The Visual Computer (2019) 35:1655–1665
https://doi.org/10.1007/s00371-018-1565-z

ORIGINAL ARTICLE

Modeling coverage with semantic embedding for image caption generation

Teng Jiang[1] · Zehan Zhang[1] · Yupu Yang[1]

Published online: 5 June 2018

Abstract
This paper presents a coverage-based image caption generation model. The attention-based encoder–decoder framework has enhanced state-of-the-art image caption generation by learning where to attend of the visual field. However, there exists a problem that in some cases it ignores past attention information, which tends to lead to over-recognition and under-recognition. To solve this problem, a coverage mechanism is incorporated into attention-based image caption generation. A sequential updated coverage vector is applied to preserve the attention historical information. At each time step, the attention model takes the coverage vector as auxiliary input to focus more on unattended features. Besides, to maintain the semantics of an image, we propose semantic embedding as global guidance to coverage and attention model. With semantic embedding, the attention and coverage mechanisms consider more about features relevant to the semantics of an image. Experiments conducted on the three benchmark datasets, namely Flickr8k, Flickr30k and MSCOCO, demonstrate the effectiveness of our proposed approach. In addition to solve the over-recognition and under-recognition problems, it behaves better on long descriptions.

Keywords Coverage model · Semantic embedding · Image caption generation · Attention-based model

1 Introduction

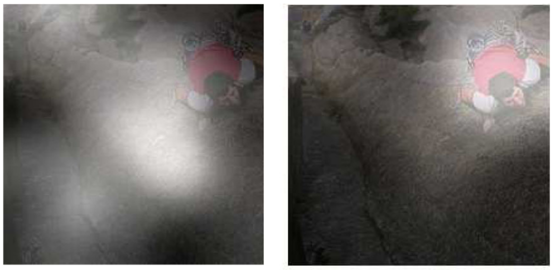
Automatically generating captions of images in sentence-level language has been a popular topic in computer vision recently. The challenge of the problem consists in not only recognition of objects in an image, but also understanding their activities, properties and relationships. Furthermore, the semantic information has to be expressed in grammatically and semantically reasonable natural language. It combines computer vision with natural language processing which are two major fields of artificial intelligence. While this task seems straightforward for humans with just a glance, it remains a challenging problem for machines over a long period of time.

The effective and widely used framework in image caption generation is attention-based encoder–decoder architecture [1–6]. The encoder–decoder framework plays an important role in many applications, vary from the most popular neural machine translation [7] and speech recognition [8] to interactive animation generation [9]. Attention mechanism has long be used in computer vision [10–12], and its integration into encoder–decoder framework has achieved many state-of-the-art results [1,13,14]. In image caption generation, the encoder phase uses a convolutional neural network (CNN) [15–17] to extract features from an image. The attention model attends to a subset of the features each step, and the decoder phase generates a corresponding word. By dynamically attending to different features of the encoder, the attention mechanism models the alignments between the objects and their corresponding descriptions.

The attention process is to decide which features contribute to the next generated word. However, after each attention step, which features having been attended is not recorded. As a consequence, the generation task might result in problems of over-recognition and under-recognition, which, respectively, means some features are unnecessarily recognized for multiple times and some features are mistakenly unrecognized but important to the semantics of an image. Figure 1 shows an example of over-recognition and under-recognition. In Fig. 1a, the words “climbing a rock” are generated twice. In Fig. 1b, the proposed coverage model alleviates the problem and recognizes the “red shirt.”

✉ Teng Jiang
jtengyp@163.com

[1] Department of Automation, Shanghai Jiao Tong University and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China



(a) A person climbing a rock climbing on a rock.

(b) A man in a red shirt is climbing a rock.

Fig. 1 Example captions of **a** over-recognition and under-recognition and **b** coverage model. The white region indicates the object recognized by the model, and the underline indicates the corresponding word

Besides, the attention process is a kind of combination of the source features. Obviously, combining all the features at each step is a complicated task. The search space increases exponentially with the number of features. To solve the problems stated above, a mechanism should be proposed to keep track of the attention process.

Inspired by Tu et al. [18], we incorporate into the coverage model, in which a sequentially updated coverage vector records the attention history. The attention model takes the coverage vector as input and focuses more on features that have not been covered. Nonetheless, the image features are assembled together without relevant information of temporal relationships. Besides, there is a large amount of redundant information in an image. Only features related to the topic of an image are supposed to be covered. Therefore, we propose semantic embedding as auxiliary guidance to coverage and attention model. The semantic embedding captures the semantics of an image. It leads the two mechanisms to concentrate more on features relevant to the subject of an image and to follow the theme of the description.

In this paper, a coverage model with semantic embedding is incorporated into attention-based image caption generation. Basically, a dynamically updated coverage vector is appended to the intermediate representations of encoder–decoder architecture to keep track of the attention history. To make sure that the coverage and attention mechanisms focus on sensible objects and ignore redundant ones, a semantic embedding is learned as extra input to both coverage and attention models. The extracted semantic information could provide global guidance for attention and coverage. In the following of this paper, we refer to the proposed model as Cov-Sem.

This paper concentrates on image caption generation with coverage and attention-based encoder–decoder architecture. The main contributions of our work are summarized as follows:

1. A coverage model is incorporated into attention-based encoder–decoder architecture to dynamically keep track of attention history. This design makes the attention mechanism ignore objects which have been recognized and focus more on uncoverd regions.
2. A semantic embedding is proposed as auxiliary input to coverage and attention model to guide their learning. Semantic embedding recognizes the image at global scale. Associated with already generated words, it provides precise understanding of the entire image.

The rest of this paper is organized as follows. In Sect. 1, related work in image caption generation is presented. In Sect. 2, the details of the Cov-Sem are described. In Sect. 3, the experiment results and examples are analyzed and finally in Sect. 4, we conclude this paper with future research directions.

2 Related work

Generating natural language descriptions for images has long been studied. Generally, the literature on image caption generation can be divided into three categories. The first are template-based methods [19,20], in which the objects, attributes and activities of an image are first detected and then filled in a pre-defined sentence template. These approaches are interpretable and can easily work with various object recognition components. However, they excessively rely on hand-designed templates and limit the expressiveness, flexibility and fluency of sentences. The second follow retrieval-based strategies [21,22], also known as transfer-based methods. They search similar images that have captions and combine these captions to generate a new one. These sort of methods require great amount of training images as retrieving library. Accordingly, they cannot generalize to images not contained in training set, i.e., lack of the ability of generalization. Moreover, if the library of images is very large, then for each image to inference, the searching time will be the main constraint to prevent these methods from utilization.

Inspired by recent progress of deep neural networks in image recognition [15–17] and machine translation [7,23], neural networks-based image caption generation has gained more interest recently. Many of these methods apply CNN for image feature extraction and RNN for caption generation. Kiros et al. [24] first introduce neural networks into caption generation, in whose method only image features are extracted by convolutional networks while the language description phrases are generated by log-bilinear model. Mao et al. [25] replace the language model with a RNN and employ a multimodal network to connect the RNN with a CNN. Donahue et al. [3] and Vinyals et al. [2] apply end-to-end neural networks to generate image descriptions. The incorporation of long short-term memory (LSTM) networks

[26,27] makes the model to learn long-term dependencies. Different from approaches above, Karpathy and Li [28] propose to project an image and its corresponding description into a joint embedding space and measure the similarity score between them. In addition to ensure the relationship between the semantics of the entire sentence and visual content, Pan et al. [29] add coherence loss to jointly learning the visual-semantic embedding and contextual relationship of the captions. Jia et al. [30] rely on canonical-correlation analysis (CCA) to build semantic representation of images as global guidance of LSTM. Zhou et al. [31] build dynamic semantic information of images based on previous generated word.

In addition to the above three main categories, deep reinforcement learning is first applied for image caption generation by Ren et al. [32], in which the caption generation is regarded as a decision-making process. It also maps the images and sentences into a joint embedding space as the ranking losses. All the previous methods distilled static image features. Xu et al. [1] incorporate attention mechanism into image caption generation. The attention mechanism encodes image features every time step according to already generated words and learned the latent alignments between image regions and words. With regard to over-recognition and under-recognition problems in attention mechanism for image caption generation, in this paper the coverage model proposed in Tu et al. [18] is introduced to keep track of attention history. Furthermore, a semantic embedding is proposed to provide global semantic information of an image. Experiment results on three benchmark datasets show the effectiveness of our proposed method, especially on descriptions with long length.

3 Coverage model with semantic embedding

3.1 Attention-based encoder–decoder architecture

Our work is built on attention-based encoder–decoder architecture. In this subsection, a brief introduction is given to this architecture.

3.1.1 Encoder: CNN for image features

The encoder takes a single raw image as input and extracts a variety of image features denoted as annotation vectors. Specifically, the commonly used convolutional neural network is applied here to extract image features. At each convolutional layer, the network extracts many channels of features and each channel is a 2-D matrix, so that convolutional neural networks possess the ability to preserve the 2-D topology architecture of images. As Zeiler and Fergus [33] mentioned, each channel in the convolutional layers corresponds to a patch of the image. The higher the layer, the more invariant of features. Therefore, the L annotation vectors produced by the encoder are machine-level representations of the image. Each vector is flattened to D dimensions as shown below.

$$\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^D, \quad i = 1, \ldots, L \tag{1}$$

Image caption generation task needs to recognize several related objects and their attributes more than just categories of an image. In order to maintain adequate information, the image features are extracted from a lower convolutional layer.

3.1.2 Decoder: attention-based LSTM

Caption generation is a sequential learning problem, on which the LSTM network is an expert with the advantage of modeling long-term dependencies. The major components of LSTM are memory cells, along which the network states propagate over time, and multiplicative gates that control the input, output of the network and the update of the cell state. The LSTM architecture applied here is built on Gers et al. [34]. An LSTM memory block is presented in Fig. 2. The memory cell and gates in an LSTM block are defined as follows:

$$\mathbf{i}_t = \sigma(W_i E\mathbf{y}_{t-1} + U_i\mathbf{h}_{t-1} + Z_i\mathbf{z}_t + \mathbf{b}_i) \tag{2}$$
$$\mathbf{f}_t = \sigma(W_f E\mathbf{y}_{t-1} + U_f\mathbf{h}_{t-1} + Z_f\mathbf{z}_t + \mathbf{b}_f) \tag{3}$$
$$\widetilde{\mathbf{c}}_t = \tanh(W_c E\mathbf{y}_{t-1} + U_c\mathbf{h}_{t-1} + Z_c\mathbf{z}_t + \mathbf{b}_c) \tag{4}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \widetilde{\mathbf{c}}_t \tag{5}$$
$$\mathbf{o}_t = \sigma(W_o E\mathbf{y}_{t-1} + U_o\mathbf{h}_{t-1} + Z_o\mathbf{z}_t + \mathbf{b}_o) \tag{6}$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{c}_t) \tag{7}$$

where $\mathbf{i}_t, \mathbf{f}_t, \widetilde{\mathbf{c}}_t, \mathbf{c}_t, \mathbf{o}_t$ and $\mathbf{h}_t$ are the input gate, forget gate, input to memory cell, memory cell, output gate and hidden state of the LSTM block at time t, respectively. The vector $\mathbf{z}_t$ is the context vector, generated by attention model. $\mathbf{y}_{t-1}$ is one-hot representation of words generated at time *t-1*, and $\mathbf{E}$ is word embedding matrix. $\odot$ represents element-wise multiplication. The nonlinearities used here are sigmoid $\sigma(\cdot)$ and hyperbolic tangent $\tanh(\cdot)$.

In addition to the LSTM network, another component of the decoder phase is attention model. It captures the relativity between the visual features and each word of generated captions. We use the soft attention model in Xu et al. [1]. Thus, the context vector at time step t is the weighted sum

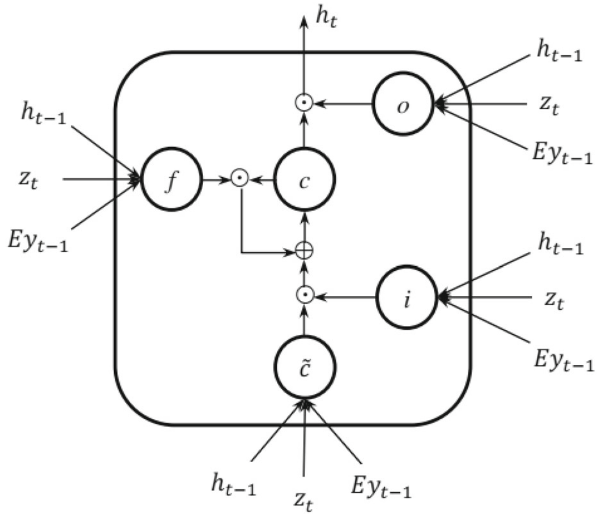


Fig. 2 An LSTM memory block

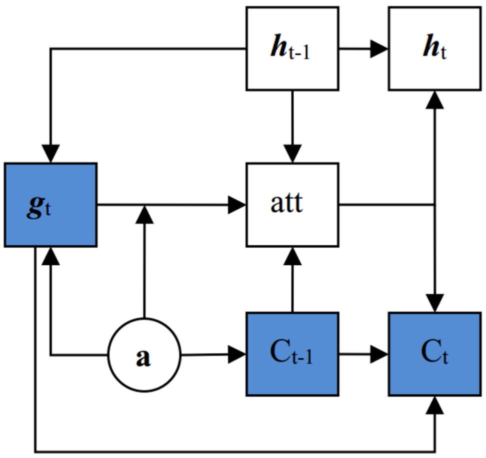


Fig. 3 Architecture of Cov-Sem model. The rectangles in blue represent coverage bedding in LSTM is unclear model and semantic embedding, respectively

of annotation vectors as following:

$$\mathbf{z}_t = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{a}_i \tag{8}$$

The weight $\alpha_{t,i}$ is a measurement of the relative importance of feature $\mathbf{a}_i$ at time t. The attention model is right for the computation of $\alpha_{t,i}$. It has two steps. For the first step, a multilayer perceptron (MLP) is applied to produce un-normalized weights based on hidden states of LSTM network at last step.

$$e_{t,i} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1}) \tag{9}$$

Here, f_{att} is an MLP model. Since there is a diversity of deep neural networks, in principle, any architecture is suitable for the attention model. As it is not the emphasis, we use the MLP for simplification. The second step normalizes the weights to sum to one.

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})} \tag{10}$$

In fact, the attention model is predictive to automatically decide which features of the image are contributable to next generated words at each time step. It focuses on what it thinks is important. Therefore, it accomplishes a sort of dynamic recognition of the original image. In summary, the attention model firstly computes the attention weights for annotation vectors and generates the context vector at this step. Then, the LSTM model takes the context vector and previously generated word as input, the hidden state at last step as recurrent, and generates a new word. This process continues until the model encounters a stopping sign, such as period. A complete sentence is generated to describe the content of the image.

3.2 Coverage model with semantic embedding

Our Cov-Sem model is attention-based encoder–decoder framework with coverage and semantic embedding, as illustrated in Fig. 3. The attention model and LSTM layer are simplified to rectangulars. There are two additional blocks besides the attention-based architecture. The details of the coverage and semantic embedding blocks will be presented in the following subsections.

3.2.1 Coverage model

The coverage model employed here is inspired by Tu et al. [18], in which the model is used for keep track of translation process. Image caption generation is similar to machine translation to a great extent, both translating one kind of representation to a natural language. Therefore, the coverage model in machine translation is incorporated into the image caption generation task.

The attention model in image caption generation is to build the correspondence of pathes of an image and the generated words. One patch should not be considered again after it has been attended. However, attention model lacks the mechanism to indicate which part of the image has been focused on. As Eq. 9 shows, the attention model only considers LSTM hidden state of last time step. The historical state is not contained in the attention model. As a result, some parts of the image may be attended more than once or may not be attended. Therefore, it is important to introduce coverage model. As the coverage model is to keep track of the attention history, a coverage vector is appended to the annotations, each dimension corresponds to one annotation vector. The value of the vector indicates the degree of each annotation being attended. The coverage vector is initialized to zero and updated after every step of the attentive recognition of the annotation vectors. Then, the coverage vector is fed to the attention model to facilitate where to attend in the next time step.

In each step, the coverage vector updates according to history coverage information and current attention results. Evidently, the coverage model is sequential. In addition, since the image has a variety of features, some important and some of no use. Therefore, it is hard to artificially design the coverage model to decide which features need to be covered to what extent. Deep neural networks are overwhelming in many complicate feature engineering applications. Consequently, we use the LSTM network to model coverage for its long-term dependency ability. For each annotation vector, there is a scalar representing the coverage of this annotation. Thus, the output of the coverage model is a L-dimensional vector. For simplicity, we omit the gates and memory cells representations and abbreviate the model as follows:

$$C_{t,j} = \mathrm{LSTM}(C_{t-1,j}, \alpha_{t,j}, \mathbf{a}_j),\ j = 1, \ldots, L \tag{11}$$

where $C_{t,j}$ is a scalar representing the coverage information of the annotation $\mathbf{a}_j$ at time t.

3.2.2 Neural network-based semantic embedding

The coverage model is used for machine translation initially. Every word in a sentence should be translated, i.e., the coverage vector should cover all the words. However, for image caption generation, not all the features contribute to the caption. There exists large amount of redundancy in an image. In addition, the attention model only captures local relationships between image patches and generated words. So it is not enough to capture the useful information of an image with just adding a coverage model. In this subsection, we propose semantic embedding as global guidance for attention and coverage model. Different from Jia et al. [30] which projected the image and sentence into a common semantic space, the semantic embedding used in this paper adopts a generative process to produce global semantic information dynamically based on currently generated words.

Given an image, there are different descriptions from different viewpoints. So the global semantic is not unique. Based on this consideration, we take into account the already generated words as input to semantic embedding. Using contextual information, the semantic guidance could avoid ambiguity. Extracting the semantic information of an image based on the image features and previously generated words is a complicated task. The semantics should be adjusted with the generation process and return to guide the generation. The embedding consists of two modules, image features embedding and contextual embedding. Convolutional neural networks are end-to-end learning architectures without hand-designing features and used in many visual recognition and image semantic parsing tasks. Thus, a CNN with fully connected output is used as the image features embedding. Its architecture is shown in Fig. 4. Concretely, the semantic embedding model has the following form:

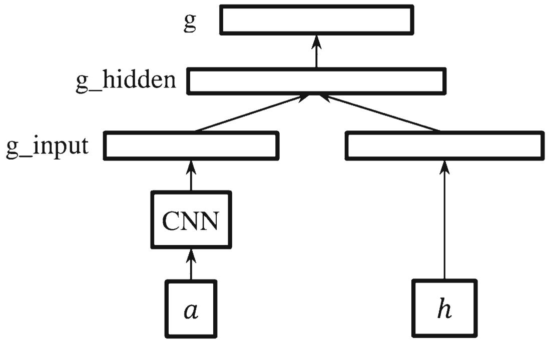


Fig. 4 Architecture of semantic embedding model

$$\mathbf{g}_t = f_g(W_{ga}\mathrm{CNN}(\mathbf{a}) + W_{gh}\mathbf{h}_{t-1} + \mathbf{b}_g) \tag{12}$$

where $\mathbf{g}_t$ is the semantic information at time t and $\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}$ represents all the annotation vectors. W_{ga} and W_{gh} are embedding matrices which project image features and contextual information into semantic space, respectively. f_g is a fully connected network that produces final semantic information.

With the semantic embedding as guidance, the attention and coverage model are redefined as follows:

$$e_{t,i} = f_{\mathrm{att}}(\mathbf{a}_i, \mathbf{h}_{t-1}, C_{t-1,i}, \mathbf{g}_t) \tag{13}$$

$$C_{t,j} = \mathrm{LSTM}(C_{t-1,j}, \alpha_{t,j}, \mathbf{a}_j, \mathbf{g}_t) \tag{14}$$

The detail of coverage model with semantic embedding is shown in Fig. 5. At each time step, the coverage model learns coverage information automatically from attention weight vector and previous coverage state with the guiding of semantic embedding. Finally, the output of LSTM network is the probability of generated words. It is computed with softmax as following:

$$p(\mathbf{y}_t|\mathbf{a}, \mathbf{y}_{t-1}) = \mathrm{softmax}(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h\mathbf{h}_t) \tag{15}$$

where $\mathbf{L}_h$ is projection matrix. Note that the previously generated word is projected to compute output probability. This is because the neighboring words have strong correlation, and the previous word is helpful for generating next word.

3.3 Training and inference

We take end-to-end training for the Cov-Sem model, which learns the attention model parameters together with cover-

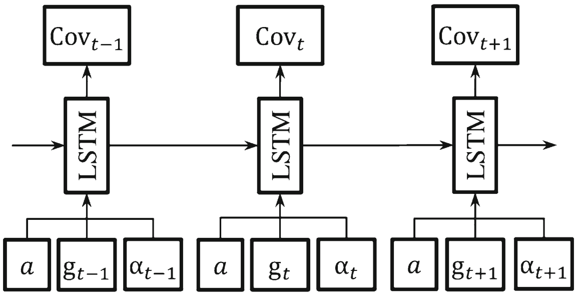


Fig. 5 Detail of coverage model with semantic embedding

age model and semantic embedding. In training the attention model, a problem is realized that the attention weight vector at each time step differed little from each other for a particular image. This indicates that the attention strength each time step was not strong enough. To make the attention weights more diversity, a penalty regularization is added to the loss function. A measure on the dissimilarity between two vectors is the cosine of their angle. For computational simplicity, the inner product of two weight vectors of different time step is chosen as the approximation for cosine similarity. Concretely, the model is trained by minimizing the negative log-likelihood of reference sentences.

$$J(\theta) = -\sum_{n=1}^{N} \log p(\mathbf{y}_n|\mathbf{a};\theta) + \lambda \sum_{t_1=1}^{C} \sum_{t_2=t_1+1}^{C} {}_{t_1}{}^{T}{}_{t_2} \tag{16}$$

where θ represents model parameters, N is batch size, $p(\mathbf{y}_n|\mathbf{a};\theta)$ is the output probability defined in Eq. 15, λ is the penalty coefficient for attention model, and C is the number of time steps, i.e., the length of sentences.

In summary, the process of generating a caption for an image is listed as follows:

1. Extract annotation vectors $\mathbf{a}$ from the image.
2. Compute global semantic embedding $\mathbf{g}_t$ according to Eq. 12.
3. Compute coverage information $C_{t,j}$ according to Eq. 14.
4. Compute attention weights $\alpha_{t,i}$ according to Eqs. 13 and 10. And then compute context vector $\mathbf{z}_t$ according to Eq. 8.
5 . Feed the context vector $\mathbf{z}_t$, previous hidden state $\mathbf{h}_{t-1}$ and previous word $\mathbf{y}_{t-1}$ to the LSTM network to select the word with maximum output probability according to Eq. 15.

The process continues until the end sign, such as period, is produced. To select an appropriate description for an image, beam search is applied following previous work. It iteratively consider the k best sentences at time t as candidates to inference sentence at next time step and the best k of them are kept. In our work, the k is set to 6 by trial and error.

4 Experiments

We describe our experimental settings and quantitative results to validate the effectiveness of our proposed model in this section.

4.1 Datasets and evaluation metrics

For evaluation, we experiment on three popular datasets Flickr8k, Flickr30k, which have 8,192 and 31,783 images, respectively, and Microsoft COCO, which has 82,783 images for training set. All the three datasets are easily available publicly. Each image in Flickr8k and Flickr30k datasets is annotated artificially with 5 sentences in English. For Microsoft COCO dataset, some images have more than 5 captions. They are kept in training set. And in validation set and test set, images with more than 5 captions are discarded for ease of computing evaluation scores. Besides, the splits of datasets for training, validation and test set matter a lot. There is a widely recognized splits for Flickr8k dataset [35]. For Flickr30k, we adopt the publicly available splits in Karpathy and Li [28]. As for Microsoft COCO dataset, since the captions for test set are not available, the strategy in Xu et al. [1] is applied to split the validation set into small validation set and test set.

Not only is a tricky problem to generate sentences for images, but also evaluating the accuracy of a description remains difficult. It concerns the semantics and syntax rather than merely matches the results with groundtruth word by word. Therefore, several evaluation metrics are applied in this paper. The most widely used metric in language generation literature is BLEU score [36], which measures the n-gram precision between generated and reference sentences with n ranging from 1 to 4. In addition to BLEU, we adopt METEOR [37] as it is highly correlated with human judgements. Furthermore, the newly proposed approach CIDEr [38] captures human consensus and is a state-of-the-art evaluation metric for image caption evaluation. Last but not the least, the classical ROUGE-L [39] measurement is added as complementary.

4.2 Training details

A few problems existed when conducting the experiments. Firstly, image recognition is a research area independent from image caption generation. It requires large amounts of images to distinguish various scenes. However, the datasets of image caption generation contain limited number of images, not enough to train a high qualified model. We use

Table 1 BLEU, METEOR, CIDEr and ROUGE-L scores for image caption generation of different models

Dataset	Model	BLEU				METEOR	CIDEr	ROUGE-L
		BLEU-1	BLEU-2	BLEU-3	BLEU-4			
Flickr8k	Google NIC [2]	63.0	41.0	27.0	–	–	–	–
	Soft attention [1]	67.1	46.1	31.3	**22.0**	20.2	49.2	49.8
	Cov-Sem	68.5	46.2	32.1	20.8	**20.7**	50.2	50.1
	Cov-Sem-FT	**69.0**	**47.1**	**32.4**	21.9	20.3	**50.7**	**50.2**
Flickr30k	Google NIC [2]	66.3	42.3	27.7	18.3	–	–	–
	Soft attention [1]	68.1	46.0	31.9	22.1	18.9	42.0	48.4
	Sem-Attention [4]	64.7	46.0	**32.4**	**23.0**	18.9	–	–
	Cov-Sem	68.3	46.1	31.8	21.9	19.1	42.3	**49.0**
	Cov-Sem-FT	**68.9**	**46.8**	31.9	22.0	**19.1**	**42.8**	48.7
MSCOCO	Google NIC [2]	66.6	46.1	32.9	27.7	23.7	–	–
	Soft attention [1]	71.8	52.5	38.5	28.3	23.9	84.5	54.8
	Sem-Attention [4]	70.9	53.7	**40.2**	**30.4**	24.3	–	–
	Cov-Sem	72.0	52.6	38.6	28.5	23.7	84.6	55.6
	Cov-Sem-FT	**72.9**	**53.8**	39.6	29.5	**24.4**	**88.3**	**56.0**

The bold numbers refer to best results across different methods

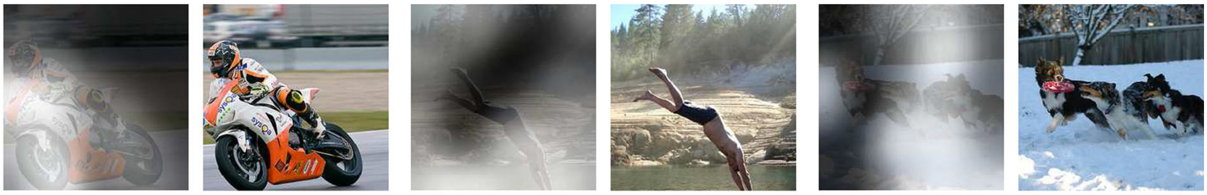
the pretrained Oxford VGGnet to extract image features. The images are centered with mean values of ImageNet 2012 dataset and then the outputs of conv5_4 layer of the VGGnet are saved as annotation vectors $\mathbf{a} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}$. Secondly, the vocabularies of Flickr30k and MSCOCO datasets contain a multitude of uncommon uncommon words and non-alphabetic characters, which lead the vocabulary size to more than 20,000. This makes the model very complex and many parameters are redundant. As a result, we limit the vocabulary size to 10,000, making the model easier to train. Thirdly, we observed that the negative log-likelihood of validation set diverged from the BLEU score in later training epochs. Since BLEU score is main evaluation metric, it was used as early stopping criterion. However, at each validation point, to generate captions of validation set consumed a lot of time, especially for Microsoft COCO dataset. To reduce training time, the BLEU score was computed when the negative log-likelihood decreased to an appropriate value. In this case, on Microsoft COCO dataset, our Cov-Sem model took about one day and a half to train on an NVIDIA GTX1080 GPU.

Our Cov-Sem model is trained using Adam algorithm with $\alpha = 0.0002$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ [40]. To alleviate overfitting, we adopted dropout [41] and early stopping as well as weight decay with value 1e−4. For Flickr8k dataset, the dimension of word embedding is set to 240, the dimension of hidden state of decoder network is 1100, and the dimension of hidden state of coverage model is 480. For Flickr30k dataset, the parameters are set to 320, 1700, 700, and for MSCOCO dataset, they are set to 400, 1800, 900. Besides, the parameters of global semantic embedding are shared by three datasets, with hidden dimension set to 1500 and output dimension set to 200.

4.3 Generation results

In addition to directly training the model, a convolutional layer is added before the annotation vectors input into decoder network to fine-tune the network. The fine-tuned model is named as Cov-Sem-FT. In Table 1, we report out evaluation results on the three datasets. We retrained soft attention model from scratch and got comparative results. As for Google NIC model, we use the results reported in Xu et al. [6]. The results show that our proposed Cov-Sem method outperforms the soft attention model for most criteria. The fine-tuned model improves the performance by 1 percentage point on MSCOCO dataset. A further analysis on the generated sentences indicates that although some sentences do not correctly describe the content of the corresponding images, they are grammatically right. This states clearly that the decoder phase is adept at generating sentences. We speculate the low accuracy is due to the fact that the encoder phase is trained for object recognition. It fails to recognize the activities or relationships of objects.

To better understand the characteristics of the Cov-Sem model, we give several examples in Figs. 6 and 7. Three over-recognition pairs are aligned in Fig. 6. For each pair, the first case is generated by soft attention model and the second case by Cov-Sem model. For visually convenient, we visualize the attention weights of over-recognized region for each image. Similarly, Fig. 7 shows three examples of under-recognition pairs. And for each pair, the attention weights of under-recognized region are visualized in the second case. As we can see, the Cov-Sem model can attend to not only salient objects but also regions to infer the attributes of objects. The over-recognized features are accurately eliminated by

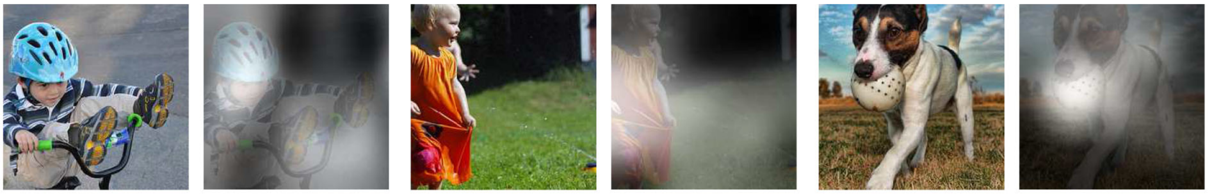


(a) A man on a motorcycle on a motorcycle.
(b) A motorcycle racer is riding a motorcycle.

(a) A man is jumping into the water into the water.
(b) A man is jumping into the water.

(a) Three dogs run in the snow in the snow.
(b) Three dogs are playing in the snow.

Fig. 6 Example captions of over-recognition. The white region indicates the object over-recognized by the attention model, and the underline indicates the corresponding word. In each pair, sentence (**a**) is the over-recognized case and sentence (**b**) is generated by Cov-Sem model

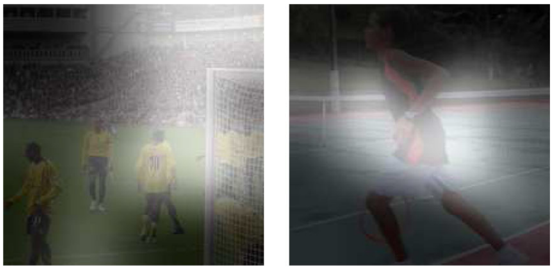


(a) A boy is riding a bike.
(b) A boy wearing a helmet is riding a bicycle.

(a) A boy is playing in a field.
(b) A boy in a yellow shirt is sitting in the grass.

(a) A brown and white dog is playing in the grass.
(b) A brown and white dog is playing with a ball.

Fig. 7 Example captions of under-recognition. The white region indicates the object under-recognized by the attention model, and the underline indicates the corresponding word. In each pair, sentence (**a**) is the under-recognized case and sentence (**b**) is generated by Cov-Sem model



(a) Two soccer players are playing soccer in a stadium.

(b) A boy in a red shirt is playing tennis.

Fig. 8 Examples of generation mistakes

the Cov-Sem model, and the under-recognized features are extracted as well.

Nonetheless, mistakes exist in the Cov-Sem model, as illustrated in Fig. 8. For the first case, the number of players is mistakenly recognized. Counting the number of objects is a higher level artificial intelligence than objects recognition, especially in scenes with complicated background. In the second case, we can see that the model takes the red region as the color of the shirt. One object possesses diverse attributes depending on the situation. Learning how to discriminate the attributes remains a difficult problem in computer vision.

4.4 Effects on length of sentences

In addition to the evaluation accuracy of descriptions, in this subsection, we investigate the influence of coverage model on the length of sentences. Since there are 5 reference sentences per image, we compute the average length of the 5 sentences as partitioning criterion. Taking the Flickr8k dataset as an example, the minimum of average length of sentences in testing set is 6.4 and the maximum 20.8. Then, the images are partitioned into 4 groups, with the interval of sentence length in each group being [6,10), [10,14), [14,18) and [18,21). For each group, we select the descriptions generated by soft attention model and Cov-Sem model, respectively, and evaluate the accuracy with the metrics BLEU, METEOR, CIDEr and ROUGE-L.

Evaluation results are reported in Table 2. We note that the maximum sentence length generated by Cov-Sem model is closer to that of reference sentences. And the average length of generate sentences of Cov-Sem is slightly greater than that of soft attention model. Furthermore, the scores of three metrics on long sentences are higher for Cov-Sem model than for soft attention model to a certain degree. Especially when sentence length is more than 18, all the scores of Cov-Sem model are higher than that of soft attention model. As an example, we consider the image in Fig. 9. The sentence generated by soft attention model is as following:

A child is playing in the water.

Our proposed coverage model describes it into:

Table 2 Effects on sentence length of Flickr8k

Sentence length	Soft attention	Cov-Sem
Max length	13	14
Min length	7	7
Mean length	10.2	10.8
[6,10)		
BLEU-1	**65.8**	63.3
BLEU-2	**44.5**	42.8
BLEU-3	**30.8**	28.7
BLEU-4	**21.2**	19.2
METEOR	**18.1**	17.1
CIDEr	**34.7**	32.2
ROUGE-L	**52.3**	51.6
[10,14)		
BLEU-1	63.9	**65.8**
BLEU-2	43.9	**44.6**
BLEU-3	30.2	**30.4**
BLEU-4	**20.6**	19.8
METEOR	**24.0**	22.8
CIDEr	**64.6**	63.4
ROUGE-L	50.4	**50.7**
[14,18)		
BLEU-1	67.1	**69.8**
BLEU-2	46.3	**47.8**
BLEU-3	32.3	**32.8**
BLEU-4	**22.2**	21.9
METEOR	20.2	**20.8**
CIDEr	**51.8**	48.1
ROUGE-L	45.0	**46.3**
[18,21)		
BLEU-1	68.3	**68.8**
BLEU-2	47.2	**48.5**
BLEU-3	33.5	**34.7**
BLEU-4	22.7	**24.8**
METEOR	16.1	**16.6**
CIDEr	19.6	**25.2**
ROUGE-L	43.3	**43.6**

The bold numbers refer to best results across different methods

A little girl in a white shirt is standing in shallow water.

As the example shows, the sentence generated by Cov-Sem model covers some primary attributes of the objects in the image, such as "yellow shirt" and "shallow water." This provides an excellent evidence that the Cov-Sem model can alleviate under-recognition and generate more informative descriptions for images.

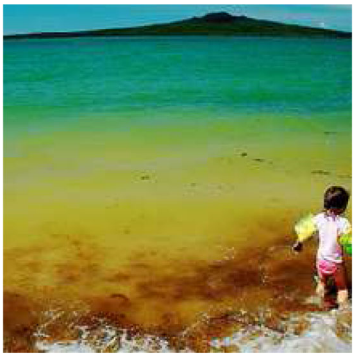


Fig. 9 Example of informative descriptions

5 Conclusion

In this paper, we introduce a coverage model with semantic embedding to attention-based encoder–decoder framework for image caption generation. The coverage model alleviates the over-recognition and under-recognition problems in attention mechanism when aligning words with image regions. The semantic embedding provides global guidance for coverage and attention model. The proposed method is evaluated on three benchmark datasets and shows the effectiveness of the coverage mechanism. Especially on complicated descriptions, Cov-Sem model shows excellent performance. Our work is complementary to attention-based model and can be applied not only to image caption generation, but also to speech recognition and other attention-based applications. Yet similar to the existing language generation approaches, the coverage model is lack of mechanisms to learning the relationships between objects in an image. Though objects recognition has gain great success in deep learning, there is few studies on interactions between objects. A potential direction is to concentrate on relation extraction task and generate its corresponding descriptions for images.

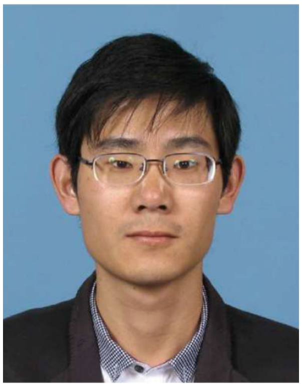
Acknowledgements This work is supported by the National Natural Science Foundation of China (Grant No. 61273161).

References

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of The 32nd International Conference on Machine Learning (ICML), pp. 2048–2057 (2015)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156–3164 (2015)
3. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2015)

4. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4651–4659 (2016)
5. Fu, K., Jin, J., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Trans Pattern Anal Mach Intell (2016). https://doi.org/10.1109/TPAMI.2016.2642953
6. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
7. Cho, K., van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
8. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964. IEEE (2016)
9. Wang, Y., Che, W., Xu, B.: Encoder decoder recurrent network model for interactive character animation generation. Vis. Comput. **33**(6–8), 971–980 (2017)
10. Mnih, V., Heess, N., Graves, A.: Recurrent models of visual attention. In: Advances in neural information processing systems (NIPS), pp. 2204–2212 (2014)
11. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. In: International Conference on Learning Representations (ICLR) (2015)
12. Wu, H., Wang, J.: A visual attention-based method to address the midas touch problem existing in gesture-based interaction. Vis. Comput. **32**(1), 123–136 (2016)
13. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J.: Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
14. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4945–4949. IEEE (2016)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. (NIPS) **25**(2), 1097–1105 (2012)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR) (2015)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning (ICML) (2015)
18. Tu, Z., Lu, Z., Liu, Y., Liu, X., Li, H.: Modeling coverage for neural machine translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 76–85 (2016)
19. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: European Conference on Computer Vision (ECCV), pp. 15–29 (2010)
20. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Lim, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: understanding and generating simple image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2891–2903 (2013)
21. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), vol. 1, pp. 359–368 (2012)
22. Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y.: Generalizing image captions for image-text parallel corpus. In: Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL), vol. 2, pp. 790–796 (2013)
23. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (ICLR) (2015)
24. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Multimodal neural language models. In: Proceedings of The 31st International Conference on Machine Learning (ICML), vol. 14, pp. 595–603 (2014)
25. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks. In: International Conference on Learning Representations (ICLR) (2015)
26. Hochreiter, S., Schmidhuber, J.: Long short term memory. Neural Comput. **9**(8), 1735–1780 (1997)
27. Greff, K., Srivastava, R.K., Koutnk, J., Steunebrink, B.R., Schmidhuber, J.: LSTM: a search space odyssey. IEEE Trans. Neural Netw. learning Syst. (2016). https://doi.org/10.1109/TNNLS.2016.2582924
28. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 664–676 (2015)
29. Pan, Y., Mei, T., Yao, T., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4594–4602 (2016)
30. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2407–2415 (2015)
31. Zhou, L., Xu, C., Koch, P., Corso, J.J.: Image caption generation with text-conditional semantic attention. arXiv preprint arXiv:1606.04621 (2016)
32. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. arXiv preprint arXiv:1704.03899 (2017)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV), pp. 818–833 (2014)
34. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: continual prediction with LSTM. Neural Comput. **12**(10), 2451–2471 (2000)
35. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: data, models and evaluation metrics. J. Artif. Intell. Res. **47**, 853–899 (2013)
36. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for computational linguistics (ACL), pp. 311–318 (2002)
37. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, vol. 29, pp. 65–72 (2005)
38. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015)
39. Lin, C.-Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL Workshop, pp. 74–81 (2004)
40. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2014)

41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)



Teng Jiang is a Ph.D. student in School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University, China. His research interests include computer vision, natural language processing and big data.

Zehan Zhang is a Ph.D. student in School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University, China. His research interests include natural language processing and fault detection and diagnosis.

Yupu Yang is a Professor in School of Electronic Information and Electrical Engineering at Shanghai Jiao Tong University, China. His research interests include deep learning, intelligent control and fault detection and diagnosis.