IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 10, 2020, accepted May 21, 2020, date of publication June 3, 2020, date of current version June 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2999568

Cross-Lingual Image Caption Generation Based on Visual Attention Model

BIN WANG[1], CUNGANG WANG[2], QIAN ZHANG[1], YING SU[1], YANG WANG[1], AND YANYAN XU[3]

[1]College of Information, Mechanical, and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China
[2]School of Computer Science, Liaocheng University, Liaocheng 252000, China
[3]Department of City and Regional Planning, University of California at Berkeley, Berkeley, CA 94720, USA

Corresponding authors: Cungang Wang (wangcungang@lcu-cs.com) and Yanyan Xu (yanyanxu@berkeley.edu)

This work was supported by the Science and Technology Innovation Action Plan Project of Shanghai Science and Technology Commission under Grant 18511104202.

ABSTRACT As an interesting and challenging problem, generating image caption automatically has attracted increasingly attention in natural language processing and computer vision communities. In this paper, we propose an end-to-end deep learning approach for image caption generation. We leverage image feature information at specific location every moment and generate the corresponding caption description through a semantic attention model. The end-to-end framework allows us to introduce an independent recurrent structure as an attention module, derived by calculating the similarity between image feature sequence and semantic word sequence. Additionally, our model is designed to transfer the knowledge representation obtained from the English portion into the Chinese portion to achieve the cross-lingual image captioning. We evaluate the proposed model on the most popular benchmark datasets. We report an improvement of 3.9% over existing state-of-the-art approaches for cross-lingual image captioning on the Flickr8k CN dataset on CIDEr metric. The experimental results demonstrate the effectiveness of our attention model.

INDEX TERMS Image caption generation, attention model, deep learning.

I. INTRODUCTION

Generating image caption automatically has become an active research topic in computer vision community. It involves both computer vision and natural language processing which are two major fields in artificial intelligence [1]. Not only must an image caption generation model be capable of determining which objects are in an image, but it also must be able to express their relationships in natural languages [2], [3]. The determination of presence and relationships of multiple objects and organizing human-like sentences to describe this information is not easy, which makes image caption generation be a challenging task.

Generally, the existing approaches for image caption generation can be mainly categorized into two paradigms: retrieval-based and generation-based. Retrieval-based image captioning methods produce a caption for an input query image by retrieving similar images from the training dataset in accordance with similarity metrics. Similarities are normally calculated between the extracted feature vectors. Then, the caption of the best candidate image is transferred to the input image. Ordonez *et al.* [4] utilize global image descriptors to retrieve images from a web-scale dataset with captions. They then re-rank the retrieved images according to semantic content similarity, and finally choose the caption of the top-ranked image as the caption of the query image. Hodosh *et al.* [5] aim to solve the problem of associating images with sentences drawn from a large and predefined pool of image descriptions. They frame image description as a ranking task. Although the outputs of retrieval-based image captioning approaches are usually grammatically correct and fluent, constraining image descriptions to sentences that have already existed cannot fit new queries well. More seriously, the generated descriptions may even be irrelevant to image contents. As a result, this kind of captioning methods have since fallen out of favor to the nowaday dominant generation-based captioning methods. Generation-based captioning methods directly generate captions from

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Zhou.

images with a learned model, which is usually a deep learning model. Recent work [6]–[11] based on this paradigm use a combination of convolutional neural networks (CNN) to obtain vectorial representation of images and recurrent neural networks (RNN) to decode those representations into natural language sentences. Most of these approaches are motivated by the sequence-to-sequence model in machine translation [6], [12]. For example, in [6], one RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder are jointly trained to maximize the conditional probability of a target sequence given a source sequence. Image caption generation is well suited to the encoder-decoder framework of machine translation, because it is analogous to ''translating'' an image to a sentence [2], [6], [13]. In addition, generation-based captioning models are able to generate novel sentences from never-seen images. They are much more flexible and have proven ability to significantly improve the image caption generation quality.

Visual attention is one of the most important mechanisms in the visual system of primates and humans [14]–[16]. The attention does not compress an entire image into a static representation, but allows for salient features to dynamically come to the forefront as needed. Particularly, people are usually inclined to pay attention on regions and objects more semantically important in an image. Motivated by human visual attention mechanism, methods using attention to guide image caption generation are first proposed in [2]. Their proposed method incorporates a form of attention with two variants: hard attention and soft attention. However, they simply merge the image feature and the hidden output of long short-term memory (LSTM) through a multi-layer perceptron network to generate attention, without considering the similarity between the semantic sentence vector and the image feature vector.

In this paper, we propose a new image caption generation method, leveraging image feature information at specific location every moment. The method generates the corresponding caption description through a semantic attention model. Traditional generation-based image captioning approaches use global image features only in the initialization phase. They cannot associate the semantic information of each word with local image features. When processing long sequences, these methods will fall into a situation that the later generated words more rely on information between sequences, ignoring image feature information. Our proposed approach can attack this issue by means of making full use of image feature information at different locations. In addition, research on image captioning has typically focused on generating caption in English for an input image, as the existing corpora for image caption generation are mostly in English. Few researches concentrate on image captioning in Chinese mainly due to the lack of corpora in Chinese. For the image captioning task, Miyazaki and Shimizu [17] and Lan *et al.* [18] present that training in a cross-lingual learning [19]–[21] manner could lead to a better performance of the model. Cross-lingual learning can be widely applied to natural language processing fields, such as cross-lingual sentiment analysis [19], cross-lingual information retrieval [21], and cross-lingual named entity identification [20]. It could bridge the semantic gap between different languages to some extent, via information complementary and sharing. Based on this, our model is designed to transfer the knowledge representation obtained from the English portion into the Chinese portion. Taking in an image, the proposed approach can simultaneously generate caption both in English and Chinese.

The contributions of this paper are summarized as follows:

(1) We propose a novel image captioning model via introducing a new attention mechanism from the perspective of vector similarity. The attention module is based on a recurrent neural network and can separate the process of generating attention from the model and the process of generating captions.

(2) The correlation between image features and semantic word features is used to express attention. In this context, the proposed model more likely focuses on the informative locations, where the image feature information has greater correlation with the descriptive information. Our attention calculation can intuitively reflect the relationship between images and caption words.

(3) Few researches on image captioning have focused on generating captions in Chinese. We create a simple yet effective image caption generation model for Chinese which benefit from exploiting the English portion of the corpus. Experimental results validate its effectiveness, indicating that a poor-resource language can benefit from a rich-resource language in image captioning.

II. RELATED WORK

In this section, we briefly present a relevance review on image caption generation and attention. As aforementioned, methods for image caption generation can be roughly categorized into two classes: retrieval-based and generation-based. Retrieval-based image captioning approaches firstly retrieve similar images from a large captioned dataset, and then modify the retrieved captions to fit the query image. These approaches typically involve an intermediate step to remove the caption specifics which are only relevant to the retrieved image [2], [22]–[24]. Generally speaking, their outputs are grammatically correct. However, the prediction ability is limited for the new queries which captions are not presented in the training sets. The generated descriptions may even be irrelevant to image contents. Retrieval-based image captioning approaches have gradually be replaced by generation-based methods.

Many of the generation-based image captioning approaches are inspired by recent advances in recurrent neural networks and the successful use of sequence to sequence training in machine translation to neural networks [25], [26]. Kiros *et al.* [9] firstly leverage deep neural networks to generate image caption. They use an image-text multimodal neural language model to jointly learn word representations

and image features by training the model together with a convolutional network. Shuang and Shan [3] present a multimodal recurrent neural network model for generating novel image captions, which directly estimates the probability distribution of generating a word given previous words and an image. The model is comprised of two subnetworks: a deep recurrent neural network for sentences and a deep convolutional network for images. The two subnetworks interact with each other in a multimodal layer to form the whole m-RNN model. Vinyals *et al.* [27] and Donahue *et al.* [28] develop a novel end-to-end recurrent convolutional architecture for large-scale visual learning. Such models do well when target concepts are complex or training dataset are limited. Pan *et al.* [29] incorporate the transferred semantic attributes learned from images and videos into the CNN plus RNN framework for video caption. Wu *et al.* [30] directly employ high-level attributes to guide the language model, and advance CIDEr and BLEU metrics in image caption. Guiguang *et al.* [31] propose an encoder-decoder framework called Reference based LSTM (R-LSTM). This model generates a more descriptive word sequence for an input image via introducing reference information from the neighboring images. Yao *et al.* [32] present LSTM with attributes, constructing variants of architectures by feeding image representations and attributes into RNNs in different ways. Liu *et al.* [33] argue that using policy gradient optimization can directly optimize for the interested metrics of the image captioning model. Other work in this category include [34]–[36]. Generation-based image captioning approaches are much more flexible for their capability of generating novel semantic word sequences from never-seen images. Our proposed approach belongs to the generation-based category. Unlike traditional generation-based image captioning approaches, our model can associate the semantic information of each word with local image features via introducing a novel attention mechanism.

Incorporating attention into neural networks is recently studied in computer vision and related area [1]. By applying attention mechanism into face recognition, Tang *et al.* [37] can robustly attend to face regions of novel test subjects upon learning images of faces. Xu *et al.* [2] incorporate a form of attention with two variants: a hard attention mechanism and a soft attention mechanism. They validate the usefulness of attention in caption generation with the state of the art performance. You *et al.* [1] propose more complex attention mechanism for image captioning. The aforementioned models generate attention without considering the correlation between image feature vector and semantic word sequence. Unlike these methods, we propose a novel image captioning method by means of introducing a new attention mechanism based on a recurrent neural network. The attention module considers the correlation between image feature sequence and semantic word sequence as the attention. Our attention-based image captioning model automatically chooses image feature information at different locations. The model generates the corresponding caption description via fully considering the correlation between image feature vector and semantic word sequence.

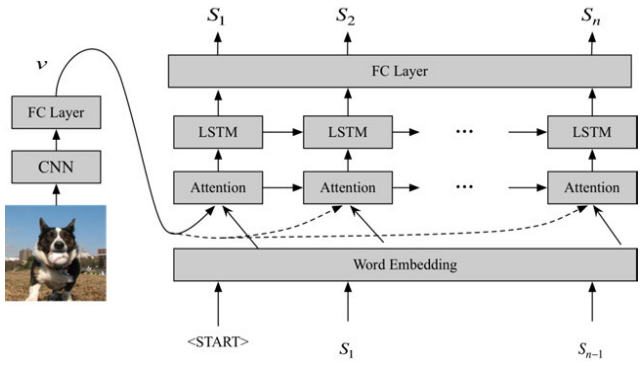


FIGURE 1. Model overview.

Besides, generating image caption is mostly in English, as most of the available datasets are in this language [31], [34]. Only few studies have been conducted on cross-lingual image captioning [17], [38], [39]. In this paper, the model is designed to perform cross-lingual image caption. We exploit the English corpus to improve the performance of image caption in Chinese. Taking in an image, the proposed approach can simultaneously generate caption both in English and Chinese.

III. PROPOSED APPROACH

In this section, we present our image captioning model with attention mechanism in three aspects. First, we show an overview of our model. Second, we describe the part of image features extraction. Finally, at the caption generation stage, we add an independent recurrent structure to our model as attention module.

A. MODEL OVERVIEW

Currently, most models applied to image caption tasks are inspired by the approach of Vinyals *et al.*, which connects CNN and RNN through a fully connected method [30]. The main advantage of this method is that it can use end-to-end training, thus avoiding the deficiencies of the model being divided into image preprocessing, feature extraction and caption generation. Following Vinyals' approach, we propose a new discriminative model based on attention mechanism for image caption generation. Fig.1 shows an overview of our model.

The framework shown in Fig.1 can be logically divided into two parts. Image feature extraction is implemented in the left part through a pre-trained convolutional neural network (introduced in Section B). Then, a recurrent structure with attention mechanism is used in right part for language modeling and generation. This part will be described in Section C in detail.

For a novel image I and its correct transcription S which is a sequence of N words $\{S_0, S_1, S_2, \ldots, S_N\}$, our model can be formalized as the maximum probability of the description

S given the image I according to Maximum Likelihood Criterion. The likelihood function can be formulated as

$$likelihood = \sum_{(S,I)} \sum_{t=0}^{N} logP\left(S_t \mid I, S_1, \ldots, S_{t-1}; \theta\right) \tag{1}$$

where θ represents parameters of the model, S_0 =<START> and S_N =<END> are two special tokens indicating the beginning and the end of the sentence S. Note that the first summation is over pairs of an image I and its description S in the training set and the second summation is over all word S_t in S. Applying the Bayes formula, the result of second summation represents the log probability of the sequence S under the condition of image I.

The parameters θ of the model can be obtained by solving the following optimization problems

$$\theta^* = \arg\min_{\theta} \sum_{(S,I)} \sum_{t=0}^{N} -logP(S_t|I, S_0, S_1, \ldots, S_{t-1}; \theta) + \lambda \|\theta\|_2 \tag{2}$$

where first term of the formula is log-likelihood, second term is regularization, λ is coefficient of regularization.

B. IMAGE FEATURE EXTRACTION WITH CONVOLUTIONAL NEURAL NETWORKS

In recent years, with the increase of computing capability, deep CNN models have made tremendous progress in various computer vision tasks. A well-known practical advantage of these models is that once they are trained on a dataset, they can be viewed as a fixed feature extractor. However, since the feature extracted from CNN's fully connected layer is usually one-dimensional vector, it only contains the globally semantic information of image and loses the positional information between objects within the image. In order to associate each feature vector with the content of two-dimensional image, we extract the feature from the pre-trained convolutional neural network' convolutional layer rather than the last fully connected layer. These features containing location information allow the model to selectively use the objects of certain locations in image to generate captions correspondingly. For each image, the feature map after multilayer convolutional network is shown in Fig.2.

Let the size of feature map be $H \times W$ and there are totally C channels, that is $CNN(I) \in R^{H \times W \times C}$. $v(i,j)$. $v(i,j)$ is defined as feature vector of this feature map at position (i,j), where $i = 1, \ldots, H$, $j = 1, \ldots, W$, $v\,(i,j) \in \hat{E}R^C$. The extractor produces $L = H \cdot W$ vectors totally, and each vector which semantically represents a certain part of image has dimension C.

$$v = [v_1, \ldots, v_L]^T, \quad v_i \in R^C \tag{3}$$

With $v(i)$ and $v(i,j)$ united, the feature map meshes the input image horizontally H and vertically W. And each position in feature map can be considered to have a one-to-one correspondence with each grid in the image.

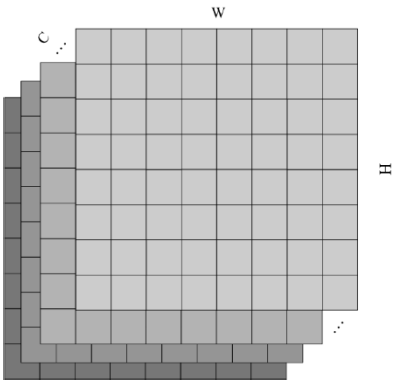


FIGURE 2. Feature map after multilayer convolutional network.

C. LANGUAGE MODELING AND GENERATION WITH ATTENTION MECHANISM

To some extent, attention model is a resource allocation model. When people are observing an image, they can generally see its entire appearance. However, the eyes of a person are often focused on a very small part when they want to observe the special contents of an image. For image caption generation, attention model is allowed to "observe" the proper position in the image at each moment, and can generate description that matches the observation content. In this paper, we propose a new attention model which uses a recurrent structure separated from the original LSTM network to obtain semantic vectors. The degree of attention is calculated by the similarity between the semantic vectors and the image feature vectors.

For the 2-dimensional image feature matrix $v = [v_1, \ldots, v_L]^T$ extracted by CNN in the previous section, we define weight $\{\omega_i\}$, $i \in [1, L]$ that can be viewed as the probability of position i of the image observed by the model. The larger weight means that the model more likely focuses on this position. According to the definition of probability, ω_i should satisfy the following formula:

$$\omega_i > 0, \quad i = 1, \ldots, L \tag{4}$$

$$\sum_{i=1}^{L} \omega_i = 1 \tag{5}$$

LSTM networks utilize the information from previous moments when generating caption at this time, and have been widely used for common caption generation models. For suitable attentions, they should be the same with caption words in number and should also consider context information in the generation process. Motivated by this inspiration, we use a similar recurrent structure for producing attentions. The attention module structure is shown in Fig.3.

Assume that u_{t-1} represents the hidden layer output of recurrent part at previous time, and S_t represents the input word encoded by one-hot at current time. Then the semantic vector u_t is calculated through inner-RNN following the

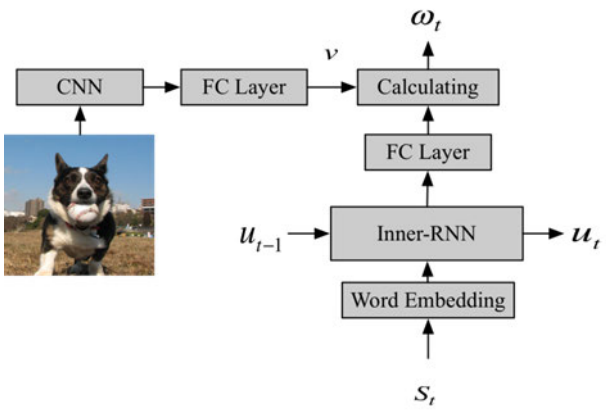


FIGURE 3. Attention module.

formulation below:

$$z_t = W_{embed} S_t \tag{6}$$
$$u_{-1} = [0, \ldots, 0]^T \tag{7}$$
$$u_t = \tanh(W_{uu} u_{t-1} + W_{zu} z_t + b_u) \tag{8}$$

where W_{embed} is an embedding matrix, u_{-1} is initialized by zero vector, W_{uu}, W_{zu} and b_u are learning weight matrices and biases.

Once we get the semantic vector u_t and image feature vectors v, we compute the model's attention at current time. However, u_t and v_i usually have different dimensions. We convert u_t and v_i to the same dimension through two different full connection layers, so that we calculate the similarity between them.

$$v'_i = W_v v_i + b_v \tag{9}$$
$$u'_t = W_u u_t + b_u \tag{10}$$

where W_v, W_u, b_v and $b_u$are learning parameters.

According to the definition of attention, a suitable measure for attention is the similarity between the semantic vector u_t and the image feature vector v_i. The larger the similarity, the greater the correlation between the descriptive information and the image feature information of the position. In other words, the model more likely focuses on the position. In natural language processing field, Cosine is the most popular one among the existing measures [40]–[42]. In mathematics perspective, Cosine similarity is perfect and can be a reasonable approximation. Inspired by this, in this paper, cosine distance is used to measure the model's attention.

$$a_{ti} = sim\left(v'_i, u'_t\right) = \frac{v_i'^T u'_t}{\left\|v'_i\right\|^2 \left\|u'_t\right\|^2} \tag{11}$$

Then we use *softmax* function to make a_{ti} satisfy that the addition equals to 1.

$$\omega_{ti} = \text{softmax}\left(a_{ti}\right) = \frac{a_{ti}}{\sum_{j=1}^{L} e^{a_{tj}}} \tag{12}$$

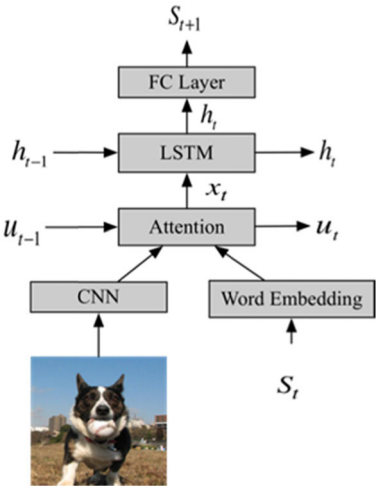


FIGURE 4. Language modeling and generation module.

With the obtained attention weight ω_{ti}, the image feature with attention information φ_t can be easily calculated by the following formula:

$$\varphi_t = \sum_{i=1}^{L} v_i \omega_{ti} \tag{13}$$

Since vector φ_t is related to time t, the input of attention model is not only word embedding vector $W_{embed} S_t$ but the combination of vector φ_t and $W_{embed} S_t$.

At every time step t, the process of language modeling and generation is shown in Fig.4. In Fig.4, x_t is a combination of vector φ_t and $W_{embed} S_t$, h_t and h_{t-1} are hidden layer output of LSTM unit at time step t and $t-1$. LSTM can address the vanishing and exploding gradients problem and handle longer dependencies well. A LSTM unit has a memory cell and various gates to control the input, the output, and the memory behaviors. At each time step t, x_t and the LSTM state c_t, h_t is as follows:

$$x_t = [W_\varphi \varphi_t + b_\varphi, W_{embed} S_t] \tag{14}$$
$$i_t = \sigma\left(W_{ix} x_t + W_{ih} h_{t-1} + b_i\right) \tag{15}$$
$$= \sigma\left(W_{fx} x_t + W_{fh} h_{t-1} + b_f\right) \tag{16}$$
$$= \sigma\left(W_{ox} x_t + W_{oh} h_{t-1} + b_o\right) \tag{17}$$
$$g_t = \tanh\left(W_{gx} x_t + W_{gh} h_{t-1} + b_g\right) \tag{18}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{19}$$
$$h_t = o_t \odot \tanh(c_t) \tag{20}$$

where i_t, f_t, c_t, o_t, g_t are the input gate, forget gate, memory state, output gate, input modulation gate of LSTM respectively, $\sigma(x) = 1/(1+e^{-x})$ is a sigmoid function, $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$ is a hyperbolic tangent function, $\odot$ denotes the elements-wise product of two vector, all of W and b are parameters for learning. Finally, the probability distribution of words can be calculated by a softmax classifier, i.e. $\text{softmax}(W_h h_t + b_h)$.

At the training stage, since the model is differentiable everywhere, the gradients of the parameters can be obtained

TABLE 1. Statistics of Flickr8K and Flickr8K CN datasets.

	Flickr8K	Flickr8K CN
Total words	7827	4739
Max length	36	15
Average length	9.87	4.95
k	5	2
Vocabulary size	2523	2805

by BPTT [43] algorithm. And at the test stage, we use Beam-Search algorithm [44] to generate captions as sequence of words.

IV. EVALUATION

A. EXPERIMENTS ON CROSS-LINGUAL IMAGE CAPTION GENERATION

1) DATASETS AND EVALUATION METRICS

In order to generate captions both in Chinese and English for an input image, we choose Flickr8K dataset and its Chinese version Flickr8K CN, which both have 5 reference sentences per image. Following most common works, we use 6000 images and 30000 captions as training set to train our model, 1000 images as validation set to tune hyper-parameters, and another 1000 images as testing set to report results. Notice that English sentences have spaces as explicit word boundary markers, but Chinese sentences do not have this characteristic. Aiming to split Chinese captions into some meaningful words as a sequence, we use common word segmentation software for Chinese text segmentation. By analyzing sentences contained in each dataset statistically, we obtain total words, max and average length of sentences. The details can be found in TABLE 1. We can easily find that the average length of English sentences is about twice larger than that of Chinese sentences. After that we discard some words that occur less than k times in the training set, which results in vocabularies with the size of 2523 and 2805 respectively.

The common metrics for evaluating the quality of image caption generation come from translation systems. In this paper, we use standard metrics, including BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L and CIDEr, to evaluate the generated Chinese sentences. We use BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR and CIDER to evaluate the quality of generated English sentences. BLEU [45], [46] was originally designed for automatic machine translation and now is widely used to evaluate image captioning. By counting n-gram co-occurrences, it rates the quality of a translated sentence given several reference sentences. ROUGE [47], [48] is an evaluation metric designed by adapting BLEU to evaluate automatic text summarization algorithms. ROUGE is based on the longest common subsequences instead of n-grams. METEOR [49] adds recall rate to remedy the fact that BLEU only considers the precision. CIDEr [50] metric measures consensus in image captions by performing a term-frequency inverse document frequency (TF-IDF) weighting for each n-gram. For all the evaluation metrics, higher scores mean better performances.

TABLE 2. Configuration of best model.

	Parameter	Value
Word Embeddings	dimensionality	256
LSTM	units	512
Dropout	probability	0.5
Attention	inner-RNN hidden layer	128
	dimensionality of image and semantic vector mapping	512
Optimizer	method	RMSprop
	learning rate	0.0005
	decay rate	0.99
	gradients clipping	5
	batch size	100

2) TRAINING

To extract image features, we employ inception-v4 version of GoogleNet pre-trained on ImageNet dataset. Then, LSTM networks is used in language modeling and generation phase, and the length of the LSTM is determined by the longest sentence. As a regularization, we also use dropout in all layers included in LSTM networks. Furthermore, early epoch stopping with waiting window of 5 epochs without improvements is applied in every experiment. We fine-tune the set of hyper-parameters by using grid search, and the final configuration of our model is shown in TABLE 2. Besides, we compare several optimization algorithms for training, including Stochastic Gradient Descent (SGD) [51], Adagrad [52], RMSprop [53], and Adam [54]. We find that Adam makes model converge fastest, but RMSprop yields a gain of above 2% than it on CIDEr metric.

3) CROSS-LINGUAL LEARNING

Cross-lingual learning is a training mode involving transferring information across different languages. For each language, we design three learning schemes including *monolingual learning*, *alternate learning*, and *transfer learning* following Miyazaki *et al.* [17] to evaluate the capability that our model can utilize cross information between different languages to get better performance.

In terms of generating English, the model trained by monolingual learning scheme only has one body for language modeling, and only the English dataset Flickr8K is used. The alternate learning demands that the model owns two separated bodies for language modeling, one for English and the other for Chinese, but the model shares the same parameters W_v and b_v in the training stage. For transfer learning scheme, the model is pre-trained completely on Chinese dataset Flickr8K CN. Then, the trained body for Chinese caption generation is removed, and the other body is added for English caption generation. The parameters W_v and b_v

TABLE 3. Learning scheme setting for each language.

Generated Language	Learning scheme	Pre-training set	Training set	No. of bodies
Chinese	Monolingual	—	Flickr8K CN	1
	Alternate learning	—	Flickr8K, Flickr8K CN	2
	Transfer learning	Flickr8K	Flickr8K CN	2
English	Monolingual	—	Flickr8K	1
	Alternate learning	—	Flickr8K, Flickr8K CN	2
	Transfer learning	Flickr8K CN	Flickr8K	2

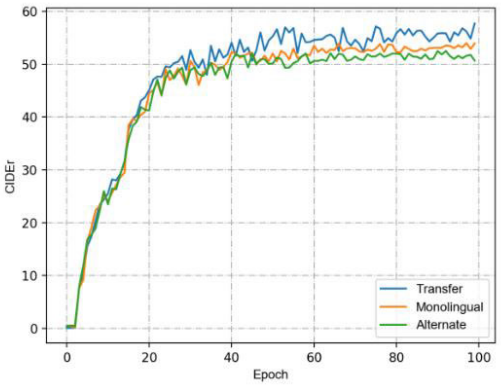


FIGURE 5. Learning curve represented by CIDEr on English caption generation.

are transferred from Chinese training to English training. The main setting of the learning scheme for each language is given in TABLE 3. The Flickr8K dataset is not particularly large. We used the 2.2 GHZ Quad-Core Intel Core i7 processor to train our model. It took about 8 hours from training to convergence for the Chinese caption generation and 11 hours for the English caption generation.

We plot the learning curves represented by CIDEr scores for English and Chinese caption generation in Fig. 5 and 6 respectively. We find that the model applied by transfer learning scheme shows better performance than the other two models both in English and Chinese. This indicates that information contained in image feature can be transferred from one language to the other, and it can help improve the model performance. Particularly, we observe that it has been greatly improved in Chinese caption generation than English by adopting transfer learning scheme, outperforming the monolingual learning approximately by 4%. As can be seen in section A, the length of Chinese sentences after word segmentation is in the range 4 to 7, but the length of English sentences is roughly two times longer. This is the reason why the performance of model for generating Chinese captions is greatly improved by using English dataset to pre-train it. However, we find that alternate learning scheme obtains the lowest evolution score among them, which suggests that there is a semantic gap between different languages.

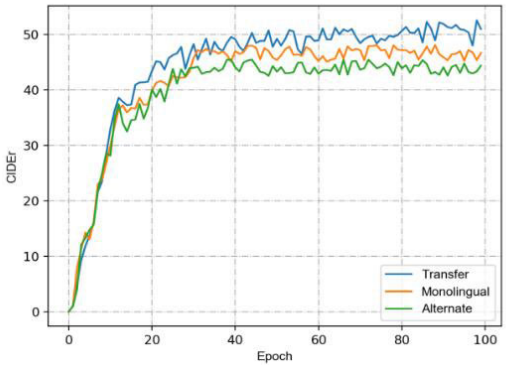


FIGURE 6. Learning curve represented by CIDEr on Chinese caption generation.

Using different languages dataset to train model alternately will lead to information conflict, and affect the model performance.

4) COMPARISON WITH RELATED WORK

We compare our proposed attention-based model with several state-of-the-art models for Chinese and English caption generation, such as CS-NIC [55], Google NIC [27], Fluency-guided Model [56], Soft-Attention [2], Hard-Attention [2], Log Bilinear [9], Multimodal RNN [34]. The baseline model uses the image feature extracted from fully connection layer of GoogleNet, and has only one LSTM for caption generation. CS-NIC [55] is a method adding Chinese caption to images. Google NIC [27] is based on a deep recurrent architecture which can generate natural sentences to describe an image. Fluency-guided model [56] proposes a learning scheme aiming to conquer the lack of fluency in the translated sentences. The most similar method to our model is Soft-Attention and Hard-Attention proposed by Xu *et al.* [2]. Different from our model where the degree of attention is calculated through the similarity between semantic vector and image feature vector, they use multi-layer fully connected networks as attention producer. Log Bilinear [9] proposes multimodal neural language models in the context of image-text learning how to jointly learn word representations and image features. Multimodal RNN [34] generates natural language descriptions of images and their regions via leveraging image datasets and their sentence descriptions to learn about the inter-modal correspondences between language and visual data.

The experimental results are shown in TABLE 4 and 5 respectively. All results of the proposed model both in Chinese and English are obtained by transfer learning scheme mentioned in previous experiment. As shown in Table 4, our proposed model outperforms the other compared methods in terms of BLEU-1,2,3,4, ROUGE-L and CIDEr for Chinese caption generation. Considering English caption generation, our proposed model significantly outperforms Baseline, Log Bilinear, multimodal RNN, Google NIC. Meanwhile, the performance of our model is higher than Soft-Attention and Hard-Attention in terms of every metric except for BLEU-1.

TABLE 4. Performance comparison for Chinese caption generation on Flickr8k CN dataset.

Generated Language: Chinese						
Model	Metric					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
Baseline	60.5	43.3	32.5	22.8	48.98	39.72
CS-NIC[55]	61.1	41.6	21.3	—	—	—
Google NIC[27]	61.8	44.8	32.6	23.7	48.62	41.90
Fluency-guided[56]	—	—	—	24.1	45.9	47.6
Soft-Attention[2]	62.9	47.3	35.9	25.9	51.82	45.45
Hard-Attention[2]	62.8	47.7	36.2	26.1	52.60	46.36
Ours	**63.7**	**49.4**	**37.2**	**28.7**	**53.34**	**51.45**

TABLE 5. Performance Comparison for English caption generation on Flickr8k dataset.

Generated Language: English						
Model	Metric					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
Baseline	57.4	39.5	27.0	17.2	18.07	41.72
Log Bilinear[9]	65.6	42.4	27.7	17.7	17.31	—
Multimodal RNN[34]	57.9	38.3	24.5	16.0	16.70	—
Google NIC[27]	63	41	27	—	—	—
Soft-Attention[2]	**67.0**	44.8	29.9	19.5	18.93	—
Hard-Attention[2]	**67.0**	45.7	31.4	21.3	20.30	—
Ours	66.8	**46.8**	**32.2**	**22.1**	**20.36**	**55.12**

FIGURE 7. Image caption generation example A.

The results demonstrate that the similarity between the semantic vector and image feature vector is a suitable measure of the model's attention. It also presents a reasonable explanation that why the model focuses on these special positions. In this sense, the attention obtained by calculating this similarity is more convincing than obtained through a multi-layer fully connected network.

5) QUALITATIVE ANALYSIS

Fig. 7 and Fig. 8 show some captions generated by our model. As transfer learning is used in our experiment, our model can generate captions both in Chinese and English for the same image. For the three test images shown in Fig. 7, the generated captions properly describe the girl'' and ''小女孩''(the corresponding Chinese caption), ''smiling'' and ''微笑''(the corresponding Chinese caption) in

FIGURE 8. Image caption generation example B.

FIGURE 9. Image caption generation example C.

the third image. Even more, we can see that some captions generated by our proposed model are better than the ground truth captions in details. For instance, the prediction caption of the first image in Fig. 8 contains ''brown dog'', which does not appear in its corresponding ground truth. This is not surprising because our approach can learn this description from the other similar images. To further validate the effectiveness of our approach on cross-lingual image caption generation, we present another two examples which are shown in Fig. 9 and Fig. 10 respectively. We can see that, compared with Fig. 7 and Fig. 8. the contents of images in Fig. 9 and Fig. 10 are more complicated, which makes the task much more challenging. Nevertheless, the captions generated for these images also properly describe which objects are in an image and what is happening in it. For example, the generated captions both in Chinese and English for the second image in Fig.9 are more precise and detailed than its ground truth captions. Additionally, by observing a large number of captions generated by our model, we find that the model has a high recognition rate for objects such as ''dog'', ''骑自行车''(ride a bike), ''a man'', and ''run'', which is consistent with high frequency of occurrence of these words in the datasets.

6) ATTENTION DISPLAYING

Since the GoogleNet contains the max pooling layers, the size of the feature map we extracted by it is usually smaller than

FIGURE 10. Image caption generation example D.

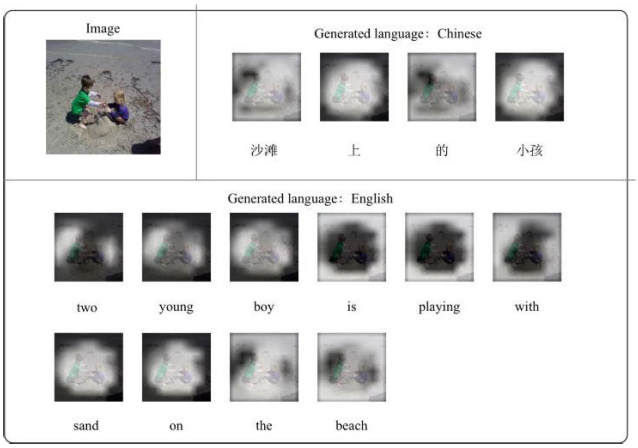


FIGURE 11. Example A of attention display.

the input image. Consequently, the size of attention map is smaller too. In order to visualize the attention weights, we applied bilinear interpolation to resize the attention map as well as the input image. A visualization of attentions for each generated word both in Chinese and English is shown in Fig. 11. As shown in Fig. 11, the white regions indicate that the similarity between the image feature at this position and the word vector feature is high. And these white regions represent where the model pays attention to at each step. For the Chinese caption generation, when generating "沙滩", the model focuses on the background of the image. When generating "小孩", the model focuses on the children in the center of the image, which is the foreground. For the English caption generation, when generating "two young boy", the model focuses on the children in the image. When generating "beach" and "sand", the model focuses on the background part gradually. We can see that, when generating a word containing specific semantic information, the model can pay "attention" to the corresponding image area according to the semantic information of the word, and when generating some words without explicit semantic information, the model sees them as a transitional part, connecting words with a clear semantic meaning. Fig. 12 shows another example of attention visualization.

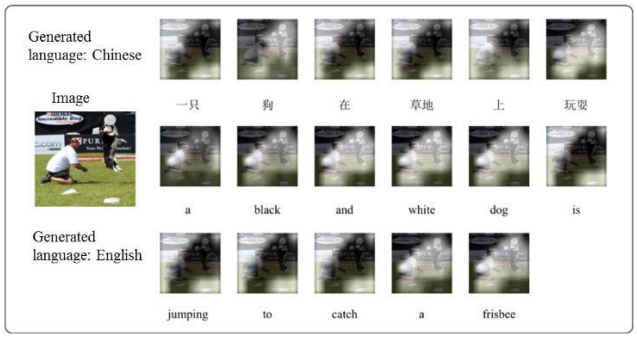


FIGURE 12. Example B of attention display.

In summary, the attentions learned by our model strongly agree with human intuition in different languages. But in some details, the attention of model cannot be described precisely. We argue that it may be due to the fact that the size of attention map is based on the smaller feature map extracted by CNNs rather than the original input image.

B. EXPERIMENTAL RESULTS ON FLICKR30k AND MSCOCO DATASETS

1) DATASETS AND EXPERIMENTAL SETTING

To further validate the effectiveness of our proposed attention model, we perform image caption generation experiments on two larger datasets:

Flickr30k [56] and MSCOCO [34]. Flickr30k dataset is comprised of 31,783 images, and each image is labeled with five captions. Following previous work [1], [2], [56], 1000 images are used for validation, 1000 images for test, and the rest are used for training. MSCOCO dataset contains 123,287 images with five captions per image. We follow the publicly available split as in [1], [2], [34] on MSCOCO dataset. That is, 5000 images are used for validation, 5000 for test, and the rest images are selected for training. Following previous work [57], we preprocess the captions by converting all the caption labels to lower case with basic tokenization and filtering out the labels which occur at least 5 times in the training set [58]. We still use the widely-used captioning metrics to evaluate the performance of image caption on Flick30k and MSCOCO datasets, including BLEU, METOR, and CIDEr. The training details can be found in section A. MS COCO is our largest dataset. Our model took less than 3 days to train on an NVIDIA Titan Black GPU on this dataset.

2) PERFORMANCE ON FLICKR30k DATASET

We compare our proposed model with several state-of-the-art approaches for image caption generation on Flickr30k dataset. The compared methods include Google NIC [27], Soft-Attention [2], Hard-Attention [2], Log Bilinear [9], PoS [59], LRCN [10], and SCA-CNN [60]. PoS [59] aims to exploit the structure information of a natural sentence and discovers the part of speech tags of a sentence are very effective cues to guide the LSTM based word generator. LRCN [10] proposes a novel recurrent convolutional

TABLE 6. Performance comparison for English caption generation on Flickr30k dataset.

Model	Metric				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Log Bilinear[9]	60.0	38	25.4	17.1	16.88
Google NIC[27]	66.3	42.3	27.7	18.3	—
Soft-Attention[2]	66.7	43.4	28.8	19.1	18.49
Hard-Attention[2]	66.9	43.9	29.6	19.9	18.46
PoS[59]	63.8	44.6	30.7	21.1	—
LRCN[10]	58.7	39	25	16.5	
SCA-CNN[60]	66.2	**46.8**	32.5	22.3	19.5
Ours	**67.8**	46.6	**33.5**	**23.1**	**20.54**

TABLE 7. Performance comparison for English caption generation on MSCOCO dataset.

Model	Metric					
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Log Bilinear[9]	70.8	48.9	34.4	24.3	20.03	—
Google NIC[27]	66.6	46.1	32.9	24.6	—	—
Soft-Attention[2]	70.7	49.2	34.4	24.3	23.9	—
Hard-Attention[2]	71.8	50.4	35.7	25.0	23.04	—
PoS[59]	71.1	53.5	38.8	27.9	23.9	88.2
LRCN[10]	66.9	48.9	34.9	24.9	—	—
ATT[1]	70.9	53.7	40.2	30.4	24.3	—
SCA-CNN[60]	71.9	54.8	41.1	31.1	25.0	—
GLA-BEAM3[61]	72.5	55.6	41.7	31.2	24.9	**96.4**
Ours	**73.1**	**55.9**	**43.4**	**32.8**	**25.49**	95.1

architecture which is end-to-end trainable, and can be compositional in spatial and temporal layers. SCA-CNN [60] introduces a novel convolutional neural network dubbed SCA-CNN, which incorporates spatial and channel-wise attentions in a CNN for the task of image captioning. The results of our model as well as other compared approaches are shown in TABLE 6. Our model shows competitive performance among the state-of-the-art approaches on Flickr30K dataset. Specifically, our method achieves the best results on BLEU-1, BLEU-3, BLEU-4, and METEOR, outperforming the recent SCA-CNN model and showing competitive performance with SCA-CNN on BLEU-2. In particular, compared with Soft-Attention [2] and Hard-Attention [2], which is the most similar to our model, we increase both BLEU-2 and BLEU-3 by nearly 3 points, BLEU-4 by more than 3 points, and METEOR by more than 2 points.

3) PERFORMANCE ON MSCOCO DATASET

To further validate the effectiveness of our model on larger dataset for the image captioning task, we present the performance of our model and other state-of-the-art approaches on MSCOCO dataset in TABLE 7. Log Bilinear [9], Google NIC [27], Soft-Attention [2], Hard-Attention [2], PoS [59], LRCN [10], ATT [1], SCA-CNN [60], and GLA-BEAM3 [61] are chosen for comparison. ATT [1] combines

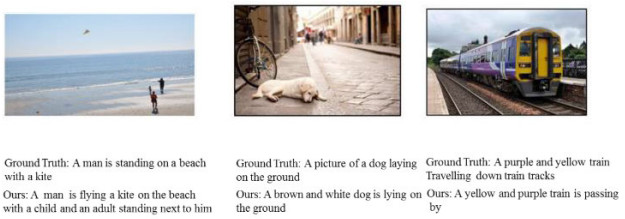


FIGURE 13. Example results for image caption generation using our approach.

both top-down and bottom-up approaches through a semantic attention model. GLA-BEAM3 [61] proposes a global-local attention (GLA) method, integrating local representation at object-level with global representation at image-level through attention mechanism. With more training data, our proposed attention model achieves even better performance. As shown in Table 7, our approach outperforms the other compared state-of-the-art approaches in terms of BLEU 1-4 and METEOR. Although the result on CIDEr of our model is a little poorer than GLA-BEAM3, we show significant superiority on BLEU 1-4 and METEOR. For example, in terms of BLEU-3, our model achieves nearly 2 points improvement. Compared with GLA-BEAM3, our method is still very competitive. Moreover, comparing to Soft-Attention and Hard-Attention model, our model shows significant improvement. Overall, our approach consistently outperforms the other compared state-of-the-art approaches on almost all the metrics. This observation further demonstrates the effectiveness and generalization capability of our model. Fig. 13 presents some caption generation results using our approach.

V. CONCLUSION

In this paper, we introduce a deep learning model with attention mechanism which involves an independent recurrent structure. The attention is measured by similarity between semantic word vector and image feature vector. The similarities can explain rationally why the model focuses on these special positions when generating caption at each step. To perform cross-lingual image caption generation, we apply three learning schemes to train our model on both Chinese and English datasets. The experimental results show that transfer learning scheme could transfer some semantic information from one language to the other. Compared with related work, our model trained by transfer learning scheme obtains the state of the art performance on both Chinese and English datasets. Experiments on both Flickr30K and MSCOCO further demonstrate the effectiveness of our attention-based model. Our method can be used under many practical conditions, such as intelligent image annotation and human-computer interaction. Since the proposed attention model is applied to the feature map after convolutional network, it may cause some details to be lost in the image. For next steps, we plan to design more complex attention-generation mechanism, which can simultaneously leverage global and

local detail information of an image, leading to generating more accurate and fluent captions.

ACKNOWLEDGMENT

The authors thank Jianfeng Wang for his contributions to this project.

REFERENCES

[1] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, vol. 2015, Feb. 2015, pp. 2048–2057.

[3] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018.

[4] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[5] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, Aug. 2013.

[6] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1–15.

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: https://arxiv.org/abs/1409.0473

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 595–603.

[10] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[11] P. Anderson, B. Fernando, and M. Johnson, "SPICE: Semantic propositional image caption evaluation," *Adapt. Behav.*, vol. 11, no. 4, pp. 382–398, 2016.

[12] S. Kuanar, V. Athitsos, N. Pradhan, A. Mishra, and K. R. Rao, "Cognitive analysis of working memory load from eeg, by a deep recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2576–2580.

[13] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*. [Online]. Available: https://arxiv.org/abs/1609.08144

[14] J. Liang, L. Jiang, L. Cao, Y. Kalantidis, L.-J. Li, and A. G. Hauptmann, "Focal visual-text attention for memex question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1893–1908, Aug. 2019.

[15] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.

[16] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, "Decoupled novel object captioner," in *Proc. ACM Multimedia Conf.*, 2018, pp. 1029–1037.

[17] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in *Proc. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1780–1790.

[18] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *ACM Multimedia*. New York, NY, USA: ACM, 2017.

[19] Q. Chen, W. Li, Y. Lei, X. Liu, and Y. He, "Learning to adapt credible knowledge in cross-lingual sentiment analysis," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 419–429.

[20] A. Zirikly, "Cross-lingual transfer of named entity recognizers without parallel corpora," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 390–396.

[21] E. Ghanbari and A. Shakery, "Query-dependent learning to rank for cross-lingual information retrieval," *Knowl. Inf. Syst.*, vol. 59, no. 3, pp. 711–743, 2019.

[22] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*. [Online]. Available: https://arxiv.org/abs/1411.2539

[23] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, C. Siming, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[24] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proc. EMNLP*, 2013, pp. 1292–1302.

[25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[26] S. Kuanar, K. R. Rao, D. Mahapatra, and M. Bilas, "Night time haze and glow removal using deep dilated convolutional network," 2019, *arXiv:1902.00855*. https://arxiv.org/abs/1902.00855

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[28] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[29] Y. Pan, T. Yao, H. Li, and T. Mei, "Video captioning with transferred semantic attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6504–6512.

[30] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.

[31] D. Guiguang, C. Minghai, Z. Sicheng, H. Chen, J. Han, and Q. Liu, "Neural image caption generation with weighted training and reference," *Cogn. Comput.*, vol. 11, pp. 763–777, Aug. 2018.

[32] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," 2016, *arXiv:1611.01646*. [Online]. Available: http://arxiv.org/abs/1611.01646

[33] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," 2016, *arXiv:1612.00370*. [Online]. Available: http://arxiv.org/abs/1612.00370

[34] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[35] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 74–93, May 2017.

[36] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," 2017, *arXiv:1704.03899*. [Online]. Available: https://arxiv.org/abs/1704.03899

[37] Y. Tang, N. Srivastava, and R. R. Salakhutdinov, "Learning generative models with visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1808–1816.

[38] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia, G. Yang, and J. Xu, "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2347–2360, Sep. 2019.

[39] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," 2018, *arXiv:1805.06297*. [Online]. Available: https://arxiv.org/abs/1805.06297

[40] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Computación Y Sistemas*, vol. 18, no. 3, pp. 491–504, Sep. 2014.

[41] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *Proc. IDEAL*. Berlin, Germany: Springer, 2013.

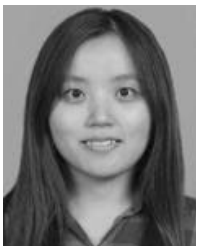
[42] A. Madylova and S. G. Oguducu, "A taxonomy based semantic similarity of documents using the cosine measure," in *Proc. 24th Int. Symp. Comput. Inf. Sci.*, Sep. 2009, pp. 129–134.

[43] J. Kasac, J. Deur, B. Novakovic, I. V. Kolmanovsky, and F. Assadian, "A conjugate gradient-based BPTT-like optimal control algorithm with vehicle dynamics control application," *IEEE Trans. Control Syst. Technol.*, vol. 19, no. 6, pp. 1587–1595, Nov. 2011.

[44] G. Mejía and K. Niño, "A new hybrid filtered beam search algorithm for deadlock-free scheduling of flexible manufacturing systems using Petri nets," *Comput. Ind. Eng.*, vol. 108, pp. 165–176, Jun. 2017.

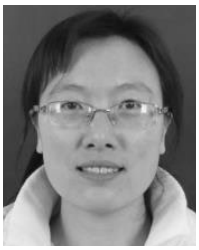
[45] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Jul. 2002, pp. 311–318.

[46] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.
[47] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[48] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.
[49] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl.*, 2007, pp. 228–231.
[50] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[51] Q. Qian, R. Jin, J. Yi, L. Zhang, and S. Zhu, "Efficient distance metric learning by adaptive sampling and mini-batch stochastic gradient descent (SGD)," *Mach. Learn.*, vol. 99, no. 3, pp. 353–372, Jun. 2015.
[52] A. T. Hadgu, A. Nigam, and E. Diaz-Aviles, "Large-scale learning with AdaGrad on spark," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Oct. 2015, pp. 2828–2830.
[53] Y. Dauphin, H. De Vries, and Y. Bengio, "Equilibrated adaptive learning rates for non-convex optimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1504–1512.
[54] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project adam: Building an efficient and scalable deep learning training system," in *Proc. Usenix Conf. Operating Syst. Design Implement.*, 2014, pp. 571–582.
[55] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 271–275.
[56] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1549–1557.
[57] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2641–2649.
[58] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1473–1482.
[59] M. Khademi and O. Schulte, "Image caption generation with hierarchical contextual visual spatial attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1943–1951.
[60] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, and W. Dong, "Image caption generation with part of speech guidance," *Pattern Recognit.*, vol. 119, pp. 229–237, Mar. 2019.
[61] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5659–5667.



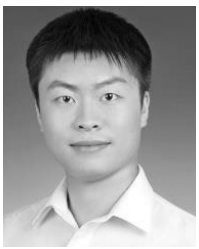
BIN WANG received the Ph.D. degree from the Department of Automation, Shanghai Jiao Tong University, China, in 2014. She is currently an Associate Professor with Shanghai Normal University, China. Her research interests include computer vision, machine learning, image processing, and urban computing.

CUNGANG WANG received the B.E. degree in educational technology from Liaocheng University, Liaocheng, China, in 2000, and the M.S. degree in computer science and technology from the Ocean University of China, Qingdao, China, in 2006. He has been an Associate Professor with Liaocheng University, since 2008. His research interests include machine learning, image processing, and pattern recognition.



QIAN ZHANG received the Ph.D. degree from Shanghai University, China. She is currently an Associate Professor with Shanghai Normal University, China. Her research interest includes video processing.

YING SU received the Ph.D. degree from Northwestern Polytechnical University, China. She had been engaged in the research of system engineering and M&S estimation. She is currently involved in intelligent information processing and applied simulation technology with Shanghai Normal University.



YANG WANG received the Ph.D. degree from the Chinese Academy of Sciences (CAS). He joined the College of Information, Mechanical, and Electrical Engineering, Shanghai Normal University, as an Assistant Professor, in 2017. He has served as a Technical Advisor in public data security for the government. He has published several research articles in reputed journals. His current research interests include big data, compressive sensing, next-generation optical processors, electric system modeling, and performance analysis.

YANYAN XU received the B.S. and M.S. degrees from the School of Information Science and Engineering, Shandong University, in 2007 and 2010, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Department of Automation, Shanghai Jiao Tong University, China, in 2015. He was a Postdoctoral Associate with the Department of Civil and Environmental Engineering, MIT, from 2015 to 2018, and a Guest Postdoctoral Fellow with the Lawrence Berkeley National Laboratory, Energy Analysis and Environmental Impacts Division, from 2017 to 2018. He has been a Postdoctoral Scholar with the Human Mobility and Networks Laboratory, Department of City and Regional Planning, University of California at Berkeley, since November 2018. His work has been published in *Nature Energy*, the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, the *Journal of the Royal Society Interface*, CEUS, IJCAI, and among others. His research interests include human mobility and urban computing, with particular emphasis placed on the use of massive trajectory data in intelligent transportation systems, urban planning, environment, and energy from the interdisciplinary perspective.