
Hui Yuan1,2, Yan Huang2, Dongbo Zhang1,3, and Liang Wang2,3,4

1. The College of Automation and Electronic Information, Xiangnan University, Xiangtan, China
2. Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA)
3. Center for Excellence in Brain Science and Intelligence Technology (CEBSIT)
4. Chinese Academy of Sciences, Artificial Intelligence Research (CAS-AIR)

A. Image Representation

We extract a set of features \( f = \{f_1, ... , f_K\} \) from each image \( i \) by the bottom-up attention mechanism(1), such that each feature \( f_k \) encodes an object or a salient region in the image. 

\( \mathbf{f}^{(k)} = \begin{bmatrix} \mathbf{f}_1^{(k)} \\ \vdots \\ \mathbf{f}_K^{(k)} \end{bmatrix} \)

B. Global Visual Semantic Similarity

- We first build up connections among image regions and perform region relationship reasoning with Graph Convolutional Networks (GCNs) to generate features with semantic relationships. 

\( \mathbf{s} = \mathbf{W} \mathbf{f} + b \)

- After that, we also use the GCNs network to perform global semantic reasoning on these features with semantic relationships to generate the global representation of the image.

\( \mathbf{S} = \mathbf{G} \mathbf{s} \)

- We use a bidirectional text-to-image encoder to map the whole text \( T \) to the same D-dimensional semantic vector \( \mathbf{a}^T \) as the text global representation \( \mathbf{S} \).

\( \mathbf{a}^T = \text{BiGRU}(T) \)

Then we adopt the cosine similarity function to measure the similarity between the global image representation \( \mathbf{S} \) and the global text representation \( \mathbf{a} \). 

\( \mathcal{L}(\mathbf{S}, \mathbf{a}) = \frac{\mathbf{S}^T \mathbf{a}}{||\mathbf{S}|| \cdot ||\mathbf{a}||} \)

C. Local Fine-Grained Correspondence

1. Image-Text: Local Cross-modal Attention 

\( \mathbf{h}_n = \text{RLU}(\mathbf{c} \cdot \mathbf{h}_n) \)

\( \mathbf{a}_n = \text{GRU}(\mathbf{c} \cdot \mathbf{a}_n) \)

\( \mathbf{d}_n = \text{RLU}(\mathbf{a}_n \cdot \mathbf{h}_n) \)

\( \mathbf{d}_n = \frac{\mathbf{d}_n}{||\mathbf{d}_n||} \)

- Obtain the final text-image similarity \( \mathcal{L}(\mathbf{S}, \mathbf{a}) \) in a locally fine-grained correspondence.

D. Model Learning Strategy

- We comprehensively take two similarity scores for global image-text matching and local region-word correspondence, as well as balance their relative importance at a certain ratio.

\( \mathcal{L}(\mathbf{S}, \mathbf{a}) = \alpha \mathcal{L}(\mathbf{S}, \mathbf{a}) + (1 - \alpha) \mathcal{L}(\mathbf{S}, \mathbf{a}) \)

- We adopt a hinge-based triplet ranking loss to learn the matching part:

\( \mathcal{L}(\mathbf{S}, \mathbf{a}) = \max(0, 1 - \mathcal{S}(\mathbf{S}, \mathbf{a}) + \mathcal{S}(\mathbf{S}, \mathbf{a})) \)

The training loss is defined as follows:

\( \mathcal{L}(\mathbf{S}, \mathbf{a}) = \alpha \mathcal{L} + (1 - \alpha) \mathcal{L} \)

In order to jointly match and generate for model learning, our final loss function is defined as follows:

\( \mathcal{L} = \alpha \mathcal{L} + (1 - \alpha) \mathcal{L} \)

E. Fine-tuning 

We implement our experiment in PyTorch framework with an NVIDIA GeForce GTX 1080TI GPU. We use the Adam optimizer to train the model with 30 epochs. And it only takes about 10 hours to finish the training.

Experimental Results

We investigate the performance of image-text matching is achieved by Visual Semantic Reasoning Network (VSR++), which measures the global alignment between images and texts. Besides existing other works (1.1) have demonstrated the effectiveness of matching whole images and full texts to a common semantic vector space for image-text matching. But they all suffer from the common drawback that they cannot well measure the local correspondences in a fine-grained manner. Especially in real-world applications, many gallery images and texts containing appearance similar regions and texts that are very difficult to distinguish with each other. But how to well measure the local correspondences between regions and texts is a challenging task, since the ground-truth annotation of local correspondence is unknown. Although some existing works (2.2) attempt to solve the local correspondence in a weakly-supervised manner, we cannot combine them into the state-of-the-art framework of global alignment is investigated.

To this end, we propose an improved Visual Semantic Reasoning model (VSR++), which can improve the global alignment by using additionally modeling the local correspondence, with the goal to improve the measurement of cross-modal similarity for fine-grained image-text matching. As an extension of VSR++, our VSR++ models the local correspondence between regions and texts in the context of global alignment by bidirectional stacked cross-modal attention(2). By incorporating the complementary advantages of global alignment and local correspondence, as well as balancing their relative importance, our VSR++ achieves the current state-of-the-art performance.