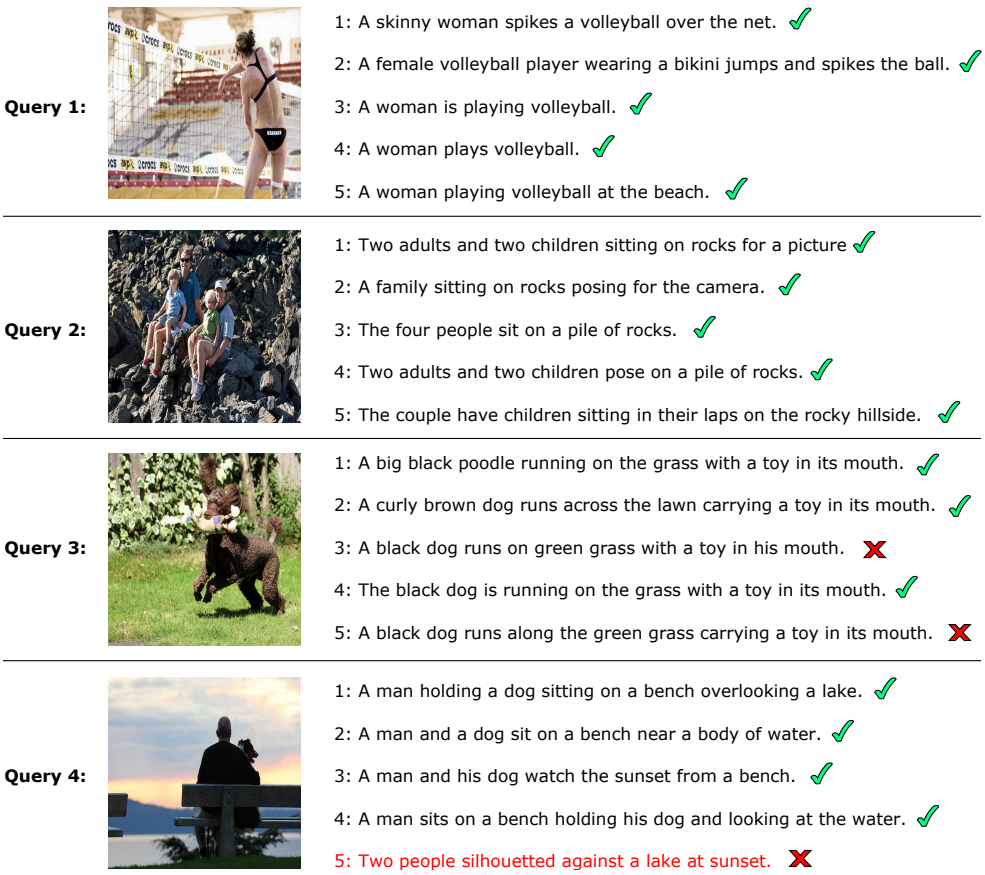

Figure 1: Additional qualitative examples of text retrieval for given image queries on Flickr30K. Incorrect results are highlighted in red and marked with red ×. Reasonable mismatches are in black but still marked with red ×. For query 4, words such as "people", "lake" and "sunset" in the incorrect sentence increase the matching score with target image. On the other hand, it is still challenging to handle counting issues for existing retrieval models.

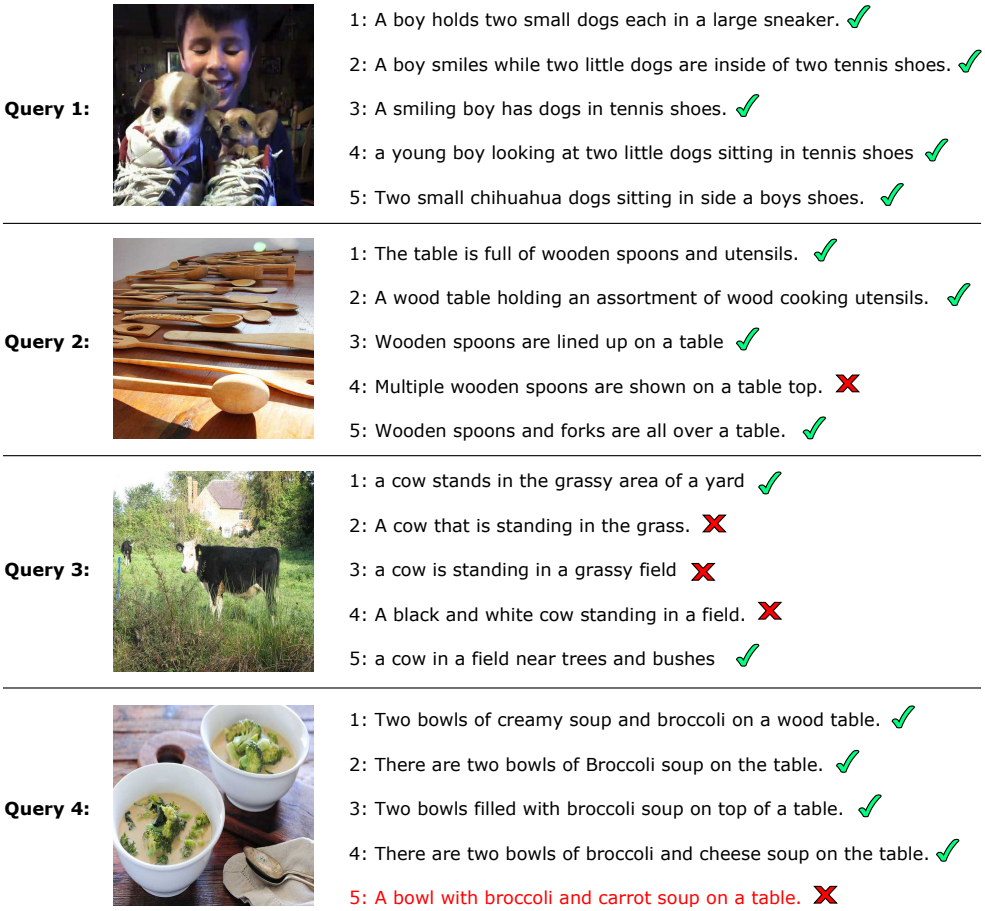

Figure 2: Additional qualitative examples of text retrieval for given image queries on MS-COCO $5K$ test set. Incorrect results are highlighted in red and marked with red $\times$. Reasonable mismatches are in black but still marked with red $\times$. For query 4, words such as "bowl", "broccoli" and "table" in the incorrect sentence increase the matching score with target image.

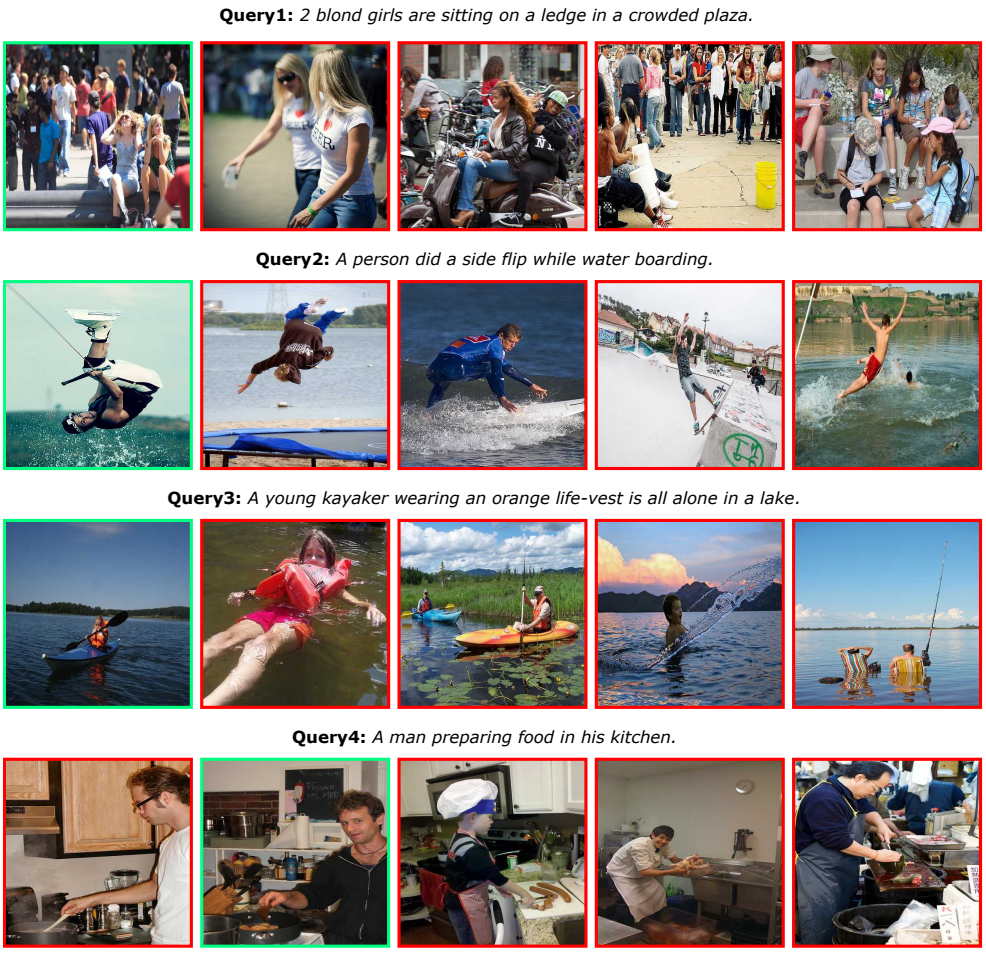

Figure 3: Additional qualitative results of image retrieval for given sentence queries on Flickr30K. Each sentence description corresponds to one ground-truth image. For each sentence query, we show the top-5 ranked images, ranking from left to right. We outline the true matches in green and false matches in red. For query 4, our model ranks one reasonable mismatch before the ground-truth.

Figure 4: Additional qualitative results of image retrieval for given sentence queries on MS-COCO $5K$ test set. Each sentence description corresponds to one ground-truth image. For each sentence query, we show the top-5 ranked images, ranking from left to right. We outline the true matches in green and false matches in red. For query 4, our model ranks one reasonable mismatch before the ground-truth.

References

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.

Pennington Jeffrey, Socher Richard, and D. Manning Christopher. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

Deng Jia, Dong Wei, Socher Richard, Li Li-jia, Li Kai, and Li Fei-fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

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

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.