Exploring Overall Contextual Information for Image Captioning in Human-Like Cognitive Style

Hongwei Ge[1], Zehang Yan[1], Kai Zhang[1], Mingde Zhao[2], Liang Sun[1]
[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, China
[2]Mila, McGill University, Montréal, Canada
{hwge,liangsun}@dlut.edu.cn, mingde.zhao@mail.mcgill.ca

Abstract

Image captioning is a research hotspot where encoder-decoder models combining convolutional neural network (CNN) and long short-term memory (LSTM) achieve promising results. Despite significant progress, these models generate sentences differently from human cognitive styles. Existing models often generate a complete sentence from the first word to the end, without considering the influence of the following words on the whole sentence generation. In this paper, we explore the utilization of a human-like cognitive style, i.e., building overall cognition for the image to be described and the sentence to be constructed, for enhancing computer image understanding. This paper first proposes a Mutual-aid network structure with Bidirectional LSTMs (MaBi-LSTMs) for acquiring overall contextual information. In the training process, the forward and backward LSTMs encode the succeeding and preceding words into their respective hidden states by simultaneously constructing the whole sentence in a complementary manner. In the captioning process, the LSTM implicitly utilizes the subsequent semantic information contained in its hidden states. In fact, MaBi-LSTMs can generate two sentences in forward and backward directions. To bridge the gap between cross-domain models and generate a sentence with higher quality, we further develop a cross-modal attention mechanism to retouch the two sentences by fusing their salient parts as well as the salient areas of the image. Experimental results on the Microsoft COCO dataset show that the proposed model improves the performance of encoder-decoder models and achieves state-of-the-art results.

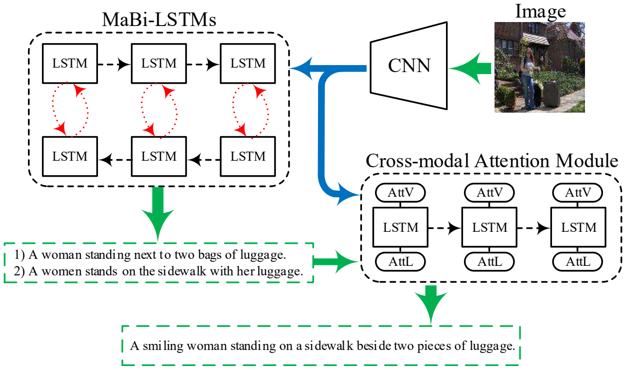

MaBi-LSTMs
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
CNN
Image
Cross-modal Attention Module
AttV
AttV
AttV
LSTM
LSTM
LSTM
AttL
AttL
AttL
1) A woman standing next to two bags of luggage.
2) A women stands on the sidewalk with her luggage.
A smiling woman standing on a sidewalk beside two pieces of luggage.

Figure 1. Overview of the proposed model. In the training process, image features extracted by CNN are input into MaBi-LSTMs to generate two sentences in an interactive manner. In the captioning process, the two sentences and image features are input into cross-modal attention module for generating the final sentence.

1. Introduction

As a multimodal task, image caption generation associates with expressing image content in a sentence accurately [17], which is called 'translation' from image to language [3, 37]. Providing an accurate description for an image is a challenging task for computers. The difficulties lie in that it not only requires the recognition of objects, attributes, and activities in an image but also the establishment of fluent sentences satisfying grammatical constraints and rules.

Extensive studies have been conducted and achieve promising results by combing advanced technologies of computer vision and natural language processing. According to the different ways of generating sentences, the existing methods generally fall into three categories [16], i.e., the template-based, the transfer-based and the deep model-based methods. The template-based methods utilize multiple classifiers to recognize objects, attributes and activities in an image by mapping from image to semantics, and then fill the identified semantic words into a manually designed template to make a sentence [34, 35, 18]. However, their performance noticeably deteriorates when dealing with more complex images due to the limitation of the number of classifiers and the lack of flexibility in templates. Different from the template-based methods, the transfer-based methods employ retrieving techniques to find visually similar images with the query image in an available database, and then transfer their captions to the query image [15, 19, 26]. The drawback lies in that the discrepancy in the similar images leads to the inaccuracy of captioning. Recently, deep learning has increasingly attracted attention due to its practical effectiveness in difficult tasks [7, 41, 30, 10]. Thus, many deep model-based methods [25, 43, 44, 20, 8] have been applied to image captioning. Generally, this kind of methods uses convolutional neural network (CNN) as a visual model to extract hierarchical features and long short-term memory (LSTM) as a language sequence model with variable length to generate descriptive sentences. Deep model-based methods not only eliminate the limitations of fixed templates but also generate original captions that are not included in available databases.

The baseline of deep model-based methods is the encoder-decoder framework where the CNN encodes raw image pixels into abstract features and the LSTM decodes abstract features into a sentence. The encoder-decoder models usually use forward LSTMs to generate words from begin to end to make a sentence [1, 2, 5]. Recently, bidirectional LSTMs have been developed to generate sentences from two directions independently, i.e., a forward LSTM and a backward LSTM are trained without interaction [38, 39]. However, there are three problems unsolved. First, only the preceding words are taken into account when predicting the next word in a sentence, namely, the information is inadequate. In fact, a sentence is made up of a sequence of words with contextual relations, where each word is related to not only its preceding words but also its subsequent words. Therefore, all explicit or implicit context information from preceding words and subsequent words will benefit the generation of current word. However, how to effectively acquire implicit information of subsequent words is challenging. Second, LSTMs may be misled during the captioning process for sentences with the same previous parts and the different remaining parts. For example, considering two sentences, one is "A little girl is happily licking an ice cream" and the other is "A little girl is happily drinking a cup of cola". Assuming that the current part is "A little girl is happily", it is difficult to determine whether the next word is "licking" or "drinking", since the verbs cannot be determined before we get the following word to describe food or drink. Thus, when the trained LSTMs encounters this kind of situation, it would "feel" confused when predicting "licking" or "drinking" without implicit context-after information in the captioning phase. Third, due to the domain adaptability of different models, there is a definite gap between the image representation obtained by CNN and the semantic generated by LSTM. It means that the information from images cannot be utilized extensively and effectively for sentence generation [42].

To address the above challenges, we conduct research inspired by a human-like cognitive style, which builds an overall cognition for the image to be described and the sentence to be constructed. More specifically, we propose a Mutual-aid network structure with Bidirectional LSTMs (MaBi-LSTMs) for exploring overall contextual information and a cross-modal attention mechanism to bride the gap between cross-domain models. Figure 1 illustrates the captioning process by using MaBi-LSTMs and cross-modal attention. The main contributions of this paper lie in:

- By designing an auxiliary structure, the forward and backward LSTMs encode the subsequent and preceding words into their respective hidden states through simultaneously constructing the whole sentence in a complementary manner. In this way, the hidden nodes contain not only the context-before information but also the context-after information.
- A cross-modal attention mechanism is proposed to dynamically focus on the salient areas of images and the salient parts of the sentences. This mechanism retouches the two sentences generated by MaBi-LSTMs into a higher quality sentence.
- The sufficient contextual information from MaBi-LSTMs and the cross-modal attention mechanism alleviate the problem that LSTM cannot seamlessly utilize image features.

2. Related works

As the proposed MaBi-LSTMs is closely related to the works utilizing deep neural networks, in this section, we review the deep models based on whether they are end-to-end trainable or phased trainable.

2.1. End-to-end trainable models

The end-to-end trainable models directly use images and corresponding captions for training. Vinyals et al. [37] use a deep CNN to encode images and use an LSTM to decode the encoded image representation into sentences. Further, in [16], semantic features of images are taken as part of the input of LSTM to produce more accurate sentences for an image. Considering that human visual attention mechanism can filter image noise based on saliency, Xu et al. [43] propose an attentive encoder-decoder model to dynamically concentrate on the local features of different image regions at each step of word sequence generation. Chen et al. [6] further extend spatial attention to spatial-channel attention considering that CNN features are naturally spatial, channel-wise and multi-layer. However, the lower and fixed resolution of CNN feature maps may lead to the loss of some important local visual information. To address this problem, Fu et al. [13] use selective search [33] to predict image regions containing objects and feed them to CNN to extract local features. Attention mechanism that only concentrates on local features cannot model global information effectively. Yant et al. [44] enhance the modelling ability of global information by a review network that plugs an LSTM in the attentive encoder-decoder. The plugged LSTM encodes local features as a series of hidden states that keep both local and global information. Although the attention mechanism can dynamically focus on salient image features, the prediction of some non-visual words (such as ‘to’, ‘the’, and ‘a’) does not require any visual information but requires contextual information only. Therefore, Lu et al. [22] propose an adaptive attention mechanism with visual sentinel. The visual sentinel is responsible for eliciting context information, while the CNN is responsible for extracting visual information. In order to fully exploit the long-term dependency of LSTM, Chen et al. [8] regularize LSTM by forcing its current hidden state to reconstruct the previous. This strategy enables LSTM’s hidden nodes to provide more information about the past. Besides, the traditional LSTM encodes its hidden state as a vector that does not contain spatial structure. For utilizing spatial structure, Dai et al. [9] adopt a two-dimensional map instead of a one-dimensional vector to represent the hidden state and validate that spatial structure can boost image caption generation.

2.2. Phased trainable models

Different from the end-to-end trainable models, the phased trainable models look for intermediary semantic attributes (usually the high-frequency words in the vocabulary) to bridge an image and its corresponding caption. This kind of models usually consists of two main components and is trained in several stages. The first component is used to produce most related word, and the second component is used to generate a sentence by utilizing the produced words as inputs. Many strategies have been proposed along this paradigm. After producing words by employing a set of attribute detectors based on CNN, You et al. [46] introduce semantic attention to focus on salient items.

Similarly, Wu et al. [42] establish a mapping to caption an image with its associate words by training a CNN in a multi-label classification senary. Fang et al. [11] adopt Multiple Instance Learning (MIL) as an alternative to yield associate words for the image to be captioned. Yao et al. [45] deliver the words from MIL to an LSTM in different ways and empirically claim that it is the optimal strategy for putting attributes into LSTM at the beginning. Instead of putting attributes into LSTM directly, Gan et al. [14] develop a semantic compositional network to exploit attributes more efficiently. The network works by extending each parameter matrix of LSTM into a set of matrices, with each being weighed based on different attributes. Besides, some other eligible features are also employed as intermediary attributes for the phased trainable models. For example, in [20], the object-level features are extracted by a pre-trained faster R-CNN [29] as intermediary attributes.

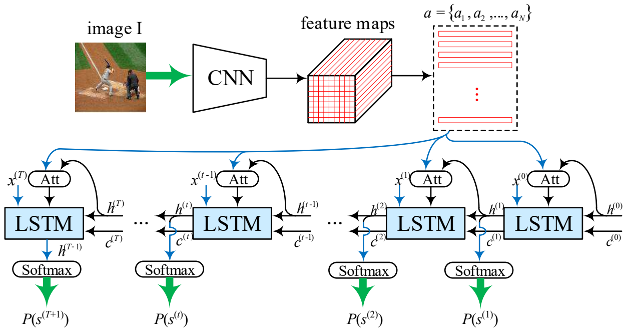


Figure 2. Attentive encoder-decoder diagram. Encoding stage: the image I is input into a CNN to extract the feature maps, and then the feature maps are split into a set of image features along spatial dimensions. Decoding stage: LSTM receives the word embedding vector of the previous moment $x^{(t-1)}$ and the salient image features $z^{(t)}$ from the attention module. Then, the hidden state of LSTM $h^{(t)}$ is input into the softmax layer to produce the probability distribution of words at the current moment $P(s^{(t)})$.

Despite significant progress, the way these models generating sentences for image captioning is different from that of human beings. When people describe pictures, they first have an overall cognition for the images to be described and the sentences to be constructed. Most existing models generate a word sequence one by one in a front-to-back manner, without considering the influence of the subsequent words on the whole sentence generation. Bidirectional LSTMs have been developed to generate sentences from two directions [38, 39] independently. Essentially, it is the same way as before since the forward and backward LSTMs are still trained without interaction. This paper proposes a mutual-aid network structure with bidirectional LSTMs for exploring overall contextual information. In MaBi-LSTMs, the forward and backward LSTMs encode the succeeding and preceding words into their respective hidden states by constructing the whole sentence in a complementary manner simultaneously. The MaBi-LSTMs generate two sentences in an interactive way. Moreover, a special module of cross-modal attention mechanism is designed to retouch the two sentences into a new sentence with higher quality. The MaBi-LSTMs work in an end-to-end trainable manner, while the stage of sentence retouching works in a phased trainable manner, where the two pre-generated sentences from MaBi-LSTMs operate as the intermediary attributes.

3. Methods

In this section, we start by briefly describing the generic encoder-decoder image captioning framework, and then we give the details of the proposed model focusing on the mutual-aid network structure with bidirectional LSTMs and the cross-modal attention for decoding.

3.1. Attentive encoder-decoder model

The attention-based encoder-decoder model consists of a CNN-based encoder and a LSTM-based decoder with the attention module.

Encoder. After feeding an image I to a CNN, the feature map from the last convolutional layer is split along spatial dimensions and yields a set of vectors $a = \{a_1, a_2, ..., a_i, ..., a_N\}$, where a_i is a D-dimensional vector representing the feature of an image region. The process means that the CNN encodes image I into N vectors:

$$a = CNN(I) \tag{1}$$

Decoder. The word sequence $S = (s^{(0)}, s^{(1)}, ..., s^{(t)}, ..., s^{(T+1)})$ represents the sentence corresponding to image I, where $s^{(t)}$ is a one-hot vector whose length is equal to the size of the dictionary. $s^{(t)}$ indicates the index of the word in the dictionary. $s^{(t)}$ can be embedded in a compact vector space by multiplying an embedding matrix E, i.e., $x^{(t)} = Es^{(t)}$. During the process of decoding $a = \{a_1, a_2, ..., a_i, ..., a_N\}$ into $S = (s^{(0)}, s^{(1)}, ..., s^{(t)}, ..., s^{(T+1)})$, the input of LSTM at time t includes not only $x^{(t-1)}$ but also the salient image feature $z^{(t)}$ computed by performing attention mechanism on all the image features:

$$h^{(t)} = LSTM(x^{(t-1)}, z^{(t)}, h^{(t-1)}) \tag{2}$$

Attention. Visual attention is usually implemented by taking the weighted sums of all image features to obtain the salient image features, which is denoted as $z^{(t)} = Att(h^{(t-1)}, a)$, where $z^{(t)}$ is correlated with $h^{(t-1)}$. The detailed computation is formulized as follows:

$$z^{(t)} = \sum_{i=1}^{N} \alpha_i^{(t)} a_i \tag{3}$$

$$\alpha_i^{(t)} = \frac{exp(MLP(h^{(t-1)}, a_i))}{\sum_{j=1}^{N} exp(MLP(h^{(t-1)}, a_j))} \tag{4}$$

Where $\alpha_i^{(t)}$ represents the weight of the feature of the i-th image region at time t, and MLP represents a multi-layer perceptron.

A softmax layer is employed to convert $h^{(t)}$ into the probability distribution of $s^{(t)}$, from which the word at time t can be sampled:

$$P(s^{(t)}|s^{(0)}, ..., s^{(t-1)}, I, \theta) = softmax(h^{(t)}) \tag{5}$$

where θ denotes the learnable parameters of LSTM.

Figure 2 shows the structure of the attentive encoder-decoder model and details the interaction between the attention module and LSTM. To illustrate more clearly, the subsequent diagrams will use an abbreviated form $LSTMA$ to represent LSTM with the attention module. Thus, Equation (2) is rewritten as:

$$\begin{aligned} h^{(t)} &= LSTM(x^{(t-1)}, Att(h^{(t-1)}, a), h^{(t-1)}) \\ &= LSTMA(x^{(t-1)}, a, h^{(t-1)}) \end{aligned} \tag{6}$$

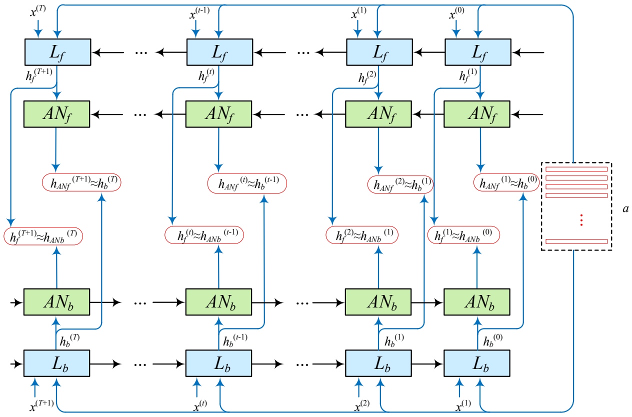

$x^{(T)}$
$x^{(t\text{-}1)}$
$x^{(1)}$
$x^{(0)}$
L_f
L_f
L_f
L_f
$h_f^{(T+1)}$
$h_f^{(t)}$
$h_f^{(2)}$
$h_f^{(1)}$
AN_f
AN_f
AN_f
AN_f
$h_{ANf}{}^{(T+1)} \approx h_b{}^{(T)}$
$h_{ANf}{}^{(t)} \approx h_b{}^{(t\text{-}1)}$
$h_{ANf}{}^{(2)} \approx h_b{}^{(1)}$
$h_{ANf}{}^{(1)} \approx h_b{}^{(0)}$
a
$h_f^{(T+1)} \approx h_{ANb}{}^{(T)}$
$h_f^{(t)} \approx h_{ANb}{}^{(t\text{-}1)}$
$h_f^{(2)} \approx h_{ANb}{}^{(1)}$
$h_f^{(1)} \approx h_{ANb}{}^{(0)}$
AN_b
AN_b
AN_b
AN_b
$h_b^{(T)}$
$h_b^{(t\text{-}1)}$
$h_b^{(1)}$
$h_b^{(0)}$
L_b
L_b
L_b
L_b
$x^{(T+1)}$
$x^{(t)}$
$x^{(2)}$
$x^{(1)}$

Figure 3. The schematic diagram of the proposed MaBi-LSTMs. The green boxes denote the auxiliary modules, and the blue boxes denote the original bidirectional LSTMs. L_f and L_b reserve two sequences of hidden states $h_f = (h_f^{(1)}, ..., h_f^{(t)}, ..., h_f^{(T+1)})$ and $h_b = (h_b^{(0)}, ..., h_b^{(t-1)}, ..., h_b^{(T)})$ in the process of generating sentences by using the image feature set a. In addition, L_f and L_b construct the hidden states of each other. AN_f models h_f in an forward order to build the new state sequence h_{AN_f} to approximate h_b for capturing the context-after information. AN_b traces h_b in a backward order to form state sequence h_{AN_b} to approximate h_f for capturing the context-before information.

3.2. Mutual-aid network structure with bidirectional LSTMs

In this section, we present the details of the proposed MaBi-LSTMs. The schematic diagram of MaBi-LSTMs is shown in Figure 3.

We introduce the auxiliary forward aid network AN_f and backward aid network AN_b in the bidirectional LSTMs. Then the auxiliary and the original LSTMs (foward LSTM L_f and backward LSTM L_b) work coordinately by mutually constructing hidden states via the auxiliary LSTMs. By introducing the auxiliary LSTMs (forward aid network AN_f and backward aid network AN_b) structure into the bidirectional LSTMs, the auxiliary and the original LSTMs (foward LSTM L_f and backward LSTM L_b) work coordinately by mutually constructing hidden states.

In MaBi-LSTMs, L_f constructs the hidden states of L_b via AN_f to capture latent context-after information and L_b builds the hidden states of L_f via AN_b to capture latent context-before information. To effectively construct the hidden states of L_f and L_b, we adopt LSTM as the specific implementation for AN_f and AN_b to model the temporal dynamics of time series data. In the process of sentence generation, L_f updates its hidden states in a front-to-back manner:

$$h_f^{(t)} = L_f(x^{(t-1)}, a, h_f^{(t-1)}) = LSTMA_f(x^{(t-1)}, a, h_f^{(t-1)}) \tag{7}$$

L_b updates its hidden states from back to front:

$$h_b^{(t-1)} = L_b(x^{(t)}, a, h_b^{(t)}) = LSTMA_b(x^{(t)}, a, h_b^{(t)})) \tag{8}$$

To make L_f and L_b aid each other, AN_f takes $h_f^{(t)}$ as input to generate hidden states $h_{AN_f}^{(t)}$ for approximating $h_b^{(t-1)}$:

$$h_{AN_f}^{(t)} = AN_f(h_f^{(t)}, h_{AN_f}^{(t-1)}) \approx h_b^{(t-1)} \tag{9}$$

Similarly, AN_b uses $h_b^{(t-1)}$ to approximate $h_f^{(t)}$:

$$h_{AN_b}^{(t-1)} = AN_b(h_b^{(t-1)}, h_{AN_b}^{(t)}) \approx h_f^{(t)} \tag{10}$$

In Figure 3, L_f and L_b construct the hidden states of each other via AN_f and AN_b, respectively. The $h_{AN_b}^{(t-1)}$ integrates the hidden states of L_b after time $t-1$ (including $h_b^{(t-1)}$ up to h_b^T). The $h_f^{(t)}$ is closely related to the first half of the sentence (from $x^{(0)}$ to $x^{(t-1)}$), which contains the context-before information implicitly. If we can make $h_{AN_b}^{(t-1)}$ close enough to $h_f^{(t)}$, $h_{AN_b}^{(t-1)}$ will contain extra knowledge about the first half sentence, which is transferred to $h_b^{(t-1)}$ implicitly. Therefore, it helps to predict the next word $x^{(t-1)}$ based on $h_b^{(t-1)}$. By analogy, $h_{AN_f}^{(t)}$ integrates the hidden states of L_f before time t (including $h_f^{(1)}$ to $h_f^{(t)}$). The $h_b^{(t-1)}$ is closely related to the latter part of the sentence (from $x^{(t)}$ to $x^{(T+1)}$), which contains context-after information implicitly . If we can make $h_{AN_f}^{(t)}$ close enough to $h_b^{(t-1)}$, $h_{AN_f}^{(t)}$ will contain extra knowledge about the latter part of the sentence. Hence, it makes sense to predict the next word $x^{(t)}$ based on $h_f^{(t)}$ with higher confidence.

Loss function. L_f and L_b offer two probability distributions bidirectionally. Therefore, the first part of the loss function should be the sum of two negative log-likelihood functions that represents the errors via supervised training. To make the AN_f and the AN_b effectively model the hidden states of L_f and L_b, respectively, the least square error that represents the error of construction is used as the second part of the loss function. The final loss function is the sum of L_1 and L_2 weighted by λ.

$$\begin{aligned} L_1 = &- \sum_{t=1}^{T} log(P_f(s^{(t)}|s^{(0)}, ..., s^{(t-1)}, I, \theta_f)) \\ &- \sum_{t=1}^{T} log(P_b(s^{(t)}|s^{(T+1)}, ..., s^{(t+1)}, I, \theta_b)) \end{aligned} \tag{11}$$

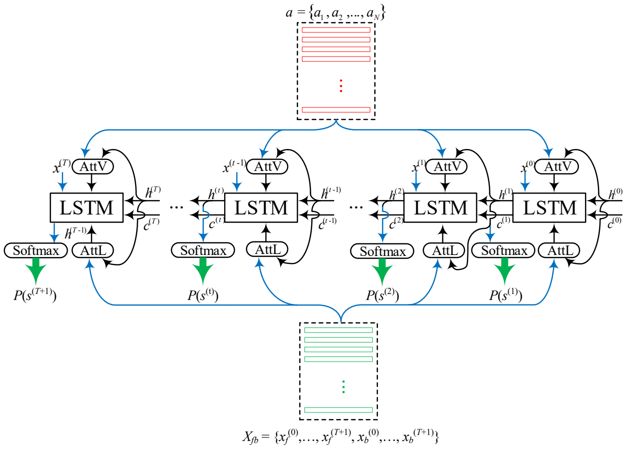


Figure 4. The cross-modal attentive decoder. Cross-modal attention incorporates semantic attention AttL and visual attention AttV. The cross-modal attention module selects salient image features and semantic features for updating the hidden states followed by the softmax to generate word probability distributions.

$$L_2 = \sum_{t=1}^{T} \|h_{AN_b}^{(t-1)} - h_f^{(t)}\|^2 + \sum_{t=1}^{T} \|h_{AN_f}^{(t)} - h_b^{(t-1)}\|^2 \tag{12}$$

$$L_{total} = L_1 + \lambda L_2 \tag{13}$$

3.3. Cross-modal attention for decoding

Cross-modal attention extends visual attention to both visual and semantic aspects. From the visual aspect, CNN can extract image feature set a. From the semantic aspect, the trained L_f and L_b can produce two sentences $S_f = (s_f^{(0)}, ..., s_f^{(T+1)})$ and $S_b = (s_b^{(0)}, ..., s_b^{(T+1)})$. S_f and S_b can be represented by their corresponding word embedding $X_{fb} = (x_f^{(0)}, ..., x_f^{(T+1)}, x_b^{(0)}, ..., x_b^{(T+1)})$. The image feature set a and the word embedding X_{fb} constitute the multimodal features in the cross-modal environment. Figure 4 shows the process of generating the final sentence by using cross-modal attention.

Cross-modal attention can be disassembled into visual attention and semantic attention and can operate on image features and word embeddings to improve the saliency of cross-modal features. The salient image features and the word embedding obtained by cross-modal attention are concatenated, and then input into LSTM to update the hidden states. Through the softmax layer, the hidden states at each time step are converted into probability distributions of words, where the whole sentence can be predicted by sampling. The loss function here is taken as the negative log-likelihood.

LSTM	λ	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
L_f	0	72.7	56.0	42.4	32.1	25.5	53.6	99.9
	0.001	73.8	57.4	43.7	33.1	25.7	54.5	102.6
	0.005	73.9	57.6	43.9	33.2	25.8	**54.9**	102.6
	0.01	**74.4**	**58.0**	**44.2**	**33.5**	**26.1**	**54.9**	**105.7**
	0.05	73.8	57.2	43.7	33.4	25.7	54.3	102.6
	0.1	72.8	56.1	42.4	32.0	25.3	53.8	100.0
L_b	0	72.9	56.0	41.8	31.0	25.3	53.5	99.7
	0.001	73.5	**57.3**	43.0	**32.2**	**25.6**	54.2	102.5
	0.005	73.2	56.7	42.7	31.8	25.5	**54.3**	102.3
	0.01	**74.2**	**57.3**	**43.9**	31.9	25.5	54.1	**103.6**
	0.05	73.2	56.7	42.8	31.9	25.4	53.9	102.2
	0.1	72.8	55.7	41.5	30.7	25.2	53.3	99.5

Table 1. Results of the attentive encoder-decoder models with different weight parameters. When $\lambda = 0$, the model is without the auxiliary structure.

4. Experiments

To validate the effectiveness of the proposed algorithm, we conduct experiments on the Microsoft COCO dataset [21]. The experimental results are analysed and compared with the state-of-the-art algorithms.

4.1. Dataset and evaluation metrics

The Microsoft COCO dataset is a popular large scale dataset for image captioning, including 82,783 images for training, 40,504 images for validation, and 40,775 images for testing. Each image in the training and validation set accompanies with five sentences and each sentence can describe the image contents. Such descriptive sentences are manually labelled by humans through the Amazon Mechanical Turk platform. To make fair comparisons with other methods, we follow a common accepted configurations in the community [17], and select 5000 images from the validation set for validating and another 5,000 images from the validation set for testing.

Commonly used metrics for measuring caption quality include BLEU-n [27], ROUGE-L [12], METEOR [4], and CIDEr [36]. BLEU-n measures the similarity between a candidate sentence against reference sentences by computing the n-gram precision. METEOR computes not only uni-gram precision but also recall and uses their weighted harmonic mean. ROUGE-L employs the longest common subsentences to measure the similarity between a candidate sentence and reference sentences at the sentence level. The above three evaluation metrics are derived from the study of machine translation tasks, while CIDEr is based on human consensus and specifically designed for image captioning. To validate the effectiveness of the proposed method, we use the four evaluation metrics: BLEU-n (n=1, 2, 3, 4), METEOR, ROUGE-L and CIDEr.

4.2. Implementation details

After using the Stanford PTB Tokenizer [24] to split sentences into words, we remove all the punctuations. The dictionary is then built by words with the frequency higher than 5. Besides those high-frequency words, the dictionary contains three particular tokens: the sentence-starting token $< start >$, the sentence-ending token $< end >$, and the unknown token $< unk >$, where $< unk >$ is used to replace words with the frequency lower than 5. For extracting compressed image features and reducing computational cost, features from CNN are input into a one-layer neural network with linear activation function for dimension reduction. To alleviate over-fitting, we stop training when the CIDEr score on the validation set begins to decline. There exists different strategies when continuously sampling words from probability distributions at each time step. The most straightforward approach is to pick the word with the highest probability at each time step to make a sentence. However, such a strategy is greedy. Beam Search differs from Greedy Search in selecting n candidate words with the highest probabilities. Beam Search degenerates into Greedy Search when $n = 1$ and transforms into Breadth-First Search when n takes the maximum value. Considering both the computation and the performance, we set the value of n as 3.

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
DeepVS [17]	62.5	45.0	32.1	23.0	19.5	-	66.0
MSR [11]	-	-	-	25.7	23.6	-	-
gLSTM [16]	67.0	49.1	35.8	26.4	22.7	-	81.3
Bi-S-LSTM^V [39]	68.7	50.9	36.4	25.8	22.9	-	73.9
Bi-LSTM$^{A,+M}$ [39]	65.6	47.4	33.3	23.0	21.1	-	69.5
Attr-CNN+LSTM [42]	74.0	56.0	42.0	31.0	26.0	-	94.0
ATT-FCN [46]	70.9	53.7	40.2	30.4	24.3	-	-
Soft-Attention [43]	70.7	49.2	34.4	24.3	23.9	-	-
Hard-Attention [43]	71.8	50.4	35.7	25.0	23.0	-	-
Review Net [44]	-	-	-	29.0	23.7	-	88.6
LSTM-A5 [45]	73.0	56.5	42.9	32.5	25.1	53.8	98.6
SCA-CNN [6]	71.9	54.8	41.1	31.1	25.0	53.1	95.2
GBVS [32]	-	-	-	28.7	23.5	21.2	84.1
Areas of Attention [28]	-	-	-	31.9	25.2	-	98.1
ARNet [8]	74.0	57.6	44.0	33.5	26.1	54.6	103.4
SCN-LSTM [14]	72.8	56.6	43.3	33.0	25.7	-	101.2
Skeleton [40]	74.2	57.7	44.2	34.0	26.8	55.2	106.9
AdaAtt [22]	74.2	58.0	43.9	33.2	26.6	54.9	108.5
GroupCap [5]	74.4	58.1	44.3	33.8	26.2	-	-
NBT [23]	75.5	-	-	34.7	27.1	-	107.2
Our MaBi-LSTMs(L_f)	77.5	59.9	46.2	36.4	27.2	56.1	110.4
Our MaBi-LSTMs(L_b)	75.8	59.4	45.1	34.9	26.8	55.8	108.5
Our MaBi-LSTMs(with attention)	**79.3**	**61.2**	**47.3**	**36.8**	**28.1**	**56.9**	**116.6**

Table 2. Comparative results of MaBi-LSTMs with different algorithms on the COCO test set.

The proposed algorithm is implemented by Python and Theano on a workstation with a Tesla P40 GPU.

4.3. Experiments for MaBi-LSTMs

To verify the effectiveness of mutual-aid bidirectional LSTMs, we conduct the comparative experiments of the attentive encoder-decoder models with or without the auxiliary structure. An Inception-V4 [31] pre-trained on the ImageNet classification dataset acts as the encoder and the parameters are kept unchanged. The loss function of the mutual-aid network structure with bidirectional LSTMs consists of two items: the cross-entropy error L_1 and the least squares error L_2. The parameter λ is important for adjusting the ratio of these two items. We empirically find that the value of L_2 is 2 or 3 orders higher than that of L_1. In order to balance these two items so that they can work on the same level, we search for an appropriate value for λ in the interval of $[0, 0.1]$. Actually, we test the performances of the proposed model when the values of λ is taken as 0, 0.001, 0.005, 0.01, 0.05, and 0.1 respectively, and the results are presented in Table 1.

When λ is taken as 0, the least squares error does not work, and only the cross-entropy error is used, that is, without auxiliary structure, the training of L_f and L_b has no interaction. When λ is taken as non-zero values except 0.1, the scores are obviously higher than those with $\lambda = 0$. The results verify the effectiveness of the auxiliary structure. When λ is taken as 0.1, the least squares error exceeds cross entropy excessively according to the difference in magnitude, which makes the training target of the model more biased towards the mutual construction and ignores the ability of word prediction. In addition, L_f generally performs better than L_b for each metric. This is consistent with our common sense because it is easier to recite a sentence from front to back than to recite it reversely. When λ is taken as 0.01, all the metrics generally achieve the best values. So, we set λ as 0.01 in the remaining experiments.

The results listed in Table 1 are obtained by only training the decoder while keeping the encoder frozen. To further improve the model performance, the model is continued to be trained by jointly fine-tuning the encoder and decoder. Table 2 presents the results obtained by L_f and L_b and the comparison results with some state-of-the-art algorithms. Notably, we determine whether to stop the training process in time for alleviating over-fitting based on CIDEr score. Without more complicated training tricks and experiences, the proposed model still achieves competitive results.

Figure 5 shows some pairs of captions generated by L_f and L_b with and without the auxiliary structure for 4 images. From the comparison results, it can be seen that MaBi-LSTMs can improve the quality of image captions. For the first image, either of the captions generated by the model without the auxiliary structure is incomplete. One only describes the person and the other only describes sheep. MaBi-LSTMs can capture both objects and express the person and sheep in a single sentence. For the second image, the initial captions do not include the important image content, ‘mirror’, and cannot accurately identify the number of sinks. The captions generated by using the auxiliary structure can describe ‘mirror’ and the number of sinks. For the third image, it is commendable that MaBi-LSTMs can still produce reliable captions even if the captions generated without auxiliary structure are completely unqualified in describing the image contents. For the fourth image, although the captions describe the main objects in image contents, they can be further polished by adding an adjective word, such as ‘pink’.

In summary, MaBi-LSTMs can not only accurately describe the objects contained in the image, but also correct some of the wrong descriptions. Moreover, the information provided by some fine qualifiers, such as colours, can also be supplemented by MaBi-LSTMs.

4.4. Experiments for the overall model

MaBi-LSTMs generate a pair of captions for an image in both forward and backward directions. Those words in the pre-generated captions are converted into word embedding vectors. Word embeddings and the image features extracted by the CNN are input into cross-modal attention module for producing image captions with higher quality. Following the same strategy as the previous experiments, we first freeze the parameters of the CNN to train only the LSTM with cross-modal attention and then train the whole model by fine-tuning jointly. The last row in Table 2 presents the results obtained by using MaBi-LSTMs and cross-modal attention together. It can be seen that cross-modal attention further enhances caption quality based on the pre-generated captions. Among all the compared methods, the overall model with cross-modal attention yields the best results.

Figure 6 illustrates two examples to visually demonstrate the effectiveness of cross-modal attention for image captioning. Because multimodal attention consists of visual attention and semantic attention, attention weights can be put on visual information, semantic information or their combination to provide more comprehensive hints for the generation of the next word.

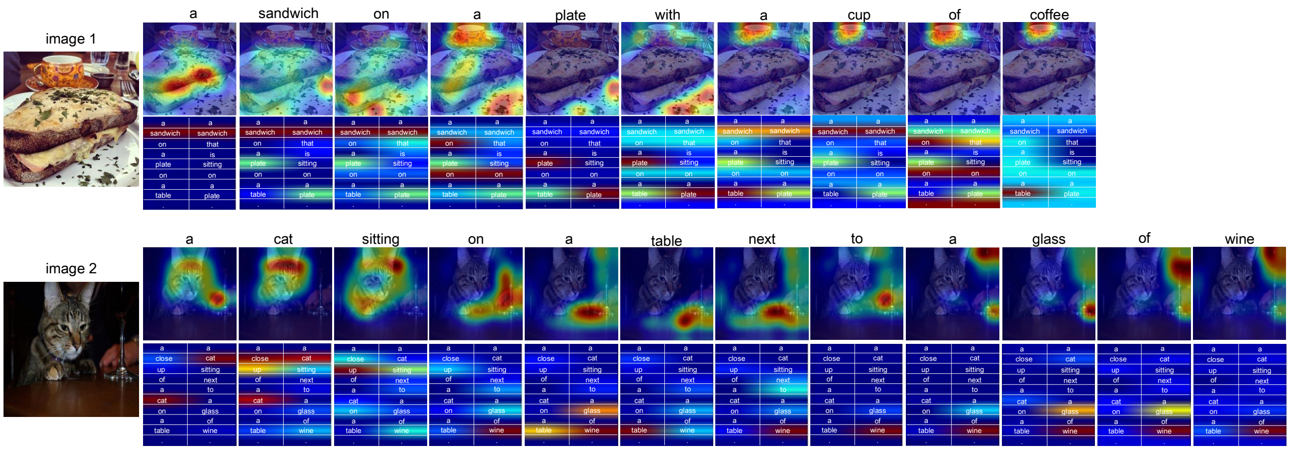
For the first example, the two pre-generated captions do not describe ‘cup’. Therefore, no useful information can be obtained from semantic attention when generating ‘a cup of coffee’. However, our model can pay attention to the image area containing the cup through visual attention. When generating ‘sandwich’, semantic attention focuses on the pre-generated ‘sandwich’ despite that visual attention does not seem to work. To generate ‘plate’, cross-modal attention focuses on the image area containing the plate and the pre-generated ‘plate’ simultaneously. In the second example, one pre-generated caption misses ‘wine glass’ and the other lacks ‘table’. The combination of two sentences contains the complete image contents. In this case, our model turns to focus on the related image regions and semantic words simultaneously so that the model can produce a caption containing the complete image contents.

5. Conclusion

Inspired from the natural expression abilities of human beings in picture descriptions, the paper proposes a mutual-aid network structure with bidirectional LSTMs for acquiring overall contextual information. By designing an auxiliary structure, MaBi-LSTMs make full use of context infor-

Images	Pairs of sentences generated by bidirectional LSTMs (without and with the auxiliary structure)	Ground truth
	A man walking down a street next to a bus. A couple of sheep walking down a street. A man walking down a street next to **sheep**. A herd of sheep walking down a street.	1) A man walks while a large number of sheep follow. 2) A man leads a large herd of sheep through town. 3) A man leading a herd of sheep down the sheep. 4) The man is walking a herd of sheep on the road through a town. 5) A man is walking a herd of sheep down a street.
	A row of white urinals in a public restroom. A row of sinks in a bathroom. A bathroom with **three sinks** and **a mirror**. A row of sinks in a bathroom.	1) This is a public bathroom with soap dispensers installed on the wall. 2) Three white sinks in a bathroom under mirrors. 3) Three sinks and a mirror in a public restroom. 4) There is a bathroom with sinks and soap dispensers. 5) There are many sinks in this public bathroom.
	A man riding a skateboard down a street. A man sitting on top of an airplane. A **green and white** plane **on display**. A man that is sitting on top of a plane.	1) That aircraft is for display not for riding. 2) A very nice looking plane on display in a big room. 3) The old single engine plane is on display under lights. 4) A small air plane on a stage on display. 5) A propeller plane sits parked inside of a building.
	A tennis player in action on the court. A woman holding a tennis racquet on a tennis court. A woman in a **pink** shirt is playing tennis. A woman holding a tennis racquet on a tennis court.	1) A woman is getting ready to strike a tennis ball. 2) A woman holding a tennis racket reaching up to catch the ball with her hand. 3) A woman in a pink dress is about to serve a tennis ball. 4) A woman that is standing on a tennis court with a racquet. 5) A woman throwing a tennis ball in the air and a racket it her other hand.

Figure 5. Pairs of captions generated by the proposed model with and without the auxiliary structure. It can be observed that MaBi-LSTMs can produce more detailed and accurate words. Those words are marked in a bold red font, such as 'green and white', 'pink', 'three', etc.



The ground truth for image 1:
1) A close up of food on a plate on a table.
2) A plate of bread with cheese on top and a fancy orange china mug beside it.
3) A plate with some meat sitting on top of it.
4) A couple of cheese covered pieces of meat on a plate.
5) A sandwich on a plate and a cup and saucer on a table.

The ground truth for image 2:
1) A cat leaning on top of a wooden table.
2) A cat with its paws on a table near a glass of wine.
3) A cat with its paws on the table next to a glass of wine.
4) A cat with its front paws on the table.
5) A cat sits with its paws on a table next to a glass of wine.

Figure 6. Cross-modal attention visualization. The visualization shows the dynamic changes of attention weights while generating captions. The first and second rows of each example visualize the attention weights on images and the pre-generated sentences. The left and right columns in the second row represent the two pre-generated sentences. The redder the greater weight, the bluer the smaller weight.

mation by mutually constructing hidden states. To the best of our knowledge, this is the first time to explore the mutual-aid structure of bidirectional LSTMs. Moreover, this paper introduces a cross-modal attention mechanism by combining visual attention and semantic attention for bridging the gap between cross-domain models. The experimental results show that the proposed algorithm has the competitive performance for image captioning. Importantly, our interests are placed on exploring the human-like cognitive style for image captioning. The methodology can be integrated into other deep learning models and applied to machine translation as well as vision question answering.

Acknowledgments. This work is supported by the National Key R&D Program of China (2018YFB1600600), the National Natural Science Foundation of China (61572104), and the Dalian Science and Technology Innovation Fund (2019J12GX035).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6077–6086, 2018.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5561–5570, 2018.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *The Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005.

[5] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1345–1353, 2018.

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6298–6306, 2017.

[7] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo. Knowledge-embedded representation learning for fine-grained image recognition. *arXiv preprint arXiv:1807.00505*, 2018.

[8] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7995–8003, 2018.

[9] Bo Dai, Deming Ye, and Dahua Lin. Rethinking the form of latent states in image captioning. In *European Conference on Computer Vision, ECCV*, pages 282–298, 2018.

[10] Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho. Learning to parse and translate improves neural machine translation. In *Annual Meeting of the Association for Computational Linguistics, ACL*, volume 2, pages 72–78, 2017.

[11] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1473–1482, 2015.

[12] Carlos Flick. Rouge: A package for automatic evaluation of summaries. In *The Workshop on Text Summarization Branches Out*, page 10, 2004.

[13] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2321–2334, 2017.

[14] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5630–5639, 2017.

[15] Ankush Gupta, Yashaswi Verma, CV Jawahar, et al. Choosing linguistics over vision to describe images. In *AAAI Conference on Artificial Intelligence, AAAI*, 2012.

[16] Xu Jia, 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, CVPR*, pages 2407–2415, 2016.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3128–3137, 2015.

[18] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1601–1608, 2011.

[19] Polina Kuznetsova, Vicente Ordonez, Tamara Berg, and Yejin Choi. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association of Computational Linguistics, ACL*, 2(1):351–362, 2014.

[20] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. Image caption with global-local attention. In *AAAI Conference on Artificial Intelligence, AAAI*, pages 4133–4139, 2017.

[21] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision, ECCV*, volume 8693, pages 740–755, 2014.

[22] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 375–383, 2017.

[23] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7219–7228, 2018.

[24] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[26] Vicente Ordonez, Xufeng Han, Polina Kuznetsova, Girish Kulkarni, Margaret Mitchell, Kota Yamaguchi, Karl Stratos, Amit Goyal, Jesse Dodge, Alyssa Mensch, et al. Large scale retrieval and generation of image descriptions. *International Journal of Computer Vision*, 119(1):46–59, 2016.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting on Association for Computational Linguistics, ACL*, pages 311–318, 2002.

[28] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *IEEE International Conference on Computer Vision, ICCV*, pages 1251–1259, 2017.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149, 2017.

[30] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection–snip. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3578–3587, 2018.

[31] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence, AAAI*, volume 4, page 12, 2017.

[32] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. Paying attention to descriptions generated by image captioning models. In *IEEE International Conference on Computer Vision, ICCV*, pages 2506–2515, 2017.

[33] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[34] Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. Efficient image annotation for automatic sentence generation. In *ACM International Conference on Multimedia*, pages 549–558, 2012.

[35] Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, and Tatsuya Harada. Common subspace for model and similarity: Phrase learning for caption generation from images. In *IEEE International Conference on Computer Vision, CVPR*, pages 2668–2676, 2015.

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4566–4575, 2015.

[37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3156–3164, 2015.

[38] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *ACM on Multimedia Conference*, pages 988–997. ACM, 2016.

[39] Cheng Wang, Haojin Yang, and Christoph Meinel. Image captioning with deep bidirectional lstms and multitask learning. *ACM Transactions on Multimedia Computing, Communications, and Applications, TOMM*, 14(2s):40, 2018.

[40] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 7378–7387, 2017.

[41] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 464–472, 2017.

[42] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 203–212, 2016.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning, ICML*, pages 2048–2057, 2015.

[44] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems, NIPS*, pages 2361–2369, 2016.

[45] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*, pages 22–29, 2017.

[46] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 4651–4659, 2016.