SuperCaptioning: Image Captioning Using Two-dimensional Word Embedding

Anonymous Authors[1]

Abstract

Language and vision are processed as two different modal in current work for image captioning. However, recent work on Super Characters method shows the effectiveness of two-dimensional word embedding, which converts text classification problem into image classification problem. In this paper, we propose the SuperCaptioning method, which borrows the idea of two-dimensional word embedding from Super Characters method, and processes the information of language and vision together in one single CNN model. The experimental results on Flickr30k data shows the proposed method gives high quality image captions. An interactive demo is ready to show at the workshop.

1. Introduction

Image captioning outputs a sentence related to the input image. Current methods process the image and text separately (Karpathy & Fei-Fei, 2015; Vinyals et al., 2016; You et al., 2016; Yao et al., 2017; Lu et al., 2017; Rennie et al., 2017; Anderson et al., 2018; Hossain et al., 2019). Generally, the image is processed by a CNN model to extract the image feature, and the raw text passes through embedding layer to convert into one-dimensional word-embedding vectors, e.g. a 300x1 dimensional vector. And then the extracted image feature and the word embedding vectors will be fed into another network, such as RNN, LSTM, or GRU model, to predict the next word in the image caption sequentially.

Super Characters method (Sun et al., 2018) is originally designed for text classification tasks. It has achieved state-of-the-art results on benchmark datasets for multiple languages, including English, Chinese, Japanese, and Korean. It is a two-step method. In the first step, the text characters are printed on a blank image, and the generated image is called Super Characters image. In the second step, the Super Characters image is fed into a CNN model for classification. The CNN model is fine-tuned from pre-trained ImageNet model.

(a) "Four men in life jackets are riding in a bright orange boat".

(b) "A woman in a black coat walks down the sidewalk holding a red umbrella".

(c) "A man in a boat on a lake with mountains in the background".

(d) "Four performers are performing with their arms outstretched in a ballet".

Figure 1. Examples of generated image captions using the proposed SuperCaptioning method.

In this paper, we address the image captioning problem by employing the two-dimensional word embedding from the Super Characters method, and the resulting method is named as SuperCaptioning method. In this method, the input image and the raw text are combined together through two-dimensional embedding, and then fed into a CNN model to sequentially predict the words in the image caption. The experimental results on Flickr30k shows that the proposed method gives high quality image captions. Some examples given by SuperCaptioning method are shown in Figure 1.

2. The Proposed SuperCaptioning Method

The SuperCaptioning method is motivated by the success of Super Characters method on text classification tasks. Super Characters method converts text into images. So it will be very natural to combine the input image and the image

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

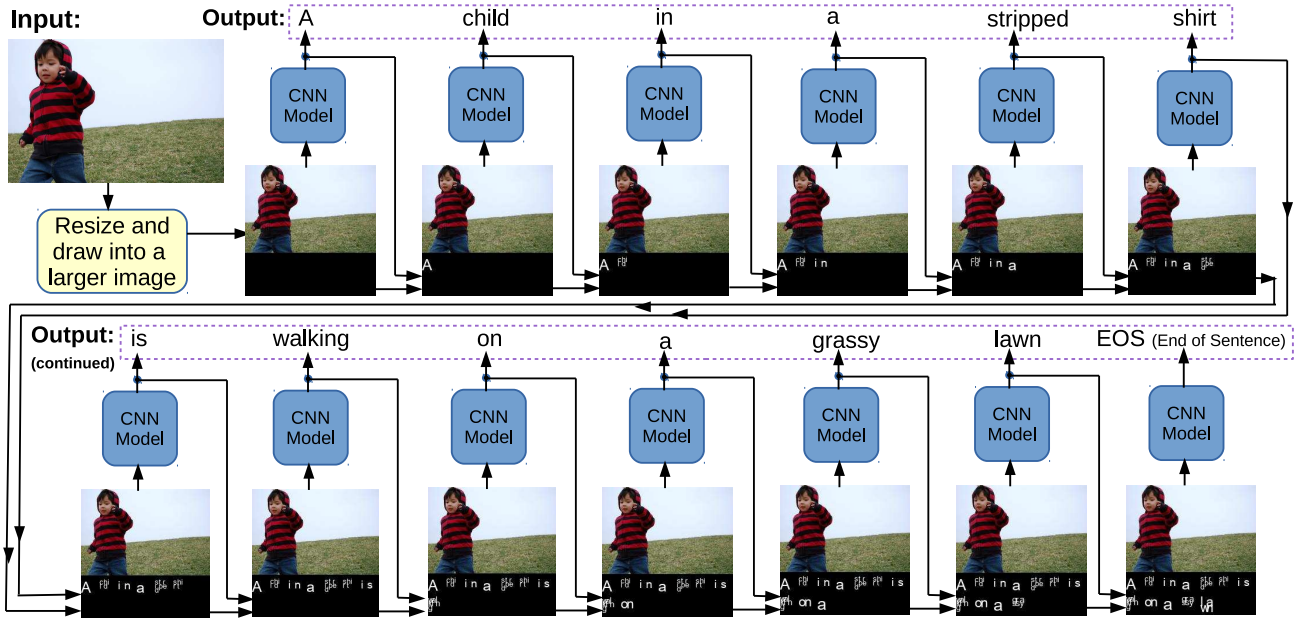

Figure 2. SuperCaptioning method illustration. The output caption is "A child in a striped shirt is walking on a grassy lawn".

of the text together, and feed it into one single CNN model to predict the next word in the image caption sentence.

Figure 2 illustrates the proposed SuperCaptioning method. The caption is predicted sequentially by predicting the next word in multiple iterations. At the beginning of the caption prediction, the partial caption is initialized as null, and the input image is resized to occupying a designed portion (e.g. top) of a larger blank image as shown in Figure 2. Then the text of the current partial caption is drawn into the the other portion (e.g. bottom) of the larger image as well. The resulting image is called the SuperCaptioning image, which is then fed into a CNN model to classify the next word in the caption. The CNN model is fine-tuned from the ImageNet pre-trained model. The iteration continues until the next word is EOS (End Of Sentence) or the word count reaches the cut-length. Cut-length is defined as the maximum number of words for the caption.

Squared English Word (SEW) method (Sun et al., 2019) is used to represent the English word in a squared space. For example, the word "child" occupies the same size of space as the word "A", but each of its alphabet will only occupies $\{1/ceil[sqrt(N)]\}^2$ of the word space, where N is five for "child" which has five alphabets, $sqrt(.)$ stands for square root, and $ceil[.]$ is rounding to the top.

The data used for training is from Flickr30k[1]. Each image in Flickr30k has 5 captions by different people, and we only keep the longest caption if it is less than 14 words as the ground truth caption for the training data. After this filtering, 31,333 of the total 31,783 images are left.

After comparing the accuracy of experimental results using different configurations for the font size, cut-length, and resizing of the input image, we finally set the font size to 31, cut-length to 14 words, and resizing the image size to 150x224 in the fixed-size SuperCaptioning image with 224x224 pixels, as shown in Figure 2.

The training data is generated by labeling each SuperCaptioning image as an example of the class indicated by its next caption word. EOS is labeled to the SuperCaptioning image if the response sentence is finished. The model used is SE-net-154 (Hu et al., 2018) pre-trained on ImageNet[2]. We fine-tuned this model on our generated data set by only modifying the last layer to 11571 classes, which is the vocabulary size of all the selected captions.

Figure 1 shows that the proposed SuperCaptioning method captions the number of objects in the image, as shown in Figure 1a "**Four men** ..."; and it also captions the colors of overlapping objects, as shown in Figure 1b "A woman in a **black coat** ... holding a **red umbrella**"; it captions the big picture of the background, as shown in Figure 1c "... with **mountains in the background**"; and it also captions the detailed activity, as shown in Figure 1d "... with their **arms outstretched** in a ballet".

[1]http://shannon.cs.illinois.edu/DenotationGraph/data/flickr30k.html

[2]https://github.com/hujie-frank/SENet

3. Conclusion

In this paper, we propose the SuperCaptioning method for image captioning using two-dimensional word embedding. The experimental results on Flickr30k shows that the SuperCaptioning method gives high quality image captions. An interactive demo is ready to show at the workshop.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Hossain, M., Sohel, F., Shiratuddin, M., and Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6):1–36, 2019.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Lu, J., Xiong, C., Parikh, D., and Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 375–383, 2017.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017.

Sun, B., Yang, L., Dong, P., Zhang, W., Dong, J., and Young, C. Super characters: A conversion from sentiment classification to image classification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 309–315, 2018.

Sun, B., Yang, L., Chi, C., Zhang, W., and Lin, M. Squared english word: A method of generating glyph to use super characters for sentiment analysis. *arXiv preprint arXiv:1902.02160*, 2019.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

Yao, T., Pan, Y., Li, Y., Qiu, Z., and Mei, T. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4894–4902, 2017.

You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.