Signal Processing: Image Communication 67 (2018) 100–107

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

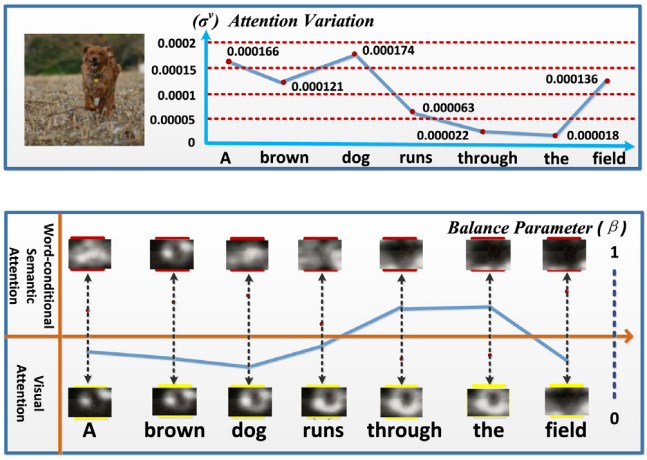

Fig. 1. (a) Attention variances σ^v in the soft attention model. Attention variance is used to measure the dispersion degree of attention weights at each step. Usually, the notional words have large attention variance value, which is easy to determine where to look in the image. On the contrary, the 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:

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

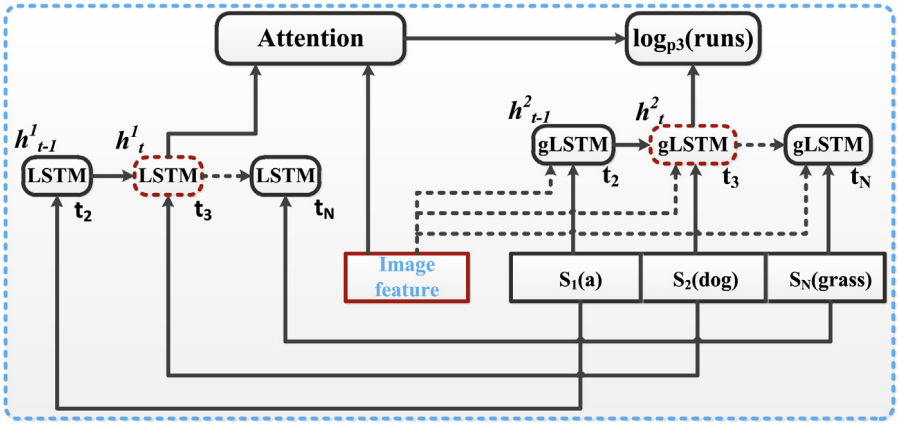

Fig. 2. Overview of the dual temporal model. Two LSTMs do not share a same time sequence. The first LSTM(without image feature) is considered to drive attention generation. The second gLSTM(with image feature) is the main process of the decoder. The red dotted box highlights the use of the current state(h_t^1 or h_t^2) which is different from the attention-based framework. The attention component contains a visual attention in the dual temporal model. It is noted that the visual attention will be replaced by the word-conditional semantic attention and visual attention which are mentioned in the following section.

where S_i represents the ith word in sentence S, S_0 is a special start word and S_N is a special stop word, $p(S_t|I, S_1, \ldots .S_{t-1})$ is the probability of the generating word S_t given image and the previous words $S_{1:t-1}$. Image is commonly represented by a CNN feature vector as the encoder, and the decoder part is usually modeled with recurrent neural networks (RNN).

As illustrated in the previous work, the Long-Short Term Memory(LSTM) achieves a better performance than vanilla RNN in image captioning. However, Jia et al. [7] pointed out that sometimes the generated sentences by the LSTM model will forget the original image content. They consider that LSTM only uses the image content at the beginning of the process and it loses a lot of visual information after a period of learning. Therefore, they propose a guiding LSTM (gLSTM), which puts the visual information into the LSTM model in every time step as an extra guidance [7]. The hidden states of LSTM and gLSTM are modeled respectively as:

$$h_t = LSTM(x_t, h_{t-1}, m_{t-1}) \tag{3}$$

$$h_t = gLSTM(x_t, g, h_{t-1}, m_{t-1}) \tag{4}$$

where x_t represents the current word, h_{t-1} is the hidden state of LSTM at time $t-1$, m_{t-1} denotes the memory cell of LSTM at time $t-1$ and g is the static image feature.

3.2. Dual temporal model

In attention-based frameworks for image captioning [4,5,1], the next generated word $word_t$ can be defined as:

$$word_t = MLP(h_{t-1}, c_t) \tag{5}$$

where h_{t-1} is the hidden state of LSTM at time $t-1$, and c_t represents the context vector, which can be obtained from Eq. (6):

$$c_t = func(h_{t-1}, V) \tag{6}$$

c_t is an important factor which provides visual attention information for caption generation [5,12,17,9]. Commonly, c_t relies on h_{t-1} and $V \in R^{d\times k}$. The visual extractor produces k vectors, each of which is a d-dimensional representation corresponding to a part of the image, where h_{t-1} is the hidden state from last time step and $V = [v_1, \ldots, v_k]$ is the image feature from CNN. Given the input vector x_t, the current hidden state can be expressed as:

$$h_t = gLSTM(x_t, c_t, h_{t-1}, m_{t-1}) \tag{7}$$

where x_t represents the current words, h_{t-1} represents the hidden state of LSTM at time $t-1$, m_{t-1} denotes the memory cell of LSTM at time $t-1$ and c_t is the weighted image context which is used to replace the static image feature g. The hidden state plays an important role in two parts: the attention generation part and the hidden-variable generation part.

Distinguished from traditional attention-based framework, we drive the two parts with two different LSTMs. The proposed model is shown in Fig. 2. The two LSTMs complement each other. The first spatial attention model stems from Xiong et al. [6] in which the current hidden state h_t is used to analyze where to look(i.e.,generating the context vector c_t) instead of the last hidden state h_{t-1}. Hence, we define h_t^1 from the first LSTM as:

$$h_t^1 = LSTM(x_t, h_{t-1}^1, m_{t-1}^1) \tag{8}$$

where h_t^1, h_{t-1}^1 and m_{t-1}^1 are the state of the first LSTM. Subsequently, given the output $h_i^1 \in R^D$ of the first LSTM, at each time step t, we generate a normalized attention weight α_i^v for each of the k image features $V \in R^{D\times K}$ as follows:

$$e_{ti}^v = w_a^T \phi(W_{va} V + (W_{ha} h_t^1)\zeta^T) \tag{9}$$

$$\alpha_{ti}^v = exp\{e_{ti}^v\} / \sum_{j=1}^{N} exp\{e_{tj}^v\} \tag{10}$$

where $w_a \in R^K$, $W_{va} \in R^{K\times D}$, $W_{ha} \in R^{K\times D}$are learned parameters, ϕ represents the $tanh$ activation function, $\zeta \in R^K$ is a vector with all elements set to 1 and $\alpha^v \in R^K$ are the attention weights. The context feature is calculated as a convex combination of all input features:

$$c_t^v = \sum_{(i=1)}^{k} \alpha_{ti}^v v_i \tag{11}$$

where v_i is ith region of the image feature V and c_t^v represents the visual context vectors.

Nevertheless, the visual information is not fully utilized in the first LSTM. In order to solve this problem, a guidance LSTM (gLSTM) is adopted to strengthen the role of visual information. Visual information is extracted from the image as extra input to each unit of the LSTM block. The main purpose is to strengthen the role of visual information. The model can be summed up as follows:

$$h_t^2 = gLSTM(x_t, V, h_{t-1}^2, m_{t-1}^2) \tag{12}$$

where h_t^2, h_{t-1}^2 *and* m_{t-1}^2 are the state of the second gLSTM. Then, the c_i^v and h_t^2 are fed into a multi-layer perceptron to generate the corresponding word:

$$word_t = MLP(h_t^2, c_t^v) \tag{13}$$

It is notable that the two LSTMs do not share the same data flow because of the data inconsistency. We call it "dual temporal model". The following works are based on the proposed dual temporal model.

3.3. Word-conditional semantic attention model

Due to the effectiveness of the attention mechanism in image captioning, soft attention [5] is adopt in the proposed dual temporal model. In fact, the soft attention plays an important role in generating the notional words (e.g. "dog" and "field"). However, in the soft attention experiments, it is found that α^v exhibits a dense distribution (with a small variance value) while the generated words are function words (e.g. "of", "on"). This phenomenon demonstrates that the algorithm has no clear clue on the generating words. In this situation, soft attention is hard to determine where to look. The main reason is that the semantic information of last word is not fully utilized. In order to solve this problem, we propose a word-conditional semantic attention model which is shown in the upper of Fig. 3. When the word-conditional semantic attention model receives an image, it redistributes the feature matrix V by the last generated words. It is achieved by:

$$V_t' = \Phi(V \odot W_c w_{t-1}) \tag{14}$$

where $V \in R^{D\times K}$ is the image feature, W_c is the word-conditional embedding matrix, $w_{t-1} \in R^L$ represents the last generated word in which L is the size of the word dictionary, and $\Phi(.)$ represents the *Relu* activation function. Then, a softmax function is performed on the mixed matrix V_t' to generate attention weights $\alpha_{ti}^s \in R^K$:

$$e_{ti}^s = w_{a'}^T \phi(W_{va'} V' + (W_{ha'} h_t^1)\zeta^T) \tag{15}$$

$$\alpha_{ti}^s = exp\{e_{ti}^s\} / \sum_{j=1}^{N} exp\{e_{tj}^s\} \tag{16}$$

where $w_{a'}^T \in R^K$, $W_{va'} \in R^{K\times D}$, $W_{ha'} \in R^{K\times D}$ are learned parameters, $\zeta \in R^K$ is a vector with all elements set to 1 and $\phi(.)$ represents the *tanh* activation function. Then, the context feature is calculated as a convex combination of all redistributed features:

$$c_t^s = \sum_{(i=1)}^{k} \alpha_{ti}^s v_i' \tag{17}$$

where v_i' is ith region of the redistributed image feature V' and c_t^s represents the semantic context vectors.

The word-conditional semantic attention allows the language model to learn the semantic information automatically through the conditional words. It acts as a complementary mechanism to visual attention.

3.4. Self-balancing attention model

To combine the advantages of the visual attention and word-conditional semantic attention, a self-balancing attention model is proposed. Generally speaking, when the weighted annotation vector α^v exhibits a large variance, it is easy to determine where to look. On the contrary, it is hard to make a correspondence between image region and word. The proposed self-balancing attention model is shown in Fig. 3. Attention Variation is introduced to measure the dispersion of α^v and α^s, which is formulated as:

$$\sigma^v = \frac{\sqrt{\frac{1}{N}\sum_1^N (\alpha_i^v - \mu^v)^2}}{\mu^v} \tag{18}$$

$$\sigma^s = \frac{\sqrt{\frac{1}{N}\sum_1^N (\alpha_i^s - \mu^s)^2}}{\mu^s} \tag{19}$$

Then, a balance parameter β is computed from σ^v and σ^s. The formula can be written as follows:

$$\beta = \frac{\sigma^v}{\sigma^v + \sigma^s} \tag{20}$$

In addition, two different approaches are considered to generate the balance parameter.

Soft Strategy: It directly uses β as the final weight to allocate the proportion of the two attentions. The two attentions are combined with a soft manner.

Hard Strategy: If β is greater than $1-\beta$, then β is equal to 1, otherwise β is equal to 0. Only one attention will be chosen.

Consequently, the context vector can be defined as:

$$c_t = \beta.c_t^v + (1-\beta).c_t^s \tag{21}$$

where c_t^v represents the context generated by σ^v, c_t^s represents the context generated by σ^s, and c_t is the balanced context vector.

3.5. Training details

In the proposed model, we initialize the model by training the model under the XE objective using ADAM [23] with a simple learning rate schedule, beginning with a learning rate of 0.0005 which is reduced to zero over 60K iterations and the hidden states of the two LSTMs are 512. We set the batch size 80 and the beam size 3. After 35 epochs, the results on COCO caption evaluation tool [26] are test every 5 epochs. It is found that the loss tends to be flatten about 37 epochs. The proposed model can be trained about 42 h on a single TitanX GPU with the COCO dataset.

4. Experimental results

4.1. Datasets and evaluation measurements

Experiments are conducted on Flickr30k [27] and Microsoft COCO [26] to evaluate the performance of the proposed model.

Flickr30k contains about 30000 images collected from Flickr. It comes with 5 reference sentences per image. Followed by [5,17], the publicly available splits containing 1000 images are used for validation and test each.

COCO is a challenging image captioning dataset, which contains 82783, 40504 and 40775 images for training, validation and test respectively. Different from Flickr dataset , images in this dataset contain complex scenes with multiple objects. Each image has 5 human annotated captions. To compare with previous methods, we follow the split from previous works [1,5,9]. For offline evaluation, we use 5000 images for validation and 5000 images for testing from the 40504 validation set. For online evaluation on the COCO evaluation server, we train the model with 82753 training dataset and 40504 validation dataset.

Pre-processing: We retain the words which appear at least 5 times, resulting in 8795 and 6359 words for COCO and Flickr30k respectively.

Compared Approaches: For offline evaluation on Flickr30k and COCO, we compare our method with **DeepVS** [1], **Berkeley LRCN** [18], **Attention** [5], **ERD** [4], **ATT-FCN** [9], **MSM** [23], **MAT** [25] and **Adaptive** [6]. For online evaluation, we compare our method with **LRCN** [18], **ATT-FCN** [9], **Attention** [5], **ERD** [4], **MSM** [23], **MAT** [25] and **Adaptive** [6].

4.2. Overall comparison

The results is reported by using the COCO captioning evaluation tool [26], which includes the following metrics: Bleu [28], Meteor [29], Rouge-L [30] and CIDEr [31]. The model is evaluated using the recent proposed metric SPICE [32], which is proved more consistent with human judgments and performs better with language generating models.

Through experiments, it is find that using **ResNet** [8] as the encoder performs better than **VGG** [33]. The methods of [6] also takes the ResNet-152 [8] in the encoder part. So, the **ResNet-152** is applied to extract the features. In addition, Google NIC, ERD and MSM use

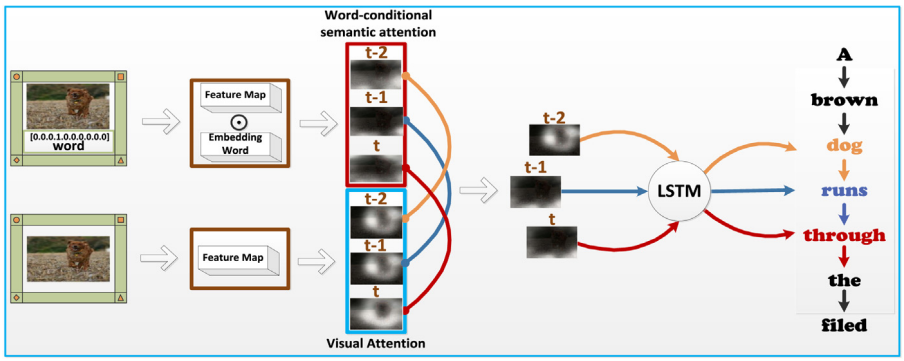

Fig. 3. Overview of the word-conditional semantic attention and visual attention model. The upper part mainly shows the forming process of semantic attention. The lower part mainly shows the formation of visual attention. A fusion strategy is performed on the two different parts to get a new context vector(two different methods). The new context is used to generate the upcoming word.

Table 1
Comparisons on MS COCO and Flickr30k. All metrics are reported using c5 references. For further comparisons, our SPICE scores are 0.162(Flickr30k) and 0.193(COCO).

Method	Flickr30k						MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr	B-1	B-2	B-3	B-4	METEOR	CIDEr
DeepVS [1]	0.573	0.369	0.240	0.157	0.153	0.247	0.625	0.450	0.321	0.230	0.195	0.660
ATT-FCN [9]	0.647	0.460	0.324	0.230	0.189	–	0.707	0.537	0.402	0.304	0.243	–
Attention [5]	0.669	0.439	0.296	0.199	0.185	–	0.718	0.504	0.357	0.250	0.230	–
ERD [4]	–	-	–	-	–	-	–	-	–	0.298	0.240	0.895
MSM [23]	–	-	–	-	–	-	0.730	0.565	0.429	0.325	0.251	0.986
MAT [25]	–	-	–	-	–	-	0.731	0.567	0.429	0.323	0.258	1.058
Adaptive [6]	0.677	0.494	0.354	0.251	0.204	0.531	0.742	0.580	0.439	0.332	0.266	1.085
DTM-SBA(ours)	**0.685**	**0.501**	**0.362**	**0.255**	**0.209**	**0.564**	**0.757**	**0.594**	**0.452**	**0.338**	**0.269**	**1.093**

Table 2
Comparisons on MS COCO Caption Challenge, using MS COCO evaluation server. All metrics are reported using c5 references.

Method	MS-COCO(c5)						
	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
LRCN [18]	0.718	0.548	0.409	0.306	0.247	0.528	0.921
ATT-FCN [9]	0.731	0.565	0.424	0.316	0.250	0.535	0.943
Attention [5]	0.705	0.528	0.383	0.277	0.241	0.516	0.865
ERD [4]	0.720	0.550	0.414	0.313	0.256	0.533	0.965
MSM [23]	0.739	0.575	0.436	0.330	0.256	0.542	0.984
MAT [25]	0.734	0.568	0.427	0.320	0.258	0.540	1.029
Adaptive [6]	0.746	0.582	0.443	0.335	0.264	0.550	1.061
DTM-SBA(ours)	**0.754**	**0.593**	**0.456**	**0.341**	**0.267**	**0.559**	**1.082**

Inception-v3 [25] as the visual feature extraction methods, which has similar classification performance compared to ResNet-152 [10].

The results on the Flickr30k and COCO datasets are shown in Table 1. Table 1 displays that the proposed model (dual temporal model + self-balancing attention)[1] exceeds all of the compared methods. A SPICE of 19.3(c5) is achieved by using public available SPICE evaluation tool. The results demonstrate the effectiveness of the proposed model. The same conclusion can also be obtained on COCO evaluation server which is shown in Table 2.

4.3. The performance of dual temporal model

To further demonstrate the effectiveness of the dual temporal model(DTM), we compare DTM with three baselines: 1) **Soft-attention model** [5]; 2)**Guiding Long-Short Term Memory model** [7]; 3)**Spatial model** [6]. DTM integrates the advantages of the above three methods. The results are shown in Table 3. DTM outperforms other methods on all of the metrics. Two conclusions can be drawn from Table 3. Firstly, it is necessary to add image feature at every time step. Secondly, the current state contains more guidance information than the previous state. Moreover, the experiment is conducted which is based on the proposed dual temporal modal without attention mechanism (DTM-NATT). The attention component in Fig. 2 is replaced by the concatenated previous hidden state of the first LSTM and the image feature. All the methods in Table 4 are not attention-based methods. As shown in Table 4, DTM-NATT also performs better than the single temporal baselines. The results in Tables 3 and 4 show that the proposed dual temporal model is effective in image captioning no matter with attention mechanism or not.

Table 3
Comparisons between dual temporal model with visual attention and three baseline methods.

Method	MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
soft attention [5]	0.707	0.492	0.344	0.243	0.239	0.960
gLSTM [7]	0.67	0.491	0.358	0.264	0.227	–
Spatial [6]	0.734	0.566	0.418	0.304	0.257	1.029
DTM(ours)	**0.742**	**0.572**	**0.426**	**0.313**	**0.263**	**1.042**

[1] DTM-SBA ranks 8th on MS COCO Caption Challenge Server by 2017-11-15 with the team name of upc-001.

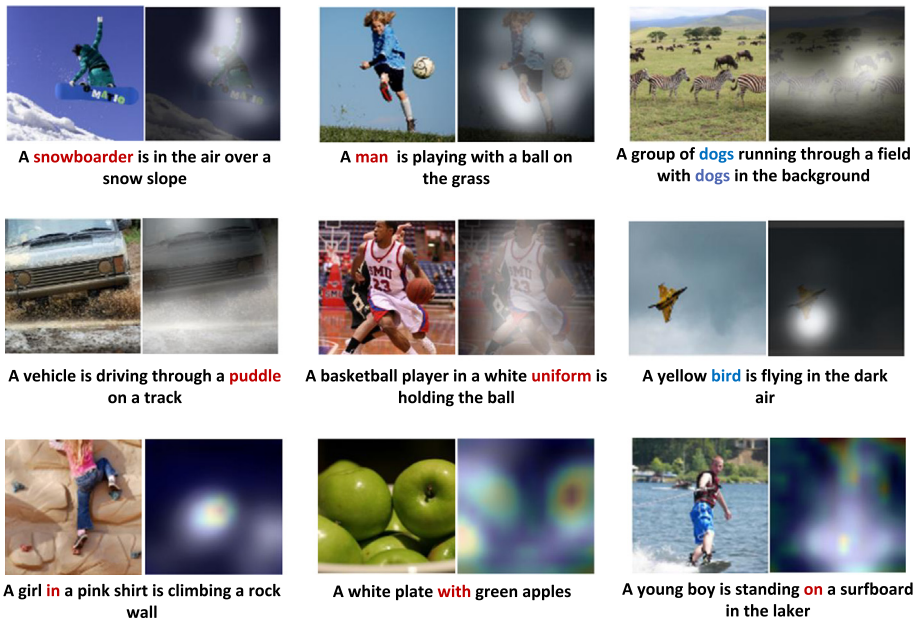

Fig. 4. Visualization of generated captions and attention maps. The blue or red words correspond to the generated white attention hotspots. The first two rows show the visualization of the visual attention model. The third row shows the visualization of the word-conditional semantic attention model.

4.4. Attention model selection

In this subsection, the performance of self-balancing attention model is discussed. Two kinds of balance strategy are mentioned in Section 3.4. DTM-SBA(soft) is a soft balance strategy, visual attention and word-conditional semantic attention are fused by a balance parameter. DTM-SBA(hard) is a hard balance strategy, only visual attention or word-conditional semantic attention will be selected. To demonstrate the effectiveness of self-balancing attention model, DTM-SBA(average) is also be discussed, which is an average strategy. The results are shown in Table 5. As shown in Table 5, DTM and DTM-SBA outperform DTM-NATT where attention mechanism is not used for image captioning. It indicates that the attention mechanism can improve the performance effectively in the proposed dual temporal modal. The direct comparison between DTM(semantic) and DTM(visual) demonstrates the advantage of the proposed word-conditional semantic model. Furthermore, self-balancing attention model (DTM-SBA) achieves better performance than visual attention model DTM(visual)and word-conditional semantic attention model DTM(semantic). It also demonstrates that word-conditional semantic attention has a complementary role with visual attention. These phenomenons not only indicate that the proposed framework has a good adaptability to the attention mechanism, but also demonstrate the balance strategy between visual and word-conditional semantic attention based on attention variation is effective for image captioning. Finally, the model with hard strategy achieves the best performance across all the evaluation metrics. On the contrary, the average strategy performs worst. This phenomenon confirms that an adaptive learning strategy is beneficial for attention-based image caption. It is in accordance with [5], in which hard attention performs better than other strategies.

4.5. Qualitative analysis

In order to make a full understanding of the proposed method, we show the captioning results and add an extra attention maps by visualizing the attention component. The results are selected from 5000 MS COCO testing images. As shown in Fig. 4, the first two lines represent the attention visualization of notional words and the last line represents the attention visualization of functional words. In particular, since the word-conditional semantic attention changes the original feature distributions extracted from image, it is not suitable to use the original picture to visualize the attention components. Therefore, we visualize the redistributed feature matrix as the background. It shows the real distribution of the word-conditional semantic attention. Through the comparison of different lines, the proposed model has a high sensitivity with word type. Besides, the columns are split as accurate and inaccurate captions. The first two columns are correct examples and the last column shows incorrect examples. The results shows that the learned alignments conform to human intuition. Though sometimes inaccurate in description (lack of high-level semantics), this model can accurately mark the corresponding locations.

Furthermore, we visualize the balance parameter of visual attention and word-conditional semantic models for each generated word. In addition, we also visualize the visual attention maps and the word-conditional semantic attention maps of each image. As shown in Fig. 5, the self-balancing model obtains an adaptive adjustment between the

Table 4
Comparisons between dual temporal model without attention mechanism (DTM-NATT) and two single temporal methods.

Method	MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
NIC [17]	0.666	0.451	0.304	0.203	–	–
gLSTM [7]	0.67	0.491	0.358	0.264	0.227	–
DTM-NATT(ours)	**0.711**	**0.512**	**0.377**	**0.279**	**0.242**	**0.913**

Table 5
Comparisons on attention model selection in DTM.

Method	MS-COCO					
	B-1	B-2	B-3	B-4	METEOR	CIDEr
DTM-NATT	0.711	0.512	0.377	0.279	0.242	0.913
DTM(visual)	0.742	0.572	0.426	0.313	0.263	1.042
DTM(semantic)	0.744	0.581	0.434	0.327	0.262	1.057
DTM-SBA(average)	0.725	0.552	0.407	0.295	0.240	1.002
DTM-SBA(soft)	0.754	0.592	0.449	0.336	0.266	1.091
DTM-SBA(hard)	0.757	0.594	0.452	0.338	0.269	1.093

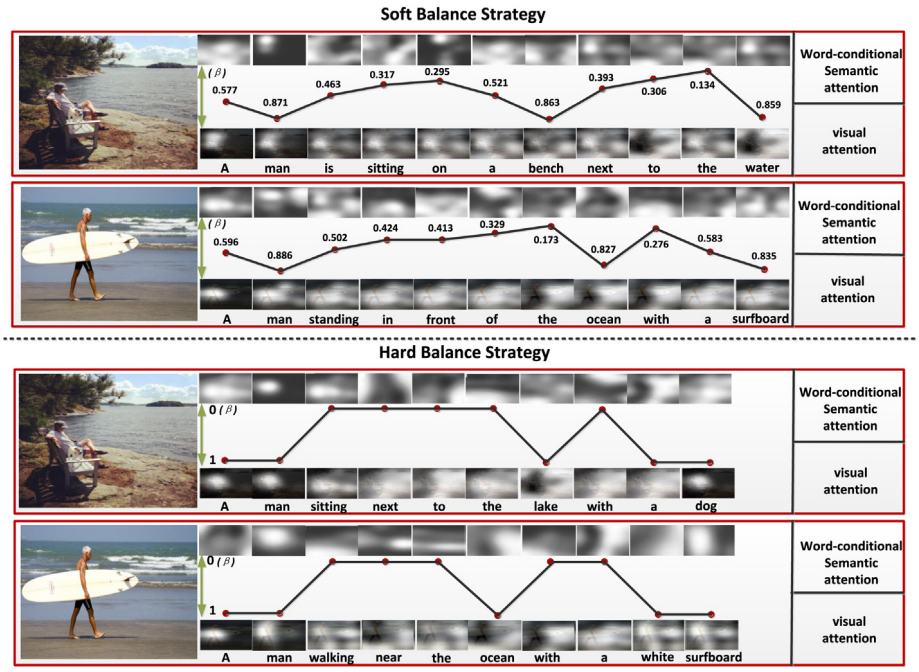

Fig. 5. Visualization of generated captions, visual attention maps, semantic attention maps and balance parameters. Two methods for generating fusion context are respectively shown in the figure.

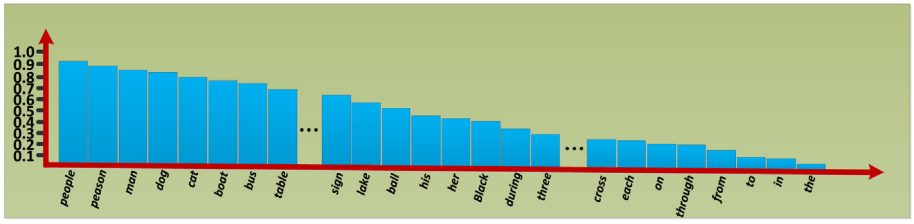

Fig. 6. The rank of the balance parameter for representative words.

visual attention and the word-conditional semantic attention on generating the upcoming word. When generating notional words such as "man", "surfboard" and "water", the visual attention possess a large balance weight. In contrast, for functional words such as "in" and "to", the word-conditional semantic attention is assigned a larger proportion.

To further illustrate the effect of the self-balancing attention model, the means of the balance parameter for each word that appears in the generated captions is shown in Fig. 6. 500 examples from COCO validate dataset are randomly sampled in the experiment. The results are shown in Fig. 6. The words are ranked by the balance parameter in a descending order. On the whole, the self-balancing model pays more attention to the visual attention when generating notional words like "man", "people", and "bus". When generating functional words such as "the" and "through" , the balance parameter leans toward the word-conditional semantic attention. The trends are developed without any priori information as input. In special cases, the same word can be assigned different weights when it appears in the different positions of a sentence. Taking "to" as an example, when the phrase is "go to", the balance parameter is less than 0.1; when the phrase is "next to", the balance parameter is more than 0.2. The reason is that the generated word relies on the different conditional words("go" and "next").

5. Conclusion

In this work, a new dual temporal model is proposed to make a full use of visual information for image captioning. A self-balancing attention model is also developed to balance the influences of visual and word-conditional semantic information. Experiments are conducted on the standard benchmarks for image captioning. Quantitative and qualitative evaluations show the effectiveness of the proposed method.

Acknowledgments

This work is supported by the grants from the National Natural Science Foundation of China (61532018, 61673396, 61671482), the Fundamental Research Funds for the Central Universities (17CX02041A, 18CX02136A, 18CX06045A).

References

[1] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: CVPR, 2015.
[2] R. Socher, A. Karpathy, Q.V. Le, C.D. Manning, A.Y. Ng, Grounded compositional semantics for finding and describing images with sentences, 2014.
[3] R. Mason, E. Charniak, Nonparametric method for data-driven image captioning, in: ACL, 2014.
[4] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, W.W. Cohen, Encode, review, and decode: Reviewer module for caption generation, in: NIPS, 2016.
[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML, 2015.
[6] C. Xiong, J. Lu, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: CVPR, 2017.

[7] X. Jia, E. Gavves, B. Fernando, T. Tuytelaars, Guiding the long-short term memory model for image caption generation, in: ICCV, 2015.
[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016.
[9] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: CVPR, 2016.
[10] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: NIPS, 2016.
[11] R. Kiros, R. Salakhutdinov, R.S. Zemel, Multimodal neural language models, in: ICML, 2014.
[12] J. Mao, Y.Y.W. Xu, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks(m-rnn), in: ICLR, 2015.
[13] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, C.L. Zitnick, Exploring nearest neighbor approaches for image captioning, Comput. Sci. (2015).
[14] K. Cho, B.V. Merrianboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder-decoder for statistical machine translation, Comput. Sci. (2014).
[15] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, Adv. Neural Inform. Process. Syst. (2014) 3104–3112.
[16] K. Cho, B.V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoderdecoder for statistical machine translation, Comput. Sci. (2014).
[17] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: CVPR, 2015.
[18] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: CVPR, 2015.
[19] L. Shiqi, Z. Cheng, F. Yan, Optimizing multi-graph learning based salient object detection, Signal Processing Image Communication, 2017.
[20] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NIPS, 2015.
[21] F.M. Muhammad, S.C. Qi, A. Gulnaz, Maximum mean discrepancy regularized sparse reconstruction for robust salient regions detection, Signal Process. Image Commun. (2017).
[22] Q. Wu, C. Shen, L. Liu, A. Dick, A. van den Hengel, What value do explicit high level concepts have in vision to language problems?, in: CVPR, 2016.
[23] T. Yao, Y. Pan, Y. Li, Z. Qiu, T. Mei, Boosting image captioning with attributes, in: ICCV, 2017.
[24] L. Zhou, C. Xu, P. Koch, J. Corso, Image caption generation with text-conditional semantic attention, in: CVPR, 2016.
[25] C. Liu, F. Sun, C. Wang, F. Wang, A. Yuille, Mat: A multimodal attentive translator for image captioning, in: IJCIA, 2017.
[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft coco:common objects in context, in: ECCV, 2014.
[27] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, in: ACL, 2014.
[28] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: ACL, 2002.
[29] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: EACL 2014 Workshop on Statistical Machine Translation, 2014.
[30] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: ACL 2004 Workshop, 2004.
[31] R. Vedantam, C.L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: CVPR, 2015.
[32] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice:semantic propositional image caption evaluation, in: ECCV, 2016.
[33] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Computer Science, 2014.

Chunlei Wu is a male associate professor in the college of computer and communication, China University of Petroleum (East China). He received the Ph.D. degree majoring in computer application technology from Ocean University of China in 2014. His current interests include image and video processing, and machine learning. He has authored and coauthored more than 30 journal and conference papers and textbooks.

YiWei Wei is a postgraduate student in college of computer and communication engineering, China University of Petroleum. His current research interests include cross modal retrieval and neural machine translation.

Xiaoliang Chu is a postgraduate in college of computer and communication engineering, China University of Petroleum. His current research interests include image caption, visual question answering and social media detection.

Fei Su is a female professor in the multimedia communication and pattern recognition lab, school of information and telecommunication, Beijing university of posts and telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.

Leiquan Wang received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT. Now he is a lecture in college of computer and communication engineering, China University of Petroleum. His current research interests include multimodal fusion, cross modal retrieval image/video caption and social media analysis.