Text Embedding Bank for Detailed Image Paragraph Captioning

Arjun Gupta, Zengming Shen, Thomas Huang
University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA
{arjung2, zshen5, huang1}@illinois.edu

Abstract

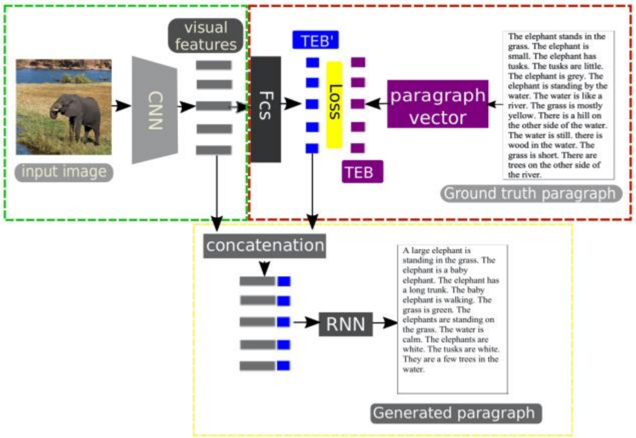
Existing deep learning-based models for image captioning typically consist of an image encoder to extract visual features and a language model decoder, an architecture that has shown promising results in single high-level sentence generation. However, only the word-level guiding signal is available when the image encoder is optimized to extract visual features. The inconsistency between the parallel extraction of visual features and sequential text supervision limits its success when the length of the generated text is long (more than 50 words). We propose a new module, called the Text Embedding Bank (TEB), to address this problem for image paragraph captioning. This module uses the paragraph vector model to learn fixed-length feature representations from a variable-length paragraph. We refer to the fixed-length feature as the TEB. This TEB module plays two roles to benefit paragraph captioning performance. First, it acts as a form of global and coherent deep supervision to regularize visual feature extraction in the image encoder. Second, it acts as a distributed memory to provide features of the whole paragraph to the language model, which alleviates the long-term dependency problem. Adding this module to two existing state-of-the-art methods achieves a new state-of-the-art result on the paragraph captioning Stanford Visual Genome dataset.

Introduction

Image paragraph captioning is the task of automatically generating multiple sentences for describing images through coherent text. Relative to the performance of single-sentence caption generating models, the performance of paragraph-length caption generating models is lower by a large margin. Paragraph captioning on images, and especially video, is a challenging task due to the requirement of both nuanced visual understanding and long-term language reasoning. Existing deep learning-based models typically consist of an image encoder to extract visual features in parallel with a Recurrent Neural Network (RNN) language model decoder to generate the sentences word by word sequentially. In the training stage, only a tiny scalar from the word level loss is available to optimize the image encoder training. This makes the visual feature extraction insufficiently detailed. To overcome this challenge, we propose the Text Embedding Bank (TEB) module. This module maps varied-length paragraphs to a fixed-length vector which we call TEB. Each unique vector in the TEB has similarity based on Euclidean distance and is indexed by the order of the word in the vocabulary. The TEB also has distributed memory: the TEB module, which holds the entire paragraph in a distributed memory model, can provide global supervision to better regularize the image encoder in the training stage. Additionally, RNNs are known to have a long-term dependency problem because of vanishing and exploding gradients which make them unable to meet long-term language reasoning; since the TEB module has distributed memory and can provide order, it is better with long-term language reasoning.

Approach

The proposed TEB module improves paragraph captioning by describing the rich content of a given image, and can be integrated with an existing image captioning pipeline. Figure 1 shows an example of this:



The elephant stands in the grass. The elephant is small. The elephant has tusks. The tusks are little. The elephant is grey. The elephant is standing by the water. The water is like a river. The grass is mostly yellow. There is a hill on the other side of the water. The water is still. there is wood in the water. The grass is short. There are trees on the other side of the river.

A large elephant is standing in the grass. The elephant is a baby elephant. The elephant has a long trunk. The baby elephant is walking. The grass is green. The elephants are standing on the grass. The water is calm. The elephants are white. The tusks are white. They are a few trees in the water.

Figure 1. Integration of TEB module

Method	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Image-Flat (Karpathy 2015)	12.82	11.06	34.04	19.95	12.20	7.71
RH (Krause et. al. 2017)	15.95	13.52	41.90	24.11	14.23	8.69
RTT-GAN (Liang et. al. 2017)	17.12	16.87	41.99	24.86	14.89	9.03
RTT-GAN+ (Liang et. al. 2017)	18.39	20.36	42.06	25.35	14.92	9.21
CapG-RevG (Chatterjee 2018)	18.62	20.93	42.38	25.52	15.15	9.43
CAE-LSTM (Wang et. al. 2019)	18.82	25.15	-	-	-	9.67
Diversity (Melas-Kyriazi 2018)	17.86	30.63	43.54	27.44	17.33	10.58
Ours (T)	15.45	23.38	41.49	23.38	11.96	6.00
Ours (T + TEB)	15.88	24.84	41.86	24.64	13.97	6.40
Ours (D + TEB)	**18.93**	**32.53**	**45.24**	**28.44**	**17.93**	**10.98**

Table 1. Our result compared with prior results on Stanford Visual Genome dataset

In Figure 1, there are three interconnected components divided into three dashed rectangular boxes. In the green box on the top left, the image encoder extracts visual features through a CNN model. In the yellow box on the bottom, an RNN based language model decoder is used to generate paragraphs. Existing deep learning-based models only contain these two components. The red box on the top right box is the TEB module: In the training stage, for an {image, paragraph} pair, the varied-length paragraph is mapped to a fixed-length vector which is called TEB through the paragraph vector framework. The visual features from the image encoder are converted to the predicted TEB (called TEB') through several fully connected layers. The TEB' is supervised by the TEB through an L1 loss, which acts as global deep supervision to regularize the visual feature extraction for the image encoder. The visual features and TEB' are concatenated and fed into the RNN as input. The generated paragraph is supervised by the ground truth paragraph through a word-level loss. In the inference stage, the TEB is not available while the TEB' acts as distributed memory to provide semantic features of the whole paragraph to alleviate the long-term dependency problem for the language model.

Results

As shown in Table 1, we conducted experiments and evaluated our TEB module on the Stanford Visual Genome image paragraph dataset, a benchmark in the field of image paragraph captioning. We compare the proposed method with a variety of strong baselines. We have three models: The "Transformer" model (T) is the Bottom-Up and Top-Down model (Anderson et. al. 2018) with the LSTM replaced by a Transformer. The "Transformer + TEB" model (T + TEB) is the "Transformer" model with our TEB module. The "Diversity + TEB" model (D + TEB) is the Diversity model (Melas-Kyriazi 2018) with our proposed TEB module.

Conclusion

In this paper, we propose the Text Embedding Bank (TEB) for image paragraph captioning, a task that requires capturing the fine-grained entities to generate a detailed and coherent paragraph. Our TEB module provides global and parallel deep supervision and distributed memory for nuanced image understanding and long-term language reasoning. Integrating the TEB module to existing state-of-the-art methods achieves new state-of-the-art results.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In Computer Vision and Pattern Recognition

Chatterjee, M., Schwing, A.G. 2018. Diverse and coherent paragraph generation from images. In Proceedings of the European Conference on Computer Vision

Karpathy, A., Li F. 2015. Deep visual-semantic align-ments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137

Krause, J.; Johnson, J.; Krishna, R.; Li F. 2017. A hierarchical approach for generating descriptive image paragraphs. In Computer Vision and Pattern Recognition, pages 3337–3345

Liang, X.; Hu, Z.; Zhang, H.; Gan, C.; Xing, E.P. 2017. Recurrent topic-transition GAN for visual paragraph generation. In Proceedings of the IEEE International Conference on Computer Vision, pages 3362–3371

Melas-Kyriazi, L.; Rush, A.; Han, G. 2018. Training for diversity in image paragraph captioning. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 757–761

Wang, J.; Pan, Y.; Yao, T.; Tang, J.; Mei, T. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation. In Proceedings of the International Joint Conference on Artificial Intelligence