IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 13, 2020, accepted March 9, 2020, date of publication March 17, 2020, date of current version April 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981513

Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards

CHUNLEI WU[1], SHAOZU YUAN[1], HAIWEN CAO[1], YIWEI WEI[2], AND LEIQUAN WANG[1]
[1]College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China
[2]School of Petroleum Engineering, China University of Petroleum-Beijing at Karamay, Karamay 834000, China

Corresponding author: Leiquan Wang (richiewlq@gmail.com)

This work was supported in part by the Key Research and Development Plan of Shandong Province under Grant 2019GGX101015, in part by the Fundamental Research Funds for the Central Universities under Grant 18CX02136A and Grant 19CX05003A-11, and in part by the National Natural Science Foundation of China under Grant 61671482.

ABSTRACT Image caption based on reinforcement learning (RL) methods has achieved significant success recently. Most of these methods take CIDEr score as the reward of reinforcement learning algorithm to compute gradients, thus refining the image caption baseline model. However, CIDEr score is not the sole criterion to judge the quality of a generated caption. In this paper, a Hierarchical Attention Fusion (HAF) model is presented as a baseline for image caption based on RL, where multi-level feature maps of Resnet are integrated with hierarchical attention. Revaluation network (REN) is exploited for revaluating CIDEr score by assigning different weights for each word according to the importance of each word in a generating caption. The weighted reward can be regarded as word-level reward. Moreover, Scoring Network (SN) is implemented to score the generating sentence with its corresponding ground truth from a batch of captions. This reward can obtain benefits from additional unmatched ground truth, which acts as sentence-level reward. Experimental results on the COCO dataset show that the proposed methods have achieved competitive performance compared with the related image caption methods.

INDEX TERMS Image caption, reforcement learning, attention mechanism.

I. INTRODUCTION

The goal of image caption is to automatically generate a natural language description of an given image. It is a challenge to transform visual information into textual languages. On the one hand, the machine has to obtain a comprehensive understanding of image content from multi-level visual features. On the other hand, the image caption algorithm is required to revise the rough semantic concept to human-like natural language descriptions step by step. Recent advancement of deep learning has significantly improved the quality of caption generation including attention mechanism and reinforcement learning.

The encoder-decoder paradigm [16] is the mainstream approach of image caption. Vinyals *et al.* [20] compress an entire image into a static representation by utilizing spatially pooled CNN feature maps to generate captions. Attention mechanisms [1], [12] improve the performance of captions by learning to focus on the regions of the image adaptively. Only single LSTM is used as the visual information handler as well as the language generator. The language generator is weakened by the simultaneous visual handler. The authors of [1] proposed the top-down architecture with two independent LSTM layers. The first LSTM layer acts as a top-down visual attention model, and the second LSTM layer as a language generator. All the image caption methods mentioned [1], [5], [12], [32] adopt high-level visual feature of CNN(last convolution layer) as image encoders. However, low-level visual features have been neglected, which is also beneficial for understanding the images. Employing multi-level features fusion can also be a better solution for image caption due to the complementariness among multi-level features. However, early fusion method has an unsatisfying performance. Meanwhile, late fusion and graph-based fusion are not satisfactory for image caption based on encoder-decoder [3]. Therefore, how to fuse multi-level visual features into image caption model is worthy of consideration.

Image caption model is typically trained to maximize the log likelihood of training set, which is also known as Cross-Entropy (XE). This makes the image caption model

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

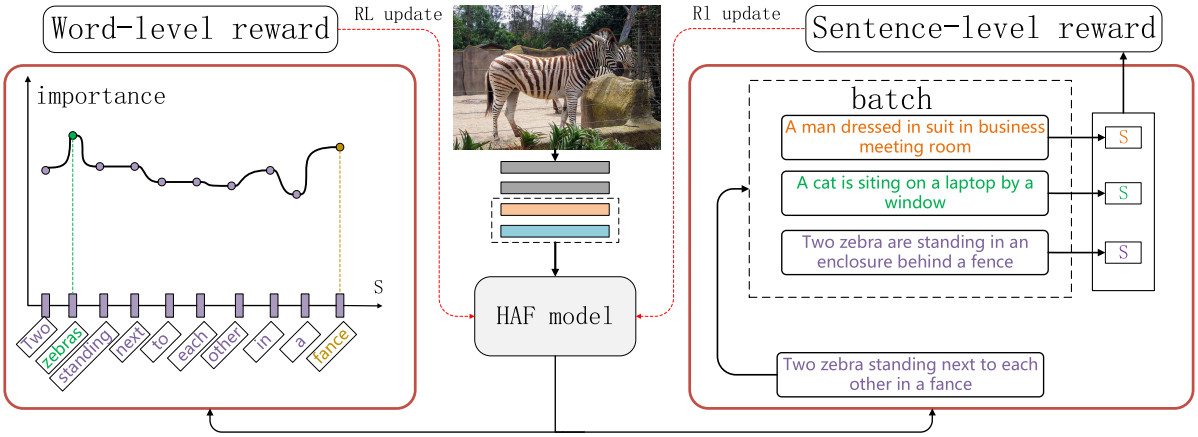


FIGURE 1. Overview of the proposed sentence-level and word-level reward respectively. On the left, word-level reward is produced by adaptively revaluating how importance the word is. On the right, sentence-level reward is constructed by sentence scores, which is calculated by scoring network.

sensitive to unusual or aberrant captions, rather than optimizing for stable output around the human consensus of what an appropriate caption would be. Moreover, caption models are usually evaluated by computing a variety of different metrics on a test set, such as BLEU [11], ROUGE [30], METEOR [2] and CIDEr [27]. The mismatch between objective function and evaluation leads an unsatisfactory effect on image caption model. This problem can be solved through the reinforcement learning (RL) [25] like Policy Gradient [21] and Actor-Critic [25]. Reinforcement learning can optimize non-differentiable, sequence-based evaluation metrics directly. The authors of SCST [22] applied CIDEr as reward when using Policy Gradient method [21], which is prone to generate more human-like captions.

In SCST [22], the same reward is given to each word as gradient weight. Nevertheless, not all words should be given equal reward in a sentence—different word may be of different importance. Reference [18], [34] utilize Monte Carlo rollouts to estimate the importance of each word. However, it has to generate abundant sentences, leading to an expensive cost on time complexity. Based on actor-critic, the author of [4] adopts value network to assess different word. Nevertheless, evaluation metrics (eg.CIDEr, BLEU) can not be optimized directly. In this paper, word-level reward is exploited to revise the image caption model based on RL training, aiming to address the different importance of each generated word (see the left part of Figure 1).

Computing the evaluation metrics (eg.CIDEr, BLEU) as reward signal is an intuitive way in RL training to generate more human-like captions. The main purpose is to approximate human judgment of appropriateness and consensus [27]. However, these evaluation metrics are not the sole criterion to judge the quality of a generated caption. The quality of a generated caption can also be evaluated by whether it matches the corresponding ground-truth in the scoring network (see the right part of Figure 1). From the perspective of information utilization, the traditional CIDEr reward makes full use of the matched ground truth information, while the scoring reward obtains benefits from additional unmatched ground truth.

In this paper, a Hierarchical Attention Fusion (HAF) model is presented for image caption. Multi-level feature maps of Resnet [13] are integrated with hierarchical attention. HAF acts as a baseline for RL-based image caption approach. Moreover, multi-grained rewards are presented in RL phase to revise the proposed HAF. Specifically, Revaluation network (REN) is exploited for reward revaluation by estimating the different importance of each word in a generating caption. The revaluated reward is achieved by weighting the CIDEr scores, where the different weights are calculated from REN. The revaluated reward can be regarded as word-level reward. To obtain benefits from additional unmatched ground truth, Scoring Network (SN) is implemented to provide a score for the generating sentence from a batch of ground truth as sentence-level reward.

To summarize, the contributions of this paper are as follows:

1) A Hierarchical Attention Fusion (HAF) model is presented as a baseline of RL training for image caption. HAF adopts multiple attention to attend hierarchical visual features of CNN, which take full advantages of multi-level visual information.
2) Revaluation network (REN) is proposed for facilitating revaluation reward calculation, which assignes different importance to the generated words in a sentence automatically during the RL training phase.
3) Scoring network (SN) is performed to provide a score as sentence-level reward. SN evaluates a generated caption from both the correspondence to the matched ground truth and the discriminativeness to the unmatched ground truth, which enforces generated captions to be the best matching of its corresponding ground truths.

II. RELATED WORK

In recent years, there is an extensive work on image caption. And we summarize some of the most relevant works below:

A. CAPTION MODELS BASED ON ENCODER-DECODER

Encoder-decoder network is wildly used in most image caption methods. Typically, the encoder is a convolutional neural network(CNN), and decoder is a recurrent neural network. Many approaches [6], [8], [10], [20], [29] apply this framework to encode an image as a pooled feature vector to represent the image, and then feed the pooled vector into decoder to generate captions. However, all these approaches utilize a static and spatially pooled representation of the image.

B. ATTENTION MECHANISMS

To dynamically lead the caption model to focus on different features at generating each word, attention mechanisms have been wildly used in many caption models. Reference [12] firstly introduces two different attention mechanisms, soft attention and hard attention, to focus on specific regions of an image automatically. Reference [38] designs Adaptive Attention for image Captioning. In [5], the authors enhance their models by detecting objects in images. Lu *et al.* [9] study when a model should attend to an image at generating a sequence of words. Some network structures, such as LSTM and CNN, are also extended to enhance attention performance. Reference [35] proposes R-LSTM to find which parts of the captions are more essential to the image. And [37] applies M-LSTM to interact with both visual and textual features to capture a high-level representation. Reference [36] integrates attention into a CNN with an emotion polarity constraint. For augmenting high-level attributes from images, [26] presents variants of attention architectures to complement image representation for sentence generation. To utilize the salient regions of the image, [1], [40] enables attention to be calculated at the level of objects. However, low-level visual features have been neglected, which is also beneficial for understanding the images.

C. REINFORCEMENT LEARNING OPTIMIZATION

As discussed in many related papers [17], [19], [23], captioning systems trained by using the cross entropy loss have both the exposure bias [19] and non-differentiable task metric issues. To overcome the limitations of maximum likelihood, the image captioning process can be formulated as a Reinforcement Learning [25] question, where the LSTM [7] language model is the agent, in which each action corresponds to predicting the next word.

Reference [33] utilizes a ''policy network'' and a ''value network'' to adjust rewards of similarity between generating captions and ground truth captions. And [39] trains caption model with discriminative loss and reinforcement learning to reduce exposure bias. Reference [24] directly optimize a linear combination of SPICE and CIDEr(namely SPIDEr reward) by using a policy gradient method. As [4] explores, Actor-critic [25] can also be unstable, for Actor-critic [25] needs to train an additional critic network to provide an estimate of the value of each generated word given the policy of an actor network. Reference [24] indicates policy gradient method, which shows high variance and instability, requires careful tuning the gradient of the expected reward. Series of recent experiment [22], [23], [29] use a learned ''baselines'' to resolve network flurry in applying RL. Reference [23], [29] use parametric functions to estimate the expected baseline reward. Instead of estimating a ''baseline'' to reduce variance, [22] utilizes the ''greedy'' output of its own test-time inference algorithm as baseline(b) to normalize the rewards it experiences.

III. CAPTION MODEL TRAINED WITH REINFORCEMENT LEARNING

Given an image I, the goal of image captioning is to generate a caption $S = w_1, w_2, ..., w_T$, where w_i denotes the i^{th} word, and we denote the ground truth captions by $G = w_1*, w_2*, ..., w_T*$.

In most captioning models, the encoder extracts features from the image, and the decoder receives image features from the encoder and generates an output sentence using a *LSTM*. Then the *LSTM* outputs a distribution over the next word w_t with softmax function:

$$h_t = LSTM(x_t, h_{t-1}) \tag{1}$$

$$p_\theta(w_t|h_t) = softmax(W_o h_t) \tag{2}$$

In the above formula, h_t maintains the information of the image and previous word, and θ denotes the parameters of the model. Traditionally, the parameters are learned by maximizing the likelihood of the observed sequence. At training time, given a target ground truth sequence $(g_1, \ldots, g_{t-1})$, the objective is to minimize the cross entropy loss (XE):

$$L(\theta) = -\sum_{t=1}^{T} \nabla p_\theta(w_t) log(g_t) \tag{3}$$

We regard this process as a RL problem [19]. In this process, the agent is the LSTM language model, which can interact with an environment of the features of words and images. And the ''action'' is the prediction of the next word. At the end of each action, the agent updates its internal 'state', which includes cells and hidden states of the LSTM, attention weights, etc. After generating the end-of-sequence (EOS) token, the agent gets ''reward'' that is the score of the generated sentence. We denote this reward by r. The score is computed by an evaluation metric (CIDEr) by comparing the generated sequence to corresponding ground-truth sequences.

$$r = CIDEr(S, G) \tag{4}$$

Mathematically, the goal of reinforcement training is to minimize the negative expected reward of model samples:

$$\Delta L(\theta) = \nabla \sum_{w_s} r p_\theta(S) \tag{5}$$

However, this direct approach runs into a problem that the reward function is non-differentiable because $r(w_s)$ is not a continuous function with respect to θ. In practice, estimating

gradients based on formula 5 are unstable and it is necessary to perform variance reduction by using a baseline estimator b [28]. In SCST, the b is CIDEr score of caption generated by model which is pre-trained by XE. Alternatively, target can differentiate from the following formula:

$$\Delta L(\theta) = -E_{p_\theta}[(r-b)\nabla log p_\theta(S)] \tag{6}$$

However, traditional reward is merely based on metric like CIDEr, which restricts the performance of captions. To address this problem, we introduce two kinds of evaluators to obtain multi-grained rewards: REN-based word-level reward and SN-based sentence-level.

IV. HAF AND MULTI-GRAINED REWARD

This section details the novel contributions of this paper.

A. HIERARCHICAL ATTENTION Fusion(HAF)

The baseline attend network we adopted is a classical structure [1], which generates a normalized attention weight α_t according to LSTM hidden state h_t at each time step t. And α_t is used to attend different spatial of the image features Att as the final representation A:

$$a_{i,t} = w_a^t tanh(W_a Att + U_a h_t) \tag{7}$$

$$\alpha_t = softmax(a_t) \tag{8}$$

$$A = \sum_{i=1}^{K} \alpha_{i,t} Att_i \tag{9}$$

where W_a, U_a, w_a^t are learned parameters.

Other than utilizing single feature of the image and adopting single attention to focus on specific regions of the image, we consider a fusion architecture of the attention model for captioning, which integrates multi-level features map as input.

As shown in Figure 2, the average features of conv5 and conv4 are as input to the first fully visual attention. Following [1], two LSTM layers are used to selectively attend to spatial image features. The input vector to the first attention LSTM at each time step consists of the previous output of the language LSTM, which is concatenated with the image features of conv4 and conv5. Notice that raw features $Att_4, Att_5 \in \mathbb{R}^{n,k}$ and pooling features $\overline{Att_4}, \overline{Att_5} \in \mathbb{R}^v$, where k stands for the numbers of regions of an image and v represents feature dimension. Thus, the formula can be given as follows:

$$A_{att} = Attend([xE, \overline{Att_4}, \overline{Att_5}], h_{t-1}^2) \tag{10}$$

where h^2 is the output of the second LSTM which consists of the image information of convolution layers and content of generated sequence. The process to produce h^2 can be given by:

$$h_t^1 = LSTM_{att}(A_{att}, h_{t-1}^2) \tag{11}$$

$$A_{conv4} = Attend(Att_4, h_t^1) \tag{12}$$

$$A_{conv5} = Attend(Att_5, h_t^1) \tag{13}$$

$$h_t^2 = LSTM_{lang}([A_{conv4}, A_{conv5}], h_t^1) \tag{14}$$

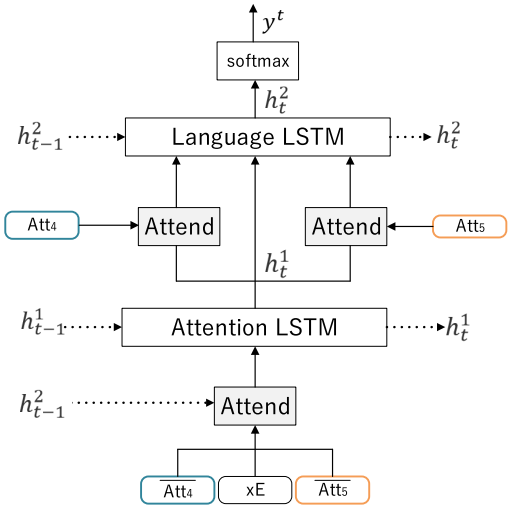


FIGURE 2. The proposed hierarchical attention fusion model. $\overline{Att_4}, \overline{Att_5}$ donate mean-pooled features of conv4 and conv5. X is one-hot encoding of input word and E is word embedding matrix of a vocabulary.

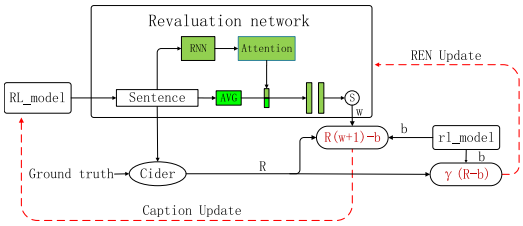


FIGURE 3. Revaluation network embeds a generating sentence and provides reward weight W. S is sigmoid and rl-model is caption model pre-trained by CIDEr.

Finally, the probability of output words is given via a non-linear softmax function:

$$p_\theta(wt|ht) = softmax(h_t^2) \tag{15}$$

B. WORD-LEVEL REWARD VIA REVALUATION NETWORK (REN)

To handle the problem that policy gradient [25] assigns the same reward to every word in a sentence, Revaluation Network (REN) is presented to revaluate metrics-based reward by estimating the importance of different word in a generating caption. And we take the metrics-based reward [22] as a base reward to guarantee stability of generating.

As illustrated conceptually in Figure 3, firstly, REN takes generated sentence S as input. Generating sentences are first processed by a RNN with an attention network and an avg-pooling layer. The embedding vector concatenated by attended sentence embedding and pooled sentence embedding is as a comprehensive representation of generated caption. Then two fully connected layer and sigmoid transformation are applied to obtain the weights W_t of different words. In particular, the caption model (rl-model in Figure 3) pre-trained by CIDEr reward for several epoch acts as baseline(b) [28] to reduce the variance considerably and lead the

optimization to expected gradient. We construct word-level reward Wr_t by formula 16. Thus, only samples from the current model that outperforms rl-model are given positive weight, and inferior samples are suppressed. Mathematically, the loss function can be formalized as formula 17:

$$Wr_t = RW_t + R - b \tag{16}$$

$$\nabla L(\theta) = -\sum_{t=1}^{T}[Wr_t \nabla log p_\theta(w_t*)] \tag{17}$$

where W_i is the output weight of REN, θ donates the parameters of image caption network, and w_t* represents different word of a generated sentence. We apply word-level reward for revising caption network after CIDEr optimizing in order to take advantage of metrics-based reward (CIDEr) and constrain the word generation space.

In addition, REN also required optimizing to provide proper weights for word reward, so we define its updating as another RL process. In this process, REN is encouraged to output the weight that is favorable to the CIDEr boosting of generating sentences. Therefor, REN has the same optimizing trend with caption model which reflects in the quality of generating sentences. Thus, REN and can be updated with reward $R - b$ that only favorable weights are promoted, and the inferior are suppressed. To tackle that REN can not get real-time feedback of caption model, the optimizing of it and caption network are asynchronous: every time after word-level updating, the caption model is frozen to generate sentences for REN to assess the quality of its output weight.

In the experiment, we observe that $R - b$ is too small which leads to weak gradients for REN. Therefore, a hyper parameter γ is set to strengthen the gradient. Similar to caption model, the REN can be updated by the following loss function through reinforcement learning algorithm:

$$\nabla L(REN) = \gamma(R - b)\nabla log P_{REN}(W) \tag{18}$$

C. SENTENCE-LEVEL REWARD VIA SCORING network(SN)

To enhance metrics-based reward (CIDEr) and take advantage from both ground truth and additional unmatched ground truth, we propose a sentence-level reward. For this purpose, a scoring network (SN) shown in Fig 4 is proposed. The SN, composed of two LSTM networks, is pre-trained by different ground truths of an image to convergence, since each image has five different ground truths. Thus, SN can be regard as a caption evaluator to enforce the constraint (reward) that the generated captions should obtain higher scores than other unmatched ground truth in sample. Here, metrics-based reward acts as a baseline to construct sentence-level reward, which ensures generating human-like sentence. Mathematically, SN encodes ground truths and generating captions into features in the same embedding space as follows:

$$s_i = LSTM(C_i) \tag{19}$$

$$g_j = LSTM(G_j) \tag{20}$$

where C and G denote generated captions and ground truth, and s_i and g_i denote its respective embedding features. The φ

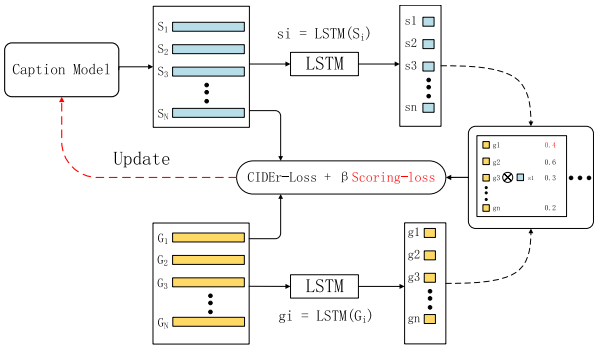


FIGURE 4. The framework of scoring network(SN). Both ground truth and unmatched ground truth are utilized to constitute sentence-level reward for RL training.

is joint representation vector between s and g:

$$\varphi_{s_i,g_i} = s_i \cdot g_i \tag{21}$$

To assign the generating captions' score higher than the score of any unmatched ground truths, we utilize hinge loss to train SN:

$$L_{sn} = max(\alpha + \varphi_{s_i,g_i'} - \varphi_{s_i,g_i}, 0) \tag{22}$$

where φ_{s_i,g_i} is matching pair, while $\varphi_{s_i,g_i'}$ is unmatching. The scores produced by SN with CIDEr acts as sentence-level reward in RL training, which encourages a generated caption from captioning model to be the best matching to the given ground truth.

$$\nabla L(\theta) = -E_{p_\theta}[(CIDEr - \beta L_{sn} - b)\nabla log p_\theta(S)] \tag{23}$$

Formula 23 is the loss function for optimizing caption model by sentence-level reward, where β is a hyper parameter to balance hinge loss and CIDEr. It is notable that the scoring process is in each mini-batch, for score in the whole dataset is time-consuming.

V. DATASET AND IMPLEMENTATION DETAILS

In this section, we introduce implementation details of the proposed model.

A. DATASET

We evaluate our proposed method on the MSCOCO 2014 caption dataset. For validation of model and offline testing, we use the ''Karpathy'' splits [14] that have been used extensively for reporting results in most previous works. The split contains 113, 287 training images with five captions each, and 5, 000 images for validation and testing respectively. We filter the vocabulary and drop any word that has counted less than 5, resulting in a vocabulary of 9, 680 words.

Flickr is also wildly used for image caption in the early works. However, it is not applied by recent related research [22], [22], [34] and the scale of flickr is smaller than COCO dataset. So this dataset is not utilized in experiment of this paper.

B. IMPLEMENTATION DETAILS

Generally, caption models are pre-trained by cross-entropy loss and then trained to maximize different RL reward. The encoder uses a pre-trained Resnet-101 [13] to get the representation of images. For each image, we extract the output of conv4 and conv5 convolutional layer from Resnet, which are mapping to a vector of dimension 1024 as input of HAF. For HAF, the dimension of image feature embedding, LSTM hidden state, and word embedding are all set to 512. The baseline model is trained under XE objective using ADAM [15] optimizer with an initial learning rate of 10^{-4}. And at each epoch, we evaluate the model and select the best CIDEr as baseline score. The reinforcement training starts from the 30 epoch to optimize the CIDEr metric with learning rate of 10^{-5}.

During the phase of word-level reward training, image caption model is pre-trained with CIDEr reward for 20 epoch, and the reward-level training for 10 epoch. In sentence-level reward training, the SN is pre-trained by different ground truth of each image for 20 epoch. As for parameter settings, the word embedding and the LSTM hidden size are set to 512. And hyper parameter margin α is set to 0.2. Besides, the caption model baseline [28] is trained for 30 epoch utilizing cross-entropy.

VI. EXPERIMENT AND ANALYSIS

A. THE PERFORMANCE OF HAF

In this subsection, the experimental results of HAF and the other compared attention methods are shown in Table 1 and Table 2. For fair comparison, the methods listed in Table 1 are trained with XE loss and Table 2 are trained with RL. It can be seen from these two tables that HAF and TopdownBU outperform other methods in all metrics. The main reason is that HAF and [1] adopt two independent LSTM networks. The first LSTM is used to process visual information, and the second LSTM is used to generate image captions. And the result of HAF is also better than Topdown-attention [1], which reveals low-level visual feature is beneficial for understanding the image. Two experimental results prove the superiority of HAF in both XE and RL training.

TABLE 1. HAF performance compared with other attention methods under caption evaluation metrics of XE training.

Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Hard-attention [12]	-	78.1	50.	35.7	25.0	23.0
Soft-attention [12]	-	70.7	49.2	34.4	24.3	23.9
ATT-FCN [31]	70.9	53.7	40.2	30.4	24.3	-
Topdown [1]	105.4	74.5	-	-	33.4	26.1
HAF(ours)	**109.0**	**75.9**	**59.5**	**45.4**	**34.4**	**26.8**

TABLE 2. Performance of HAF trained with RL training. The model is trained by using self-critical sequence training (SCST), thereby optimizing the CIDEr metric.

Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Att2in+SCST [22]	111.4	-	-	-	31.3	26.0
Att2all+SCST [22]	114.0	-	-	-	30.0	25.4
Topdown+SCST [1]	111.1	76.6	-	-	34.0	26.5
HAF+SCST(ours)	**115.4**	**79.2**	**62.5**	**47.4**	**35.3**	**27.3**

TABLE 3. Performance of REN reward trained with different parameter β. HAF+REN(1)* is directly trained by word-level reward with $\beta = 1$. The others is under pre-training by CIDEr and then optimized by word-level reward.

Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Topdown+SCST [1]	111.1	76.6	-	-	34.0	26.5
HAF+SCST	115.4	79.2	62.5	47.4	35.3	27.3
HAF+REN(1)*	113.0	79.0	61.8	47.0	35.0	27.2
HAF+REN(1)	115.2	79.1	62.3	47.1	35.2	27.2
HAF+REN(5)	115.7	79.9	62.6	47.7	35.4	27.2
HAF+REN(10)	115.9	80.2	**63.0**	47.6	35.5	27.2
HAF+REN(15)	**116.2**	**80.3**	**62.9**	**47.8**	**35.5**	27.2
HAF+REN(20)	115.8	80.2	62.8	47.7	35.4	27.2

B. THE EFFECTIVENESS OF WORD-LEVEL REWARD VIA REN

Revaluation network (REN) is exploited to construct word-level reward. In table 3, experiments are conducted by RL training with word-level reward. Generally, word-level reward achieves better performance than CIDEr reward. This demonstrates the effectiveness of word-level reward and it can address the different importance of each generated word. Hyper parameter γ is employed to enhance the gradient for REN learning. We report the performances with varied γ. Specifically, HAF+REN achieves best performance when $\gamma = 15$. A proper γ can provide gradient signal that is beneficial for REN's updating.

C. THE EFFECTIVENESS OF SENTENCE-LEVEL REWARD VIA SN

SN is utilized to obtain scoring reward and sentence-level reward is constructed by both scoring reward and CIDEr reward. In order to demonstrate the effectiveness of the scoring reward, experiments are performed with different parameter β to balance scoring reward and CIDEr reward in Table 4. We investigate $\{0.3, 0.5, 0.8\}$ as the weight of scoring reward, and the results indicate that $\beta = 0.3$ leads to the best performance. But too much emphasis on scoring reward will harm the model performance, because it weakens the role of CIDEr reward. This shows that both CIDEr and our proposed scoring reward are effective. However, the relationship between CIDEr reward and scoring reward needs to be properly balanced. From Table 4 we may conclude

FIGURE 5. Captions from different models describing the top target images. Ours refers to HAF+SN{0.3}+REN{15} and GT is ground truth.

TABLE 4. Performance of SN reward trained with different parameter γ.

Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Topdown+SCST [1]	111.1	76.6	-	-	34.0	26.5
HAF+SCST	115.4	79.2	62.5	47.4	35.3	27.3
HAF+SN(0.3)	**116.0**	**80.0**	**62.6**	**47.6**	**35.4**	27.3
HAF+SN(0.5)	115.6	79.9	62.5	47.5	35.3	27.3
HAF+SN(0.8)	115.6	79.7	62.5	47.4	35.2	27.3

TABLE 5. Performance comparison of the proposed method with other methods on MS COCO dataset.

Method	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Att2all+SCST [22]	99.4	-	-	-	30.0	25.4
Att2in+SCST [22]	101.3	-	-	-	31.3	26.0
TD-Multinomial-FC [34]	109.8	75.9	59.5	44.6	33.1	26.0
Topdown+SCST [1]	111.1	76.6	-	-	34.0	26.5
TD-Multinomial-ATT [34]	111.6	76.5	60.3	45.6	34.0	26.3
HAF+SCST	115.4	79.2	62.5	47.4	35.3	27.3
HAF+SN{0.3}+REN{15}	**116.4**	**80.5**	**62.9**	**47.7**	**35.5**	27.3

that HAF+SN outperforms SCST+HAF. It illustrates that the image caption model can benefit from the unmatched ground truth with scoring reward.

D. HAF WITH MULTI-GRAINED REWARDS

Table 5 shows the result of HAF (HAF+SCST) with Multi-grain rewards (HAF+SN+REN). The image caption model is firstly trained with sentence-level reward and then fine-tuned with word-level reward based on RL. Compare with metric-based reward [1], [22], our model improves the quality of captions significantly. Specially, TD-Multinomial [34], which is trained by the monte-carlo reward, preforms better than metric-based reward. That is because monte-carlo methed assigns each word with different reward. And the compared result also shows our two models outperform the monte-carlo reward, which is for two reasons: (1) HAF improve the metric of baseline model significantly. (2) Both word-level and sentence-level reward we proposed can boost the baseline model to generate high quality captions.

In the end, we show a sample of test set images in Fig 5. The captions of ours (HAF with Multi-grained rewards) are generated by HAF+SN{0.3}+REN{15} and each image follows with a human caption GT. We highlight the elements of HAF captions in green and ours in red, which (subjectively) seems to aid fluency and diversity. We can see that HAF generates better descriptions in many aspects than Topdown [1], such as quantity, colors, etc. Surprisingly, HAF also generates captions that is not in ground truth, such as ''red stop'' in the third picture. This reveals that HAF takes full use of hierarchical visual information and has a better understanding of the image. Furthermore, image caption model trained with multi-grained rewards produces fluent and diverse captions. For example, it generates phrase ''filled with'' in the second image compared with HAF and Topdown model. Different from generating the similar sentence pattern of ground truth, HAF+SN{−0.3}+REN{15} provides a diverse description: ''a red stop sign on the side of a road with woods in the background'' in the third image.

VII. CONCLUSION

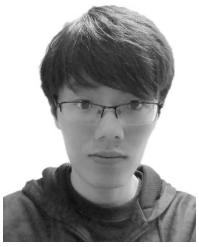
In this paper, we present Hierarchical Attention Fusion(HAF) framework for image caption. HAF, which integrate multi-level visual features with hierarchical attention, boost

the caption performance significantly. This reveals low-level visual feature is beneficial for understanding the image. Furthermore, two different network are introduced as reward evaluator for reinforcement training. Via REN and SN, we obtain two different rewards: word-level reward and sentence-level reward. Both rewards lead to improvements across most metrics.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. CVPR*, Jun. 2018, pp. 6077–6086.
[2] S. Banerjee and A. Lavie, "Meteor: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl.*, 2005, pp. 65–72.
[3] C. Wu, Y. Wei, X. Chu, S. Weichen, F. Su, and L. Wang, "Hierarchical attention-based multimodal fusion for video captioning," *Neurocomputing*, vol. 315, pp. 362–370, Nov. 2018.
[4] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, *An Actor-Critic Algorithm for Sequence Prediction*. Cambridge, MA, USA: MIT Press, 2016.
[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, "From captions to visual concepts and back," in *Proc. CVPR*, Jun. 2015, pp. 1473–1482.
[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. ECCV*, 2010, pp. 15–29.
[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. CVPR*, Jun. 2015, pp. 2625–2634.
[9] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. CVPR*, Jun. 2017, pp. 375–383.
[10] Y. J. J. Yang Wang Mao, W. Xu, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (M-RNN)," in *Proc. ICLR*, 2015.
[11] K. Papineni, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2002, pp. 311–318.
[12] X. Kelvin, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. ICML*, Jun. 2015, pp. 2048–2057.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[14] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. CVPR*, Jun. 2015, pp. 3128–3137.
[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[16] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoderdecoder for statistical machine translation," in *Proc. EMNLP*, 2014, pp. 1–15.
[17] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proc. ECCV*, 2016, pp. 3–19.
[18] L. Yu, W. Zhang, J. Wang, and Y. Yu, "SeqGAN: Sequence generative adversarial nets with policy gradient," in *Proc. CVPR*, Jun. 2017, pp. 2852–2858.
[19] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proc. ICLR*, 2016.
[20] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, Jun. 2015, pp. 3156–3164.
[21] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. NIPS*, 1999, pp. 1057–1063.
[22] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. CVPR*, Jul. 2017, pp. 7008–7024.
[23] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *Proc. ECCV*, 2017, pp. 873–881.
[24] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of SPIDEr," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 873–881.
[25] A. Barto and R. S. Sutton, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[26] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. CVPR* , Oct. 2017.
[27] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. CVPR*, Jun. 2015, pp. 4566–4575.
[28] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, 1992.
[29] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proc. ICCV*, Dec. 2015, pp. 2407–2415.
[30] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out 2004*, pp. 74–81.
[31] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. CVPR*, Jun. 2016, pp. 4651–4659.
[32] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Proc. NIPS*, 2016, pp. 2361–2369.
[33] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *Proc. CVPR*, Jul. 2017, pp. 290–298.
[34] H. Chen, G. Ding, and S. Zhao, "Temporal-difference learning with sampling baseline for image captioning," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6706–6713.
[35] M. Chen, G. Ding, S. Zhao, H. Chen, Q. Liu, and J. Han, "Reference based LSTM for image captioning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 3981–3987.
[36] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, observe and tell: Attribute-driven attention model for image captioning," in *Proc. IJCAI*, Jul. 2018, pp. 606–612.
[37] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.
[38] L. Gao, X. Li, J. Song, and H. T. Shen, "Hierarchical LSTMs with adaptive attention for visual captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[39] L. Gao, K. Fan, J. Song, X. Liu, X. Xu, and H. T. Shen, "Deliberate attention networks for image captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 8320–8327.
[40] Y. Cheng, F. Huang, L. Zhou, C. Jin, Y. Zhang, and T. Zhang, "A hierarchical multimodal attention-based neural network for image captioning," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2019, pp. 889–892.

CHUNLEI WU received the Ph.D. degree in computer application technology from the Ocean University of China, in 2014. He is currently an Associate Professor with the College of Computer and Communication, China University of Petroleum (East China). He has authored or coauthored more than 30 journal and conference papers and textbooks. His current interests include image and video processing, and machine learning.



SHAOZU YUAN is currently pursuing the master's degree with the College of Computer and Communication, China University of Petroleum (East China). His current research interests include image caption, video question and answering, and reinforcement learning.

HAIWEN CAO is currently pursuing the master's degree with the College of Computer and Communication Engineering, China University of Petroleum (East China). Her current research interests include video-based action recognition and object detection.

YIWEI WEI received the master's degree in software engineering from the China University of Petroleum (East China). He is currently a Lecturer with the College of Petroleum, China University of Petroleum-Beijing at Karamay. His current research interests include cross modal retrieval, neural machine translation, and image/video caption.

LEIQUAN WANG received the Ph.D. degree in communication and electrical systems from BUPT. He is currently a Lecturer with the College of Computer and Communication Engineering, China University of Petroleum (East China). His current research interests include multimodal fusion, cross modal retrieval, and image/video caption.