Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 9562587, 11 pages
https://doi.org/10.1155/2020/9562587

Research Article

Gated Object-Attribute Matching Network for Detailed Image Caption

Jing Yun,[1,2] **ZhiWei Xu**,[3] **and GuangLai Gao**[1]

[1]*Department of Computer Science, Inner Mongolia University, Hohhot 010021, China*
[2]*College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China*
[3]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China*

Correspondence should be addressed to Jing Yun; yunjing_zoe@163.com

Received 17 June 2019; Revised 29 October 2019; Accepted 11 November 2019; Published 13 January 2020

Guest Editor: Marco Perez-Cisneros

Image caption enables computers to generate a text description of images automatically. However, the generated description is not good enough recently. Computers can describe what objects are in the image but cannot give more details about these objects. In this study, we present a novel image caption approach to give more details when describing objects. In detail, a visual attention-based LSTM is used to find the objects, as well as a semantic attention-based LSTM is used for giving semantic attributes. At last, a gated object-attribute matching network is used to match the objects to their semantic attributes. The experiments on the public datasets of Flickr30k and MSCOCO demonstrate that the proposed approach improved the quality of the image caption, compared with the most advanced methods at present.

1. Introduction

One of the artificial intelligence (AI) dreams is making computers to be able to see and understand the rich visual world around us and endowing them with the ability to communicate with us in natural language. These aspirations are motivated by offering valuable practical applications, such as early childhood education and visual dysfunction assistance. The task may be very simple for humans, but it is very difficult for computers.

Image caption is an approach that enables computers to express what is seen in natural language. It requires detecting the objects in an image and describing what they are and what their attributes are. Recently, the image caption has made remarkable progress, especially with the frameworks based on convolution neural network (CNN) and long short-term memory network (LSTM) [1–4]. In these frameworks, CNN encodes the image into visual feature vectors. Then, LSTM maps the visual feature vectors to a sentence.

The attention mechanism is indispensable for humans when handling the visual problems. Humans pick out the key points and discover the essential information involved in the image by scanning the image and focusing on the salient aspects. Inspired by the attention mechanism of humans, researchers propose the visual attention-based neural networks [5–7] for image captioning, which have been developed with significant improvements. These models can selectively focus on specific objects instead of scanning the whole image. Therefore, the visual attention-based neural network can find the location information of the focused objects by locating the object regions but lacks the information about the object current state such as *sitting*, *lying*, and *redorblack*.

Additionally, the semantic attention-based neural networks [8–10] utilize the semantic attributes to focus on semantically important concepts about the image, to improve the accuracy of the generated caption. The semantic attributes include the state of the objects in an image, such as *sitting* and *lying* and *color*. All of these semantic attributes are the object properties detected from the image. It is apparent that attributes are valuable for improving the accuracy of the generated captions. However, the semantic

attention-based networks fail on locating the objects in the image [9].

As shown in Figure 1, the semantic attention-based network evaluates the relations of the objects and the related image region, as well as the relations of the attributes and the related image region. Limited by the significance evaluation of semantic attention-based networks, they can only include the relation between *motorcycle* and *track* in the image caption process but miss the relation between *motorcycle* and *yellow*. Consequently, the generated caption misses the attributes describing the color of the motorcycle. Actually, these attributes with low significance values are also indispensable for guaranteeing the accuracy of the generated caption.

Either visual attention-based neural networks or semantic attention-based neural networks for describing objects' attributes cannot utilize the mutual relations of objects and their attributes. A specific image caption is expected to tell not only what the objects are but also how these objects are. The significant challenge is how to accurately incorporate semantic attributes into visual attention-based networks.

To generate more detailed and accurate captions, we propose a model for addressing the aforementioned challenge by utilizing the mutual relations of objects and attributes. By leveraging a gated object-attribute matching network, the model discovers the essential relations of objects and attributes obtained by the visual attention module and the semantic attention module, respectively, and maps all meaningful attributes to the corresponding objects. In this way, a more detailed and accurate caption can be obtained.

As illustrated in Figure 2, the gated object-attribute matching network for the detailed image caption generates a detailed description of objects in the image. The proposed model consists of three steps: (1) The visual features are extracted by the object detectors that locate image regions of the objects. The semantic features are extracted by the attributes extractors which obtain descriptive words from the image captions in training sets to collect all types of attributes. (2) The first LSTM layer encodes the visual features of objects and the semantic features of the attributes. Then, we match the objects with their attributes by involving a gated attention-based recurrent network in order to mapping all meaningful attributes to the corresponding objects. (3) The second LSTM layer, *i.e.*, language LSTM, integrates the objects with their attributes to generate the corresponding caption.

Our main contributions are as follow:

(i) We study two types of image caption models, *i.e.*, the visual attention-based neural network and the semantic attention-based neural network. A visual attention-based network will focus on each specific image region and evaluate the relations of the related objects and this region. In this way, we can locate the objects but cannot match the corresponding descriptive attributes to these objects. On the contrary, a semantic attention-based network will evaluate the relations of the related objects and this image region, as well as the relations of the related attributes and this image region. However, with various limitations, semantic attention-based networks cannot accurately match each attribute to the corresponding object too. The significant problem is how to match each attribute to the corresponding object while using semantic attention-based networks.

(ii) We propose a gated object-attribute matching network that incorporates the semantic attributes into the visual attention-based network. This obtains additional information to get not only location information about objects but also their attributes for generating a more detailed image caption. Our work improves the image captioning by constructing a detailed presentation out of the visual representation and the semantic representation and improving the description model that exploits more information from the image. Furthermore, the usage of attributes benefits an elegant model which generates sentences from an open lexicon, making the description of the image more realistic.

(iii) Specifically, we implement the proposed model on the gated object-attribute matching network and conduct the experiments on public datasets Flickr30k and MSCOCO with metrics BLEU, METEOR, and CIDEr, respectively. The experimental results demonstrate that our model outperforms the state-of-the-art approaches in the accuracy of the generated captions.

We organize the paper into five parts: Section 1 is the introduction to an overview of the proposed model. Section 2 reviews the related work about the image caption. Then, Section 3 announces particulars of the proposed gated object-attribute matching network model, and Section 4 evaluates the effectiveness of the model on far-reaching experiments. Finally, Section 5 summarizes work in the paper.

2. Related Work

Image captioning is a very important part of artificial intelligence which leverages a natural language sentence to describe the content of the image automatically. Recently, the related work thrives for research interests.

2.1. Basic Image Caption Framework. The modeling methods discussed up to now in this paper are in use of the encoder-decoder framework for machine language translation. By using the successful application to machine translation [11, 12], these methods are considered to apply to the image caption which is the similar task to translate the image into the corresponding text by reasonable innovation. Kiros et al. [13] took the lead in using the encoder-decoder algorithm to generate text to describe the image, and a log-bilinear model with multiple modes as an input was proposed. Vinyals et al.

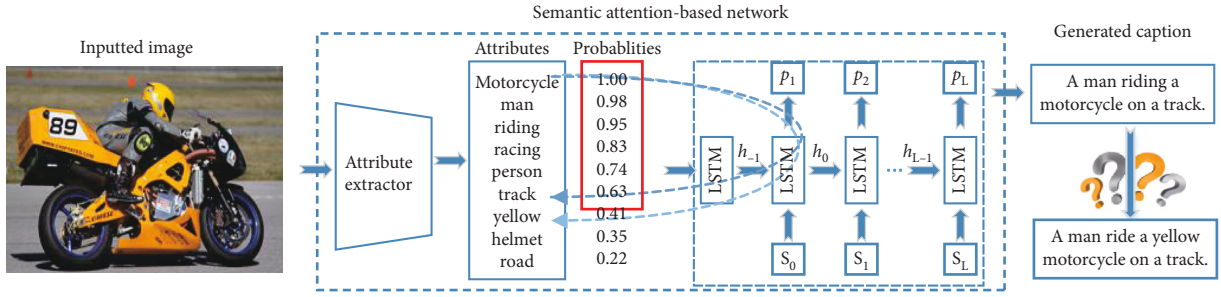

Figure 1: The semantic attention-based neural network.

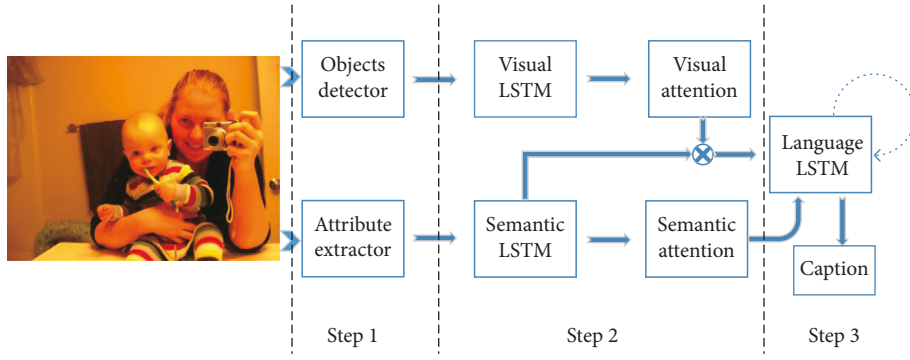

Figure 2: An illustration of the proposed gated object-attribute matching network model.

[1] only took the image as the input of RNN at the beginning. Then, Kiros et al. [14] improved the model using LSTM to encode sentences. Donahue et al. [15] applied the LSTM model to video description generating. All the above methods replaced the encoder by a pretrained CNN on ImageNet and used the output of its last layer as the encoder result. Then, the result of the encoder was put to the language model which is regarded as the decoder to generate text. In contrast, Karpathy and Fei-Fei [2] adopt the result of object detection from R-CNN and output of a bidirectional RNN to learn a joint embedding space for caption ranking and generation.

However, most of these methods take the entire image feature as the encoded vector. When the decoder generates words in each step, it gets the same context from the encoded vector. It should "pay attention" to different image regions when generating different words of a sentence.

Generated captions by using these approaches are always too simple to describe the essential content of the image [1, 16]. We can generate richer descriptions by exploiting compositional structures.

2.2. Visual Attention-Based Approach. The human visual cognitive system selectively focuses on specific regions or objects in its sight, which is the primary reason why the visual attention mechanism is important in image caption generation.

When the visual attention mechanism is applied to the image caption generation, different image regions have different impacts on the caption. Therefore, researchers propose diverse attention mechanisms that guide the model to focus on different areas of the image when generating each word in the caption, such as [5–7, 17, 18].

The attention-based approach [6] was intended to achieve the purpose of paying attention to the salient region that "should" be paid attention to when generating a word. Jin et al. [5] used the selective search [19] to locate the salient region in the image to attend. Then Lu et al. [7] introduced an adaptive attention model considering only visual words require information from the image. Afterwards, Anderson et al. proposed the object detector to get basic attention units [17]. Lu et al. [18] enforced the generative model to predict a template-like sentence and make notable advances.

2.3. Semantic Attention-Based Approach. Another branch of the image caption model is semantic attention approaches [8–10, 20] which detect semantic concepts from images to guide the generation of description. Especially, these approaches focus on the interesting semantic attributes in the image, by paying attention to semantic attributes with different weights [9]. Wu et al. [8] investigated the effect of adding advanced semantic concepts to the image caption model on generating image captions, which can significantly improve the performance of image captions. In addition, Fang et al. [21] utilized multiple instance learning to extract attributes related to different semantic information, which was feed as input to the LSTM language model to generate text description. The proposed long short-term memory with attributes (LSTM-A) [20] outperforms the state-of-the-art model at that time. Gan et al. [10] extended the weight

matrix for LSTM which dependents on each semantic tag for making the corresponding caption.

2.4. Reinforcement Learning in Sequence Generation. To deal with the loss-evaluation mismatch problem, the reinforcement learning is utilized to train deep end-to-end systems directly on nondifferentiable metrics. Therefore, researchers use reinforcement learning to directly optimize the evaluation metrics [22–26]. Ren et al. [24] integrated a policy network and a value network to collaboratively generate text description. The policy network was used to predict a sentence, while the value network served as a global predictor for the sentence in evaluation. Both networks are enhanced by reinforcement learning. Additionally, Liu et al. [23] incorporated metric SPICE [27] and metric CIDEr [28] for the policy gradient algorithm to improve text generation accuracy, instead of directly estimating the reward in reinforcement learning.

Rennie et al. [25] used the generated sentence at the test time as the baseline in the objective function, when they trained text generation model directly on nondifferentiable metrics. Guo et al. [26] suggested restricting the action space by an n-gram language prior in order to reduce the learning bias which leads to favorable readability of captions and model stability. The implicit optimization towards the target metric improves the results. All these methods of reinforcement learning aforementioned have been proved to be effectively improving the performance of the image caption.

2.5. Discussion. Models based on visual attention can predict which regions of the image to be focused on but lack the related semantic attributes to describe the attended regions. On the contrary, models based on semantic attention can only learn the attributes in the image but cannot map these attributes to the corresponding objects.

To generate more detailed and accurate captions, we investigate how to exploit the mutual relations between visual features and semantic features, *i.e.*, the mutual relations between objects and their attributes, and finally propose a model to map the attributes to the corresponding objects.

3. Gated Object-Attribute Matching Network

In this section, we present the proposed image captioning model, gated object-attribute matching network, in order to obtain more detailed image captions. In detail, we propose a gated object-attribute matching network to discover the potential relations of objects and high-level semantic attributes and thus incorporate the visual attention module and the high-level semantics attention module into image caption generation.

The entire framework of the caption generation network is as follows:

Step 1: the visual features and semantic features, *i.e.*, objects and the corresponding attributes, are extracted by the object detector and the attribute extractor, respectively

Step 2: firstly, the visual LSTM encodes the extracted objects to h_t^v, while the semantic LSTM encodes the attributes of semantic features to h_t^a
Secondly, the visual attention module maps the objects h_t^v, to their context vector c_t^v, as well as the semantic attention maps the attributes h_t^a, to their context vector c_t^a
Thirdly, a determinant gate ($\otimes$) is involved to map the objects to the corresponding semantic features, and the incorporated representation serves as the attribute-aware visual features V_t^g

Step 3: the second LSTM layer, namely, language LSTM, integrates the attribute-aware visual features V_t^g with the semantic attention c_t^a to obtain description of different objects and generate a more detailed image caption

We first propose our entire framework of the detailed caption generation network, as illustrated in Figure 3. In Section 3.1, we introduce the involved visual features and the corresponding semantic features, *i.e.*, objects and the corresponding attributes. The proposed network has two LSTM layers. The first LSTM layer includes a visual LSTM, a semantic LSTM, and a gated attention-based recurrent network, which is proposed to map the visual features to the corresponding semantic features. The second LSTM layer is language LSTM. In Section 3.2, we introduce the visual LSTM and the semantic LSTM. The visual attention and the semantic attention are in Section 3.3. In Section 3.4, we introduce the gated object-attribute matching. The language LSTM is in Section 3.5. The objective function is explained in Section 3.6.

3.1. Visual and Semantic Features. The original image captioning models [1, 13–15] simply analyse the entire image instead of identifying and locating visual features. In order to generate an accurate image caption, we need to discover the essential visual features in images, *i.e*, objects. We leverage bounding boxes used in faster R-CNN [29] to identify and locate different objects in an image.

Specifically, we select the top-k region proposals from a specific image. For each selected region j containing an object, its corresponding visual feature v_j is composed of region proposals and their locations, through which we can identify and locate different objects in the image. We define the set of concatenation feature vectors v_j as the visual features of an image, $V = \{v_j \mid 1 < j < k\}$, $v_j = [v_j^v, v_j^l]$, $v_j \in \mathbb{R}^D$. The region proposal feature v_j^v is the ROI pooling

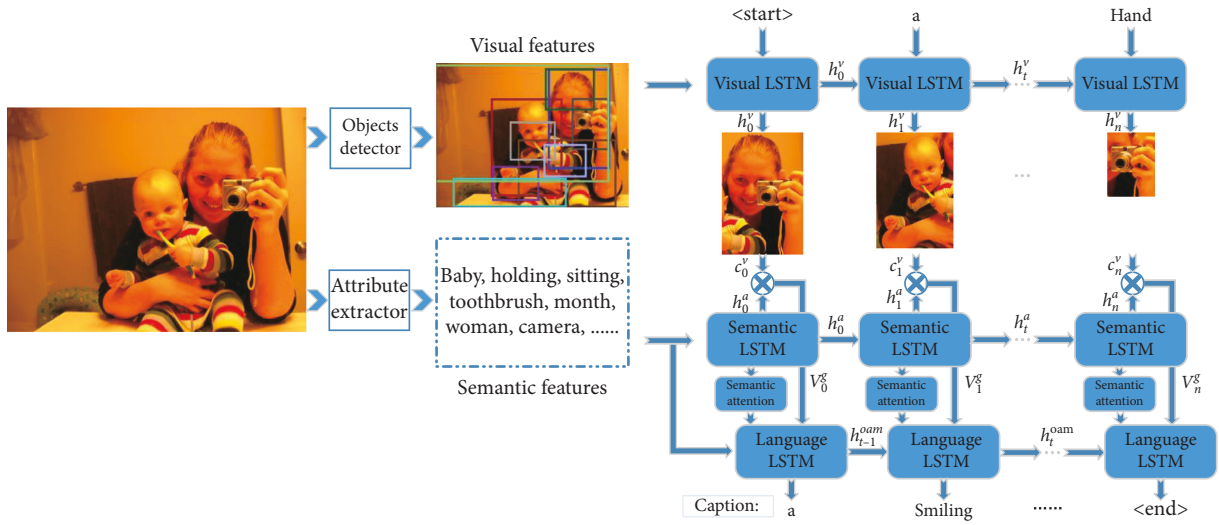

Figure 3: The framework of the proposed gated object-attribute matching network in more detail.

feature, and v_j^l is the location coordinate of the region proposal. And D is the dimension of the visual feature vectors which is the sum of region feature dimension and location coordinate dimension.

One of the limitations of the top-k paradigm is that it is hard to attend to fine details which may be important in terms of describing the image.

To handle this problem, more detailed information about the image should be involved in analysis. As the primary semantic features, attributes in an image contain more detailed information and thus can be used to generate specific description of the image, which do not only include object names but also include motions and properties. Considering that some descriptive attributes such as *beautiful* may not be easily identified with bounding boxes in an image, the classical training method cannot be used as attributes extractor. Actually, the multi-instance learning (MIL) [21] is feasible to be used to construct attribute extractors. To achieve a comprehensive attribute extractor, we first build an attribute vocabulary that contains the most usual words found in the image captions. Therefore, the attributes can describe not only object names but also object motion and properties. Additionally, we use bags to discover attributes in images. Bag $b(i)$ stands for image i, while each instance in this bag stands for region proposal j in image i. Firstly, we let bag $b(i)$ be positive if the attribute is in the caption of image i, and negative otherwise when traversing the attribute vocabulary. Then we train the extractor by iteratively selecting instances from the positive bags, followed by retraining using the updated positive labels. The retraining process calculates the probability of an attribute word w_c in bag i as follows:

$$p^{w_c} = 1 - \prod_{j \in \text{image}(i)} \left(1 - p_j^{w_c}\right), \tag{1}$$

where $p_j^{w_c}$ is the probability of the finding word w_c in the subregion j of image i. We compute the sigmoid function on the dense layer [29] to get the probability $p_j^{w_c}$. The probability $p_j^{w_c}$ is calculated as follows:

$$p_j^{w_c} = \frac{1}{1 + \exp\left[-\left(w_w b_{ij} + b_w\right)\right]}, \tag{2}$$

where b_{ij} is the dense layer representation for the subregion j in image i and the hyperparameter w_w and b_w are updated by cross-entropy loss function. We use the cross-entropy loss as the objective function and optimize the training process with the stochastic gradient descent.

Let c be the number of attributes, and the semantic feature set A can be defined as the probability distribution vector:

$$A = \{p^{w_1}, p^{w_2}, \ldots, p^{w_c}\}. \tag{3}$$

3.2. Visual LSTM and Semantic LSTM. As illustrated in Figure 4, our image captioning model is composed of two LSTM layers as follows [17]: the first LSTM layer is the most important component of the model, including a visual attention-based LSTM, a semantic attention-based LSTM, and a determinant gate for matching objects and attributes. The second LSTM layer is a language module to generate a caption. In this section, we begin to introduce the visual attention-based LSTM and the semantic attention-based LSTM used in the first LSTM layer.

A visual attention-based LSTM is involved in our image captioning model to discover the potential relations among objects. As depicted in Figure 4, at step t, the input vector of the visual attention-based LSTM comprises its last step output h_{t-1}^v, the average value of visual features $\overline{v}$, $\overline{v} = (1/k)\sum_{i=1}^{k} v_i, v_i \in V$, and the input word x_t. Word x_t has been represented as word embedding of dimension equal to the size of word dictionary. In the training stage, x_t is the t^{th} word that comes from the image caption at step t. In the testing stage, x_t is the word generated by LSTM at the last step. The concatenate input vector is fed into the visual LSTM to produce the vector h_t^v according to the following formulas:

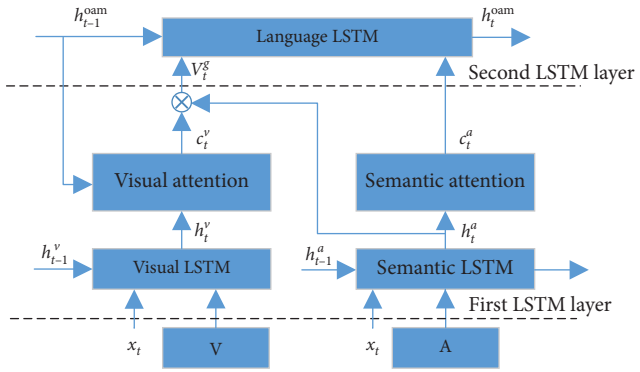

FIGURE 4: An illustration of the proposed gated object-attribute matching network model.

$$
\begin{aligned}
I_t^v &= \sigma\left(x_t W_{xi}^v + \overline{v} W_{vi}^v + h_{t-1}^v W_{hi}^v + b_i^v\right),\\
F_t^v &= \sigma\left(x_t W_{xf}^v + \overline{v} W_{vf}^v + h_{t-1}^v W_{hf}^v + b_f^v\right),\\
O_t^v &= \sigma\left(x_t W_{xo}^v + \overline{v} W_{vo}^v + h_{t-1}^v W_{ho}^v + b_o^v\right),\\
\widetilde{C}_t^v &= \tanh\left(x_t W_{xc}^v + \overline{v} W_{vc}^v + h_{t-1}^v W_{hc}^v + b_c^v\right),\\
C_t^v &= F_t^v \odot C_{t-1}^v + I_t^v \odot \widetilde{C}_t^v,\\
h_t^v &= O_t^v \odot \tanh\left(C_t^v\right).
\end{aligned}
\tag{4}
$$

Correspondingly the gates are defined as follows: the input gate is I_t^v, the forget gate is F_t^v, and the output gate is O_t^v. $W_{xi}^v, W_{xf}^v, W_{xo}^v$, $W_{vi}^v, W_{vf}^v, W_{vo}^v$ and $W_{hi}^v, W_{hf}^v, W_{ho}^v$ are the weight parameters, and b_i^v, b_f^v, b_o^v are the bias parameters. The candidate memory cell is $\widetilde{C}_t^v$, and the memory cell is C_t^v. Here, $W_{xc}^v, W_{vc}^v, W_{hc}^v$ are the weights, and b_c^v is a bias. The hidden state h_t^v is then fed into visual attention at the next step.

In addition, a semantic attention-based LSTM is involved in our image captioning model to identify the relations of attributes. At step t, the input of the semantic LSTM consists of its last step output h_{t-1}^a, the attributes vector A, and the input word embedding vector x_t. Its output h_t^a is calculated by

$$
\begin{aligned}
I_t^a &= \sigma\left(x_t W_{xi}^a + A W_{ai}^a + h_{t-1}^a W_{hi}^a + b_i^a\right),\\
F_t^a &= \sigma\left(x_t W_{xf}^a + A W_{af}^a + h_{t-1}^a W_{hf}^a + b_f^a\right),\\
O_t^a &= \sigma\left(x_t W_{xo}^a + A W_{ao}^a + h_{t-1}^a W_{ho}^a + b_o^a\right),\\
\widetilde{C}_t^a &= \tanh\left(x_t W_{xc}^a + A W_{ac}^a + h_{t-1}^a W_{hc}^a + b_c^a\right),\\
C_t^a &= F_t^a \odot C_{t-1}^a + I_t^a \odot \widetilde{C}_t^a,\\
h_t^a &= O_t^a \odot \tanh\left(C_t^a\right).
\end{aligned}
\tag{5}
$$

The computation in LSTM_a is similar to LSTM_v. Here, $W_{xi}^a, W_{xf}^a, W_{xo}^a, W_{xc}^a$, $W_{ai}^a, W_{af}^a, W_{ao}^a, W_{ac}^a$ and $W_{hi}^a, W_{hf}^a, W_{ho}^a, W_{hc}^a$ are the weight parameters, and $b_i^a, b_f^a, b_o^a, b_c^a$ are the bias parameters.

3.3. Visual Attention and Semantic Attention. A visual attention module and a semantic attention module are applied to discover the objects with high significance values, *i.e.*, the primary elements with significant relations with the others. In detail, the encoded visual content h_t^v is utilized to identify the relevant objects of the word x_t by using a soft attention mechanism. Additionally, the context vector of the objects at step t is obtained based on h_t^v and the visual feature set V:

$$
\begin{aligned}
\alpha_{t,i}^v &= \mathrm{soft\,max}\left(e_{t,i}^v\right)\\
&= \frac{\exp\left(e_{t,i}^v\right)}{\sum_{t'=1}^k \exp\left(e_{t,t'}^v\right)},\\
e_{t,i}^v &= W_s^a \tanh\left(W_v^v v_i + W_h^v h_t^v\right),\\
c_t^v &= \sum_i \alpha_{t,i}^v v_i, \quad v_i \in V,
\end{aligned}
\tag{6}
$$

where c_t^v is the context vector, $\alpha_{t,i}^v$ is the weight of the object i at step t, $t' = 1, \ldots, k$, and W_s^v, W_h^v, W_v^v are the hyper-parameters to be learned.

Similarly, the context vector of semantic attention, c_t^a, is used to identify the significant attributes of the image, which is calculated in terms of h_t^a and the semantic feature set A, as follows:

$$
\begin{aligned}
\alpha_{t,i}^a &= \mathrm{soft\,max}\left(e_{t,i}^a\right)\\
&= \frac{\exp\left(e_{t,i}^a\right)}{\sum_{t'=1}^k \exp\left(e_{t,t'}^a\right)},\\
e_{t,i}^a &= W_s^a \tanh\left(W_v^a v_i + W_h^a h_t^a\right),\\
c_t^a &= \sum_i \alpha_{t,i}^a a_i, \quad a_i \in A.
\end{aligned}
\tag{7}
$$

3.4. Gated Object-Attribute Matching. In order to map the attributes to the corresponding objects and provide the detailed description of the objects in an image, we match each object with its salient context elements, including the related objects and attributes, by using a gated object-attribute matching network. This gated object-attribute matching network is based on an attention gate [30]. This attention gate is based on both of the objects and their semantic attributes, instead of only relying on one type of elements, just like what GRU [31] does. The attention gate V_t^g bases on the current context vector of visual c_t^g and the encoded semantic features h_t^g:

$$
V_t^g = \mathrm{sigmoid}\left(W^g\left[h_t^a, c_t^v\right]\right) \otimes \left[h_t^a, c_t^v\right],
\tag{8}
$$

where V_t^g is the attribute-ware objects, regarded as the high-level visual features after matching the corresponding sematic features. $\otimes$ is an element-wise product operation. W^g is a weight matrix to be learned.

3.5. Language LSTM. The second LSTM layer is a language LSTM for generating image captions. The language LSTM utilizes the high-level visual features V_t^g and the context vector of semantic attention c_t^a, to generate the image

captions. The objects in the image at step t and their relevant attributes are encoded to vector h_t^{oam} as follows:

$$h_t^{\text{oam}} = \text{LSTM}\left(h_{t-1}^{\text{oam}} v_t^g c_t^a\right). \tag{9}$$

This final representation h_t^{oam} contains the entire information of the location of the salient objects and their corresponding properties. Finally, while generating a caption of an image, each word in the generated caption is picked up by using a softmax classifier, in terms of the input word x_t and the high-level representation h_t^{oam} in the last step:

$$p\left(w_t \mid w_{1:t-1}, I\right) = \text{soft max}\left(W_x x_t + W_h h_t^{\text{oam}} + b\right), \tag{10}$$

where W_x and W_h are the hyperparameters to be learned.

3.6. Objective. The aim of an image captioning model is to maximize the probability that the generated description is similar to a realistic image caption. Therefore, the objective function of our network minimizes the difference between realistic image captions and the generated image captions, *i.e.*, to minimize the following cross-entropy function:

$$£(\theta) = -\sum_t \log p\left(w_t \mid w_{1:t-1}, (V, A)\right), \tag{11}$$

where θ is the set of all the hyperparameters to be learned in the training stage. V is the visual feature set, $V = \{v_{1:k}\}$. And A is the semantic feature set of image attributes. $p(w_t \mid w_{1:t-1})$ is the probability that the word w_t appears behind the word sequence $w_{1:t-1}$.

Considering the significant performance of reinforcement learning (RL) to deal with the loss-evaluation mismatch problem, we optimize our image captioning network by involving an RL-based training process.

According to the cross-entropy training model initialized in the previous step, we obtain the negative expectation score and try to minimize it:

$$£(\theta_r) = -E_{w1:T\sim\theta_r}(\gamma(w_{1:T})), \tag{12}$$

where γ is an evaluation score of some metric, *e.g.*, CIDEr, which serves as a reward of policy gradient for reinforcement learning. We update the gradient of the objective function as follows:

$$\nabla_{\theta_r} £(\theta_r) \approx -(\gamma(w_{1:t}) - gt)\nabla_{\theta_r} \log p_{\theta_r}(w_{1:T}, (V, A)), \tag{13}$$

where gt is the ground truth that serves as a reference reward. Then the policy gradient for reinforcement learning can calculate this reward. The policy gradient for reinforcement learning improves the evaluation metric score of the generated description of the image captioning model and improves the accuracy of the generated captions.

3.7. Framework for Image Captioning. The gated object-attribute matching network is the end-to-end model. The training procedure is illustrated in Figure 3. Given an image and its corresponding caption, our model maximizes the probability of word sequence:

$$\theta = \arg\max_\theta \sum_{(I,y)} \log p(y \mid I; \theta), \tag{14}$$

where θ represents the model parameters and I is the image and is the word sequence of corresponding caption. Based on the chain rule, the log likelihood of the joint probability distribution over y is comprised of T conditional probabilities:

$$\log p(y) = \sum_{t=1}^{T} \log p\left(y_t \mid y_{t-1}, \ldots, y_1, I\right), \tag{15}$$

where T is the total length of the caption. Here, the dependency on model parameters θ is removed for convenience. During the training stage, (I, y) is a training image caption pair, and the overall optimization objective is the sum of log probabilities over all training pairs in the training set. Firstly, the visual features are extracted by the object detectors from the image. The object detectors also locate image regions of the objects. So, we get visual feature vector $V = \{v_j \mid 1 < j < k\}, v_j = [v_j^v, v_j^l], v_j \in \mathbb{R}^D$ by the object detectors. Meanwhile, the semantic features are extracted by the attributes extractors which obtain descriptive words from the image captions in training sets to collect all types of attributes. Then, we get semantic features vector $A = \{p^{w_1}, p^{w_2}, \ldots, p^{w_c}\}$ by the attributes extractors.

Secondly, the visual LSTM encodes the visual features V to h_t^v; meanwhile, the semantic LSTM encodes the semantic features A to h_t^a. Then, the visual attention maps h_t^v to context vector c_t^v, as well as the semantic attention maps h_t^a, to context vector c_t^a. Next, a determinant gate ($\otimes$) is involved to map the objects to the corresponding semantic features, and the incorporated representation servers as the attribute-aware visual features V_t^g.

Thirdly, the language LSTM integrates the attribute-aware visual features V_t^g with their semantic context vector c_t^a to generate the corresponding caption. To minimize the difference between realistic image captions and the generated image captions, we use the cross-entropy function as the objective function. We also optimize the objective function by involving an RL-based training process.

4. Experiments

4.1. Datasets. In order to evaluate the effectiveness of the proposed model, we select the most popular datasets for the image captioning, including Flickr30k [32] and MSCOCO [33]. The Flickr30k dataset contains about 31,000 images, each with 5 sentences annotated to describe the image as a reference, while the MSCOCO dataset contains about 123,000 images and each with 5 sentences also. We follow the public available splits [2], remaining 1,000 images for validation, 1,000 for testing, and others for training on Flickr30k, while 5,000 images remain for validation as well as 5,000 images for testing on MSCOCO.

For representing the sentence in word embedding, we fixed vocabulary size to 10,000 for both datasets including special start sign $< \text{BOS} >$ and end sign $< \text{EOS} >$.

Table 1: The performance of the state-of-the-art image captioning models on the Flickr30k and MSCOCO testing splits.

Model	Flickr30k						MSCOCO					
	B@1	B@2	B@3	B@4	METEOR	CIDEr	B@1	B@2	B@3	B@4	METEOR	CIDEr
DeepV-SAlign [7]	0.573	0.369	0.240	0.157	0.153	0.247	0.625	0.450	0.321	0.230	0.195	0.660
Soft-Attention [26]	0.667	0.434	0.288	0.191	0.185	—	0.707	0.492	0.344	0.243	0.239	—
Hard-Attention [26]	0.669	0.439	0.296	0.199	0.185	—	0.718	0.504	0.357	0.250	0.230	—
Attribute-FCN [15]	0.647	0.460	0.324	0.230	0.189	—	0.709	0.537	0.402	0.304	0.243	—
Adaptive-Attention [11]	0.677	0.494	0.354	0.251	0.204	0.531	0.742	0.580	0.439	0.332	0.266	1.085
Attribute-CNN + LSTM [14]	—	—	—	—	—	—	—	0.56	0.42	0.31	0.26	0.94
NBT [13]	0.720	—	—	0.285	0.231	0.575	0.759	—	—	0.349	0.274	1.089
Up-Down [12]	—	—	—	—	—	—	0.802	0.641	0.491	0.369	0.276	1.179
Ours (Box proposed)	0.711	0.507	0.393	0.266	0.211	0.630	0.753	0.592	0.462	0.341	0.266	0.954
Ours (RL)	0.735	0.522	0.401	0.297	0.219	0.674	0.772	0.620	0.476	0.352	0.270	1.098

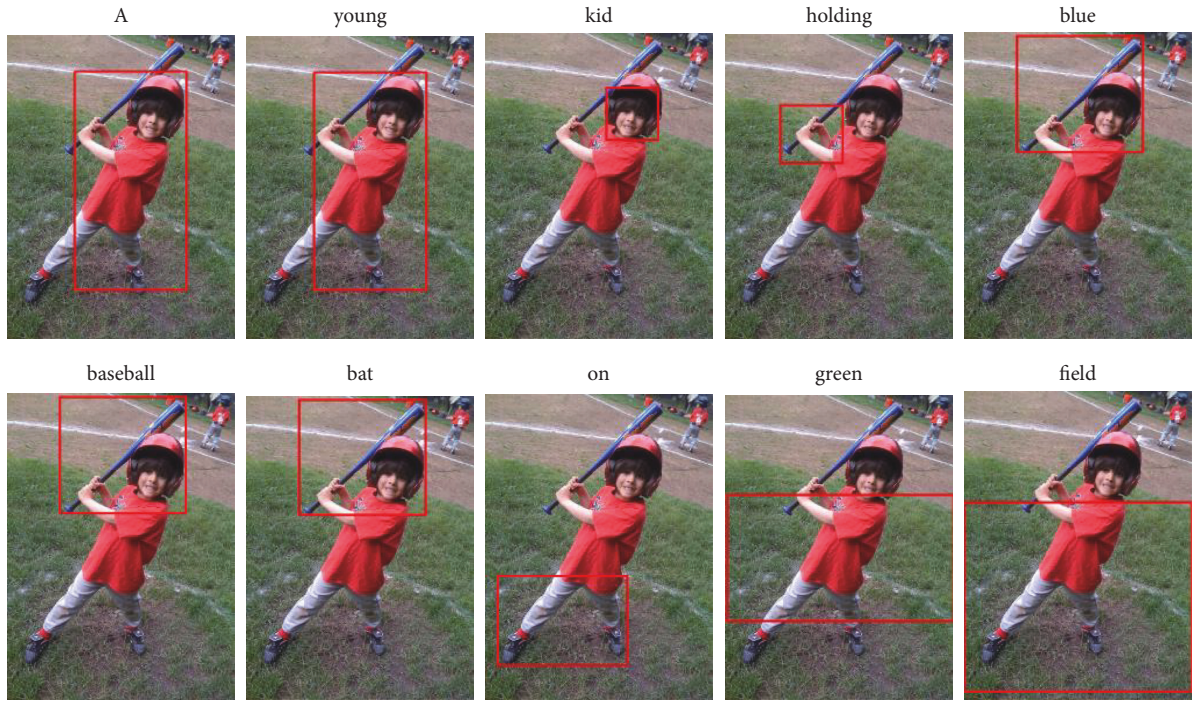

Figure 5: A visualized example of the localized objects corresponding to the detected semantic features.

4.2. Implementation Details. While implementing our model, ResNet-101 [34] is used as the pretrain model for feature map layers. Then we choose faster R-CNN [29] as the object detector to get the proposal regions in an image. To get the primary objects and their location, we fix the IoU threshold on 0.7 for nonmaximum suppression and finally select the top-k region proposals. For each selected region i, its corresponding visual feature $v_i = [v_i^v, v_i^l]$, where v_i^v is a region feature obtained with a max-pooling feature map from ROI align and v_i^l is the coordinates of the bounding box, $x_{\text{left}}, y_{\text{top}}, x_{\text{right}}$, and y_{bottom}. Let the width and height of the image be W and H, respectively. The normalized coordinates of the bounding box is $x_{\text{left}}/W, y_{\text{top}}/H, x_{\text{right}}/W$, and y_{bottom}/H.

In addition, the number of hidden nodes is 1024, the word embedding size is 512, the dropout ratio is 0.5, the minibatch size is 50, and the model training process lasts 50 epochs.

4.3. Evaluation

4.3.1. Metrics. BLEU [35] is a metric used for evaluating machine translation algorithms based on n grams. We report BLEU score on BLUE@1, BLEU@2, BLEU@3, and BLEU@4. METEOR [36] evaluates the word alignment between the generated captions and the realistic captions by considering accuracy and recall rate. Different from other metrics, CIDEr is specially proposed for evaluating the accuracy of different image captioning models by using TF-IDF (term frequency-inverse document frequency). We compare our model with the state-of-the-art methods in terms of all the above metrics.

4.3.2. Compared Methods. We evaluate the accuracy of our proposed image captioning model by comparing our model with the state-of-the-art methods, including

(a) (b) (c)

Figure 6: A visualized example of the localized objects corresponding to the detected semantic features. (a) GT: a group of sheep are standing together in the grass. Ours: a group of sheep are gazing on the green grass field. (b) GT: a cat is sitting on the bench. Ours: a black cat laying down on a stone bench. (c) GT: a woman with an umbrella is walking down the street. Ours: A woman walks down the street holding a green umbrella.

DeepV-SAlign [2], Soft-Attention [6], Hard-Attention [6] and recently proposed Attribute-FCN [9], Adaptive-Attention [7], Attribute-CNN + LSTM [8], NBT [18], and Up-Down [17].

4.3.3. Results Analysis: Quantitative Analysis. The comparison results of our model and the compared methods on Flickr30k and MSCOCO are listed in Table 1. Since it can improve the quality of the generated captions to involve RL into our model, we present our results without RL and with RL, respectively, for a fair comparison.

As listed in Table 1, our proposed model outperforms all compared models. Since Attribute-FCN and Attribute-CNN + LSTM used semantic attributes in image captioning, the two models achieve significant performance improvements. Compared with Attribute-CNN + LSTM, our model improves BLEU-4 from 0.31 to 0.366, METEOR from 0.26 to 0.279, and CIDEr from 0.94 to 1.185 on COCO. Our model also outperforms the Up-Down model, improving METEOR from 0.276 to 0.279, and CIDEr from 1.179 to 1.185 on COCO. The Up-Down model combines bottom-up and top-down attention mechanisms and obtains the best performance than other compared models. Similarly, on Flickr30k, our model obtains the best evaluation results on all metrics.

4.4. Quantitative Analysis. In Figure 5, we visualize the process that the model focuses on different objects when generating the corresponding caption. Our model finds the most salient region proposals for the corresponding attributes with the highest significance value.

To demonstrate the performance of our image captioning model, we present some images and the corresponding generated captions in Figure 6, which are selected from the public dataset, MSCOCO. GT stands for the ground truth caption. As illustrated in Figure 6, we can see that our proposed model can detect more detailed attributes for the image. As for the first image, our model generates a caption "A group of sheep are gazing on the green grass." Compared with the caption provided in MSCOCO, gazing and green are included in the generated caption. A more detailed image caption is generated.

This indicates that by localizing and describing objects in a gated object-attribute matching network, and the generated caption can be more detailed as well as specific.

5. Conclusion

In this paper, we study the two types of image caption models, *i.e.*, the visual attention-based neural network and the semantic attention-based neural network. To address the challenge of accurately incorporating semantic attributes into visual attention-based networks, we explore the mutual relations of objects and attributes and improve the quality of the generated captions. In detail, we propose a novel gated object-attribute matching network, which is used to match the objects to their semantic attributes.

The experiments on the public datasets of Flickr30k and MSCOCO demonstrate our model can obtain more detailed image captions compared with many other methods.

Data Availability

The datasets used in our research are the public datasets. We get the datasets from the following sites: (1) The dataset of MSCOCO http://cocodataset.org/. (2) The dataset of Flickr30k https://forms.illinois.edu/sec/229675.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Natural Science Foundation of Education Department Inner Mongolian Autonomous Region (NJZY19083).

References

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Boston, MA, USA, October 2015.

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pp. 3128–3137, Boston, MA, USA, June 2015.

[3] R. Bernardi, R. Cakici, D. Elliott et al., "Automatic description generation from images: a survey of models, datasets, and evaluation measures," *Journal of Artificial Intelligence Research*, vol. 55, pp. 409–442, 2016.

[4] G. Kulkarni, V. Premraj, V. Ordonez et al., "Babytalk: understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[5] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," 2015, https://arxiv.org/abs/1506.06272.

[6] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *Proceedings of the International Conference on Machine Learning*, pp. 2048–2057, Lille, France, July 2015.

[7] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 375–383, Honolulu, HI, USA, July 2017.

[8] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 203–212, Las Vegas, NV, USA, June 2016.

[9] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659, Las Vegas, NV, USA, June 2016.

[10] Z. Gan, C. Gan, X. He et al., "Semantic compositional networks for visual captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5630–5639, Honolulu, HI, USA, July 2017.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," 2014, https://arxiv.org/abs/1406.1078.

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

[13] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on Machine Learning*, pp. 595–603, Beijing, China, June 2014.

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

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634, Boston, MA, USA, June 2015.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," 2014, https://arxiv.org/abs/1412.6632.

[17] P. Anderson, X. He, C. Buehler et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.

[18] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7219–7228, Salt Lake City, UT, USA, June 2018.

[19] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4894–4902, Venice, Italy, October 2017.

[21] H. Fang, S. Gupta, F. Iandola et al., "From captions to visual concepts and back," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482, Boston, MA, USA, June 2015.

[22] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," 2015, https://arxiv.org/abs/1511.06732.

[23] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *Proceedings of the IEEE international Conference on Computer Vision*, pp. 873–881, Venice, Italy, October 2017.

[24] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 290–298, Honolulu, HI, USA, July 2017.

[25] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, Honolulu, HI, USA, July 2017.

[26] T. Guo, S. Chang, M. Yu, and K. Bai, "Improving reinforcement learning based image captioning with natural language prior," 2018, https://arxiv.org/abs/1809.06227.

[27] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: semantic propositional image caption evaluation," in *European Conference On Computer Vision*, pp. 382–398, Springer, Berlin, Germany, 2016.

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, Boston, MA, USA, June 2015.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural information Processing Systems*, pp. 91–99, London, UK, 2015.

[30] S. Wang and J. Jiang, "Machine comprehension using match-lstm and answer pointer," 2016, https://arxiv.org/abs/1608.07905.

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, https://arxiv.org/abs/1412.3555.

[32] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[33] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, Berlin, Germany, 2014.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pp. 770–778, Vegas, NV, USA, June 2016.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, Association for Computational Linguistics, Philadelphia, PA, USA, July 2002.

[36] M. Denkowski and A. Lavie, "Extending the meteor machine translation evaluation metric to the phrase level," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 250–253, Association for Computational Linguistics, Stroudsburg, PA, USA, 2010.