Graph Structured Network for Image-Text Matching

Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, Yongdong Zhang
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

Siddhant Shingade 19D070057 Mayur Ware 19D070070

Introduction and Motivation

In this paper the authors attempt to solve the problem of image text matching by using a unique graphical approach thereby learning fine grain correspondence.The main contributions from the author include :
(1) First to propose a framework that performs image-text matching on heterogeneous visual and textual graphs by using a Graph Structured Matching Network(GSMN) that explicitly constructs the graph structure for image and text, and performs matching by learning fine-grained phrase correspondence. **(2)** First method that infers fine-grained phrase correspondence by propagating node correspondence using a graph convolutional layer. **(3)** Conducting extensive experiments on Flickr30K and MSCOCO, for proving that their method is better and more useful than the existing ones.

Approach

Textual Graph Construction : An undirected sparse graph,G1 = (V1, E1) is constructed and for each text, a matrix A represents the adjacent matrix of each node and self loops are added. The semantic dependency within the text is identified using Stanford CoreNLP. Each word represents the graph node, and there exists a graph edge between nodes if they are semantically dependent. The similarity matrix S of word representations u is

$$s_{ij} = \frac{exp(\lambda u_i^T u_j)}{\sum_{j=0}^{m} exp(\lambda u_i^T u_j)}$$

[where the sij indicates the similarity between i-th and j-th node and λ is a scaling factor. The weight matrix We is obtained by a Hadamard product between similarity matrix and adjacent matrix, followed by L2 normalization]
Visual Graph Construction : G2 = (V2, E2). In this visual graph, each image is represented as an undirected fullyconnected graph, and each node is set as salient regions detected by Faster-RCNN. Polar coordinate (ρ, θ) based on the centres of the bounding boxes of pair-wise regions are used to compute edge weights and the edge weight matrix We is set as pair-wise polar coordinates. Using the polar coordinate to model the spatial relation of each image, disentangles the distance of pair-wise regions.
Node Matching : Each node in the textual and visual graphs will match with nodes from another modality graph to learn node correspondence. We first depict the node-level matching on textual graph in details, and then roughly describe that on visual graph since this operation is symmetric on two kinds of graphs. Concretely, we first compute similarities between visual and textual nodes, denoted as $U_\alpha V_\beta^T$, followed by a softmax function along the visual axis.

$$C_{t \longrightarrow i} = softmax_\beta \lambda (U_\alpha V_\beta^T) V_\beta$$

Structural Level Matching : To be specific, the matching vector of each node is updated by integrating neighborhood matching vectors using GCN. The GCN layer will apply K kernels that learn how to integrate neighborhood matching vectors, formulated as

$$x_i = ||_{k=1}^{K} \sigma(\sum_{j \in N_i} W_e W_k x_j + b)$$

[where Ni is the neighborhood of i-th node, We indicates the edge weight, Wk and b are the parameters to be learned of k-th kernel]
Objective Function :
When using the text T as query, we sample its matched images and mismatched images at each mini-batch, which form positive pairs and negative pairs. The similarity in positive pairs should be higher than that in negative pairs by a margin γ. We focus on optimizing hard negative samples that produce the highest loss, that is,

$$L = \sum_{(I,T)} [\gamma - g(I,T) + g(I,T')]_+ + [\gamma - g(I,T) + g(I',T)]_+$$

where I', T' are hard negatives, the function $[.]_+$ is equivalent to max[·,0], and g(·) is the global similarity of an image-text pair

Experimental Observations

To validate the effectiveness of our proposed method, we evaluate it on two most widely used benchmarks, Flickr30K and MSCOCO. The commonly used evaluation metrics for image-text matching are Recall@K (K=1,5,10), denoted as R@1, R@5, and R@10, which depict the percentage of ground truth being retrieved at top 1, 5, 10 results, respectively.
The higher Recall@K indicates better performance. Additionally, to show the overall matching performance, we also compute the sum of all the Recall values (rSum) at image-to-text and text-to-image directions, that is

$$rSum = \{R@1 + R@5 + R@10\}_{Imageasquery} + \{R@1 + R@5 + R@10\}_{Textasquery}$$

i) Image-text matching results on Flickr30K : Best results were obtained using GSMN(sparse+dense) method where Faster R-CNN is the Image Backbone and ft Bi-GRU is the Text Backbone. For this method, Image-to-Text matching accuracies for R@1, R@5 and R@10 are 76.4%, 94.3% and 97.3% respectively. Whereas, Text-to-Image matching accuracies for R@1, R@5 and R@10 are 57.4%, 82.3% and 89.0% respectively. Value of rSum for this method comes to be 496.8
ii) Image-text matching results on MSCOCO : Here also, same method is used. For this method, Image-to-Text matching accuracies for R@1, R@5 and R@10 are 78.4%, 96.4% and 98.6% respectively. Whereas, Text-to-Image matching accuracies for R@1, R@5 and R@10 are 63.3%, 90.1% and 95.7% respectively. Value of rSum for this method comes to be 522.5

Advantages

1) This method is unique and takes into account the finer details in image-text matching. **2)** This design can learn correspondence of relation and attribute, which are mostly ignored by previous works. **3)** Experimental results show it's superiority over previous works.

Scope of Improvement

1) Improvement in object correspondance can be done by increasing the complexity of the graph by including more relations and attributes.The use of directed acyclic graphs in this method can be exciting to research on. **2)** This method can be improved to include correspondence with more than one sentence i.e. a description or a paragraph.