A Novel Adaptive Attention Model for Image Captioning

Donglin Liang\textsuperscript{1, a}, Jinzhao Wu\textsuperscript{2, *}, Anping He\textsuperscript{3, b}, Ming Ding\textsuperscript{4, c}

\textsuperscript{1}Guangxi university for nationalities Nanning, Guangxi, China
\textsuperscript{2}Guangxi university Nanning, Guangxi, China
\textsuperscript{3}Lanzhou University Lanzhou, Gansu, China
\textsuperscript{4}Guangxi university for nationalities Nanning, Guangxi, China

*Corresponding author e-mail: 1125858775@qq.com, *847132754@qq.com, 
\textsuperscript{a}847132754@qq.com, \textsuperscript{b}heap@lzu.edu.cn, \textsuperscript{c}603269048@qq.com

Abstract. Image captioning is an extremely challenging work, which involves both computer vision and natural language processing. However, in recent years, with the introduction and improvement of various new technologies, it has made a great advancement. In this paper, a novel attention-based image caption model is proposed. We use Resnet harnessed by outstanding performance in convolutional neural network, as the encoder. For the decoding, LSTM network equipping with fitness of dealing with problems highly related to time series, is adopted. At each time step, the attention mechanism guides the language model to synthesize context vectors adaptively. The experimental results show that the model has good performance in various quantitative metrics as well as in human evaluation.

1. Introduction
In recent years, image captioning has become a hot research topic in academia and industry. It has been applied in many fields, and brings a lot of convenience to people. It can generate relevant descriptions for the image library, so that users can quickly and efficiently search for the images they want by descriptions or keywords. The core technology it uses can also be combined with voice technology to create specific products that help the visually impaired to identify what is in front of them. Obviously, it is not a simple matter to generate the corresponding description for an image in terms of a machine. Because it requires capturing all kinds of semantic information in the image, such as the attributes and location of the object. In addition, in order to train a high-precision target model, a large enough training set is needed. However, with the release of large classification data sets and the development of deep learning technology, there has been a breakthrough in the automatic generation of image captioning. Recently, attention-based encoder-decoder models [1, 2, 3] have been explored, where the attention mechanism typically produces a spatial map highlighting image regions relevant to each generated word. Common attention mechanisms use the hidden state of the LSTM network [4] to determine which part of the image to focus on. However, since the hidden state is only sensitive to short-term inputs, the attention distribution calculated by this method may be less correlated with long-term information. That is to say, this has a potential drawback of losing information which could be useful for richer, more descriptive captions.

In this paper, we propose a novel attention model, called Long Short-Term Adaptive Attention model (LSTAA). Unlike most attention models, it determines which region of the image to focus on...
by the hidden state and the cell state of the LSTM network. At each time step, two attention weight vectors are generated, which are then respectively subjected to a dot product operation with the extracted image feature vectors, and the result is used as an input to the regulator. The main role of the regulator in the model is to synthesize context vectors [1] adaptively. In order to get a better model, we also use some methods commonly used in neural networks, such as weight normalization [5], dropout [6].

Our key contributions can be summarized as follows: (1) We introduce a new adaptive attention encoder-decoder model for image captioning, and evaluate it on the challenging MSCOCO [7] dataset. Experimental results show that our model can improve the performance of image captioning systems. (2) We also study the influence of some important factors on the system performance, such as weight normalization, dropout and the number of LSTM hidden layers. Experiments show that they have a significant positive benefit for image captioning.

2. Related Work

Early methods of image captioning were usually temple-based [8, 9, 10], which typically use the output of target detection, attribute classification and scene recognition to fill in the blank of the caption templates. Farhadi et al. [8] proposed a triplet of scene elements which is converted to text using templates. Li et al. [9] proposed to create an image description by piecing together the concepts detected. Kuznetsova et al. [10] also proposed a more powerful language model. The drawback of these methods is that the generated descriptions are not vivid enough and human-crafted templates do not work for all images. Inspired by the successful application of neural networks in machine translation, quite a few scholars have used neural network-based methods [1, 3, 11, 12, 13] to generate image captions in recent years. Generally, the convolutional neural network (CNN) is used to encode the input image into a set of feature vectors, which are then decoded into a sequence of words by the recurrent neural network (RNN). Xu et al. [1] first introduced the attention mechanism into the image captions generation algorithm, enabling the language model to selectively focus on different areas of the image. Later, Lu et al. [3] proposed an adaptive attention mechanism, which enables the model to determine whether image information needs to be used in generating each word adaptively. In addition, Jia et al. [11] take semantic information as additional input of LSTM, making the model generate a more appropriate description of the image content. The model proposed by You et al. [12] makes use of both global feature vectors of images and attribute vectors representing high-level semantic concepts, and combines attention mechanism to further improve the effect of image description. Wang et al. [13] proposed the CNN+CNN model for image captioning. Our approach uses both the previous hidden state \( h_{t-1} \) and the previous cell state \( m_{t-1} \) of the LSTM network to guide the model to synthesis context vectors.

3. Encoder-Decoder for Image Captioning

The model we propose is based on the encoder-decoder framework commonly used in image captioning [1, 15]. Given an image and the corresponding caption, the encoder-decoder model directly maximizes the following objective:

\[
\theta^* = \arg \max_{\theta} \sum_{(I, y)} \log p(y | I; \theta)
\]  

(1)

Where \( \theta \) are the parameters of the model, \( I \) is the image, and \( y = \{y_1, ..., y_t\} \) is the corresponding caption. According to the chain rule, the log likelihood of the joint probability distribution can be decomposed into ordered conditionals:

\[
\log p(y | I) = \sum_{t=1}^{T} \log p(y_t | y_1, ..., y_{t-1}, I)
\]  

(2)
There are two common frameworks for image captioning: general encoder-decoder framework [16] and attention-based encoder-decoder framework [1, 3, 12]:

For the general framework (see Fig. 1 (a)), the context vector $c_t$ is only dependent on the encoder, a convolutional neural network (CNN). When each word in the sentence is generated, it keeps constant.

For the attention-based framework (see Fig. 1 (b)), the context vector $c_t$ is related to the attention weighting factor $a_{ist}$, $i=1,...,k$, which is calculated by the image features and the hidden state of the LSTM network. And many researchers have shown that attention mechanism can significantly improve the performance of image captioning. The basic form of attention is as follows:

$$e_{t,i} = f_{att}(v_i, h_{t-1})$$  \hspace{1cm} (3)

$$a_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{k} \exp(e_{t,j})}$$  \hspace{1cm} (4)

$$c_t = \sum_{i}^{k} a_{t,i} v_i$$  \hspace{1cm} (5)

Where $e_{t,i}$ is the attention distribution, $f_{att}$ is the attention function, $h_{t-1}$ is the previous hidden state of the LSTM network, $V= [v_1,...,v_k]$, $v_i \in R^d$ is the image features and $c_t$ is the context vector.

4. The Proposed Adaptive Attention Model

We used Resnet-152 [17], which was pre-trained on ImageNet, as an encoder for extracting image features. Compared with the general convolutional neural network, this kind of network is deeper and easier to optimize. At the same time, its complexity is lower than VGGnet [18]. The deep residual network can greatly increase depth and produce higher accuracy results, which is essentially better than the previous networks. Specifically, the outputs of the last convolutional layer of ResNet-152 are used, which have a dimension of 2048*7*7. It can be expressed as $V= [v_1,...,v_k]$, $v_i \in R^{2048}$, where $k$ represents the location of the $k$th grid.
In the task of automatically generating image captioning, the attention mechanism has attracted great interest due to its good performance. Our proposed adaptive attention model consists of an adaptive module and two attention modules. We define the attention module associated with the previous hidden state $h_{t-1}$ as Short-Term Attention module (STA), whose output is defined as $\alpha_t^{(h)}$. The attention module associated with the previous cell state $m_{t-1}$ is defined as Long-Term Attention module (LTA), and its output is defined as $\alpha_t^{(m)}$.

For the short-term attention module, at each time step, the image features $V$ and the previous state $h_{t-1}$ of the LSTM are fed into a single layer neural network followed by a softmax function to generate the attention distribution $z_t^{(h)}$:

$$z_t^{(h)} = w_o^T \tanh(W_v^T V + (W_h^T h_{t-1}))$$

$$\alpha_t^{(h)} = \text{softmax}(z_t^{(h)})$$

where $1 \in \mathbb{R}^k$ is a vector with all elements set to 1. $W_v \in \mathbb{R}^{k \times d}$, $W_h \in \mathbb{R}^{k \times n}$ and $w_o \in \mathbb{R}^k$ are parameters to be learnt. $\alpha_t^{(h)} \in \mathbb{R}^k$ is the attention weight which is normalized by $z_t^{(h)}$. Therefore, the context vector $c_t^{(h)}$ can be obtained by:

$$c_t^{(h)} = \sum_{i=1}^{d} \alpha_t^{(h)} v_i$$

Note that $v_i$ does not change over time. Similarly, for the long-term attention module, the same method is used to calculate $c_t^{(m)}$. Motivated from Lu et al. [3], we introduce an adaptive method for synthesizing $c_t^{(h)}$ and $c_t^{(m)}$ into a new context vector. It is defined as $c_t$:

$$c_t = \beta_h c_t^{(h)} + (1 - \beta_h) c_t^{(m)}$$

Where $\beta_t$ is the regulator factor, which is a scalar in the range [0,1]. A value of 1 indicates that only the output of the short-term attention module is used, and 0 indicates that only the output of the long-term attention module is used. The regulator factor $\beta_t$ is calculated by the previous hidden state $h_{t-1}$ and can be expressed as:

$$\beta_t = \sigma(W_f h_{t-1} + b_f)$$
Where $W_f \in \mathbb{R}^n$ are weight parameters to be learned, $b_f$ is bias, and $\sigma$ is the logistic sigmoid activation.

When decoding, the context vector $c_t$ and the embedded vector $e_t$ are concatenated into a new vector $x_t = [c_t; e_t]$ as an input to the LSTM network. Where $[\cdot;\cdot]$ indicates concatenation. Initial memory state and hidden state of LSTM as follows:

$$m_o = \tanh(W_p \left( \frac{1}{k} \sum_i v_i \right) + b_p)$$  \hspace{1cm} (11)

$$h_o = \tanh(W_q \left( \frac{1}{k} \sum_i v_i \right) + b_q)$$  \hspace{1cm} (12)

Where $W_p, W_q \in \mathbb{R}^{n \times d}$, $b_p$ and $b_q$ are biases. The word probability at time $t$ can be calculated as:

$$p(y_t | V, y_{<t}) = \text{soft max}(W_o (W_p e_t + W_q h_t + W_c c_t))$$  \hspace{1cm} (13)

where $W_o \in \mathbb{R}^{m \times n}$, $W_p \in \mathbb{R}^{m \times m}$, $W_q \in \mathbb{R}^{m \times n}$ and $W_c \in \mathbb{R}^{m \times d}$ are weight parameters to be learned. And $l$ is the size of the vocabulary, $m$ and $n$ denote the embedding and LSTM dimensionality respectively.

5. Experiments

We first briefly discuss the datasets and settings used in the experiments, and then analyze the results of these experiments.

5.1. Datasets and settings

We choose the popular data set MS-COCO [7] to evaluate the performance of our model. MS-COCO contains 82,783, 40,504 and 40,775 images for training, validation and testing, respectively. Each image has 5 human annotated captions. We report all the results using Microsoft COCO caption evaluation metrics, including BLEU [19], Meteor [20], Rouge-L [21] and CIDEr [22]. In our experiments, the dimensions of the word embedded vector and the hidden layer of the LSTM network were set as 512 and 1024 respectively. We discarded the captions of length greater than 18 and built a vocabulary of words that occurred at least 3 times in the training set, resulting in 11163 words was finally constructed. The input of the convolutional network is a center cropped 224x224 image. The Adam algorithm was applied to optimize the network.
5.2. Results and Analysis

The performance of different models are provided in Table 1. STA indicates short-term attention, LTA indicates short-term attention, and LSTAA indicates long short-term adaptive attention. On the whole, STA performs better than LTA, and LSTAA performs better than the former two. We added weight normalization (WN) [5], dropout (Dp) [6] incrementally and found that performance improves with every addition.

We also studied the effect of the number of hidden layers on the overall performance of LSTM network. With the increase of the number of hidden layers, the performance would be better, but the variation was small, and the training time of the model was significantly increased.

| Methods   | BLEU−1 | BLEU−2 | BLEU−3 | BLEU−4 | METEOR | ROUGE−L | CIDEr |
|-----------|--------|--------|--------|--------|--------|---------|-------|
| STA       | 70.9   | 52.3   | 39.6   | 29.1   | 24.2   | 57.2    | 89.4  |
| LTA       | 70.7   | 51.8   | 39.5   | 28.9   | 23.9   | 57.0    | 89.2  |
| LSTAA     | 71.3   | 52.5   | 40.1   | 29.2   | 24.3   | 57.4    | 90.7  |
| LSTAA+WN  | 71.4   | 52.6   | 40.2   | 29.5   | 24.5   | 57.7    | 90.9  |
| LSTAA+WN+D| 71.6   | 53.1   | 40.4   | 29.8   | 24.6   | 57.5    | 91.0  |

Table 2. The performance of different LSTM hidden layers

| Hidden Layers | BLEU−1 | BLEU−2 | BLEU−3 | BLEU−4 | METEOR | ROUGE−L | CIDEr |
|---------------|--------|--------|--------|--------|--------|---------|-------|
| 1             | 71.6   | 53.1   | 40.4   | 29.8   | 24.6   | 57.5    | 91.0  |
| 2             | 71.7   | 53.4   | 40.7   | 30.0   | 24.7   | 57.6    | 91.2  |
| 3             | 71.8   | 53.5   | 40.5   | 30.2   | 24.7   | 57.7    | 91.3  |
| 4             | 72.2   | 53.7   | 40.8   | 30.3   | 24.8   | 57.8    | 91.4  |
| 5             | 72.3   | 53.8   | 40.9   | 30.4   | 24.9   | 57.8    | 91.6  |

6. Conclusion

It is often the case that only a small region is relevant at any given time while generating a description. Then in this paper, we proposed a novel adaptive attention encoder-decoder framework which utilizes both the previous hidden state \( h_{t-1} \) and the previous cell state \( m_{t-1} \) of the LSTM networks to determine the region of the image feature to focus on. Extensive experiments performed on the MS COCO dataset verify the advantages of our approach. Though our model is evaluated on image captioning, it can be potentially applied to other works involving vision and language.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61772006, the Science and Technology Program of Guangxi under Grant No. AB17129012, the Science and Technology Major Project of Guangxi under Grant No. AA17204096, the Special Fund for Scientific and Technological Bases and Talents of Guangxi under Grant No. AD16380076, the
Special Fund for Bagui Scholars of Guangxi. and the basic ability improvement project for young and middle-aged teachers of Guangxi colleges and universities (2017KY0174).

References

[1] R. Kiros K. Cho A. Courville R. Salakhutdinov R. Zemel K. Xu, J. Ba and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. 2015.

[2] Zhilin Yang, Yuan Ye, Yuexin Wu, Ruslan Salakhutdinov, and William W. Cohen. Encode, review, and decode: Reviewer module for caption generation. 2016.

[3] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. 2016.

[4] Alex Graves. Long Short-Term Memory. 2012.

[5] Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. 2016.

[6] I. Sutskever W. Zaremba and O. Vinyals. Recurrent neural network regularization. 2014.

[7] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. 2014.

[8] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, and David A. Forsyth. Every picture tells a story: Generating sentences from images. 2010.

[9] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. 2011.

[10] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. 2012.

[11] Jia Xu, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. Guiding the long-short term memory model for image caption generation. In IEEE International Conference on Computer Vision, 2016.

[12] Quanzeng You, Hailin Jin, Zhaowen Wang, Fang Chen, and Jiebo Luo. Image captioning with semantic attention. 2016.

[13] Qingzhong Wang and Antoni B. Chan. Cnn+cnn: Convolutional decoders for image captioning. 2018.

[14] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Fethi Bougares, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014.

[15] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. Computer Science, 2014.

[16] Junhua Mao, Xu Wei, Yang Yi, Wang Jiang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). Eprint Arxiv, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. Computer Science, 2014.

[19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. Bleu: a method for automatic evaluation of machine translation. 2002.

[20] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. 2014.

[21] Carlos Flick. Rouge: A package for automatic evaluation of summaries. In Workshop on Text Summarization Branches Out, 2004.

[22] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. 2015