International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 16 (2019) pp. 3500-3502

The Significance of self-attention over LSTM in image captioning

Sreela S R

Department of Computer Science,
Cochin University of Science and Technology ,
Kerala, India

Sumam Mary Idicula

Department of Computer Science,
Cochin University of Science and Technology,
Kerala, India.

Abstract

Image captioning is a cherishing topic in two critical fields such as computer vision and natural language processing. It generates the description for the image. Image captioning system follows encoder-decoder architecture. Our paper proposed a deep learning model for image captioning with self-attention. Self- attention is a process of identifying the dependencies of words to capture the internal structure of the sentence and producing the next word in the sentence. Self-attention is implemented in different ways such as it is incorporated before and after LSTM. The model produces the best results when self-attention is implemented after LSTM. Self-attention enhances the accuracy of the image captioning system.

Keywords: Image captioning, Self attention, Deep learning

INTRODUCTION

Image captioning is a process of analyzing objects, actions, spatial relationships and identifying the semantic content of the image and produces the human-readable grammatically correct sentences. Image captioning system helps to improve the life of visually impaired people, effective searching of image collections and enhance the image searching capability of search engines.

Attention plays an important role in image captioning. The attention process is helped to find the important features needed for predicting the next word. Attention can be done in two ways such as visual attention and self-attention. The significant part of the image is identified in visual attention. The choice for a particular prediction of a word can place constraints on future prediction decisions. It is done using self-attention. The common sequential architecture used in image captioning is the recurrent neural network such as LSTM or Gated Recurrent Unit(GRU). Self-attention in LSTM deals with memory problems. Current image captioning architectures calculate a new internal state using the previous states. In self-attention, if the input is a sentence, then each word in the sentence needs to compute attention variables. The goal of self-attention is to understand the dependencies between the words in the sentence and find the internal structure of the sentence using that information. It is capable of learning the distant dependencies within the sentence. In our paper, we concentrated on the importance of self-attention in image captioning system.

Main objective of the work is to develop an image captioning system with self-attention. Self-attention improves overall image captioning system. To incorporate the objectives, an image captioning system is developed with the following features.

- An encoder-decoder framework is used for implementing image captioning.
- Self-attention module is implemented for finding the important word in the partial captions for predicting the next word.

RELATED WORKS

Image captioning methods are classified into the retrieval based method, template-based method, and end-to-end learning based image captioning[1]. The retrieval based method finds the similar images and transferring the caption of similar images with query image. The method in the research article[4] follows retrieval based image captioning system. The template method follows sentence generation using keywords. BabyTalk[2], Midge[3] systems etc. are template based image captioning systems. End-to-end captioning is a very effective method which also produces correct and fluent sentences. Nowadays most of the image captioning works follows end-to-end learning.

A subtype of end-to-end learning captioning system is attention based image captioning model. Attention model is a kind of model which mimics the human visual system. Most of the attention based image captioning system follows the visual attention mechanism. Another important kind of attention is self attention[5].

Image description model follows merge architecture. The image features and caption features are merged to form next word in the caption. The model is described in figure 1.

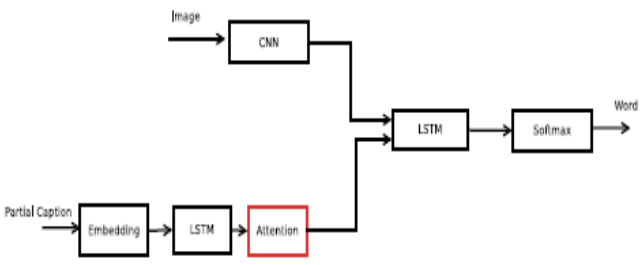

Figure 1: Proposed Model

The images are encoded as CNN features. Here Residual neural network(Resnet)[6] is used for encoding images. Caption features are extracted using LSTM. The important caption words are obtained during the attention phase. The attention model is explained in the next section. Softmax function is used for predicting the next word.

The image captioning model is organized as three sub models such as image feature model, caption embedding model, and caption extraction model.

Image feature model

CNN generates the feature map of the input image. This feature map is used for finding the caption of the image.

Caption embedding model

This model produces the abstract representation of partial captions. Initially, the partial caption is the start symbol <*START*> . Partial captions are generated by appending the previously generated words of the whole model. This model encodes partial caption as a fixed size vector. The model contains the embedding layer, LSTM and attention layer. Embedding layer produces the fixed-size representation of the partial caption. LSTM finds the long-term dependencies in the partial caption. Attention module extracts the most important part of the partial caption for predicting the next word. Attention is done using self-attention mechanism. Self-attention model is described in the next section.

Caption extraction model

It predicts the probabilities of the generated word. This model contains LSTM and softmax layer. Softmax layer produces the probabilities of the output word.

Self attention module

Self-attention is implemented in the LSTM. The following equations define it. Let x be the input to the attention model

$$x_p = Permute(x) \quad (1)$$

$$z = Softmax(W_x x_p + \varphi) \quad (2)$$

Where *Permute* is a function which interchanges the dimension of x, W_x is the weight vector, x^T is the transpose of x, and φ represents bias term.

$$z_p = Permute(z) \quad (3)$$

$$z_o = x.z_p \quad (4)$$

EXPERIMENTS

Dataset

Flickr8k is used for implementing the work. It is a collection of 8000 images with captions. 6000 images are used for training and 1000 images each for validation and testing.

Training details

The pre-trained Resnet model is used for extracting image features. The Resnet model is trained with ImageNet. Resnet has 50 layers. The vocabulary of flickr8k is used. The vocabulary contains 8236 words. The images are resized to 224 X 224 X 3, and the pixel values are normalized in the range of 0 and 1. Our model is optimized using Rmsprop algorithm with learning rate 0.001 and rho 0.9.

RESULTS

The model efficiency is determined using the metrics such as accuracy and loss. The loss is computed using categorical cross entropy. The categorical cross entropy is computed using the equation 5.

$$loss = -\sum_{c=1}^{M} y_{p,c} \log P_{p,c} \quad (5)$$

Where p is the predicted class and c is the original class.

The attention model is placed in different places in the model. The accuracy and loss of model under different conditions is plotted in figures 2 and 3. The model get best accuracy and loss when the self attention is placed after first LSTM. So we adopted this architecture for image captioning.

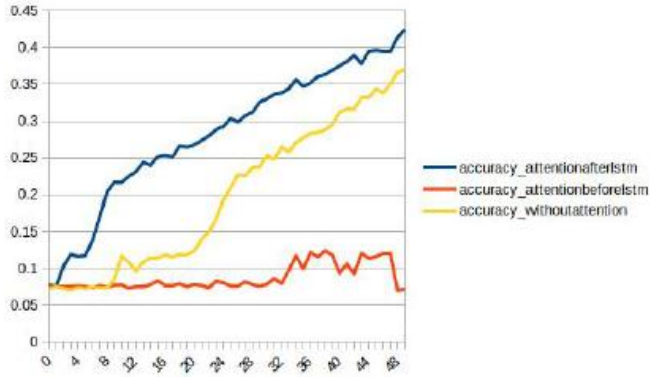

Figure 2: Comparison of accuracies over different methods

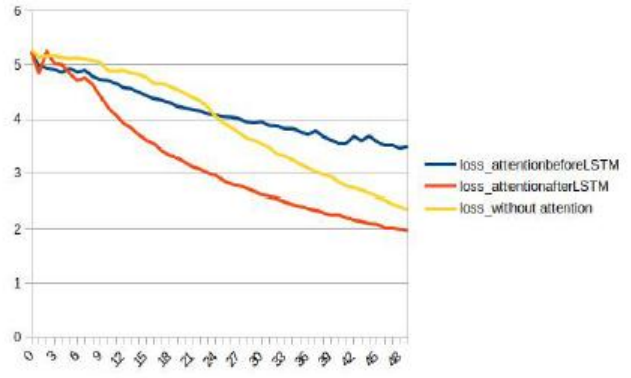

Figure 3: Comparison of losses over different methods

Our model is evaluated using BLEU score. BLEU score is estimated using the equation

$$BLEU = \min\left(1, \frac{hyplength}{reflength}\right)\left(\prod_{i=1}^{4} Precision_i\right)^{1/4} \quad (6)$$

Where *hyplength* is the generated caption length and *re f length* is reference caption length.

$$Precision = \frac{Overlapped\ n-grams}{Total\ no\ of\ n-grams\ in\ reference} \quad (7)$$

Our model achieves a BLEU score of 68.5. The output of the image captioning system is depicted in figure 4. The output contains generated description and ground truth description.

Figure 4: Generated Description: A brown dog is running in the snow .

Ground Truth: A dog run through the snow .

CONCLUSION

We proposed an image captioning model with self-attention. The self-attention module studied the dependencies of the partial captions and predicted the next word using the previously generated important words. The model produced the best accuracy when the self-attention module is placed after the LSTM. The model is evaluated using BLEU metric and the benchmark dataset Flickr8k. We hope that in future new self-attention methods can be used in image captioning models for improving the results.

REFERENCES

[1] Liu, Xiaoxiao, Qingyang Xu, and Ning Wang. "A survey on deep neural network-based image captioning." The Visual Computer (2018): 1-26.

[2] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

[3] Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[4] Kuznetsova, Polina, et al. "Collective generation of natural image descriptions." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers- Volume 1. Association for Computational Linguistics, 2012.

[5] Lin, Zhouhan, et al. "A structured self-attentive sentence embedding." arXiv preprint arXiv:1703.03130 (2017).

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.