Spatial Cross-Attention for Transformer-based Image Captioning

Khoa Anh Ngo and Byonghyo Shim
Department of Electrical and Computer Engineering, Seoul National University
Email : {ngoak, bshim}@islab.snu.ac.kr

I. Introduction

Image captioning is task of describing an image in human language. In solving this task, Transformer-based networks [1] has been widely adopted for their ability to dynamically align each word to different image patches [2, 3, 4]. To generate a caption, Transformer selects relevant image patches and translates them into word. The process of finding relevant image patches is known as self-attention where correlation between image patches and words is used as the attention score. For example, for the word 'frisbee', the image patch contains the object would have a high attention score. However, the locations between 'frisbee' and 'person' are not being considered, thus resulting in a loss of spatial information.

In this paper, we propose a novel cross-attention layer named **Spatial Cross-Attention (SCA)**. The key idea of the proposed method is to utilize the relative distance between image patches as the relative distance between objects. More specifically, to describe a spatial word, we calculate the distance of relevant patches. In doing so, we only extract the relevant spatial information and ignore others spatial relationships of other objects. We call this process a *soft-selection* mechanism.

IV. Spatial Cross-Attention

We employ a Transformer-based network for image captioning. An encoder of the Transformer divides an image into patches and arrange them in a sequence. The input sequence is processed by multiple self-attention layers. Each self-attention layer projects each element of the input sequence into query, key, and value vectors. The attention scores are calculated via dot-product between key and query. The output of a self-attention layer is a processed sequence whose element is a weighted sum of a value vector with the calculated attention scores.

Similar to the encoder, a decoder takes input as a sequence of word embedding. The word sequence is also process by multiple self-attention layer with look-ahead mask to follow auto-regressive property.

We model the spatial information as the relative distance and orientation between image patches. Formally, we measure the 2D coordinate differences and project to an embedding of the same dimension with the value vector. As mentioned in Section I, we employ a *soft-selection* mechanism to dynamically incorporate relevant spatial information into the value vector. Formally, we weight

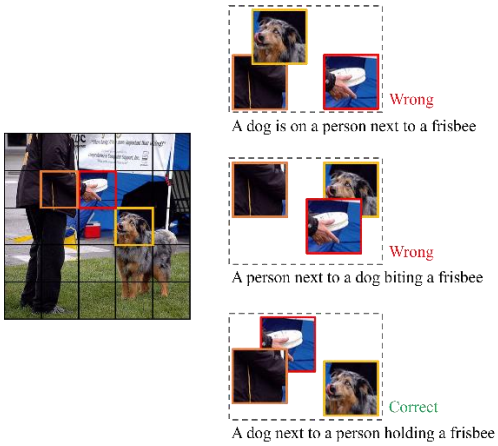

Fig 1. Without knowing the spatial relationship between objects, the model cannot correctly describe their spatial arrangement (1st and 2nd row)

the relative embedding with its corresponding image patches and sum all weighted relative embeddings into a spatial feature namely l_{ts},

$$l_{ts} = \sum_{\{i=1\}}^{S} \alpha_{ts}\alpha_{ti}R_{\{h_s - h_i, w_s - w_i\}}, (1)$$

where α_{ts} and $\alpha_{ti}$are the attention scores of t-th word to the s-th and i-th image patch, respectively. $R_{\{x,y\}}$ is the relative embedding of image patches with relative distances of x- and y-axis, respectively. The spatial feature is added into the value vector, thus resulting in a spatial-visual value vector.

V. Conclusion

We expect to achieve a high quality image captioning. Sp ecifically, the proposed method should help Transformer network to generate more precise spatial description.

Reference

[1] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz, and Polosukhin Illia. 2017. Attention is all you need. In NeurIPS

[2] Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image Captioning: Transforming Objects into Words. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alche-Buc, ´ F.; Fox, E.; and Garnett, R., eds., Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

[3] Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2019. Meshed-Memory Transformer for Image Captioning.

[4] Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework.