CAMP: Cross-modal Adaptive Message Passing for Text-image Retrieval
Supplementary Material

Zihao Wang[1*] Xihui Liu[1*] Hongsheng Li[1] Lu Sheng[3] Junjie Yan[2] Xiaogang Wang[1] Jing Shao[2]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]SenseTime Research [3]Beihang University
zihaowang@cuhk.edu.hk {xihuiliu, hsli, xgwang}@ee.cuhk.edu.hk
lsheng@buaa.edu.cn {yanjunjie, shaojing}@sensetime.com

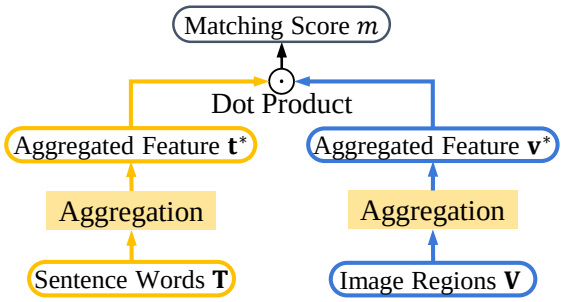

Figure 1. The architecture of the base model.

1. Architecture of Models in Ablation Study

In the ablation study in our paper, we compare our model with several ablation models to demonstrate the effectiveness of our design. We explain the architectures of the first three ablation models mentioned in Sec.4.5, denoted as "Base model", "w/o cross-attn" and "w/o fusion".

1.1. Base Model

As shown in Fig. 1, the base model attends to region features $\mathbf{V} \in \mathbb{R}^{d\times R}$ and word features $\mathbf{T} \in \mathbb{R}^{d\times N}$ separately to extract visual and textual features, and compare their similarities in the joint space. We adopt the same attention-based aggregation strategy as in Sec.3.3 of the paper,

$$\mathbf{a}_v = \text{softmax}\big(\frac{\mathbf{W}_v\mathbf{V}}{\sqrt{d}}\big)^\top, \qquad \mathbf{v}^* = \mathbf{V}\mathbf{a}_v. \tag{1}$$

$$\mathbf{a}_t = \text{softmax}\big(\frac{\mathbf{W}_t\mathbf{T}}{\sqrt{d}}\big)^\top, \qquad \mathbf{t}^* = \mathbf{T}\mathbf{a}_t. \tag{2}$$

and the final matching score is calculated as $m = \mathbf{v}^{*\top}\mathbf{t}^*$. This model is trained by ranking loss with hardest negatives.

*The first two authors contributed equally to this work.

1.2. w/o Cross-attn

To illustrate the importance of the cross-modal attention mechanism for cross-modal message aggregation, we replace the cross-modal attention with average pooling over words or regions to aggregate messages, as shown in Fig. 2. To take the textual branch as an example, every row of the aggregated textual message $\tilde{\mathbf{T}} \in \mathbb{R}^{N\times d}$ has the same value $\tilde{\mathbf{t}} \in \mathbb{R}^{1\times d}$,

$$\tilde{\mathbf{t}} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{t}_i^\top \tag{3}$$

where $\mathbf{t}_i$ denotes the ith column of the original textual feature $\mathbf{T}$, which is the feature of the ith word. The aggregated visual messages $\tilde{\mathbf{v}}$ are obtained in the same way.

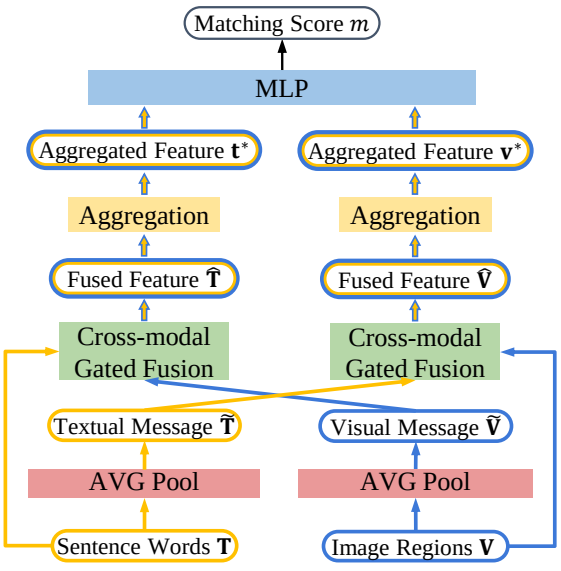

Figure 2. The architecture of the w/o cross-attn model.

1.3. w/o Fusion

We implement a cross-modal attention model without fusion to demonstrate the effectiveness of cross-modal fusion in incorporating deeper cross-modal interactions. As shown in Fig. 3, after obtaining the aggregated messages $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{V}}$, we directly aggregate them without fusing them with the original features. The final matching score m is also calculated by cosine distance and the model is trained by ranking loss with hardest negatives.

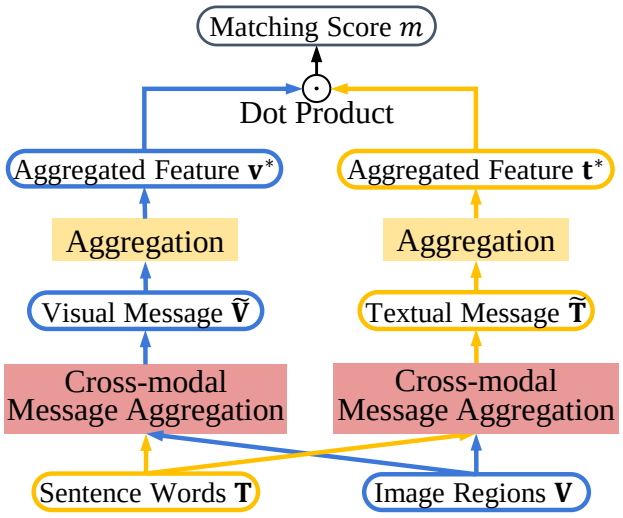

Figure 3. The architecture of the w/o fusion model.

2. Visualization of Attention Weights and Gate Values

To demonstrate the effectiveness of the proposed model, we visualize the attention weights and gate values in our cross-modal adaptive message passing module. We visualize more representative examples for the two image-caption pairs which have been shown in our paper. The original images and sentences are shown in Fig. 4.

For the cross-modal message aggregation, our model both performs word-level attention based on the cues from region features, and region-level attention based on the cues from word features. The cross-modal aggregated messages are then obtained from the attention weights and fused with the original features in our proposed cross-modal gated fusion module. We show the attention weights and the gate values for both two branches respectively in Fig. 5 and Fig. 6. To show how our model handle the negatives, we construct negative examples by swapping the captions of those two positive pairs and also visualize the cross-modal attention weights and gate values in Fig. 7. It is shown that if a word/region matches well with the image/sentence, it would receive a high value, encouraging the fusion operation. On the contrary, for negative pairs without clear correspondences, the gate value would be low, suppressing the fusion operation.

Figure 4. The original image-sentence pairs.

Method	Time Complexity	Single Query Runtime	Overall Test Runtime
Base Model	$O(N^2D)$	0.61 ms	3.387 s
CAMP	$O(N^2MD^2)$	87.24 ms	263.497 s
Rerank (top-10%)	–	4.85 ms	13.164 s

Table 1. Inference time complexity and runtime for different approaches on Flickr30K test split.

Method	Caption Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CAMP	**68.1**	**89.7**	**95.2**	**51.5**	**77.1**	**85.3**
Base Model	63.5	87.1	93.1	46.2	74.2	83.4
Rerank (top-10%)	66.4	89.5	94.5	50.6	75.2	83.9

Table 2. Results of additional experiments on Flickr30K test split.

3. Inference Time Complexity Analyzing

Theoretical inference time complexity. Suppose there are N_1 images and N_2 sentences (denoted as N for simplicity), each image/sentence has M_1 regions and M_2 words (denoted as M for simplicity), and the dimension for word/region features is D. The computational complexity of the baseline model is $O(N^2D)$, which is the same as VSE++ [5]. The complexity of the cross-modal message aggregation is $O(N^2M^2D)$, which is the same as SCAN [30]. The complexity of the cross-modal gated fusion is $O(N^2MD^2)$. Since $D \gg M$, the overall time complexity of the CAMP model is $O(N^2M^2D + N^2MD^2) = O(N^2MD^2)$.

Test runtime. Our implementation is based on Pytorch 0.3.1 on 8 GTX-1080 GPUs. We report the runtime of different approaches in Table 1. The *single query runtime* refers to the time for comparing a single sentence query with all the images in Flickr30K test split. The *overall test runtime* denotes the time of testing all images and all captions from Flickr30K test split.

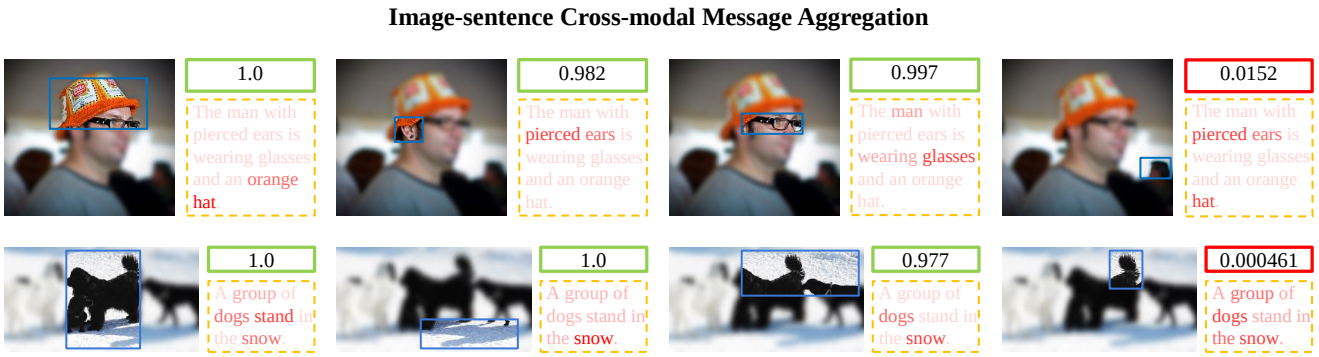

Figure 5. The cross-modal image to sentence attention weights for the corresponding pairs.

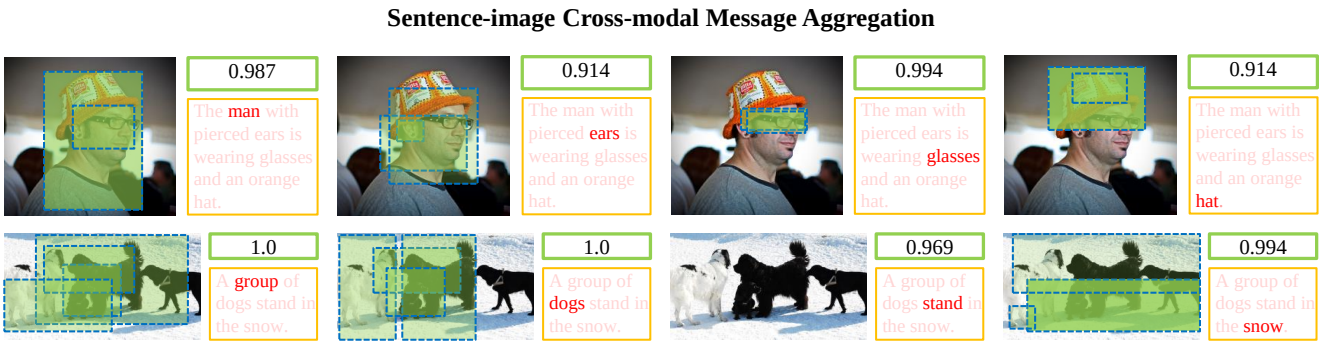

Figure 6. The cross-modal sentence to image attention weights for the corresponding pairs.

Reranking for efficiency in real applications. For real applications, in order to balance the time complexity and retrieval performance, we can employ a reranking approach which firstly uses an efficient model (e.g., VSE++) to roughly rank the images or captions, and then performs fine-grained retrieval by our CAMP model for only the top-k% ranked images or captions (top-10% in our implementation). As shown in Table 1 and Table 2, the reranking approach achieves similar performance compared to our final results, while significantly reduces the inference time.

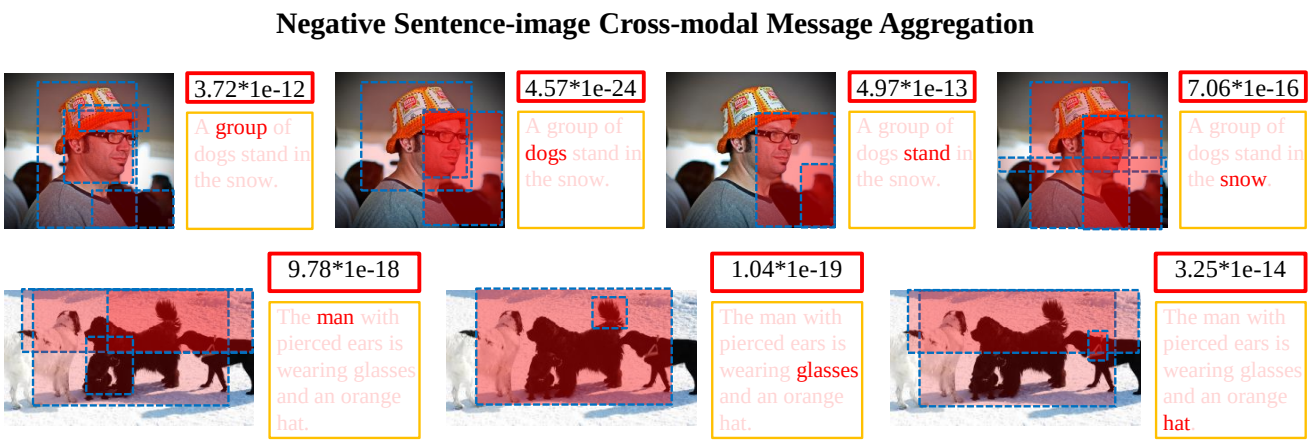

Figure 7. The cross-modal sentence to image attention weights for the negative pairs.