Adaptive Attention-based High-level Semantic Introduction for Image Caption

XIAOXIAO LIU and QINGYANG XU, School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, China

There have been several attempts to integrate a spatial visual attention mechanism into an image caption model and introduce semantic concepts as the guidance of image caption generation. High-level semantic information consists of the abstractedness and generality indication of an image, which is beneficial to improve the model performance. However, the high-level information is always static representation without considering the salient elements. Therefore, a semantic attention mechanism is used for the high-level information instead of conventional of static representation in this article. The salient high-level semantic information can be considered as redundant semantic information for image caption generation. Additionally, the generation of visual words and non-visual words can be separated, and an adaptive attention mechanism is employed to realize the guidance information of image caption generation switching between new fusion information (fusion of image feature and high-level semantics) and a language model. Therefore, visual words can be generated according to the image features and high-level semantic information, and non-visual words can be predicted by the language model. The semantics attention, adaptive attention, and previous generated words are fused to construct a special attention module for the input and output of long short-term memory. An image caption can be generated as a concise sentence on the basis of accurately grasping the rich content of the image. The experimental results show that the performance of the proposed model is promising for the evaluation metrics, and the captions can achieve logical and rich descriptions.

CCS Concepts: • **Computing methodologies** → **Vision for robotics**; **Scene understanding**;

Additional Key Words and Phrases: Image caption, high-level semantic, adaptive attention, CNN, LSTM, visual sentinel

ACM Reference format:
Xiaoxiao Liu and Qingyang Xu. 2020. Adaptive Attention-based High-level Semantic Introduction for Image Caption. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 4, Article 128 (December 2020), 22 pages.
https://doi.org/10.1145/3409388

This work was supported by the National Natural Science Foundation of China under Grants No. 61573213, No. 61803227, No. 61603214, and No. 61673245; the National Key Research and Development Plan of China under Grant No. 2017YFB1300205, the Shandong Province Key Research and Development Plan under Grants No. 2018GGX101039 and No. 2016ZDJS02A07, and the China Postdoctoral Science Foundation under Grant No. 2018M630778.
Authors' addresses: X. Liu and Q. Xu (corresponding author), School of Mechanical, Electrical & Information Engineering, Shandong University, Weihai, Shandong, 264209, China; emails: sdulxx@163.com, qingyangxu@sdu.edu.cn.

1551-6857/2020/12-ART128 $15.00
https://doi.org/10.1145/3409388

1 INTRODUCTION

Image captions comprise a research hotspot in the field of image understanding and can be used for subtitle translation, robots, and so on. Humans can effortlessly detect objects and the connections between two objects in an image and convert the image to natural language, which depends on a human's strong semantic understanding capability of the surrounding environment and visual concepts. Despite the rapid development of artificial intelligence techniques, especially various deep learning algorithms, the conversion between visual task and linguistic task remains a challenge in the field of computer vision. To describe an image by natural language, the objects in the image should be captured, and how the objects are related as well as their respective attributes should be described. Therefore, an image caption algorithm always includes image recognition and natural language processing modules. In a conventional image model, the image features are considered the source of words generation and a language sequence will be output.

The study was motived by the problem of comprehensively describing an image by language. Our study involves the application of integrating attention with high-level semantics in image caption. Sematic information application has attracted an increasing amount of attention in recent years [1, 6]. How to extract and introduce high semantics to an image caption model and guide language generation is challenging work. Some researchers have connected semantic information with the region in an image caption model [7] and assigned different levels of attention to different regions of an image [8]. For example, Wu et al. verified the importance of concepts in image caption and Visual Question Answering in Reference [1]. The introduction of semantics directly explains the contents of an image, which can guide a language model to generate a more comprehensive description. These studies have inspired us to introduce adaptive attention mechanism for high-level semantics to improve the performance of an image caption model. The attention mechanism is a unique human vision sensing mechanism that enables humans to quickly and efficiently perceive important information about the surrounding environment and adjust the distribution of attention in the next step based on previous information. Therefore, the attention mechanism can select the salient elements of high-level semantics that represent the image concept. Some information that was originally disregarded by an image recognition module can be presented to a language module in the form of high-level semantics to generate a more comprehensive description. High-level semantic information has stronger abstractedness and generality about the image. The images in our study are processed by a convolutional neural network (CNN). The high-level sematic information derives from the refinement of dense captions. In image caption generation, image features and high-level semantic information are processed by adaptive attention module and then sent to language model to guide the word generation, which provides more intuitive information that can be used with the image feature vector for image caption generation. The total structure of the model is shown in Figure 1. In the image caption generation process, the image feature vector, different attention levels of high-level semantics, and previous descriptions are gathered by the model. Therefore, the image content can be fast and accurately mastered by the image caption model.

Our study introduces an adaptive attention mechanism to high-level semantics to improve image caption model. The purpose is to investigate the specific effects of high-level semantic elements by attention mechanism in image caption generation. The contributions of this study are described as follows:

(i) High-level semantic information is a summary description of images and is more intuitive than an image feature vector. High-level semantics can be considered as information to be used in addition to image feature vectors to guide image caption generation. Therefore, high-level semantic information is extracted by refining dense captions. Additional

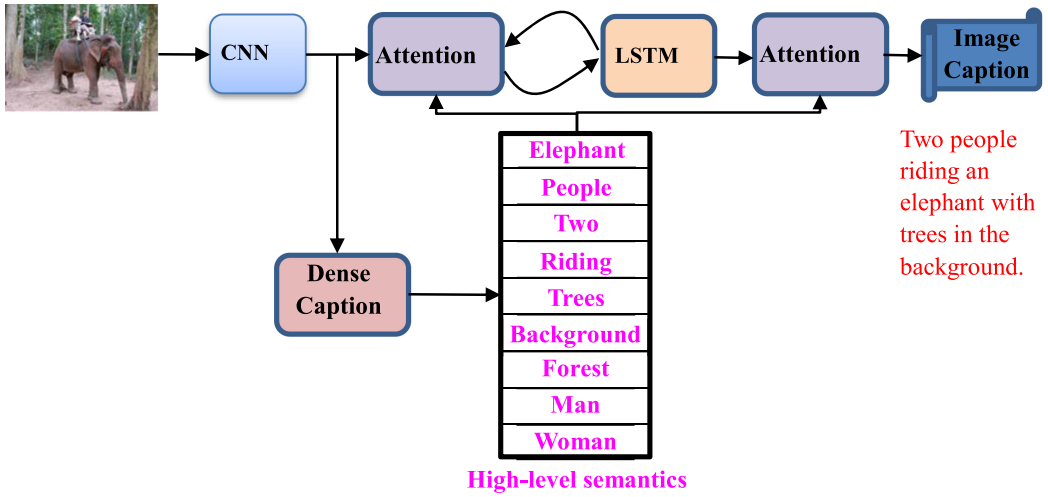


Fig. 1. Image caption generation based on adaptive attention mechanism.

abstractedness information inputs an image caption model for comprehensive image caption generation.

(ii) An adaptive attention mechanism is adopted for high-level semantics, which is the input and output of long short-term memory (LSTM). The attention distribution of high-level semantics can dynamically determine the salient semantic elements according to image features and previously generated words, which can reflect the concern issue of the model.

(iii) The language sentence always contains some non-visual words and common fixed collocation. A visual sentinel is adopted to determine the reliance information of the model and guide the switch between using new visual information and using a language model to generate the caption.

Compared with other image caption models, the performance of image caption generation is improved, and the description of an image is more comprehensive. This study can provide readers with as much information as possible to introduce attention mechanism to high-level sematic information to improve the performance of image caption models.

The remainder of this article is organized as follows: Section 2 introduces the motivation and related work on the image caption model design. In Section 3, attention-based high-level semantic introduction is described. Section 4 provides the simulation experimental settings and results. In Section 5, we conclude this article.

2 RELATED WORKS

Image Caption. With the development of deep learning techniques and the emergence of a variety of sophisticated image caption data sets, image caption algorithms have rapidly developed. Deep neural networks are extensively employed in the automatic generation of image captions due to their powerful approximation ability. With the application of deep learning, CNNs and LSTM have been successfully applied to image caption generation, which can be divided into three categories: retrieval-based image caption, template-based image caption, and end-to-end model-based image caption [9]. Previously, an image caption was implemented by the retrieval-based method. For the retrieval-based method, the visual words can be obtained according to features, including objects

(nouns), attributes (adjectives), actions (verbs), and backgrounds, and then the detected words are sent to an LSTM language model to generate a description [8, 10, 15]. The model is trained by a large amount of online resources to extract the semantic information of the seed images in advance, and then an image similarity retrieval method is adopted to derive the semantics of the new image. These methods always include a generalization strategy to attenuate or remove some details, which are only associated with the retrieved image. However, the training data set should be complete. A new objects description that does not exist in the training dataset is almost impossible.

Due to the limitations of retrieval-based image captions, the template-based image caption method is proposed, which divides an image caption into two subtasks: an object detection task that is based on a computer vision algorithm to gather the words of an image, such as objects, attributes, actions, and relationships, which may appear in a caption, and then use the language model to describe the detected objects by natural language [16, 19]. The object detection and language modules included in the image caption model are independently trained. Template-based methods for image caption generation are highly flexible and can improve the grammatical accuracy of sentences. However, the expression of a sentence is rigid, because the sentences strictly adhere to the grammatical rules of a template. Since an object is independently detected, image caption generation that is based on a language model disregards the relationship between two objects, and some logical confusions always occur.

The two-step method cannot realize the end-to-end training and prediction of image captions. Therefore, image caption generation loses sight of the image and is only based on the language model. The description always includes wrong logic for the image. Therefore, the end-to-end method is proposed. The end-to-end model aims to generate a description of an image in one step. An image feature vector is extracted by a convolutional neural network and sent directly to an LSTM language model to generate the description [16, 23]. This model is inspired by machine translation and a sequence-to-sequence model, and a CNN is used as an encoder to extract an image feature as fixed-length vectors and then send it to the LSTM decoder to generate sentences. The end-to-end learning method can directly realize image-to-sentence mapping. The parameters of image processing and language modules can be directly trained by an algorithm. This image caption model is similar to the machine translation model. In 2015, Google researchers proposed an end-to-end image caption model—neural image captioning (NIC)—based on the principle of machine translation [11]. The NIC model combines CNNs and RNNs (recurrent neural networks) to directly generate a caption based on an image. The deep CNN (DCNN) is used as an encoder to read an image and extract the feature of an image as the input of the RNN for image understanding, which is similar to the machine translation model. The decoder RNN uses this vector as the initial value of the hidden layer for the sequence generation. Kiros et al. constructed a joint multimodal mapping space using deep neural networks and LSTM [24]. Mao et al. proposed a multimodal recurrent neural network (mRNN) to directly process the visual input at each time step [25]. The probability distribution of the next generated word is estimated based on the image features extracted by the deep CNN and previous word. Vinyals et al. [11] and Donahue et al. [14] employed a more powerful LSTM as a language model and improved the performance of the image caption model. The end-to-end image caption model combining a CNN and RNN, which are the two core technologies of computer vision and natural language processing, has substantially improved the performance of the image caption model and become the most commonly employed image caption model. With further requirements of accuracy improvement, Li et al. proposed the concept of a dense caption [26]. A complete convolution location network is used to detect additional details of an image in this model, and then the region description is generated. The end-to-end model implements a direct mapping of an image to caption and achieves a multimodal information transformation. This model is the most prevalent and best-performing model; however, some important

details of an image can be easily disregarded, which produces an incomprehensive caption. With advancements in research, various improved end-to-end models, such as attention mechanism and context fusion, have emerged to improve the accuracy of description.

Attention and high-level-based image caption. Attention mechanisms are becoming a popular strategy in deep-learning techniques, which enable a model to dynamically focus on parts of the input. In 2015, Xu et al. [8] introduced an attention-based image caption model that can generate the words for the focused parts of an image. Pedesoli et al. [27] proposed a region attention-based image caption model that associated the regions of an image with the words of a caption. Lu et al. [28] proposed an adaptive attention-based image caption model, and a visual sentinel was introduced to determine the attention allocation for visual and language model information. Tavakoli et al. [29] proposed a bottom-up attention-based image caption model to focus on the important objects first and then focus on the remaining objects. These image caption models adopt the spatial attention of image as the input. Visual and sematic information are always complementary. Visual information can strengthen the visual words that appear in a caption. The sematic information introduction can balance visual words with non-visual words. Recently, a series of researches have incorporated semantic information and attention mechanisms into image caption, which achieves a remarkable improvement in the quality of captions. Wu et al. [1] incorporated the high-level concepts into CNN-RNN instead of visual features at 2016, and the concepts came from the classification results of sub-regions. Wu et al. [3] also proposed a similar sematic concept-based video caption model that had different high-level concepts extraction rule. The high-level information came from the detection of image, image frame captions, and an external knowledge. The introduction of sematic information to a CNN-LSTM-based encoder-decoder has achieved substantial improvement in the area of image captions. However, the style of sematic information introduction and the gathering of sematic information are also challenging. Zhang et al. [30] made used of object detection module to predict local descriptions of these regions, which were used as the semantic feature to guide image caption generation. Jin et al. [31] proposed an attention-based image caption model to extract abstract information according to the semantic relationship between visual information and textual information. You et al. [2] proposed a semantic-based image caption model. The visual concepts are gathered by an image retrieval technique. The image features are extracted and then converted to words. Karpathy et al. [32] proposed a multimodal image caption model for the image regions and sentences modalities. The alignments of the modalities improve image caption generation. Yao et al. [33] introduced the attributes of an image as semantic information to improve the performance of an image caption model. Wang et al. [6] proposed a new concept to bridge the affective gap between image caption and emotions. Wu et al. [7] proposed a spatial attention optimization approach based on sematic-related regions to overcome inaccurate semantic guidance, and a Consensus Selection (CS) strategy based on semantic similarity was adopted for caption selection. Zhang et al. [34] proposed a sematic-guided image caption model, and the sematic information is obtained from the pixel-level labels of an FCN (fully convolutional networks). The visual fused with semantic information to generate joint context information for word generation. Ding et al. [35] proposed two types of attention mode stimulus and concept driven attention for visual question answering (VQA). The visual features concatenated with embedding vector of question as the joint embedding information for output classifier. Shen et al. [36] proposed multimodal stochastic recurrent neural networks (MS-RNNs) to model the uncertainty observed in the video caption using latent stochastic variables to improve the performance of video caption. Another research aLSTMs [37] integrated attention mechanism with LSTM to capture salient features of the video for video caption generation. However, most of the studies convert the semantic information to a static representation without consider attention mechanism, and some studies lacks of adaptive attention mechanism for visual and non-visual words

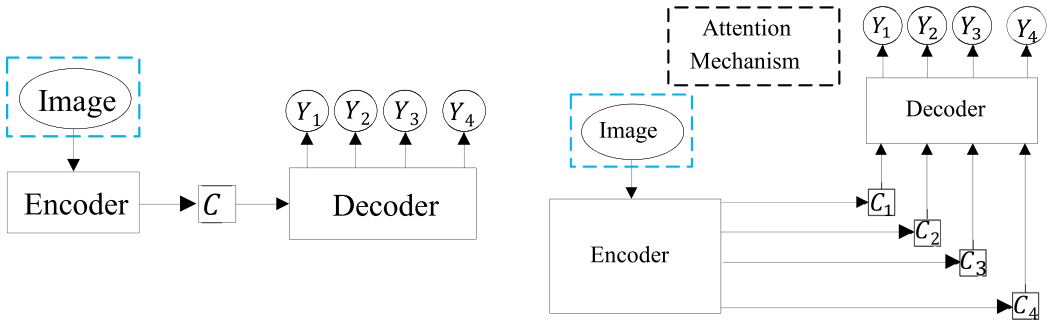


Fig. 2. Diagram of end-to-end model.

generation. A recent study hLSTMat proposed by Shen et al. [38]. A hierarchical LSTMs structure adaptive attention mechanism enabled the model to generate image and video captions with a remarkable performance. Spatial or temporal attention is adopted to select specific regions or frames to predict the related words as context information for the word generation. He et al. [5] proposed a double attention model for sub-region image feature extraction and semantic features distilling, and the visual and semantic information input LSTM simultaneously. These studies introduce the semantic information or integrate the attention mechanism to the image caption model, and have improved the performance of conventional image caption model. Inspired by the studies, the attention mechanism is introduced for salient semantic elements selection of high-level semantic information. An adaptive attention mechanism is integrated to fuse the visual and high-level semantic information for the LSTM input, and a similar mechanism is carried out for the LSTM output. A visual sentinel is adopted to control the visual and non-visual words generation reliance switching between image information and language model.

3 ADAPTIVE ATTENTION-BASED HIGH-LEVEL SEMANTIC INTRODUCTION FOR IMAGE CAPTION

3.1 Attention Mechanism

The majority of attention mechanisms adopt sub-region image feature vector that is associated with t to guide the image caption generation instead of a fixed feature vector extracted by a CNN, as shown in Figure 2. The image feature vector C is replaced by the feature vectors $C_1, C_2, \cdots C_t$ at different time steps. The attention distribution is calculated based on a previously generated description and image feature vector [39]. Therefore, different positions of the image have different levels of attention at different time steps. The specific steps are listed as follows:

(i) Extract the region feature from the input image;
(ii) Calculate the attention coefficient;
(iii) Send the focused features to the LSTM with predicted words; and
(iv) Calculate the output of the hidden layer of LSTM.

The previously predicted words can be used as reminders to determine which part of the image should receive more attention rather than focus on the entire image without purpose [8, 11, 40, 41]. The attention mechanism not only imitates the cognitive process of the human visual system but also reduces the amount of calculation. The accuracy of the model can be improved, because the meaningless content is excluded at each time step. Compared with the conventional image caption model, only the image features extracted by the convolutional neural network are sent to

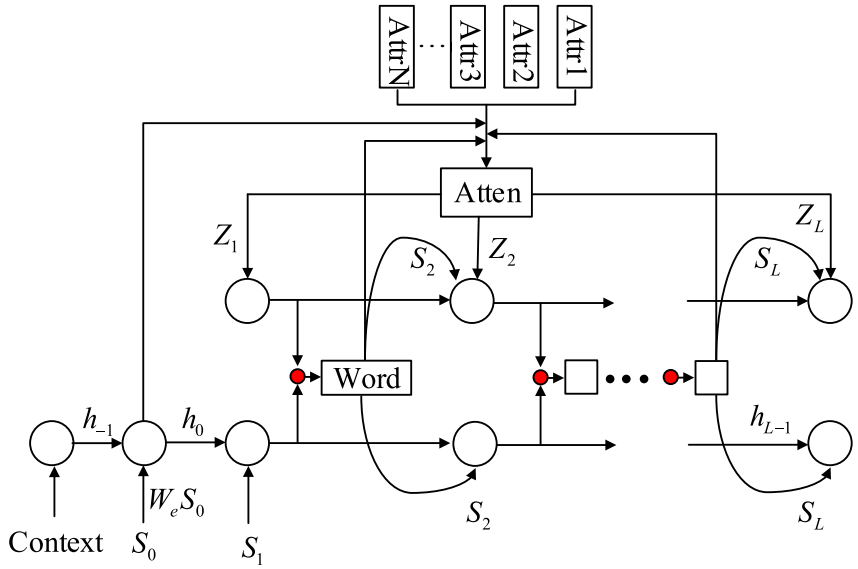


Fig. 3. High-level semantics introduction based on attention mechanism.

the language model. This method inputs different region features at different times to guide the image caption generation, which is more reasonable.

3.2 Total Architecture

High-level semantics is beneficial to guide the caption generation. Therefore, high-level semantics are introduced in the image caption model in this article, and an attention mechanism is used for the salient semantic elements selection. The attention mechanism has different levels of attention for the high-level semantics in the process of description generation instead of only focusing on the image feature vectors of different regions. The feature context extracted by a convolutional neural network is fused with high-level semantic information, which represents the rich content of an image. Additionally, the guidance information is switched between the image information and the language model information by using a visual sentinel in the image caption generation. The image feature context extracted by the convolutional neural network inputs the LSTM only at time step -1 as the perception of the whole image by the decoder. The high-level semantic AttrN will guide the decoder to generate the description in the subsequent time step according to the attention coefficient. The total structure of image caption model is shown in Figure 3.

The combination of image and text features, and the combination of global features (image feature vectors) and local details (image details represented by high-level semantics) jointly guide image caption generation, which renders image caption generation more accurate and comprehensive. The image feature vector and the high-level semantics jointly affect the hidden layer state of the LSTM language model, and the word generation probability is predicted by the hidden layer state of the LSTM. The output word will be the feedback of input for the next time step to affect the hidden layer state of the LSTM. The visual sentinel is introduced to save the state of the language model. The language model predicts the word in the next time step switch between the visual sentinel that represents the state of the language model and the new image content Z_t, which determines the input of the LSTM at the next time step. The calculation of the image caption

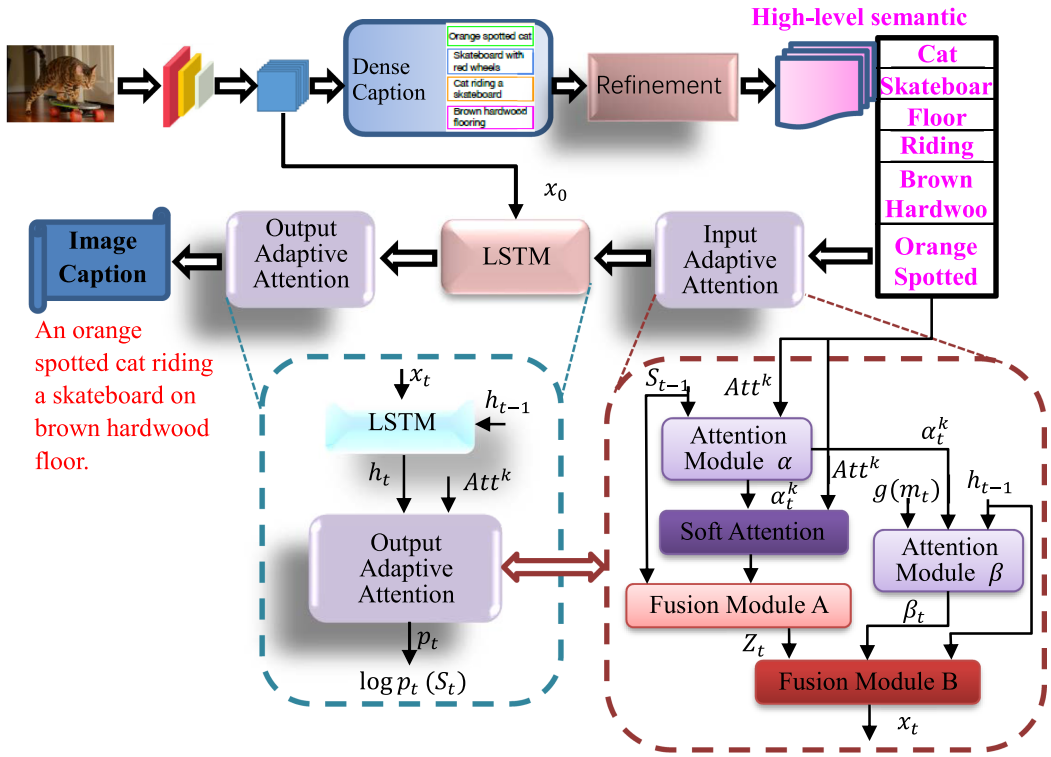


Fig. 4. Architecture of proposed model.

model is shown in Equation (1),

$$\begin{cases} x_{-1} = CNN(I) \\ x_0 = W_e S_0 \\ h_t = LSTM(h_{t-1}, x_{t-1}) \\ S_t \sim p_t = \varphi(h_t, Att(I)) \\ Z_t = \phi(S_{t-1}, Att(I)) = \phi(W_e S_{t-1}, W_z Att(I)), \\ x_t = f(Z_t, g(m_t)) = f(\phi(S_{t-1}, Att(I)), h_{t-1}) \end{cases} \tag{1}$$

where W_e is the mapping of words; the one-hot word vector S_0 is mapped as a 512-dimensional vector; W_z represents the learnable mapping, which is used to map the high-level semantics to a 512-dimension vector; S_0 is the start of sentence; S_{t-1} is the one-hot vector representation of the generated word at the previous time step; and $Att(I)$ is the high-level semantic information of the input image I. $Att(I)$ includes the objects (noun) in the image, attributes of the objects (adjective), and the position, action, and relationship (verb and preposition) information; this term is also presented in the form of a one-hot word and shares the vocabulary of S with a size of γ, ϕ, and φ, thus representing the attention mechanism for assigning different levels of attention to different high-level semantics. The attention mechanism ϕ is used to calculate the new input Z_t according to the different concerns of high-level semantic by the attention mechanism for the new image. x_t and h_t are the input of LSTM and the output of LSTM, respectively. m_t is the memory content of LSTM, and $g(m_t)$ is the partial memory content output by the control visual sentinel gate. f is the fusion function of the image and language model.

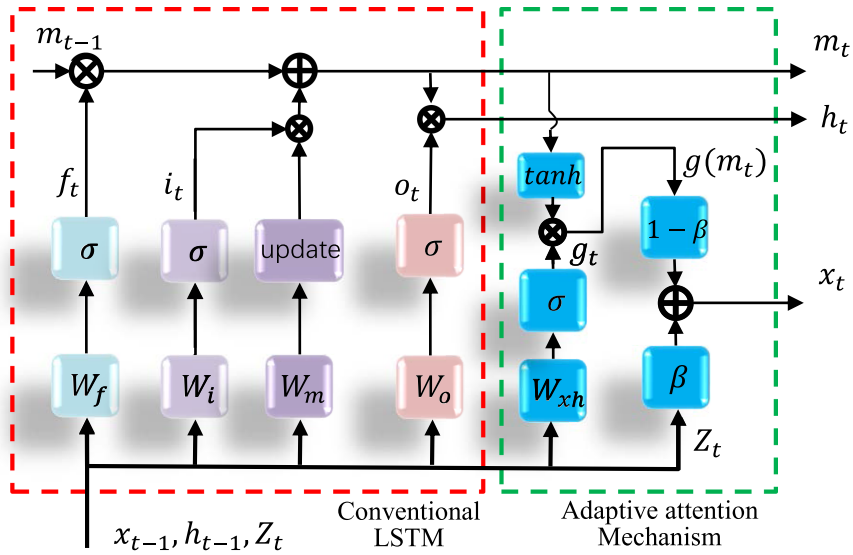


Fig. 5. Composite diagram of LSTM and adaptive mechanism.

The specific calculation architecture of our model is shown in Figure 4. High-level sematic information is derived from the dense caption. The image feature and high-level semantic information are used for the image caption generation. Two adaptive attention modules are designed for the input and output of the LSTM module. The two adaptive attention modules have similar structures, which contain a soft attention module and a visual sentinel module. The functional parts will be stated in the following subsections.

3.2.1 Adaptive Attention-based LSTM Decoder. LSTM [42] is employed in this article. LSTM is a memory cell for sequential data modeling. The output updating h_t of LSTM is shown in Equation (2):

$$\begin{cases} f_t = \sigma(W_{fx}x_{t-1} + W_{fh}h_{t-1}) \\ i_t = \sigma(W_{ix}x_{t-1} + W_{ih}h_{t-1}) \\ o_t = \sigma(W_{ox}x_{t-1} + W_{oh}h_{t-1}) \\ m_t = f_t \otimes m_{t-1} + i_t \otimes \tanh(W_{mx}x_{t-1} + W_{mh}h_{t-1}) \\ h_t = o_t \circ m_t \end{cases}, \tag{2}$$

where f_t, i_t, m_t, and o_t are the forget state, input state, memory state and output state, respectively, of LSTM; W_{ij} represents the trainable parameters; h_t is the hidden layer output of LSTM; $\otimes$ denotes the elements multiplication; and σ is a sigmoid function.

In this article, the adaptive attention mechanism is composite with the LSTM model. Figure 5 is a composite diagram of LSTM and the adaptive mechanism. The diagram has two parts: conventional LSTM and the adaptive attention mechanism. The conventional LSTM part corresponds to the calculation of Equation (2). The adaptive attention mechanism is implemented for focused information gathering. Z_t is the fusion information of generated words and focused high-level semantic information. g_t is a gate signal for the switch of Z_t and $g(m_t)$, which is referred to as the visual sentinel.

3.2.2 Adaptive Attention Mechanism for LSTM Input. According to Equation (1), ϕ is the attention function. After extracting the global image features by the convolutional neural network as the perception of image for the language model at time step -1, different attention levels $Att(I)$ for

the higher-level semantics can be provided by the attention function ϕ. The high-level semantics $Att(I): Att1, Att2, \cdots AttN$ of the image can better represent the content of the image. At time step t, the attention function ϕ assigns coefficients α_t^k for high-level semantics $Att(I)$ according to the predicted words S_{t-1} at the previous time step. The calculation of the attention coefficient is shown as Equation (3):

$$\alpha_t^k \propto softmax\left(S_{t-1}^k U Att^k\right) \propto softmax\left(S_{t-1}^T W_e^T U W_z Att^k\right), \tag{3}$$

where S_{t-1} and Att^k are one-hot vectors. α_t^k is the attention coefficient for every high-level semantic Att^k at every time step. α_t^k represents the association between high-level semantics and the previously generated words S_{t-1}. The attention coefficient is calculated by the softmax function. U represents the mapping between the previously generated words S_{t-1} and high-level semantics with the size $\gamma * \gamma$. Since the vocabulary is too large, the dimensionality reduction mapping W_e and W_z are introduced to map S_{t-1} and Att^k to a lower dimensional space with the one-hot type. The size of the image feature vector acquired by CNN is 512. Therefore, U is a matrix with the size 512*512 for mapping. The attention coefficient is assigned to different high-level semantics. The weighted summation of all high-level semantics is mapped to the Z_t with the word vector generated at the previous time step. Z_t is a compound output of previously generated words and high-level semantics with different attention levels at time step t, which will be used as the input of the LSTM language model at time step $t+1$ to affect the hidden layer state of LSTM and predict the word at the step $t+1$ (if the visual sentinel is not being added to the model). The calculation formula is shown in Equation (4):

$$Z_t = \phi(S_{t-1}, V_{att}(I)) = W^{Z,S}\left(W_e S_{t-1} + diag(W^{Z,A}) \sum_k \alpha_t^k W_Z Att^k\right), \tag{4}$$

where $W^{Z,S} \in R^{m*d}$ is the mapping matrix, $diag(w)$ is the diagonal matrix of w, and $W^{Z,A} \in R^d$, $d = 512$ is the mapping at the word vector space.

Inspired by the adaptive attention mechanism, the high-level semantic of the image does not have to be considered at every time step, because some non-visual words (of, in, the, ...) and fixed collocation ("phone" is always behind "taking a cell") can be completely predicted by the language model without paying attention to images or high-level semantic information. The fusion mechanism was designed, as shown in Figures 4 and 5, and a visual sentinel is introduced. The LSTM is applied as the language model decoder, which stores long-term and short-term visual and linguistic information [30]. The image caption model switches to a language model to guide the image caption generation when the model does not need to pay attention to the image content. This new part is referred to as the visual sentinel. The gate in the model that determines whether to focus on the image content or language model is referred to as the sentinel gate. The visual sentinel used by the LSTM decoder aims to store language information in his memory cell [28] and can be calculated by Equation (5):

$$\begin{cases} g_t = \sigma(W_x x_{t-1} + W_h h_{t-1}) \\ g(m_t) = g_t \tanh \otimes (m_t) \end{cases}, \tag{5}$$

where W_x and W_h are trainable parameters, x_{t-1} is the input of LSTM at time step $t-1$ (the image feature at x_0 and will be the composite information of image and language), m_t is the memory cell of LSTM, g_t is the gate signal, $\otimes$ indicates the elements multiplication, and σ is the sigmoid function.

By using visual sentinels, the adaptive context feature vector x_t is the fusion of the image content Z_t and visual sentinels information $g(m_t)$. Z_t is the visual information from high-level semantics that represent image content, and $g(m_t)$ is the visual sentinel that represents the language model

information. Therefore, the input of LSTM is not a single Z_t but is a fusion of Z_t and $g(m_t)$. The calculation formula is shown in Equation (6):

$$x_t = \beta_t Z_t \otimes (1 - \beta_t) g(m_t), \tag{6}$$

$\beta_t \in [0, 1]$ is the sentinel gate at time step t, which determines the attention level of the language model output state m_t and high-level semantic information Z_t and will affect the hidden layer state of LSTM. The calculation formula is shown in Equation (7):

$$\begin{cases} \beta_t = W_s g(m_t) + W_g h_{t-1} \\ \alpha_t^{k+1} = softmax[S_{t-1}^T W_e^T U W_z Att^k; \beta_t] \\ x_t = \beta_t Z_t \otimes (1 - \beta_t) g(m_t) \\ h_t = LSTM(h_{t-1}, x_t) \end{cases}, \tag{7}$$

where $[;]$ represents concatenation in tandem, and α_t is the joint attention coefficient between high-level semantics and visual sentinel $g(m_t)$. The sentinel gate β_t is the last element of α_t^{k+1}. The attention mechanism is switched between the high-level semantic information and the language model. At the time $\beta_t = 1$, the mechanism only uses the information of the language model to predict the words for the next time step; that is, the image information is not acquired, the next word is only generated according to the language model, and $\beta_t = 0$ is the opposite. The design of the adaptive attention mechanism can adaptively determine the switch of attention between the image content and the language model. Therefore, the LSTM model can obtain more reasonable input.

3.2.3 Adaptive Attention Mechanism for LSTM Output. After the introduction of the high-level semantic by the adaptive attention mechanism and the control of the visual sentinel, the LSTM model obtains a reasonable input. The output of LSTM is also followed by the adaptive attention mechanism φ. The structure of the output attention mechanism is similar to the input attention model. First, the attention function assigns different attention coefficients to high-level semantics according to the hidden layer state of LSTM, and then the fusion information y_t is calculated based on the adaptive attention and the prediction probability of the output word gathered by the softmax function. The specific calculation formula is shown in Equation (8):

$$\begin{cases} \hat{\alpha}_t^k = softmax(h_t^T V tanh(W_z Att^k)) \\ \hat{Z}_t = w^{\hat{Z},h}(h_t + diag(w^{\hat{Z},A}) \sum_k \hat{\alpha}_t^k \tanh(W_z Att^k)) \\ \hat{\beta}_t = W_s g(m_t) + W_g h_t \\ \hat{\alpha}_t^{k+1} = softmax[h_t^T V tanh(W_z Att^k); \hat{\beta}_t] \\ y_t = \hat{\beta}_t \hat{Z}_t \otimes (1 - \hat{\beta}_t) g(m_t) \\ p_t = softmax(W_p(y_t + h_t)) \end{cases}, \tag{8}$$

where h_t is the hidden layer output of LSTM; $V \in R^{n*d}$ is the bilinear parameter matrix; W_z is introduced to map Att^k to a lower dimensional space with the one-hot type; $w^{\hat{Z},h} \in R^{m*d}$ is the mapping matrix, $diag(w)$ is the diagonal matrix of w, and $w^{\hat{Z},h} \in R^d$, $d = 512$ is the mapping at the word vector space; $\hat{Z}_t$ is a compound output of h_t and high-level semantics with different attention levels at time step t; $\hat{\beta}_t \in [0, 1]$ is the sentinel gate at time step t; $[;]$ represents concatenation in tandem, the sentinel gate $\hat{\beta}_t$ is the last element of $\hat{\alpha}_t^{k+1}$; $tanh()$ is the activation function. The attention coefficients $\hat{\alpha}_t^k$ will be calculated as the weights that correspond to high-level semantics Att^k, and the weighted sum of the activation values will be used as a complement of the hidden layer state of the language model to determine the distribution of p_t, which can be used for loss calculation.

3.2.4 Loss Function. The training dataset I contains images with the high-level semantic representation $Att(I)$ of the image, and the natural language label S_t, which corresponds to the image, is acquired. The training of the model continuously updates the parameters of the attention mechanism model and language model by minimizing the loss function [43]. The loss function is defined as the negative logarithmic mean of the probability of a predicted word. The final words output is determined by the high-level semantics and generated words, as expressed in Equation (9):

$$\begin{cases} S_t \sim p_t : \log p(S|V_{att}(I)) = \sum_1^L \log(S_t|S_{1,t-1}, Att(I)) \\ Loss = \frac{1}{N}\sum_{i=1}^N \log p(S^i|V_{att}(I^i)) + \lambda_\theta \|\theta\| \end{cases}, \tag{9}$$

where L is the length of the sentence, N is the number of images, $\lambda_\theta\|\theta\|$ is the regularization item, and θ represents the learnable parameters.

Regularization term $\lambda_\theta(\theta)$ for the attention coefficient is introduced to prevent overfitting, and θ is the learnable parameter of attention mechanism. The specific calculation of $\lambda_\theta(\theta)$ is shown in Equation (10):

$$\lambda_\theta(\theta) = \|\theta\|_{1,p} + \theta^T{}_{q,1} = \left[\sum_i \left[\sum_t \theta_t^i\right]^p\right]^{1/p} + \sum_t \left[\sum_i (\theta_t^i)^q\right]^{1/q}, \tag{10}$$

The design principle of the regularization function γ depends on the penalty factor: $p > 1$ means that the attention assigned to single high-level semantics at a given time step is excessive, and $0 < q < 1$ means that the attention assigned to all attributes of high-level semantics at any time step is excessive. $p = 2,\ q = 0.5$ are adopted in the experiment [2] and serve as a 2-norm function. By the penalty of Equation (11), the model requires the attention function to focus on every high-level sematic information and sparsity of attention.

3.3 High-level Semantic Gathering

High-level semantic information is obtained by the dense caption in this article. The dense caption model can gather the description of an image with short phrases that correspond to the local feature of an image [26]. High-level semantic information consists of the attribute and position of an object and more detailed and general information about an image. Therefore, high-level semantic information can be obtained by the refinement of the dense caption. The process of refinement includes eliminating duplicates, deleting the functional word, and then selecting 5∼10 descriptions with higher confidence to construct high-level semantic information. Figure 6 shows a diagram of high-level information extraction.

4 EXPERIMENTAL STUDIES

4.1 Parameter Setting

To verify the validation of the improved model, some simulations are carried out. The specific parameters of the training are set as follows: all hidden layer size, word vector size, and input layer size are set to 512 dimensions. To facilitate comparison with other models, the convolutional neural network adopts ResNet [44]. The stochastic gradient descent algorithm is adopted in the training with a learning rate of 0.001, and a random inactivation rate of 0.5 is utilized to reduce overfitting. The Batch_size is set to 10, the training iteration is 150,000 times, and the trained model can be obtained by training for 16 hours with the configuration of E5-2678V3-8G-GTX1080TI. To verify the validity of the model, the Microsoft COCO and Flickr30k image caption assessment tool were adopted for evaluation. COCO is a large image dataset designed for object detection, segmentation, person keypoints detection and stuff segmentation. It contains the object instances and object keypoints for object detection. COCO is also the largest dataset for image caption,

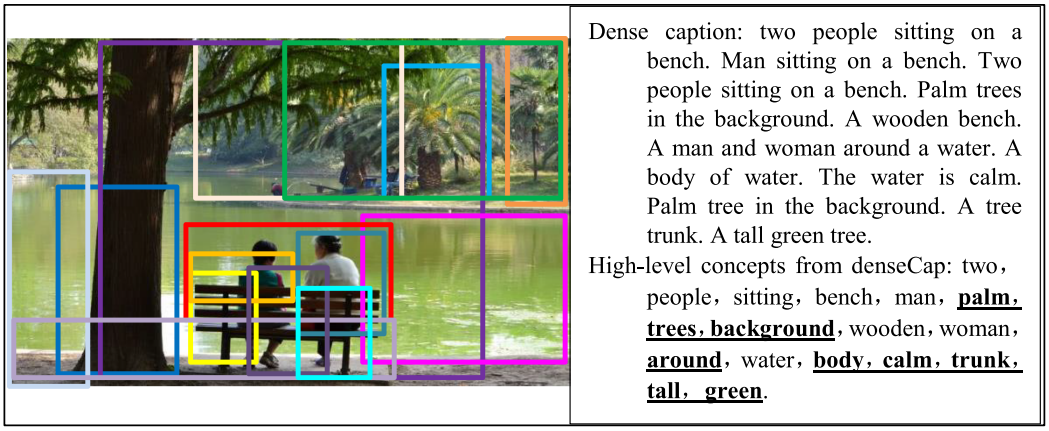


Fig. 6. High-level semantic information extraction.

which has 82,783 training images, 40,504 validation images and 40,775 resting images in coco2014 version. In this dataset, each image is labeled by 5 manual captions. Flickr30k consists of 31,783 images collected by Flickr. Most of the images are human activities concern and each image has five descriptions. Two types of evaluations are carried on the MSCOCO dataset. An offline evaluation is conducted with the way of "Karpathy" data split rule, which has been widely used by researchers, namely, 113,287 images for training, 5,000 images for validation, 5,000 images for testing on COCO dataset and 29,000 images for training, 1,000 images for validation and 1,000 images for testing on Flickr30k dataset. An online evaluation is carried out further on the COCO Image Captioning Challenge sets c5 and c40 [45].

Automatic evaluation indicators, such as bilingual evaluation understudy BLEU [46], metric for evaluation of translation with explicit ordering METEOR [47], recall-oriented understudy for gisting evaluation ROUGE [48], and consensus-based image description evaluation CIDEr [49], are adopted to evaluate the performance of the proposed method with the Microsoft COCO dataset and Flickr30k.

4.2 High-level Semantics Gathering

To obtain the high-level semantic information, the dense caption model is adopted as a baseline. The high-level semantic acquisition results are shown in Figure 7. Compared with multi-label classification, some attribute information for the object, as well as more detailed information, position information and general information of the images, are obtained. For example, the words "wooden" and "bench" in Figure 7(a) and the words "palm" and "trees" in Figure 7(b) are detailed information. The words "watching" and "game" are the summary information in Figure 7(c), and the words "blue" and "shirt" are details. Because the high-level semantics obtained by the dense caption are derived from the regional description, they can contain more detailed information and some objects without a visual bounding box. Some positional relations or generalized words, such as "background," "next," "ground," "watching," and "game," as well as adjectives and adverbs such as the word "calm" in the description of water in Figure 7(b) and the word "wearing" in Figure 7(c), are generated by the language model.

4.3 Qualitative Analysis

An example is carried out to qualitatively analyze the performance of the model as shown in Figure 8. NIC model can generate a brief and general description. However, the description is

Multi-label classification: riding, elephant, people, man, elephants, dirt, horse, forest, standing, two, woods, walking, field, trees, tree.
High-level concepts from denseCap: riding, elephant, **sitting**, people, **wooden**, **bench**, back, man, woman, two, **background**, trees, forest. (a)

Multi-label classification: people, bench, sitting, water, lack, benches, man, river, boat, group, park, wooden, two, woman.
High-level concepts from denseCap: two, people, sitting, bench, man, **palm**, **trees**, **background**, wooden, woman , **around**, water, **body**, **calm**, **trunk**, **tall**, **green**. (b)

Multi-label classification: tennis, court, ball, player, man, racket, woman, baseball, playing, holding, game, crowd, people, bat, standing, racquet, hit, swinging.
High-level concepts from denseCap: woman, playing, tennis, people, **watching**, **game**, court, man, **wearing**, **blue**, **shirt**, red, player, **white**, **black**, **shoes**, racket, hand, **match**, lines, ground, ball, air. (c)

Multi-label classification: people, woman, market, man, group, bus, train, standing, street, bananas, food, table, truck, sitting, vegetables, around
High-level concepts from denseCap: two, people, standing, **next**, **each**, **other**, man, **wearing**, **red**, **jacket**, **green**, **plant**, **outdoor**, market, bag, white, plastic, blue, truck, **parked**, **background** (d)

Fig. 7. High-level semantics that obtain results for different images.

always incomplete and is often only a general description of the main part or a part of the image. The adaptive attention model generates a better description than NIC model. However, it is still incomplete. Compared with the NIC and adaptive attention models, the proposed model can generate more detailed captions. For Figure 8(a), NIC model depicts the right half of the image as "a group of people standing around a large truck," which disregards "buy vegetables in market." The adaptive attention model generates more detailed local features of the image via the adaptive attention mechanism; thus, the generated description is relatively more specified. The visual sentinel is introduced as a switching gate to guide image caption generation, which reduces the influence of non-visual vocabulary in the training and is more reasonable in the generation of some related words and prepositions, for example, the word "next to" in the sentence "a man standing next to a woman" in Figure 8(a). In some phenomena, the description is incomplete, and some marginalized regions, such as "truck" in Figure 8(a), "trees" in Figures 8(b) and 8(c), and "audience" in Figure 8(d), are always disregarded. The image body represents the most intuitive semantic information about the image, that is, foreground information. The intuitive feeling of the image is

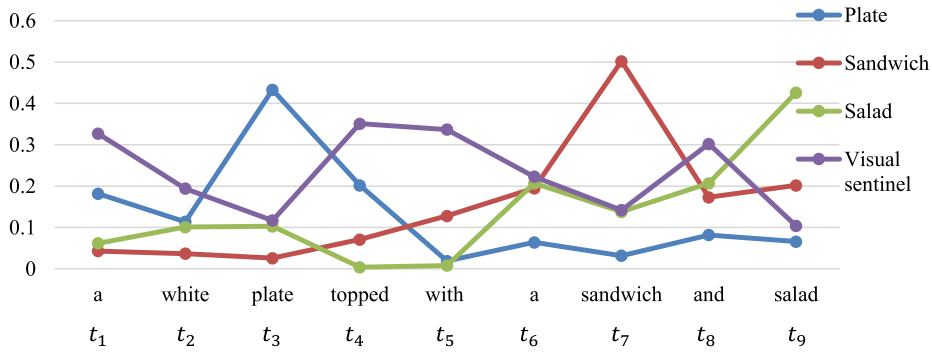
	NIC: a group of people standing around a large truck. Adaptive attention: a man standing next to a woman in a market. Our model: a group of people standing in an outdoor market in front of a lot of trucks.	NIC: Inaccurate, lack of concrete information Adaptive attention: lack of scene information	(a)
	NIC: two people riding on the back of an elephant. Adaptive attention: two people riding on the back of an elephant. Our model: two people riding an elephant with trees in the background.	NIC: lack of scene information. Adaptive attention: lack of scene information.	(b)
	NIC: a couple of people sitting on a bench near a river. Adaptive attention: two people sitting on a bench near a lake. Our model: two people sitting on a wooden bench near a lake with palm trees in the background.	NIC: lack of some details of bench and neglect tree. Adaptive attention: lack of some details of bench and neglect tree	(c)
	NIC: a tennis player swinging a racket at a ball. Adaptive attention: woman is playing tennis on a clay court. Our model: a woman in blue shirt is playing tennis on a tennis court.	NIC: lack of some details, such as playing, shirt, scene. Adaptive attention: lack of shirt information.	(d)

Fig. 8. Image caption results for different images.

the image content that occupies most of the image area and is located at the front of the human visual experience, as shown in Figure 8(d). The NIC model describes the image as "a tennis player swinging a racket at a ball," but the "audience" is disregarded. By introducing high-level semantic information in the proposed model, this easily neglected information can be input in the model by the adaptive attention mechanism in the form of high-level semantics. These words will be considered when generating a description; thus, the generated description is more detailed and specific.

To visualize the function of high-level semantic information and visual sentinels, the attention coefficient for different high-level semantic information and visual sentinels at different time steps are exhibited by visualization technique. High-level semantic information consists of 5–12 semantic words extracted from an image with highest scores. As shown in Figure 9, visualizing the attention coefficient of different high-level semantic at each time step, it can be seen that the relationship between the attention distribution of the high-level semantic and the generated word is obvious. The generated word has bigger attention coefficient. For some non-visual words, the distribution of attention is more scattered, but the highest score of attention mostly appears in

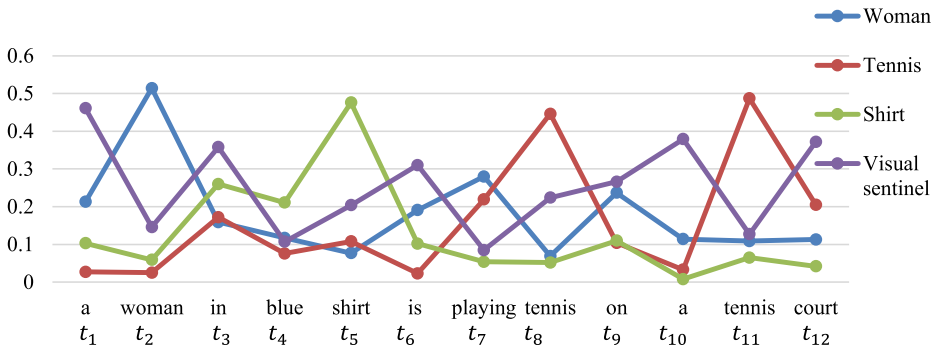
Note: 'Plate', 'Sandwich' and 'Salad' are high–level semantics. t_x is the moment of words generation, and the basis of word generation is exhibited. At t_1 moment, visual sentinel obtains bigger attention, and 'a' is generated almost based on language model instead of image feature and high-level semantic. At t_2 moment, visual sentinel also gathers bigger focus, then the word 'white is generated mostly based on language model. At t_2 moment, 'plate' obtains more attention, and then the word 'plate' is generated. The visual words and non-visual words are generated based on different mechanism.



(a) Attention coefficient distribution visualization for image a

Note: 'Woman', 'Tennis', and 'Shirt' are high-level semantics. t_x is the moment of words generation, and the basis of word generation is exhibited. At t_1 moment, visual sentinel gets more attention, and 'a' is almost generated based on language model. At t_2 moment, 'woman' gathers bigger attention coefficient, then the word 'woman' is generated according to the attention. At t_3 moment, the visual sentinel gets more attention, and the word is generated according to language model.

At t_4 moment, 'shirt' gets more attention, and then the word 'shirt' is generated. The attention coefficient decides the basis of words generation.



(b) Attention coefficient distribution visualization for image b

Fig. 9. Visualization of attention coefficient for high-level semantics.

Table 1. Comparison Among Models on MSCOCO

Researches	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
NIC at 2015 [11]	66.6	45.1	30.4	20.3	25.4	49.0	94.3
Soft attention at 2015 [8]	70.7	49.2	34.4	24.3	23.9	–	77.3
Hard attention at 2015 [8]	71.8	50.4	35.7	25.0	23.0	–	–
(RA+SS)-Ensemble [31]	72.4	55.5	41.8	31.3	24.8	53.2	95.5
Stimulus and concept driven	74.8	52.5	36.5	23.5	23.5	50.5	104.1
SCN-LSTM Ensemble of 5 [50]	74.1	57.8	44.4	34.1	26.1	–	104.1
AG-SAM [6]	73.7	54.2	39.2	28.3	25.5	56.8	98.3
LSTM-A3 [33]	73.5	56.6	42.9	32.4	25.5	53.9	99.8
Attention-joint context [34]	71.2	51.4	36.8	26.5	24.7	–	88.2
Adaptive Attention [28]	74.2	58.0	43.9	33.2	26.6	–	108.5
VSDA [5]	75.3	55.9	42.1	31.9	25.3	53.4	97.4
EE-LSTM-P (VGGNetFT) [30]	75.7	59.4	45.3	34.6	26.8	56.0	109.6
SGA-BR [7]	76.7	60.4	46.0	34.7	26.6	55.6	108.7
CSF [51]	76.4	60.2	46.1	35.0	27.1	56.1	110.2
Att-RegionCNN-LSTM [3]	74.0	56.0	42.0	31.0	26.0	94.0	104.9
Att-GT-LSTM [3]	80.0	64.0	50.0	40.0	28.0	96.0	107.0
hLSTMat-RF [38]	79.9	–	–	37.5	28.5	58.2	125.6
Ours	**84.1**	**70.2**	**55.9**	**45.4**	**36.0**	**66.5**	**132.3**

Note: BLEU, METEOR, ROUGE, and CIDEr are commonly used for fair and thorough performance evaluation. BLEU-N (B-1, B-2, B-3, B-4) metric is good at short sentence evaluation. ROUGE has different types for evaluating different types of texts. METEOR can perform an evaluation on various segments of a caption.

the visual sentinel. This experiment verifies the guidance role of high-level semantic and visual sentinels in image caption generation.

4.4 Comparing with State-of-the-Art

4.4.1 Performance Comparison Among Different Models. High-level semantics differ from the image feature vectors extracted from a pixel at a fixed position by convolutional neural networks, which can be the summary of the pixel features at any position in an image and can be the concept without direct visual representation. Therefore, high-level semantics can intuitively provide accurate semantic prompts for an image caption model, including the objects, object attributes, object actions, positional relationships and other information in the image. In addition, high-level semantics can be used as an external input with image features to guide image caption generation. In this article, an attention mechanism is used for high-level semantics to select the salient semantic elements, and high-level semantics and image features are combined to guide image caption generation.

The performance of different models was evaluated by automatic evaluation indicators BLEU, METEOR, and CIDEr. The results of models evaluation on the COCO data set are shown in Table 1. NIC [11]is an earlier famous end-to-end encoder-decoder neural network for image caption early. CNN is used as encoder and LSTM is adopted as decoder. No other special technology is involved. Therefore, the performance (66.6 B-1, 45.1 B-2, 30.4 B-3, 20.3 B-4, 25.4 METEOR, 49.0 ROUGE-L, 94.3 CIDEr) is not as good as subsequent models. It is always adopted as the baseline.

Soft and hard attention [8] were proposed for focusing sub-regions of the image instead of the whole image. (RA+SS)-ENSEMBLE [31] also adopted the similar spatial attention for image caption. Ding et al. [35] proposed two types of attention mode stimulus and concept driven attention.

The performances (74.8 B-1, 52.5 B-2, 36.5B-3, 23.5 B-4, 23.5 METEOR, 50.5 ROUGE-L, 104.1 CIDEr) of these models are better than NIC due to the usage of attention mechanism. However, improved only with attention mechanism, the performance shows a limited improvement.

SCN-LSTM [50] introduced semantic concept for image caption generation, and a Semantic Compositional Network (SCN) was proposed for concept extraction. LSTM-A3 [33] also made used of the attribute information of image for image caption generation. AG-SAM [6] made use of emotional high-level concepts and the attention mechanism was used for controlling the degree of concepts usage. Attention-joint context [34] adopted a fine-grained attention mechanism based on image segment, and also utilized the semantic joint text for image caption. This study adopts the segmentation results as semantic information. However, the attention mechanism is not integrated to capture the salient semantic elements for image caption generation. Adaptive Attention [28] was proposed to decide the word generation reliance on image information or language model in 2017. VSDA [5] adopted sub-region of image and semantic attention for the LSTM input. This study involves two attention mechanisms for image feature and semantics. The idea of attention mechanism for semantics is similar with ours. However, we adopt an adaptive attention for visual and non-visual word generation instead of visual attention. EE-LSTM-P (VGGNetFT) [30] made used of object detection module to predict categories of local regions, which were used as semantic concepts for image caption generation. This study doesn't adopt attention mechanism. A baseline model is used for the whole image and sub-region descriptions generation, and then extracts the high-level semantics according the two types descriptions. Semantic Guidance Attention (SGA) [7] utilizes semantic word representations to provide an intuitive semantic guidance that focuses accurately on sub-regions for image caption generation. Therefore, the semantic information is not used for image caption generation directly. Cascade semantic fusion (CSF) architecture [51] aimed to mine the representative features to encode image content through attention mechanism, object-level and image-level semantic attention were adopted to enrich the context information for image caption generation. Att-RegionCNN-LSTM [3] also made use of high-level concepts for image caption. However, the high-level semantic concepts were not explicit instead of progress directly from image features, and the Att-GT-LSTM extracted the high-level concepts from the label data. Although, this study adopted the ground truth as high-level concepts, it verified the help of high-level concepts for image caption generation very well, and got a good performance of 80.0 B-1, 64.0 B-2, 50.0 B-3, 40.0 B-4, 28.0 METEOR, 96.0 ROUGE-L, and 107.0 CIDEr. However, attention mechanism is not used here. hLSTMat [38] was a latest research that utilized spatial or temporal attention for selecting specific regions or frames to predict the related words, and the adaptive attention mechanism was also adopted. A hierarchical LSTMs structure enabled the model to generate image and video captions with a remarkable performance. Although, the performance (79.9 B-1, 37.5 B-4, 28.5 METEOR, 58.2 ROUGE-L, 125.6 CIDEr) of hLSTMs is not superior to Att-GT-LSTM due to the usage of ground truth in Att-GT-LSTM, hLSTMat is superior to Att-RegionCNN-LSTM.

These studies integrated spatial attention with semantic concepts, or made use of adaptive attention mechanism, and got a better performance than earlier models. However, the semantic concepts are always static representations. For the proposed model, the attention mechanism is adopted for salient high-level elements selection, and an adaptive attention is also adopted to control the visual and non-visual word generation reliance. The semantics attention, adaptive attention and previous generated words have a comprehensive fusion to construct a special attention module. This module has applied to the input and output of LSTM. Therefore, proposed model has a promising performance than other models for the BLEU, METEOR, and CIDEr automatic evaluation indicators.

Table 2. Comparison Among Models on Flickr30k

Model	B-1	B-2	B-3	B-4	METEOR	CIDEr	ROUGE-L
NIC [11]	66.3	42.3	27.7	18.3	–	–	–
(SS+RA)-Ensemble [31]	64.9	46.2	32.4	22.4	19.4	47.2	45.1
Soft attention [8]	66.7	43.4	28.8	19.1	18.49	–	–
Hard attention [8]	66.9	43.9	29.6	19.9	18.46	–	–
SCN-LSTM Ensemble of 5 [50]	74.7	55.2	40.3	28.8	22.3	–	–
Att-reginCNN+LSTM [3]	73.0	55.0	40.0	28.0	–	–	–
Stimulus and concept driven [35]	66.3	43.7	29.2	21.1	–	–	–
VSDA [5]	68.1	49.8	35.7	25.6	20.8	53.2	47.4
SGA-BR [7]	69.0	50.9	36.8	26.4	21.5	–	–
Adaptive attention [28]	67.7	49.4	35.4	25.1	20.4	53.1	–
hLSTMat [38]	73.8	55.1	40.3	29.4	23.0	66.6	–
Ours	**86.4**	**68.3**	**51.5**	**38.2**	**26.9**	**83.4**	**57.5**

Table 3. The Evaluation Results on COCO Dataset

	B-1		B-2		B-3		B-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
Human	66.0	88.0	47.0	74.0	32.0	63.0	22.0	47.0	20.0	34.0	48.0	63.0	85.0	91.0
Hard Attention [8]	70.5	88.1	52.8	77.9	38.3	65.8	27.7	53.7	24.1	32.2	51.6	65.4	86.5	89.3
Google NIC [11]	71.3	89.5	54.2	80.2	40.7	69.4	30.9	58.7	25.4	34.6	53.0	68.2	94.3	94.6
MSR Captivator [52]	71.5	90.7	54.3	81.9	40.7	71.0	30.8	60.1	24.8	33.9	52.6	68.0	93.1	93.7
ATT [2]	73.0	90.0	57.0	82.0	42.0	71.0	32.0	60.0	25.0	34.0	54.0	68.0	94.0	96.0
Att-RegionCNN+LSTM [3]	73.0	89.0	56.0	80.0	41.0	69.0	31.0	58.0	25.0	33.0	53.0	67.0	92.0	93.0
Adaptive Attention [28]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
TD [53]	75.7	91.3	59.1	83.6	44.1	72.6	32.4	60.9	25.9	34.2	54.7	68.9	105.9	109.0
SGA-BR [7]	76.2	93.1	59.9	85.9	45.4	76.0	34.1	65.1	26.6	35.7	55.4	70.6	105.5	107.7
LSTM-A3 [33]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	24.0	35.4	56.4	70.5	116.0	118.0
SCST: Att2all [54]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	1.147	116.7
Semantic constructor [55]	73.1	90.6	56.3	82.0	42.4	71.3	32.0	60.4	25.7	34.4	53.7	68.6	96.9	97.7
hLSTMat [38]	79.4	94.4	**63.5**	88.0	**48.7**	**78.4**	**36.8**	**67.4**	**28.2**	**37.0**	**57.7**	**72.2**	**120.5**	**122.0**
Our model-semantics	72.4	89.0	55.4	80.1	41.3	69.9	30.7	58.6	24.8	32.2	53.0	67.5	93.1	95.2
Our model-adaptive	**79.9**	**97.2**	63.0	**89.4**	47.1	77.8	34.4	654	26.4	36.3	55.5	71	104.9	107.5

Note: c5 and c40 indicate the label number of each image.

Another dataset Flickr30k is adopted for the evaluation, and the result is shown in Table 2. Our model performs best. These results (Tables 1 and 2) show the advantage of proposed model, which generates the best results on all the two datasets over all evaluation metrics. Especially, our model gets a 20% improvement compared with the baseline model (NIC) for all evaluation metrics, and gets a 5% improvement than State-of-the-Art.

4.4.2 Performance Comparison with State-of-the-Art. We further conducted an online evaluation on the COCO Image Captioning Challenge sets c5 and c40. These results are shown in Table 3. The "our model-semantic" model is an ablation experiment of removing adaptive attention mechanism. The "our model-adaptive" model is a complete model. According to the result of Table 3, we can see the adaptive mechanism plays an important role in enhancing the

performance of the image caption model. Compare our model with state-of-the-art models, we achieve 0.799/0.972 on B-1(c5/c40) and 0.894 on B-2(c40), and surpass human performance on the 14 metrics reported. Other state-of-the-art published methods are also shown for comparison. The performance of hLSTMat [38] is remarkable and has better performance for most of the indicators.

5 CONCLUSIONS

The high-level semantics that most directly represent the image content are introduced as new information and are used in conjunction with the image features to guide the image caption generation, which enables the image caption model to better notice the details of an image and generate a more comprehensive and accurate description. In this article, an attention mechanism is introduced to select the salient high-level semantics for the LSTM module. The adaptive attention mechanism, attention high-level semantic module integrates with visual sentinel, aims to achieve guidance information switching to generate a more comprehensive and smooth natural language description. The image caption model proposed in this article can describe the image as accurately and comprehensively as possible. However, this image caption model has a strong dependence on the accuracy of high-level semantic acquisition. Therefore, future work will be the accurate semantic information capturing; integrate object detection and recognition, mutual relationship detection, phrase description, and regional description; fully utilize information about different semantic levels of an image to complement each other; and be expected to enhance the performance of the image caption model.

REFERENCES

[1] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. 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*. 203–212.

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

[3] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van denHengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1367–1381.

[4] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van denHengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 6 (2018), 1367–1381.

[5] Chen He and Haifeng Hu. 2019. Image captioning with visual-semantic double attention. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1 (2019), 1–16.

[6] Anqi Wang, Haifeng Hu and Liang Yang. 2018. Image captioning with affective guiding and selective attention. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 3 (2018), 1–15.

[7] Jie Wu, Haifeng Hu, and Yi Wu. 2018. Image captioning via semantic guidance attention and consensus selection strategy. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 4 (2018), 1–19.

[8] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[9] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surveys* 51, 6 (2019), 118–154.

[10] Xinli Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2422–2431.

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

[12] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-RNN). In *International Conference on Learning Representations*. 1–15.

[13] Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 664–676.

[14] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2017. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 677–691.

[15] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L. Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *IEEE International Conference on Computer Vision*. 2533–2541.

[16] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision*. 15–29.

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

[18] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piortr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*. 1473–1482.

[19] Ning Wang, Jing-Chao Sun, Meng Joo Er, and Yan-Cheng Liu. 2016. Hybrid recursive least squares algorithm for online sequential identification using data chunks, *Neurocomputing* 2016, 174: 651–660.

[20] Siming Li, Girish Kulkarni, Tamara L. Berg, C. Alexander Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale N-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning*. 220–228.

[21] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 359–368.

[22] Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1292–1302.

[23] Remi Lebret, Pedro O. Pinheiro, and Ronan Collobert. 2014. Simple image description generator via a linear phrase-based approach. ArXiv Preprint arXiv:1412.8419.

[24] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*. 595–603.

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

[26] Johnson Justin, Karpathy Andrej, and Fei-Fei Li. 2016. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

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

[28] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. 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*. 3242–3250.

[29] Rezazadegan Tavakoli, Hamed R. Tavakoliy, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision*. 2506–2515.

[30] Xiaodan Zhang, Shengfeng He, Xinhang Song, Rynson W. H. Lau, Jianbin Jiao, and Qixiang Ye. 2020. Image captioning via semantic element embedding. *Neurocomputing* 395, 212–221.

[31] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. 2015 Aligning where to see and what to tell: Image caption with region-based attention and scene factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 12 (2015), 2321–2334.

[32] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2407–2415.

[33] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4894–4902.

[34] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. 2019. High-quality image captioning with fine-grained and semantic-guided visual attention. *IEEE Trans. Multimedia* 21, 7 (2019), 1681–1693.

[35] Songtao Ding, Shiru Qu, Yuling Xi, and Shaohua Wan. 2020. Stimulus-driven and concept-driven analysis for image caption generation. *Neurocomputing* 39, 520–530.

[36] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. 2019. From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 10 (2019), 3047–3058.

[37] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. 2017 Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia* 19, 9 (2017), 2045–2055.
[38] Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. 2020 Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 5 (2020), 1112–1131.
[39] Meng Joo Er, Yong Zhang, Ning Wang, and Mahardhika Pratama. 2016 Attention pooling-based convolutional neural network for sentence modelling. *Info. Sci.* 373, 388–403.
[40] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* 27. 2204–2212.
[41] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 MSCOCO Image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 652–663.
[42] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (1997), 1735–1780.
[43] Rawaa Dawoud Al-Dabbagh, Saad Mekhilef, and Mohd Sapiyan Baba. 2015 Parameters' fine tuning of differential evolution algorithm. *Comput. Syst. Sci. Eng.* 30, 2 (2015), 125–139.
[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770–778.
[45] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Doll A R and C. Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. ArXiv Preprint, arXiv:1504.00325.
[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.* 311–318.
[47] Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the 2nd Workshop on Statistical Machine Translation.* 228–231.
[48] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL2004.* 74–81.
[49] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 4566–4575.
[50] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 5630–5639.
[51] Shiwei Wang, Long Lan, Xiang Zhang, Guohua Dong, and Zhigang Luo. 2019. Cascade semantic fusion for image captioning. *IEEE Access* 7, 66680–66688.
[52] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.* 100–105.
[53] Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* 6706–6713.
[54] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7008–7024.
[55] Jinsong Su, Jialong Tang, Ziyao Lu, Xianpei Han, and Haiying Zhang. 2019. A neural image captioning model with caption-to-images semantic constructor. *Neurocomputing* 367, 144–151.

Received December 2019; revised April 2020; accepted July 2020