Multi-level Visual Fusion Networks for Image Captioning

Dongming Zhou [a], Canlong Zhang* [a], Zhixin Li [a] and Zhiwen Wang [b]

[a]*Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, China, GuiLin 541000*
[b]*College of Computer Science and Communications Engineering, Guangxi University of Science and Technology, China, Liuzhou 545006*

***Abstract*—Image captioning is a multi-modal complex task in machine learning. Traditional methods focus only on entities in visual strategy networks, and can't reason about the relationship between entities and attributes. There are problems of exposure bias and error accumulation in language strategy networks. To this end, this paper proposes a multi-level visual fusion network model based on reinforcement learning. In the visual strategy network, multi-level neural network modules are used to transform visual features into feature sets of visual knowledge. The fusion network generates function words that make the description more fluent, and is used for the interaction between the visual strategy network and the language strategy network. The self-criticism strategy gradient algorithm based on reinforcement learning in language strategy networks is used to achieve end-to-end optimization of visual fusion networks. We evaluated our model on the Flickr 30K and MS-COCO datasets, and verified the accuracy of the model and the diversity of model learning subtitles through experiments.Our model achieves better performance over state-of-the-art methods.**

***Index Terms*—Image Captioning, Visual Fusion, Reinforcement Learning, Policy Network**

I. Introduction

Image captioning can be understood as giving a picture a piece of text that is described in a natural language. Image captioning and Visual Question Answering (VQA) [16] system is an intersection of computer vision and natural language processing (NLP), which is a more challenging task than target detection, image classification and semantic. In the extraction of image entities and attributes, we also infer the relationship between entities and attributes.

Inspired by machine translation, the encoder-decoder [1], [10], [11]frameworks are widely used in image captioning. The encoder side extracts image features using a convolution neural network (CNN). The decoder end inputs the extracted image features into a Long Short-Term Memory (LSTM) network, then outputs a sequence describing the image. However, CNN does not recognize the relationship between scenarios based on context when extracting visual features. When using visual attention, only one visual area can be fixed at each step, and there is no interaction between different visual areas. When dealing with complex scene combinations, the sequence captioning error occurs over time.

***Corresponding author: zcltyp@163.com**

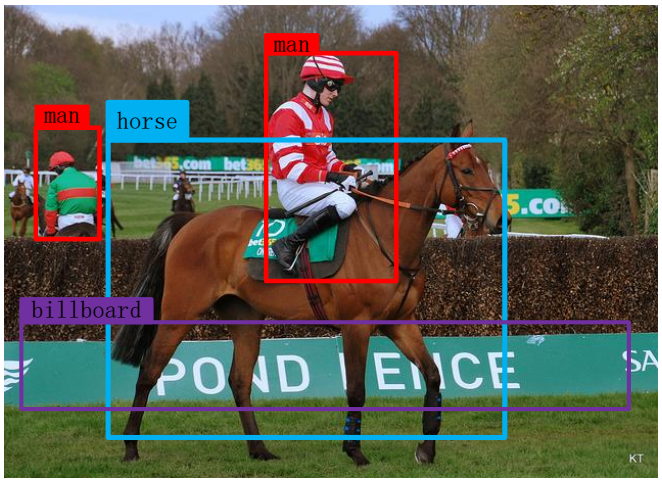

Fig. 1. MVF first detects targets present in the image.

The human visual system is not a simple scene overlay when describing a colorful world. Not only will the visual fusion be based on the context, but also multi-step inference based on the received visual signals. This is a human talent, but it is a challenging task for the machine. Although visual representation learning [13] and language modeling [17] have made great progress in their respective ways in recent years, how to establish cross-modal connections between vision and language is still an urgent problem to be solved. The multi-level visual fusion networks proposed in this paper does not only fix the current visual attention when generating the sequence, but interprets the visual information of the previous time step as a scenario, and then uses the current visual attention perception determines whether the situation is conducive to the generation of the next word. Compared with the traditional attention model [19], which encoder context into the hidden state of LSTM, the method presented in this paper shows the role of context in predicting sequences. As shown in Figure 1 , the target detection module first detects entities in the image (*such as*: "man", "horse") and considers their relevant regions. The attribute module converts the attributes extracted by CNN into entity attribute knowledge feature sets,

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

and generates adjectives describing the entities. (*eg*: ”red”, ”brown”). Then the relational module infers the relationship between the entity and the entity, the entity and the attribute in multiple steps. For example, according to the entities ”man” and ”horse”, according to the context, ”is riding”. Finally, when generating sequences, the fusion network can make the generated description more coherent and consistent with grammatical rules.

In summary, the main contributions of this paper are as follows:

i) An end-to-end MVF model is proposed, which fusion multi-level visual features and enriches visual language tasks through the matching of different neural modules.

ii) On this basis, the attention module is redesigned. The adaptive attention in this paper has different attention strategies for different parts of speech, which reduces the interference of non-visual gradients on visual information.

iii) Finally, the design module with a controller has used to fine-grain the tasks of each neural module described in the image. The experimental results show that the CIDEr score of the MVF model is significantly improved. MVF is a general framework that supports potential improvements.

II. Related Work

A. Image captioning

The study of image caption has a long history. Early image captioning methods used template-based filling first constructing a sentence pattern when generating a captioning, and then filling the word into a fixed sentence pattern. This method generates a captioning of the template and the word function is not joint training, so performance and evaluation indicators do not perform well. Inspired by machine translation, the attention-based encoding and decoding framework [1], [3], [10], [21] has recently been used to achieve superior performance. Among them, Lu *et al*. [10] found that the gradient of the non-visual word would mislead or reduce the effectiveness of visual information, and proposed an adaptive gating mechanism, and decoders had different language strategies for different words. Chen *et al*. [3] proposed channel attention, studied the effects of visual attention model on space and channels, and applied the attention mechanism to the coding end. Anderson *et al*. [1] applied target detection technology to image captioning and proposed a bottom-up attention mechanism, which could make the image captioning more natural, but could not deduce the relationship between entities and attributes in the image. The above model only focuses on the visual attention of the current time step and ignores the consideration of visual context over time, which we believe is a key factor leading to the obscurity of the captioning sequence. For this reason, we introduce the situational awareness fusion network, which integrates the previous time step content with the current time step attention. The attention network generates more efficient feature vectors, which are then inputted into the language strategy network.

B. Strategy Gradient

The image describes the use of cross entropy loss as a loss function during training, using backpropagation to maximize the likelihood of a ground-truth. As discussed in [2], the training LSTM network uses Teacher forcing—the input of the actual word of the image captioning label at each step, but at the time of testing, the input to the next moment of the LSTM network is the output of a moment is not a real word, which leads to exposure bias. The word at the next moment in the test depends on the word generated by the previous time. If the word generated at a certain moment is not good, the error will be accumulated, and the word generated later will be affected. Cross entropy is used during training, but there are problems in the evaluation when using BLEU, CIDEr, and ROUGE. The concept of sequence decision-making is introduced in reinforcement learning, and the problem of exposure bias is well solved during training [9], [14], [22]. The decision forces the agent to think about the next move, state, and reward. In the image captioning, the reward can select the score of CIDEr, the state is the image, the generated word and the scene, and the action is to select the visual feature vector and the next word. Ranzato *et al*. [12] proposed a strategy gradient method using Monte Carlo search technique to train sequence-like tasks, and solved the problem of exposure bias and non-differentiable measurement of test sequences. Rennie *et al*. [15] proposed a self-critical training method. Since the biased baseline can be an arbitrary function and does not depend on actions, the model uses the reward of generating words in the test phase as a baseline. Inspired by the above work, this article has made some improvements based on the reinforcement learning framework. At each time step, the output of one time on the language policy network and the visual policy network is simultaneously input to the next moment. Therefore, when generating a visual captioning sequence, not only the current time step vision but also the context awareness can be used for multi-step reasoning.

III. MULTI-LEVEL VISUAL Fusion NETWORK

This section will introduce the architecture diagram of multi-level visual fusion network in detail. As shown in Figure 2, it is a MVF structure diagram based on codecs. MVF mainly includes three parts: visual network, fusion network, and language network. In section III-A, the visual network will be introduced. We define the form of the problem, the image captioning the task as a continuous decision making process. The visual network includes a CNN and three neural modules, which generate feature vectors for language decoding. In section III-B, we will introduce the fusion network, which includes adaptive attention module, module controller, and multimodal attention. Section III-C will introduce the language network. LSTM inputs part of the accumulated situational awareness into the module controller and the fusion network for multi-step reasoning.

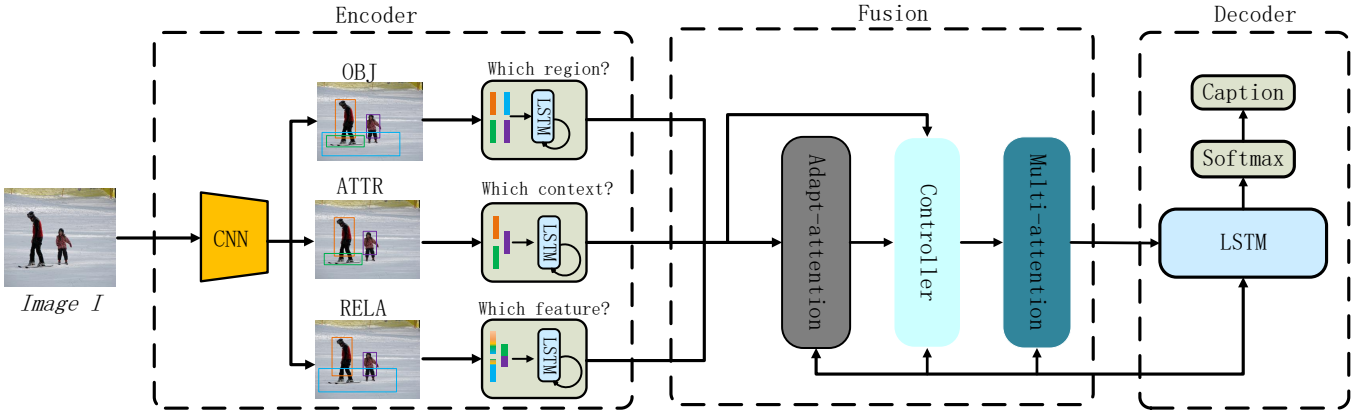

Fig. 2. MVF structure diagram with a loop fusion process added after the to better represent decoding. The dashed line from the LSTM module to the control module and the fusion module indicates the context in which the sentence needs to be observed.

A. Visual Network

In this paper, Given image I, the local features $\{v_1, \cdots, v_k\}$ of the image are extracted using Faster R-CNN [13], and v_i is the feature of the localized area of the mean, and the first k ROIs are selected for each image. In the meantime, the visual representation of the image is $x = \{x_1, x_2, \cdots, x_T\}$, where x_t represents the action of the visual network at time t, the corresponding description sequence is $Y = \{y_1, y_2, \cdots, y_T\}$, y_t is the action of language network at moment$t$. At time step t, each policy network acts as a agent and receives the state f_t of the environment, then produces a series of $(a_t|a_{1:t-1}, \theta)$. At time t the visual context $\{x_{i<t}\}$, generates the sequence $\{y_{i<t}\}$, and evaluates the score obtained from the real sequence ground-truth and the generated sequence Y. Input $P_t = \{p_{t,1}, p_{t,2}, \cdots, p_{t,d}\}$ is a set of attribute features. The network selects a visual representation y from the input according to this action. In the visual network, LSTM is used to encode the environment state, and φ is used as the symbol of the network. Then y is formalized as:

$$y = \sum_{i=1}^{k} \varphi_{i,t}, p_{t_i} \tag{1}$$

When calculating the action probability distribution, follow the attention mechanism:

$$a_{i,t} = W_a^T \tanh\left(W_h v_i + W_p h_t\right) \tag{2}$$

$$\varphi_{i,t} = \text{softmax}\left(a_t\right) \tag{3}$$

Where,W_a,W_h,W_p are trainable hyperparameters. the hidden state of LSTM can be calculated as:

$$h_t = \text{LSTM}\left(x, h_{t-1}\right) \tag{4}$$

We note that at time step t the agent in reinforcement learning is a network in the visual network, and the action is to select the next visual feature v_t and visual representation y_t. The reward in reinforcement learning comes from the feedback of the language network evaluation indicators. Therefore, it is only necessary to determine the environmental state f_t of each network and the input attribute feature p_t at time t.

1) Object Detection Module: The target detection module f_t^o is composed of three parts at time t. The hidden state h_{t-1}^l at a time on the language policy network LSTM, hidden state h_{t-1}^l at the moment of the language strategy network LSTM, the regional feature $\bar{v} = \frac{1}{k}\Sigma_i v_i$ of the mean pooling, and the word embedding matrix at the moment y_{t-1}. Then the environment code f_t^o of the target detection module at time t can be expressed as:

$$f_t^o = \text{LSTM}\left[h_{t-1}^l, \bar{v}, W_p \Pi_t\right] \tag{5}$$

$W_p \in \mathbb{R}^{|\Sigma| \times M}$ is a word embedding matrix learned from scratch, $\prod$ is one-hot encoding matrix. The area feature detected at each time step is $p_t^c = \{\bar{v}_1, \bar{v}_2 \ldots \bar{v}_k\}$, the output of the target detection module at time t is a single feature v_t^o , which will be used in the fusion network.

2) Attribute Module: The attribute module is used to detect the attributes of the entity. It is designed to transform the attributes of the entity into a feature set of attribute knowledge and generate adjectives such as "black" and "red". This module takes the last fully connected layer of ResNet101 and Faster R-CNN as the feature set of attribute knowledge. However, not all attributes are helpful for word generation at the moment. At time t the output of the attribute module is $\{v_{i<t}\}$. The attribute module combines it with the detected regional features, and selects the feature with the most information as the input of the LSTM in the visual strategy network. The environment state of the attribute module is defined as:

$$f_t^a = \text{LSTM}\left[f_t^o, \frac{1}{k}\sum\nolimits_i v_i, W_e \Pi\left(y_{t-1}\right)\right] \tag{6}$$

Input feature $p_t^a = \{v_1, \cdots, v_{t-1}\}$ then merge the regional feature with v_t^o to get feature $\delta_{t,i}$ of the attribute module:

$$\delta_{t,i} = \frac{e^{p_{t,i}} \cdot w_c^T}{\sum_{i=1}^{N} e^{p_{t,i}}} \tag{7}$$

Among them, W_c^T is a projection of the attribute feature as a region feature to the original dimension, and $e^{p_{t,i}}$ is the logarithm of the eigenproperty vector. Attribute characteristics are used in the relationship module.

3) Relationship Module: The relation network represents attributes as feature sets of potential interactions between two objects, which helps to generate verbs like preposition "in" or "riding". The module hidden state of the language policy network $t-1$, the mean pooled region feature, and the words generated at the previous moment into the environment state:

$$f_t^r = \text{softmax}\left(W_a^p f_t^o + W_c^p f_t^a + b^p\right) \tag{8}$$

The input features of the relational network come from the attribute module network:

$$p_t^r = p\left(\delta_{t,i} | r, \mathbf{y}_{1:t-1}\right) p\left(r | \mathbf{y}_{1:t-1}\right) \tag{9}$$

Then the output of the relation network at time t is regarded as the relation eigenvector v_t^r.

B. Fusion Network

There are three modules in the fusion network. Adaptive attention module, module with controller and multimodal attention. Adaptive module is used to reduce the effectiveness of non-visual word gradient on visual information, module collocation controller is used to match modules in visual network and adaptive module to generate complete description sentences, and multi-modal attention is used for visual description output. The function of the adaptive attention module is to generate non-visual information words that make the description sequence smoother, such as "a" or "an". The decoder should have different attention strategies for words of different parts of speech, and the generation of non-visual information words depends more on semantic information than visual information.At each time step t, h_t can be known from Eq 2, so the standardized attention weight $\alpha_{i,t}$ can be calculated as:

$$\alpha_{i,t} = W_a^T \tanh\left(W_{va} v_i + W_{ha} h_t\right) \tag{10}$$

Among them, parameters that can be learned in training of $W_{va} \in \mathbb{R}^{H \times V}$,$W_{ha} \in \mathbb{R}^{H \times M}$ and $W_a \in \mathbb{R}^H$ At each moment, $\widehat{C}_t$ decides whether the function word "a" or "an" is generated by the model or by the language network. $\widehat{C}_t$ can be expressed as:

$$\widehat{c}_t = \beta_t c_t + (1 - \beta_t) h_t^l \tag{11}$$

When $\beta_t = 0.5$ is selected during the experiment, the experimental effect is the best. As can be seen from the adaptive feature vector $\widehat{C}_t$ and the standardized attention weight $\alpha_{i,t}$, the environment encoding of the adaptive attention module is:

$$f_t^s = \text{LSTM}\left[\sum_{i=1}^{k} \alpha_{i,t} \widehat{c}_t, \bar{v}, W_e \Pi_{t-1}\right] \tag{12}$$

It will be used in each step of the language strategy network and will be considered part of the context in subsequent time steps. We noticed that although the input characteristics of each module are different, the coding environment is partially the same. In the experiment, in order to avoid the occurrence of overfitting and reduce the complexity of the model, this article shares the LSTM parameters.

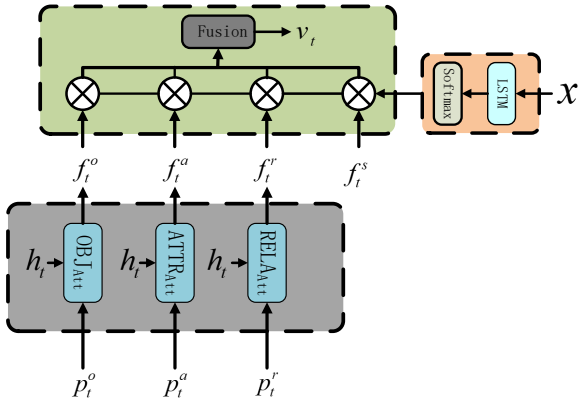

Fig. 3. Module controller structure, which soft-merges the environmental states encoded by four modules into a fusion feature.

The module controller will be used to describe the word collocation of the sequence. Its structure is shown in Figure 3 , although the existing image description methods based on slot filling can realize simple visual reasoning, how to combine simple visual neural modules to define a complete set of visual reasoning in the face of complex scenes. Before visual features are fused, the input attribute features are transformed into information features:

$$\begin{cases} f_t^o = \text{Att}_{Obj}\left(p_t^o, h_t\right) \\ f_t^a = \text{Att}_{Atr}\left(p_t^a, h_t\right) \\ f_t^r = \text{Att}_{\text{Rela}}\left(p_t^r, h_t\right) \end{cases} \tag{13}$$

Where, h_t is the state of LSTM in the module controller at time$t$. And input features p_t^o, p_t^a and p_t^r of object detection module, attribute module and relational module are generated by three visual modules in section III-A. According to Eq 13, the three transformed features can be obtained, and the adaptive attention module can get f_t^s. The module controller generates four soft weights for them, and the process of generating soft weights is expressed as:

$$\begin{cases} x = \text{Concat}\left(p_t^o, p_t^a, p_t^r, h_t\right) \\ w = \text{Softmax}(\text{LSTM}(x)) \\ v_t = \text{Concat}\left(w_o f_t^o, w_a f_t^a, w_r f_t^r, w_f f_t^s\right) \end{cases} \tag{14}$$

Among them, x is a concatenation of three visual input vectors, and w is a four-dimensional soft attention vector. The resulting vector vt is sent to the language strategy network for decoding. Visual cues and language context are indispensable conditions for generating soft weights and modules. In addition, the layout of the neural module at the next moment is highly similar to the layout at the previous moment. Therefore, LSTM is used to accumulate this knowledge and then update to generate new soft weights.

C. Language Network

At each time step, MVF generates a visual representation of the merged situation, selecting the word that best fits the current situation. The language strategy network takes the multimodal visual feature vector and the implicit state h_t^c of

TABLE I
DEMOGRAPHIC PREDICTION PERFORMANCE COMPARISON BY THREE EVALUATION METRICS.

Methods	MS-COCO datasets							
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Adative [10]	74.7	58.4	43.9	34.4	25.7	54.8	105.6	-
SCST [15]	77.6	61.8	47.4	34.9	25.2	56.6	116.4	18.1
SCA-CNN [3]	70.5	49.6	43.7	31.6	25.1	-	97.3	17.8
Top-Down [1]	73.8	54.6	44.2	36.2	26.1	55.4	120.4	19.1
NBT [11]	75.5	56.2	43.8	34.9	26.4	-	108.9	20.4
RFNet [6]	78.1	60.5	47.4	37.5	26.3	57.3	125.7	19.7
GHA [18]	73.3	56.4	46.2	35.1	25.5	55.8	99.8	18.9
CGO [23]	78.5	60.2	48.7	34.4	25.7	55.8	125.2	-
SGAE [20]	77.6	63.8	48.4	36.9	25.2	57.6	126.4	19.1
MVF	**80.5**	**64.1**	**49.3**	**38.5**	**28.2**	**58.1**	**128.1**	**22.1**

the work as input, and then updates the implicit state of LSTM:

$$h_{t+1}^{l} = \mathrm{LSTM}\left(\left[h_t^c, v_t\right], h_t^l\right) \tag{15}$$

When calculating the distribution of words in the vocabulary, this paper uses the fully connected layer as the hidden state of LSTM. The probability of each word after normalization by the softmax function can be expressed as:

$$\varphi\left(y_t | y_{1:t-1}\right) = \mathrm{softmax}\left(W_p h_t^l + b_p\right) \tag{16}$$

Where b_y is the bias value and W_y is the weight parameter, both of which are learned in training. The complete captioning sequence is the product of all time step conditional distributions, which can be expressed as:

$$\varphi\left(y_{1:T}\right) = \prod_{t=1}^{T} \varphi\left(y_t | y_{1:t-1}\right) \tag{17}$$

IV. EXPERIMENTS

In this section, we will first introduce the data set used in the experiment and the hyperparameters set in the experiment. Then compare the different methods and discuss some details during the experiment in detail. Finally, quantitative and qualitative analysis of the experimental results.

A. Dataset

In selecting the experimental dataset, the image described herein is the most popular MS-COCO dataset [8]. There are 82,783 pictures in the data set for training, 40,504 pictures for evaluation, and 40,775 pictures for testing, each with 5 sentences. Since the official test set is not publicly available, this article follows the Karpathy split [7] used in previous work, 113,287 images for training, 5,000 images for evaluation, and 5,000 images for testing.

B. Parameter Setting

In order to make Faster R-CNN faster in convergence, this paper uses ResNet-101 [5] to initialize network parameters.In the cross entropy training process, the Adam optimization algorithm is used for optimization. The initial learning rate is set to 5e-4, and the index is reduced by 0.8 every 5 cycles. The total training period is set to 100 cycles, and after 40 cycles, intensive learning training is started, the learning rate is set to 5e-5, and the contraction is 0.1 every 10 cycles. The batch size is set to 100 pictures at a time, and the beam width of the beam search is set to 5. The number of layers of the LSTM is set to 1, the number of hidden cells is set to 1300, the number of hidden cells in Eq 12 is set to 1024, and the word embedding size is set to 1000.

C. Comparison with State-of-the-Art Methods

In the comparison of experimental methods, this paper compares the image description based on reinforcement learning with the traditional image description based on codec framework. The methods of using reinforcement learning are: SGAE [20],CGO [23],Top-Down [1],NBT [12], RFNet [6], GHA [18]. The results of the experiment using the MS-COCO dataset are shown in Table I. MVF achieved state-of-the-art results on multiple evaluation indicators. As shown in Figure 4, if RL-based fine-tuning is used, the experimental results will be further improved. Traditional image description methods ignore the influence of context on the description sequence, and some important visual information will be lost as the time step increases. The MVF model proposed in this paper alleviates this problem through the fusion of visual features at different stages. It can be seen from the results that the MVF model proposed in this paper achieves better performance than traditional methods in almost every metric, which clearly shows the effectiveness of the proposed MVF model.
As shown in Table II, it is the MVF online test result. Please note that this model exhibits good performance compared to industrial companies with extensive computing resources. With the fine-tuning of parameters and the improvement of hardware resources, we believe that the MVF model will have a lot of potential improvement. In order to better understand MVF, as shown in Figure 6, this paper visualizes the output prediction of the language policy network. Perceptually fused networks

TABLE II
COMPARISON OF MVF ON THE MS-COCO TEST SERVER AND SEVERAL OTHER METHODS EVALUATION INDICATORS.

Methods	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
PG-CIDEr [9]	75.1	88.6	59.1	84.2	44.5	73.8	33.6	63.7	27.5	33.9	55.1	69.4	104.2	107.1
SCST [15]	78.2	90.7	61.9	86.0	47.0	75.9	34.6	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Google-NIC [19]	71.2	87.6	53.9	80.2	41.2	69.7	30.9	58.7	25.4	34.6	52.9	68.4	94.2	94.7
Adative [10]	74.4	92.0	58.4	84.5	43.9	74.6	33.5	63.7	26.4	36.0	55.1	71.2	104.2	105.9
SCA-CNN [3]	71.2	88.9	54.5	81.4	41.0	69.1	31.6	57.9	23.8	33.2	53.1	67.4	91.2	92.1
Top-Down [1]	80.2	93.7	64.2	87.9	49.1	80.2	36.9	68.5	27.0	35.9	56.4	70.8	117.5	124.1
NBT [11]	69.1	94.6	59.2	86.7	48.7	**81.3**	34.7	67.5	27.3	34.6	55.6	71.6	108.9	112.3
GHA [18]	72.9	93.7	56.0	81.8	41.9	70.8	31.3	59.8	25.2	34.1	53.3	68.3	95.4	96.3
RFNet [6]	78.4	94.0	64.9	88.2	50.1	80.1	38.0	68.0	28.2	37.2	58.2	73.1	122.9	125.1
SCNT [4]	77.6	93.1	61.3	86.1	46.5	76.0	34.8	64.6	26.9	35.4	56.1	70.4	112.6	115.3
CGO [23]	77.8	92.9	61.2	85.5	45.9	74.6	33.4	62.5	26.4	33.4	55.4	69.1	110.2	112.1
SGAE [20]	78.6	93.3	64.5	88.9	50.7	80.4	37.5	68.7	28.2	37.2	**58.6**	73.6	124.8	126.5
MVF	**81.0**	**95.3**	**65.0**	**89.3**	**51.2**	79.6	**38.5**	**69.7**	**28.7**	**38.0**	57.3	**74.4**	**127.6**	**128.4**

can not only focus on single entity objects in the graph, such as dogs, teddy bears, and beds. And you can generate a combined word placement that connects the physical teddy bear and the bed. The combined generated words make the captioning more human-like, and in the case of a deep understanding of the scene, it is possible to avoid generating rigid captioning sentencesfluctuates less. Compared with the NBT model, it fully reflects the advantages of multi-level visual fusion. Adaptive attention reduces the interference of non-visual gradients on visual information, and has different attention strategies for different parts of speech. Modules and controllers enrich visual semantic information, define a complete set of visual reasoning steps, and complete complex description tasks through the combination of neural modules.

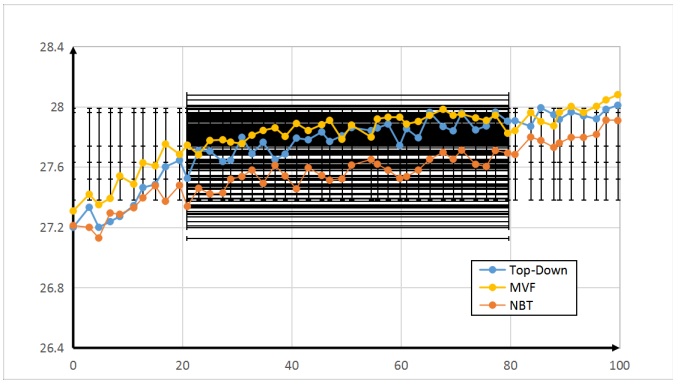

Fig. 4. Fine-tuning RL parameters and METEOR evaluation index have been improved to some extent.

In addition to the BCMR indicator, the SPICE evaluation indicator of MVF has also improved to some extent. SPICE subdivides semantic categories and has the greatest correlation with human visual judgment results. In Figure 5, compared with Top-Down and NBT, MVF has improved the classification index of semantic categories. MVF fully considers the relationship between entities and attributes, and performs knowledge reasoning on visual context.

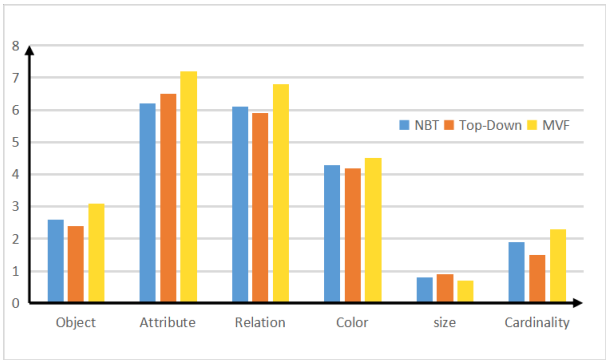

Fig. 5. Comparison diagram of SPICE semantic classification results of MVF.

D. Ablation study

The collocation between different neural modules has different variants. In this section, the MVF model is studied in detail through the ablation comparison experiments. MVF/$_{O+Self}$ means using the target detection module and adaptive attention module, MVF/$_{O+Cont}$ means using the target detection module and module with the controller, MVF/$_{O+A+Self}$ means using the target detection module, attribute module and adaptive attention module, MVF/$_{O+R+Self}$ means using target detection module, attribute module and adaptive attention module, MVF/$_{O+R+Cont}$ means target detection module, relationship module and module with controller. The experimental results are shown in Table III. Compared with MVF/$_{O+Self}$ and MVF/$_{O+Cont}$, the MVF/$_{O+Self}$ has a certain degree of improvement, which

Fig. 6. Visually describes the sequence generation process.

indicates that the function words are not the key factors affecting the final description sequence, but the function words in the proper position make the description more natural. MVF/$_{O+A+Self}$ is significantly improved compared to MVF/ $_{O+R+Self}$, indicating that the rich visual fusion information comes from the last few time steps. The convolutional neural network extracts the primary image features in the shallow layer and the relationship network can effectively use visual context visual information. Compared with MVF/$_{O+A+Self}$ and MVF/$_{O+R+Cont}$, it shows that the module with the controller can handle the matching between neural modules very well, avoiding the over-fitting phenomenon of the entire network.

TABLE III
PERFORMANCE OF ABLATION EXPERIMENTS ON MS-COCO DATASET.

Modle	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
MVF/$_{O+Self}$	36.9	26.6	56.6	123.1	20.9
MVF/$_{O+Cont}$	37.8	26.8	57.0	123.6	21.2
MVF/$_{O+A+Self}$	37.2	27.4	57.4	126.7	21.1
MVF/$_{O+R+Self}$	37.6	27.4	57.4	124.7	21.1
MVF/$_{O+R+Cont}$	**38.5**	**28.3**	**57.9**	**128.3**	**21.6**

When training a generated sequence using a strategy gradient, the reward function can choose CIDEr, BLEU, METEOR, ROUGE, and SPICE. As shown in Table IV , the horizontal axis represents the evaluation index at the time of training, and the vertical axis represents the evaluation index at the time of evaluation, and optimization of different evaluation indexes may result in different results. Through experimental comparison, it is known that the specific evaluation index is optimized during training, and the index can obtain the best performance when tested. Optimizing the overall performance of BLEU and CIDEr is the best, but the cost of computing is more than CIDEr when using BLEU as a reward, so this article uses CIDEr as an evaluation index.

TABLE IV
PERFORMANCE OF ABLATION EXPERIMENTS ON MS-COCO DATASET.

Metric	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
BLEU-4	**38.5**	36.6	37.4	37.9	37.3
METEOR	26.5	**28.4**	27.3	27.2	26.8
ROUGE	56.7	57.6	**58.1**	56.4	57.0
CIDEr	114.5	113.4	124.0	**128.4**	122.3
SPICE	20.2	19.4	19.8	20.6	**21.1**

V. CONCLUSION

This paper proposes an image description framework based on visual fusion network. MVF makes full use of the advantages of the visual environment for visual reasoning when generating description sequences, so it can process complex visual combinations over time. Based on extensive comparative experiments and ablation experiments, the validity of the MVF model is tested on the MS-COCO dataset. Compared with the deep learning-based image description model, the experimental results of the MVF model are significantly improved. In future work, 1) apply the MVF model to other visual reasoning tasks—visual question answering systems, and 2) migrate the exploration scenarios into scene graph generation and video description.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Nos. 61866004, 61663004, 61966004, 61962007, 61751213), the Guangxi Natural Science Foundation (Nos. 2018GXNSFDA281009, 2017GXNSFAA198365, 2019GXNSFDA245018, 2018GXNSFDA294001), the Innovation Project of Guangxi Graduate Education(Nos. XYCSZ2020071), the Guangxi "Bagui Scholar" Teams for Innovation and Research Project, and Guangxi Collaborative Innovation Center of Multi-source Information Integration and Intelligent Processing.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 1171–1179, 2015.

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[4] Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. Self-critical n-step training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6300–6308, 2019.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018.

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[9] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017.

[10] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard 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*, pages 375–383, 2017.

[11] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

[12] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[14] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298, 2017.

[15] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[16] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[18] Qingzhong Wang and Antoni B Chan. Gated hierarchical attention for image captioning. In *Asian Conference on Computer Vision*, pages 21–37. Springer, 2018.

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

[20] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019.

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

[22] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

[23] Yue Zheng, Yali Li, and Shengjin Wang. Intention oriented image captions with guiding objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8395–8404, 2019.