Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Modeling visual and word-conditional semantic attention for image captioning

Chunlei Wu [a], Yiwei Wei [a], Xiaoliang Chu [a], Fei Su [b,c], Leiquan Wang [a,*]

[a] College of Computer & Communication Engineering, China University of Petroleum (East China), Qingdao, China
[b] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[c] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China

ARTICLE INFO

Keywords:
Image captioning
Word-conditional semantic attention
Visual attention
Attention variation

ABSTRACT

Extensive efforts have been focused on attention-based frameworks for image captioning, which have achieved good performances when the generated words have an explicit corresponding with the image region. However, the generation of functional words, such as "on", "of", have not been investigated. In this paper, a dual temporal modal is first proposed for image captioning to address the role of visual information on every time step. Based on the dual temporal modal, word-conditional semantic attention is also proposed to solve the problem of functional words generation. Finally, a balance strategy is adopted on the basis of the attention variation to make a trade off between visual attention and word-conditional semantic attention. Extensive experiments are conducted on Flickr30k and COCO dataset to validate the effectiveness of the proposed method.

1. Introduction

Generating descriptions for images has been a challenging task in computer vision [1–3]. Recent attempts [4,5,1] mainly focus on the advances of attention-based model in machine translation. The attention-based image captioning model is developed from encoder–decoder framework, which transforms the visual feature (CNN decoder) to the target caption (LSTM decoder). The key insight of attention-based model is to make the highlighting spatial feature map an explicit correspondence to the generated words [5,6].

Attention based model has been proved to be effective for image captioning. However, it still suffers from the following two concerns. On the one hand, it loses track of the typical visual information. The generated sentence is prone to deviate from the original image content. An extension of LSTM (called gLSTM) that is guided by visual information of image should be beneficial to generating image captions [7]. On the other hand, the context vector for attention is correlated with the current hidden state [8]. Traditional attention methods use the last hidden state (h_{t-1}) as guidance. Recently, Xiong et al. [6] successfully performs current hidden state to generate image caption. The original visual information, however, is not fully considered which makes the generated caption lack personalities.

A highly qualified image caption generator should not only reflect the contents presented in the image, but also conform to the grammar rule. The attention based model generates context vector based on the visual feature at each time step, no matter what the upcoming word is [9,5,10]. This model mainly focuses on the accuracy of the notional words (e.g. "dog", "field"), which can be recognized from the image. However, it does little on the functional words (e.g. "the", "through"). Fig. 1(a) shows the distributions of soft attention weights over visual features. The variance of attention weight vector differs a lot when generating different words. A large variance indicates the upcoming word has an explicit correspondence with visual region. On the contrary, a small variance means that the word is puzzled on finding the corresponding visual signal. These variances illustrates that not all the words in the generated caption rely on visual information, such as the words "the" and "through". In fact, the semantic context plays an important role in generating the above two words. Both visual attention and semantic attention should be considered in image captioning. The authors of [6] use information preserved in memory cell as semantic information. However, utilizing the last generated word for semantic attention is much more flexible for image captioning.

In this paper, a new dual temporal model is proposed by simultaneously using two different LSTMs. The first LSTM is used to preserve the accumulated visual information. The other LSTM is applied to prevent the loss of visual information on each time step in the learning process. Both accumulated and original visual information are combined

[*] Corresponding author.
E-mail addresses: wuchunlei06@163.com (C. Wu), 360976808@qq.com (Y. Wei), 772063390@qq.com (X. Chu), sufei@bupt.edu.cn (F. Su), richiewlq@gmail.com (L. Wang).

https://doi.org/10.1016/j.image.2018.06.002
Received 1 December 2017; Received in revised form 28 May 2018; Accepted 4 June 2018
Available online 15 June 2018

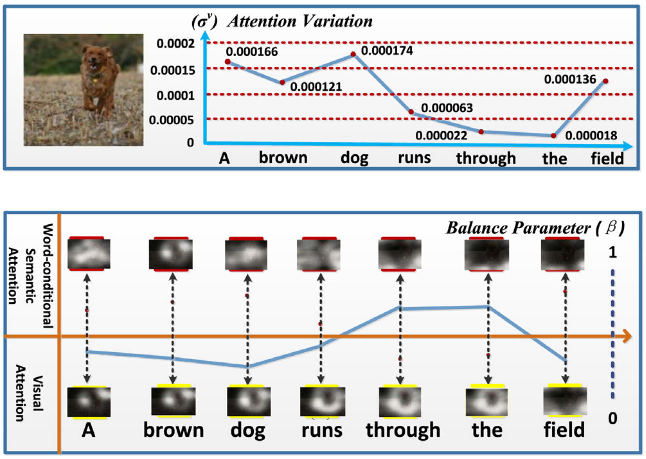

Fig. 1. (a) Attention variances σ^v in the soft attention model. Attention variance is used to measure the dispersion degree of attention weights at each step. Usually, the notional words have large attention variance value, which is easy to determine where to look in the image. On the contrary, the functional words are always with small attention variance value. (b) An example of visual and semantic attention for image captioning. β is the balance parameter of visual attention and semantic attention.

to diminish the uncertainty and enhance the flexibility of the next word prediction. Moreover, a self-balancing attention framework which contains visual attention and word-conditional semantic attention is also proposed. The visual attention aims to combine each generated word with the relevant image region, while the word-conditional semantic attention jointly learns how to focus on an image feature given the generated word. Then, attention variation is introduced to measure the dispersion of the distribution of balance parameter generated by the two attention vectors. Finally, a fusion of the visual attention and word-conditional semantic attention is performed to generate the corresponding words (see Fig. 1(b)).

To summarize, the main contributions of this paper are as follows:

- A new dual temporal model is proposed for image captioning, which contains two LSTMs in parallel. The two different LSTMs ensure the utilization of image information to strengthen the accuracy of attention model and diminish the uncertainty of the next word prediction respectively.
- Word-conditional semantic attention is proposed to solve the functional-words-generation problem by redistributing visual features with word-conditional guidance.
- Attention variation is introduced to measure the dispersion of visual context vector and semantic context vector. A self-balancing attention model is proposed to balance the influences of visual attention and semantic attention.
- Comprehensive experiments are conducted to empirically analyze the proposed method. The experimental results on COCO and Flickr30k datasets validate the effectiveness of this method.

The remainder of this paper is organized as follows. Section 2 discusses the most relevant work. In Section 3 the main frameworks and the training details are discussed. Section 4 demonstrates the experimental results. The last section is the conclusion.

2. Related work

Image caption generation is becoming important both in computer vision and machine learning communities. Recently, the neural network-based approaches [11–14] have become main stream in image captioning fields. Generally, the neural network-based literatures on image captioning can be divided into three categories: CNN + RNN based methods, attribute based methods and attention based methods.

CNN + RNN based captioning is inspired by the success of sequence-to-sequence encoder–decoder frameworks in machine translation [15,16]. The combination of CNN and RNN is the fundamental method, where CNN is used to extract the visual feature, and RNN is performed to construct the language model [2]. For predicting the next word given the image and previous words, Kiros et al. [11] first proposed a feed forward neural network, which is a multimodal log-bilinear model. However this method was gradually replaced by some novel ideas. For example, Vinyals et al. [17] used a LSTM instead of a vanilla RNN as the decoder. Mao et al. [12] presented a m-RNN model, where the CNN feature of the image is fed into the multimodal layer after the recurrent layer rather than the initial time step. But, the main drawback of m-RNN is the image represented with a static input. The visual feature extracted by CNN can well represent an image; however, the visual information will gradually diminish with the cells of RNN increased. To solve this problem, Donahue et al. [18] developed a strategy to feed the image feature to the RNN at each time step.

Attribute based captioning utilizes the high-level concepts or attributes [19–21] and then injects them into a neural-based approach as semantic attention to enhance image captioning. Yang et al. [4] put an intermediate attribute prediction layer into the predominant CNN–LSTM framework and implemented three attribute-based models for the tasks of image captioning. Wu et al. [22] proposed a method of incorporating high-level concepts into the successful CNN–RNN approach. Furthermore, Yao et al. [23] presented variants of architectures for augmenting high-level attributes from images to complement image representation for sentence generation.

Attention based captioning makes the image captioning more intelligent. Attention based captioning models incorporate an attention mechanism to learn a latent alignment from scratch when generating corresponding words [4,5,1]. Inspired by the traditional attention-based framework, Wei et al. [24] put forward a semantic attention mechanism for image caption generation which allows the caption generator to automatically learn which parts of the image feature to focus on when given previously generated text. Chang et al. [25] introduced a sequential attention layer, which takes all encoding hidden states into consideration when generating each word. Xiong et al. [6] initiated an adaptive attention model with a visual sentinel which can decide when and where to attend to the image. The method this paper proposed is also built on the attention framework. However, it is quite different from all the above attention-based models. In this paper, a fusion of visual and word-conditional attention based on coefficient of variation is explored to balance the influences of visual attention and semantic attention.

3. Proposed method

In this section, the previous encoder–decoder frameworks for image captioning is described in Section 3.1, then we the proposed model is presented in Section 3.2 and Section 3.3. In Section 3.4, the training details of the proposed model is stated.

3.1. Encoder–Decoder for caption generation

Encoder–Decoder [17,5,21]framework is widely used in image captioning. Its essential idea is to maximize the following formula with image and the corresponding sentence:

$$\theta^* = argmax_\theta \sum_{(I,y)} \log p(S|I;\theta) \tag{1}$$

where θ represents the parameters of the model, I is the image and S is the generated sentence. Applying the Bayes chain rule, the log of the distribution can be decomposed into the following structure:

$$\log p(S|I) = \sum_{t=1}^{N} \log p(S_t|I, S_1, \dots .S_{t-1}) \tag{2}$$