IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 10, 2021, accepted March 13, 2021, date of publication March 22, 2021, date of current version May 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3067607

Multi-Gate Attention Network for Image Captioning

WEITAO JIANG[1], XIYING LI[2], (Member, IEEE), HAIFENG HU[1], (Member, IEEE), QIANG LU[2], AND BOHONG LIU[1]

[1]School of Electronic and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[2]School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China

Corresponding author: Xiying Li (stslxy@mail.sysu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1601101 and Grant 2018YFB1601100, and in part by the National Natural Science Foundation of China under Grant U1611461.

ABSTRACT Self-attention mechanism, which has been successfully applied to current encoder-decoder framework of image captioning, is used to enhance the feature representation in the image encoder and capture the most relevant information for the language decoder. However, most existing methods will assign attention weights to all candidate vectors, which implicitly hypothesizes that all vectors are relevant. Moreover, current self-attention mechanisms ignore the intra-object attention distribution, and only consider the inter-object relationships. In this paper, we propose a Multi-Gate Attention (MGA) block, which expands the traditional self-attention by equipping with additional Attention Weight Gate (AWG) module and Self-Gated (SG) module. The former constrains the attention weights to be assigned to the most contributive objects. The latter is adopted to consider the intra-object attention distribution and eliminate the irrelevant information in object feature vector. Furthermore, most current image captioning methods apply the original transformer designed for natural language processing task, to refine image features directly. Therefore, we propose a pre-layernorm transformer to simplify the transformer architecture and make it more efficient for image feature enhancement. By integrating MGA block with pre-layernorm transformer architecture into the image encoder and AWG module into the language decoder, we present a novel Multi-Gate Attention Network (MGAN). The experiments on MS COCO dataset indicate that the MGAN outperforms most of the state-of-the-art, and further experiments on other methods combined with MGA blocks demonstrate the generalizability of our proposal.

INDEX TERMS Image captioning, self-attention, transformer, multi-gate attention.

I. INTRODUCTION

Image captioning is a challenging task of automatically generating a fluent and reasonable sentence to describe visual contents of an image. As an interdisciplinary field involving computer vision and natural language processing, it has attracted numerous attentions in the past several years and has great potential in human-machine interaction, supporting the visually impaired and intelligent assistant.

Most existing image captioning methods that apply the widely used attention-based encoder-decoder framework [1]–[4] have achieved excellent results. Specifically, for a given image, a set of feature vectors are encoded by a Convolutional Neural Network (CNN) first, and a caption is then decoded via a Recurrent Neural Network (RNN) with these vectors. In between, the attention mechanism plays a pivotal role by guiding the decoding process with dynamic attended vector. Since the attention mechanism is introduced by [1], it has become a basic module of practically all image captioning models. Later on, other variants of attention mechanism are proposed, such as semantic attention [5], stacked-attention [6], adaptive attention [2] and channel attention [7]. Recently, [8] shows the excellent performance of self-attention, and the intention of stacking self-attention layer has been expanded by some works [4], [9]–[11].

Despite the accomplishment that the aforementioned attention-based approaches have achieved, attention mechanism still remains some unsolved problems. Firstly, the attention weights will be assigned to all candidate vectors (*e.g.* object region features) when calculating the attention vector. Hence, it may extract some unrelated or even misleading information. This is particularly evident in

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

self-attention-based encoder for image captioning, which calculates pairwise similarity between all input feature vectors and implicitly assumes that all vectors are relevant. We argue that, for an object in image, it will not have relationship with all other objects. Secondly, most attention mechanisms concentrate on the object-wise attention distribution while neglecting the significance of feature-wise information. Specifically, as for a set of image feature vectors (object or patch feature), most attention mechanisms simply treat all information equally in one feature vector and multiply it by the same attention weight, which ignores the noise that may exist in the feature vector itself. Furthermore, the self-attention-based transformer architecture [8] is designed for natural language processing task (*e.g.* machine translation), where the inputs are mainly text features, to capture the long-term dependencies among all input text features. Whereas in image captioning, most recent works [4], [10] adopt the transformer as feature enhancer in the encoder, where the inputs are usually image features.

To tackle the problems mentioned above, in this paper, we propose a Multi-Gate Attention (MGA) block with pre-layernorm transformer architecture for image captioning. It extends the vanilla self-attention by modifying the architecture and adding multiple gate mechanisms. Compared with the original transformer architecture (Figure 1(a)), the presented pre-layernorm transformer (Figure 1(b)) places the layer normalization before self-attention module and further removes the feed-forward layer and subsequent layers to simplify the model, making it more efficient for image caption task. Then, based on the self-attention, an Attention Weight Gate (AWG) module and a Self-Gated (SG) module are incorporated to constrain the attention mechanism to concentrate on the most relevant information, and consider the intra-object attention distribution.

Technically, before conducting self-attention, layer normalization on the input vectors is implemented first. Then utilize SG module to eliminate noise and other irrelevant information in the normalized feature vectors. After SG module, self-attention module is applied to model relationships among all input feature vectors, where the similarity scores between query and key vectors are calculated first, and then passes these scores through a softmax layer to generate attention weights. However, the softmax layer in the self-attention will assign a corresponding weight to every value vector, even if the query and key vector have no relevancy. Therefore, we present the AWG module to relieve this by restraining the outputs of softmax layer. More specifically, by passing the attention weights and value vectors through AWG module, the output attended vectors maintain only the most contributive elements while eliminating the trivial components. Afterwards we utilize GLU [12] to cope with the concatenation of outputs of AWG module and SG module for the purpose of further filtering irrelevant information. Finally, a residual connection around GLU and layer normalization is employed to generate the final outputs.

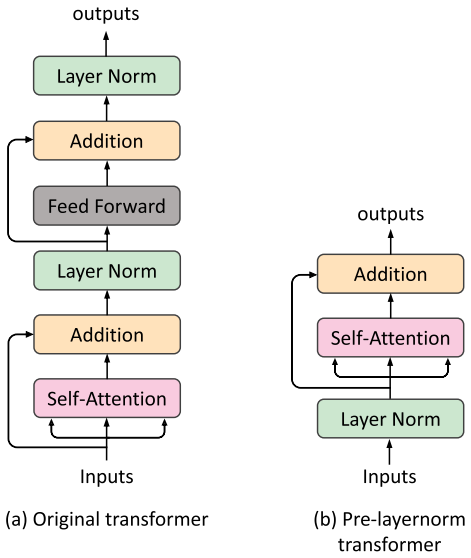

FIGURE 1. The structure of original transformer and pre-layernorm transformer.

By integrating MGA block with pre-layernorm transformer architecture into the image encoder and likewise AWG module into the language decoder, a novel Multi-Gate Attention Network (MGAN) is proposed to refine the feature representation of the image encoder and provide more accurate information for the language decoder. Main contributions of this paper are as follows:

- We propose a pre-layernorm transformer for image feature enhancement in the image encoder. By only stacking layer normalization and self-attention layer, it can model relationships among image features efficiently.
- We propose a Multi-Gate Attention (MGA) block which contains an Attention Weight Gate (AWG) module and a Self-Gated (SG) module to constrain the attention mechanism and consider the intra-object attention distribution respectively.
- By applying MGA block with pre-layernorm transformer architecture to the image encoder and AWG module to the language decoder, a Multi-Gate Attention Network (MGAN) is proposed. Extensive experiments on MS COCO dataset indicate the effectiveness and generalizability of our proposal.

II. RELATED WORKS

A. IMAGE CAPTIONING

Image captioning is an active and challenging research area. Before the advent of neural network-based methods [9], [10], [13], [14], earlier attempts to image captioning are mainly template-based [15]–[17] and retrieve-based [18]–[20]. The former produces templated image caption with slots which are filled by the outputs of object detection or attribute prediction. The latter retrieves the most similar captions, and generates descriptions by recomposing and generalizing the retrieved results. The generated sentences of these early methods are unsatisfying in terms of accuracy and diversity, highly limited by the templates or the retrieved results.

At present, neural network-based methods [3], [4], [9], [11] dominate the image captioning community through superior achievement. Inspired by the advances in machine translation, [13] first adopts the neural encoder-decoder framework to solve the image captioning problem and achieves tremendous improvement than previous methods. Different from [13], which only inputs image features at initialization, Yao *et al*. [21] incorporate attributes information into the decoder in different ways. Furthermore, retrieved captions [22], reinforcement learning [23] are introduced to boost image captioning.

B. ATTENTION MECHANISM

Attention mechanisms are widely used in current image captioning models, where a weighted summation on candidate vectors is generated at each time step for word reasoning. Based on the neural network, [1] first integrates the spatial attention mechanism into image captioning task. On the basis of spatial attention, numerous variants of attention mechanism have been proposed by researchers. Reference [7] proposes a spatial and channel-wise attention, which applies the channel-wise attention to assess the importance of each channel before spatial attention. Instead of forcing attention to be active constantly, [2] presents an adaptive attention mechanism which automatically decides whether to utilize visual features or not. In addition, [3] proposes a combination of bottom-up and top-down spatial attention which enables attention to be conducted at object level rather than equally-sized image regions. Different from spatial attention, [5] presents a semantic attention where visual features in the spatial attention mechanism are replaced by semantic concept proposals. Reference [14] integrates both spatial and semantic attention into the language decoder, as well as exploits Graph Convolutional Network (GCN) to model spatial and semantic object relationships in the image encoder. Moreover, lots of complex attention mechanisms are proposed. Reference [24] presents a multimodal attention network to manage information from different modalities. Reference [25] proposes a multistage attention mechanism which operates in a coarse-to-fine manner to ensure global consistency and local accuracy. And in [26], a multiscale self-attention is introduced to capture features from different scales.

C. TRANSFORMER BASED METHODS

Recently, in the field of machine translation, [8] proposes the transformer architecture and demonstrates that remarkable results can be attained by merely using self-attention mechanism. Following this, several works extend the self-attention mechanism to image captioning task. Reference [4] stacks multiple layers of self-attention to refine visual object features via modeling relationships between objects in the encoder. Then it applies an attention gate to filter out irrelevant attention results. Later on, [10] introduces X-Linear attention block which utilizes bilinear pooling and exponential linear unit to model high-order feature interactions across multimodal inputs. Different from the aforementioned methods, which apply a transformer-like encoder and an LSTM decoder [11] takes advantage of full attentive model that introduces the benefits of normalization inside self-attention and proposes geometry-aware self-attention to calculate the pairwise geometry relations of objects. Reference [9] presents an EnTangled Attention to utilize visual and semantic information simultaneously in a tangled way.

Inspired by the advancement of transformer [8] and dimensionality reduction [27], [28], in this paper, we adopt the transformer-like encoder to refine the feature representations and an LSTM decoder to inference the appropriate word.

III. METHOD

In this section, the framework of our method will be presented first. Then, we introduce the pre-layernorm transformer architecture. Subsequently, the proposed Multi-Gate Attention (MGA) block as well as its components, Self-Gated (SG) module and Attention Weight Gate (AWG) module, will be elaborated. Finally, we describe how to construct the Multi-Gate Attention Network (MGAN) for image captioning.

A. FRAMEWORK

Most existing image captioning methods adopt the encoder-decoder paradigm. Recently, with the introduction of transformer [8], some works [4], [10] extend this paradigm by adding an additional transformer-like encoder. In this paper, we strive for improving the effectiveness of using transformer-like encoder to enhance the feature representation.

Given an image I, the encoder first extracts a set of image feature vectors (patch or object region feature) $V_r = \{v_i\}_{i=1}^{k}$, where $v_i \in \mathbb{R}^d$, k is the number of image regions, and d represents the dimension of each vector. Then, the feature V_r will be enhanced by exploring relationships among all feature vectors. In this process, the MGA blocks are applied in a stacked manner. Finally, a sentence $w = \{w_1, w_2, \ldots, w_T\}$ to describe the image is generated by the decoder.

$$V_r = CNN(I) \tag{1}$$
$$V_r^N = MGAs(V_r) \tag{2}$$
$$S = Decoder(V_r^N, w) \tag{3}$$

where $CNN(\cdot)$ is the image encoder and $MGAs(\cdot)$ is a stack of MGA blocks with pre-layernorm transformer architecture (N times). We omit some details of the decoder for simplification.

Following previous works, our model is trained by the cross-entropy (XE) loss first:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log p_\theta(w_t^* \mid w_{1:t-1}^*) \tag{4}$$

where θ represents the parameters of our model, and $w_{1:T}^*$ is the ground-truth caption.

Then the model is further optimized with CIDEr reward using the Self-Critical Sequence Training [23]:

$$L_{RL}(\theta) = -E_{w\sim p_\theta}[CIDEr(w)] \tag{5}$$

where $CIDEr(w)$ is the CIDEr reward for the random sampled sentence. The gradients can be approximated as follows:

$$\nabla_\theta L_{RL}(\theta) \approx -(CIDEr(w^s) - CIDEr(\hat{w}))\triangle\theta \log p_\theta(w^s) \tag{6}$$

where $w^s = (w_1^s, \ldots, w_T^s)$, w_t^s is the word sampled at time step t. The $CIDEr(\hat{w})$ represents the reward obtained by greedy sampling.

B. MULTI-GATE ATTENTION BLOCK

Based upon the self-attention mechanism, we propose a novel attention module, namely MGA block, which integrates the pre-layernorm transformer architecture, AWG module and SG module.

1) PRE-LAYERNORM TRANSFORMER

Transformer-like architectures are widely used in sequence modeling tasks such as language modeling [29] and machine translation [30]. It also represents the state-of-the-art in image captioning [10], [11]. Nevertheless, the original transformer architecture is designed for capturing the long-term dependencies among all input text features for natural language processing task. Inspired by the recent progress on modified transformer [31] and image captioning [4], we present a pre-layernorm transformer to explore a more effective way to utilize transformer for image captioning.

We first review the original transformer architecture, as shown in Figure 1(a), which mainly consists of two sub-layers: a self-attention layer and a feed-forward network layer. Besides, residual connection and layer normalization are applied for both sub-layers respectively.

As for the pre-layernorm transformer, as illustrated in Figure 1(b), which removes the feed-forward network sub-layer and puts the layer normalization layer before self-attention. Compared with the original transformer, our pre-layernorm transformer is more simplified and requires fewer parameters. More importantly, it gets rid of the warm-up stage essential for transformer-based image captioning methods [9]–[11] and won't degrade the performance.

2) ATTENTION WEIGHT GATE MODULE

In conventional self-attention mechanism, which calculates the similarity scores between queries Q and keys K first, and then passes the scores through a softmax layer to generate the attention weights W. Finally, it takes a weighted sum of values V on the basis of attention weights. This process can be defined as:

$$Attention(Q, K, V) = WV = softmax(\frac{QK^T}{\sqrt{d}})V \tag{7}$$

where Q, K, V are three different linear projections of image feature vectors V_r, and d is the dimension of Q. Similar to transformer [8], the self-attention in this paper is conducted in a multi-head fashion (with 8 heads).

Nevertheless, the softmax layer will assign a corresponding weight to each value vector, even if the query and key vector have no relevancy. To solve this problem, the AWG module is proposed, which directly constrains the output of softmax layer (attention weights) since the exponential operation in softmax will enlarge the numerical difference between elements. As shown in Figure 2(b), based upon the attention weights W, AWG first masks the weight value W_{ij} below the threshold value by multiplying the mask M. Then, each row of the output matrix is normalized to 1. The details of AWG are as follows:

$$M_{ij} = \begin{cases} 1 & \text{if } W_{ij} \geq g_i \\ 0 & \text{if } W_{ij} < g_i \end{cases} \tag{8}$$

$$W_{ij}^m = Norm(W_{ij}M_{ij}) = \frac{W_{ij}M_{ij}}{\sum_{j=1}^{k} W_{ij}M_{ij}} \tag{9}$$

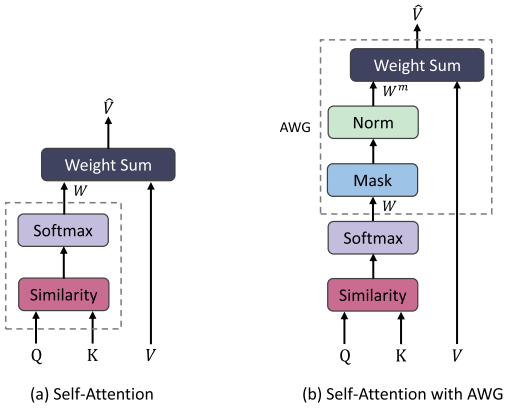

FIGURE 2. Comparison between self-attention mechanism and self-attention with AWG. The AWG module directly constrains the output W of softmax by masking the values below threshold value.

where i, j represent the row and column of matrix respectively, and $G = (g_1, \ldots, g_k)$, g_i is the l-th largest value or mean of the i-th row of W. Eq. (9) denotes the normalization operation. The final result is calculated as follows:

$$\hat{V} = V \odot W^m \tag{10}$$

where $\odot$ denotes element-wise multiplication, and W^m is the masked attention weights. The throughout process of AWG can be defined as:

$$AWG(W, V) = V \odot Norm(WM) \tag{11}$$

Note that AWG module can be seamlessly integrated into conventional self-attention mechanism without increasing any parameters.

3) SELF-GATED MODULE

Most recent image captioning methods apply the pre-trained Faster R-CNN [32] to extract image region features since it is introduced by [3]. However, the detected object regions in

image are usually located by bounding box, and thus these regions may contain some irrelevant information (*e.g.* background or parts of other objects). In addition, the extracted feature vectors are directly used by language decoder or attention mechanism. Therefore, we propose a SG module to consider the intra-object attention distribution for eliminating the irrelevant information in object feature vector.

Specifically, given an image feature vector set V_r, the intra-object attention distribution is produced by projecting V_r via an embedding layer, followed by a sigmoid activation function. Then, SG module generates the gated feature vectors by accumulating another linear projection of V_r with intra-object attention weight:

$$SG(V_r) = \sigma(W_{v1}V_r) \odot W_{v2}V_r \tag{12}$$

where W_{v1} and W_{v2} are learnable matrices. SG module does not change the dimension of feature vectors V, and thus can be integrated into self-attention mechanism.

In order to incorporate SG module into pre-layernorm transformer, we design three variants: Pre-SG, Post-SG and Parallel-SG, as illustrated in Figure 3. If not mentioned particularly, in this paper, SG refers to Pre-SG.

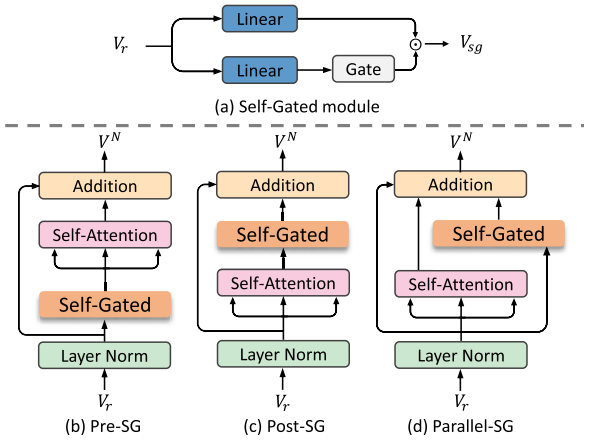

FIGURE 3. The structure of SG module and three variants of combining SG with pre-layernorm transformer.

C. MGAN FOR IMAGE CAPTIONING

Based upon the encoder-decoder structure, we present the model, MGAN, which integrates MGA blocks into the image encoder for feature enhancement and AWG module into the language decoder for feature selection.

1) ENCODER WITH MGA

There are two image encoders in MGAN. One is the pre-trained Faster R-CNN which extracts the image region features V_r, and the other is MGAs which refines the features V_r by modeling relationships among all the regions. MGAs is composed of a stack of N MGA blocks ($N = 6$), as shown on the left of Figure 4.

For a set of image feature vectors $V_r = \{v_1, \ldots, v_k\}$ provided by the CNN encoder, MGAs is utilized to strengthen the representation of each feature vector v_i via exploring relationships between all feature vectors. Take the first block of MGA (MGA_1) as an example, layer normalization on the input vectors is performed first. Then a SG module is applied to filter out some unrelated information in the feature vectors. The output of SG module is fed to the subsequent self-attention and AWG module to explore the interactions between all input vectors. After that, the outputs of AWG and SG module are concatenated and then input to a GLU for further eliminating the irrelevant attention results. Finally, same to [8], residual connection is used to generate the final output. This process can be defined as:

$$V_{ln} = LayerNorm(V_r) \tag{13}$$
$$V_{sg} = SG(V_{ln}) \tag{14}$$
$$Q, K, V = Linears(V_{sg}) \tag{15}$$
$$\hat{V} = AWG(W, V) \tag{16}$$
$$V^1 = V_{ln} + GLU([\hat{V}, V_{sg}]) \tag{17}$$

where $Linears(\cdot)$ denotes three different linear projections of V_{sg}, and [,] represents the concatenation operation. W is attention weights defined in Eq. (7).

Note that the aforementioned MGA block doesn't change the dimension of its input. Therefore, in the image encoder, N MGA blocks can be stacked in sequence to generate the enhanced feature vectors V^N.

2) DECODER WITH MGA

The language decoder is designed to generate a sentence w with the enhanced image feature vectors. Since the image features used by the decoder have been refined by the encoder, we discard the SG module in the decoder. As illustrated in Figure 4, conditioned on the word embedding vector $W_e w_{t-1}$, mean-pooled global image feature $\bar{v} = \frac{1}{k}\sum_{i=1}^{k} v_i^N$, where v_i^N is the i-th vector of V^N, and previous context vector c_{t-1}, a $LSTM$ is utilized to generate the hidden state h_t:

$$h_t = LSTM([W_e w_{t-1}, \bar{v} + c_{t-1}], h_{t-1}) \tag{18}$$

where W_e is a word embedding matrix. Afterwards, self-attention mechanism and subsequent AWG module are applied to calculate the attended image feature vector $\hat{V}_t$:

$$W = softmax((W_q h_t) \cdot (W_k V^N)^T / \sqrt{d}) \tag{19}$$
$$\hat{V}_t = AWG(W, W_v V^N) \tag{20}$$

where W_q, W_k and W_v are learnable matrices. Next, we obtain the context vector c_t from a GLU:

$$c_t = GLU([\hat{V}_t, h_t]) \tag{21}$$

Such a context vector c_t is finally utilized to predict the probability distribution of the word through a linear layer with softmax activation:

$$p(w_t \mid w_{1:t-1}) = softmax(W_p c_t) \tag{22}$$

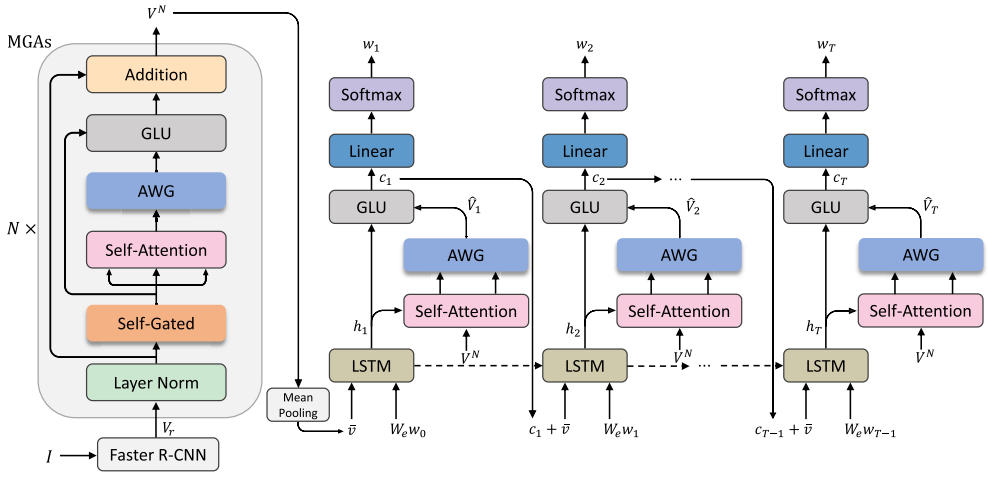

FIGURE 4. The overall framework of the proposed Multi-Gate Attention Network. The Faster R-CNN first detects a set of object region feature vectors V_r. Then, a stack of MGA blocks are utilized to enhance the feature representation by modeling relationships between them. Based on the enhanced features V^N, language decoder with AWG module is applied to generate plausible image captions.

IV. EXPERIMENT

A. DATASET

To evaluate the effectiveness of our proposal, extensive experiments are carried out on the MS COCO dataset [33]. Following most image captioning methods, we utilize the Karpathy data split [34] for performance comparisons which contains 113, 287 images for training, 5, 000 for validation and 5, 000 for testing. Each image has 5 human annotated captions. For the dataset, all sentences are converted to lower case, and the words that appear less than 5 times are dropped. We trim each caption to a maximum of 16 words, which results in a vocabulary of 10, 369 words.

B. IMPLEMENTATION DETAILS

We utilize the Faster R-CNN [32] pre-trained on ImageNet [35] and Visual Genome [36] to extract image object region features [3]. Each original object feature is a 2, 048-dimensional vector. Before being used by other modules, it is transformed into the dimension of 512. The hidden size of LSTM is also set as 512, and the dimension of word embedding layer is set as 1, 024. In the training phase, Adam [37] optimizer is used, and we first train the model under cross-entropy loss for 30 epochs with a batch size of 32. Then we further optimize it with CIDEr reward for additional 30 epochs. The learning rate for the first stage is set as $2e$-4 and decays by a factor of 0.8 every 3 epochs. In addition, the warm-up step is not required. As for the CIDEr-optimization phase, the learning rate is fixed at $2e$-5. At the inference stage, the beam search strategy is adopted and the beam size is set as 2. Five different metrics are used, including BLEU@N [38], METEOR [39], ROUGE-L [40], CIDEr [41] and SPICE [42], to evaluate the performance of our model.

C. ABLATION STUDY

The key components of the proposed MGAN include the pre-layernorm architecture, AWG module and SG module. We first design the Base model which only adopts the visual features V_r extracted by Faster R-CNN in the encoder, and discards the AWG module in the decoder as well as replaces the GLU with a linear transformation. Based on the Base model, ablation experiments are designed as below:

1) PRE-LAYERNORM TRANSFORMER

To explore the performance of stacking pre-layernorm transformer in the image encoder, we apply it to the Base model. As shown in Table 1, the models with additional transformer-like encoder surpass the Base model by a large margin, which indicates the significance of exploring relationships among objects. The training process of "+transformer" is unstable as we dropped the warm-up step. Therefore, following [9], we set the warm-up steps as 20, 000 to train this model only. Comparing to "+transformer", "+pre-layernorm" is more lightweight and does not hurt model performance.

2) AWG MODULE

The selection of threshold value may affect the performance of attention mechanism in the encoder and decoder. Reserving too many or too few attention weight values may degrade the performance of self-attention by either keeping irrelevant information or filtering out some important information. Table 2 illustrates the effect of the threshold value g_i in the encoder and the decoder respectively.

We start from the definition of the enhanced BaseLine (BL) for subsequent experiments. As it can be seen from the first row of Table 2, on the basis of Base model with pre-layernorm transformer (Pre-layernorm), adding a GLU in the pre-layernorm transformer (incomplete MGA block) can boost the performance. Then, integrating the same GLU into the decoder, a further performance improvement can be observed, and we define this model as the BL. Similarly, via adding AWG and GLU to the pre-layernorm transformer and decoder respectively, the models have competitive results,

TABLE 1. Ablation study about the Base model with different additional encoders on MS COCO Karpathy's test split under XE loss, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE scores. The Time refers to the average running time (seconds) per epoch.

Model	B@1	B@4	M	R	C	S	Time
Base	76.7	35.7	27.3	56.5	111.8	20.5	391.3
+ Enc: pre-layernorm	**77.5**	**36.6**	27.7	**57.1**	115.5	**21.1**	468.2
+ Enc: transformer	77.3	**36.6**	**27.8**	**57.1**	**115.7**	21.0	559.6

TABLE 2. Ablation study about the use of AWG module in the image encoder and language decoder on MS COCO Karpathy's test split under XE loss.

Model	B@1	B@4	M	R	C	S	Time
Pre-layernorm①	77.5	36.6	27.7	57.1	115.5	21.1	468.2
①+Enc: GLU②	77.6	36.8	27.9	57.4	116.3	21.2	491.1
②+Dec: GLU (BL)	77.5	36.9	28.1	57.3	117.1	21.4	498.9
①+Enc: AWG(10) ③	77.7	37.1	27.7	57.3	116.1	21.0	481.3
③+Dec: GLU	77.4	36.8	27.9	57.2	116.9	21.3	488.1
BL +Enc+AWG(1)	76.9	36.6	**28.2**	57.1	117.3	21.4	513.0
+ AWG(5)	**78.0**	37.1	27.8	57.5	117.6	21.1	513.2
+ AWG(10)	77.8	37.2	**28.2**	**57.6**	**118.0**	21.4	513.4
+ AWG(15)	77.8	**37.3**	28.1	57.4	117.3	21.3	512.5
+ AWG(mean)	77.6	36.9	28.1	57.3	116.9	**21.5**	504.5
BL +Dec+AWG(1)	77.0	36.5	27.9	57.1	115.5	21.1	505.0
+ AWG(5)	77.5	37.0	**28.1**	**57.4**	117.3	21.3	504.9
+ AWG(10)	**77.6**	36.9	**28.1**	57.3	117.6	**21.4**	505.0
+ AWG(15)	77.2	36.6	**28.1**	57.3	116.7	21.3	505.1
+ AWG(mean)	**77.6**	**37.1**	28.0	57.3	**118.1**	21.3	503.6

which indicates the effectiveness of AWG module and GLU. Furthermore, the model with AWG doesn't introduce any parameters, but has almost the same results obtained from the model with GLU.

Based on the BL, we add the AWG module with different threshold value to the attention module in encoder and decoder respectively. In Table 2, the AWG(l) and AWG(mean) represent the l-th largest value and mean of an attention weight vector respectively. When incorporating AWG module in the encoder, all models outperform BL or have comparable performance. Compared with BL, continuing to increase the value of l will not bring positive or negative effects to the model, since too large l value means that the AWG module hardly works. Applying AWG module to the decoder, inappropriate l value tends to degrade the performance. In summary, AWG(10) works best in the encoder while AWG(mean) is more suitable for the decoder, which is also the default setting for subsequent experiments, if not mentioned specifically. We argue that this difference may be caused by the change of the query vector (visual features V_r to contextual feature h_t).

3) SG MODULE

As shown in Figure 3, three variants of integrating SG module into image encoder are proposed. The experiments are conducted on the basis of BL, therefore, we introduce the GLU into these variants by placing GLU behind self-attention. From Table 3, it can be observed that the Pre-SG model is more effective than the Post-SG and Parallel-SG because it eliminates irrelevant information in the input vectors at the beginning. Compared with BL, the Post-SG slightly degrades the BLEU scores, which indicates that stacking gate in an inappropriate position will not bring benefit for performance even decline the performance. The Parallel-SG has no obvious performance gain than the BL. We infer this may be caused by the addition operation of the outputs of SG module, self-attention and residual connection, which will introduce redundant information.

TABLE 3. Ablation study about SG module on MS COCO Karpathy's test split under XE loss.

Model	B@1	B@4	M	R	C	S	Time
BL	77.5	36.9	28.1	57.3	117.1	21.4	468.2
Pre-SG	**77.8**	**37.1**	**28.2**	**57.5**	**117.9**	**21.5**	528.0
Post-SG	77.2	36.4	28.1	57.3	117.1	**21.5**	527.1
Parallel-SG	77.7	36.9	28.1	57.4	117.4	**21.5**	529.8

4) RUNNING TIME

In order to compare the training efficiency of different models more intuitively, we also report the average running time per epoch for all models as well, which is presented in Table 1, Table 2 and Table 3. Note that all the models are trained on a single NVIDIA GTX 1080Ti GPU.

From Table 1, we observe that the additional transformer-based encoders increase the running time significantly. These two models have comparable performance, whereas the "+transformer" has a much heavier computation load due to its feed-forward network layer. It can be seen from Table 2 that the GLU and AWG slightly raise running time, while also bringing performance gain. Compared with BL model, the added time of integrating AWG module in the encoder and decoder is distinct, which can be attributed to the difference in the size of query vector. In Table 3, we can find that the time complexity of the three SG models is similar. Moreover, the running time of the full model (MGAN) is 546.1 seconds, which is still lower than the "+transformer" model. The MGAN outperforms the original "+transformer" model by a large margin and maintains a shorter running time. In summary, the comparison results demonstrate the effectiveness of our proposal in improving model performance and reducing computational overhead.

D. PERFORMANCE COMPARISON

We compare the proposed MGAN, where MGAs is applied in the encoder and AWG(mean) is adopted in the decoder as shown in Figure 4, with the existing state-of-the-art models on the widely used Karpathy's test split. The models we compared include: NIC [13], SCST [23], LSTM-A [21], Up-Down [3], RFNet [43], GCN-LSTM [14], ETA [9], AoANet [4], Sub-GC [45], MT [44] and NG-SAN [11]. In all the above models, except NIC and LSTM-A, the other models all employ attention mechanism. Moreover, Sub-GC and MT leverage extra Graph Convolutional Network (GCN) to model the relationships among visual features, while ETA, AoANet and NG-SAN are transformer-based methods.

TABLE 4. Performance comparison with state-of-the-art methods on MS COCO Karpathy's test split under XE loss. *h* and *v* represent the hidden size of LSTM and the embedded image feature size. - indicates that the metric is not provided. * denotes the results come from our reimplementation where the changes are *h*, *v* and batch size only.

Model	h/v	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
NIC [13]	512/512	-	-	-	29.6	25.2	52.6	94.0	-
SCST [23]	512/512	-	-	-	30.0	25.9	53.4	99.4	-
LSTM-A [21]	1024/1024	75.4	-	-	35.2	26.9	55.8	108.8	20.0
Up-Down [3]	1000/512	77.2	-	-	36.2	27.0	56.4	113.5	20.3
RFNet [43]	512/-	76.4	60.4	46.6	35.8	27.4	56.8	112.5	20.5
GCN-LSTM [14]	1000/512	77.3	-	-	36.8	27.9	57.0	116.3	20.9
ETA [9]	none/512	77.3	-	-	37.1	28.2	57.1	117.9	21.4
AoANet [4]	1024/1024	77.4	-	-	37.2	**28.4**	57.5	**119.8**	21.3
AoANet*	512/512	77.3	61.6	48.0	37.1	28.2	57.4	117.1	21.4
Sub-GC [45]	1024/1024	76.8	-	-	36.2	27.7	56.6	115.3	20.7
MT [44]	1000/1000	78.1	-	-	**38.4**	28.2	**58.0**	119.0	21.1
MGAN	512/512	**78.4**	**62.8**	**48.9**	37.5	28.2	57.8	118.8	**21.6**

TABLE 5. Performance comparison with state-of-the-art methods on MS COCO Karpathy's test split under CIDEr reward optimization.

Model	h/v	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
NIC [13]	512/512	-	-	-	31.9	25.5	54.3	106.3	-
SCST [23]	512/512	-	-	-	34.2	26.7	55.7	114.0	-
LSTM-A [21]	1024/1024	78.6	-	-	35.5	27.3	56.8	118.3	20.8
Up-Down [3]	1000/512	79.8	-	-	36.3	27.7	56.9	120.1	21.4
RFNet [43]	512/-	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
GCN-LSTM [14]	1000/512	80.5	-	-	38.2	28.5	58.3	127.6	22.0
ETA [9]	none/512	80.7	-	-	39.1	28.6	58.6	125.0	22.4
AoANet [4]	1024/1024	80.2	-	-	38.9	29.2	58.8	129.8	22.4
AoANet*	512/512	79.5	63.9	49.5	37.5	28.5	58.1	128.3	22.2
MT [44]	1000/1000	**80.8**	-	-	38.9	28.8	58.7	129.6	22.3
NG-SAN [11]	none/2048	-	-	-	**39.9**	**29.3**	**59.2**	**132.1**	**23.3**
MGAN	512/512	80.6	**65.2**	**50.7**	38.5	28.9	58.6	129.6	22.7

As the transformer-based models which stack multiple layers of encoder have heavy computational overhead, we applied a lower-dimensional image feature size (512) to alleviate this. Additionally, in order to verify the impact of the feature size, we use the publicly available codes provide by [4] to reimplement the AoANet where the only changes compared to the original implementation are batch size (10 to 32), hidden size and image feature size (1, 024 to 512). Meanwhile, we report the hidden size of LSTM and embedded image feature size of all compared methods in Table 4 and Table 5.

For the XE loss training stage, the experimental results are reported in Table 4, from which we can observe that the proposed MGAN has the best results among all compared methods in terms of BLUE-1 to BLEU-3 and SPICE, as well as performs on par with AoANet and MT in METEOR and ROUGE. In particular, MGAN exceeds AoANet by 1.0 points on the BLEU-1 score. Besides, MGAN outperforms AoANet* (our reimplementation) in all metrics. The performance boost demonstrates the effectiveness of our model.

For the CIDEr reward optimization stage, the comparison results are reported in Table 5. From the table, we can find that MGAN outperforms most competing methods. MGAN has comparable results with AoANet, while comparing favorably to AoANet*. We infer that this difference may be caused by the low-dimensional image feature size and hidden size of LSTM. Furthermore, MGAN performs on par with MT but inferior than NG-SAN. We speculate that this is because NG-SAN utilizes high-dimensional image feature size and additional relative position and geometry relationships among objects to boost image understanding. Nevertheless, considering that our motivation is to simplify the transformer-based encoder and reduce computational overhead, the performance of MGAN is competitive.

E. RESULTS ON OTHER MODELS COMBINED WITH MGAs

The proposed MGAs, which stacks N layers of MGA block, is an additional image encoder and is flexible to be integrated into image caption models adopting encoder-decoder framework. Therefore, to completely verify the generalizability of MGAs, we combine it with three existing models: Up-Down, AoANet and X-LAN. For the Up-Down, the MGAs is applied in the same way as our MGAN, just plugging MGAs into the image encoder. As for the AoANet and X-LAN, which already have an extra transformer-like encoder, so we replace it with MGAs. Note that we do not make any modifications to the other modules of these two models.

As illustrated in Table 6, it can be observed that the MGAs boosts the performance of Up-Down, AoANet and X-LAN. Especially for Up-Down, the CIDEr score increases by 4.2 points. The improvement of AoANet and X-LAN is not as significant as Up-down. This can be attributed to that the visual features used in the decoder of AoANet and X-LAN have been refined by a transformer-like encoder, while the

TABLE 6. Performance of other models combined with MGAs on MS COCO Karpathy's test split under XE loss. * indicates that the results come from our reimplementation.

Model	B@1	B@4	M	R	C	S
Up-Down*	77.0	36.0	27.3	56.7	112.3	20.5
Up-Down + MGAs	**77.4**	**36.6**	**27.9**	**57.1**	**116.5**	**21.3**
AoANet*	77.3	37.1	28.2	57.4	117.1	21.4
AoANet + MGAs	**77.5**	**37.2**	**28.4**	**57.5**	**118.6**	**21.6**
X-LAN*	77.7	37.8	28.2	57.6	119.1	**21.5**
X-LAN + MGAs	**78.3**	**38.2**	**28.3**	**58.0**	**119.9**	**21.5**

Up-Down: a bathroom with a toilet and a sink.
Up-Down + MGAs: a bathroom with two sinks and a trash can.
AoANet: a bathroom with two sinks and a trash can.
AoANet + MGAs: a public restroom with two sinks and a trash can.
X-LAN: a bathroom with two sinks and a trash can.
X-LAN + MGAs: a bathroom with two sinks and a trash can.

Up-Down: a clock on the side of a building.
Up-Down + MGAs: a man standing in front of a large clock.
AoANet: a man standing next to a bunch of clocks.
AoANet + MGAs: a person standing in front of a clock on a wall.
X-LAN: a person standing next to a wall with clocks.
X-LAN + MGAs: a group of clocks that are on a wall.

Up-Down: a bench sitting on the side of a river.
Up-Down + MGAs: a boat is in the water near a bench.
AoANet: a boat is in the water near a bench.
AoANet + MGAs: a small boat in a river next to a bench.
X-LAN: a small boat on a body of water.
X-LAN + MGAs: a boat is on the water near a bench.

Up-Down: a train is on the tracks at a station.
Up-Down + MGAs: a train is parked at a train station.
AoANet: a train is parked at a train station.
AoANet + MGAs: a blue and yellow train is pulling into station.
X-LAN: a blue and white train at a train station.
X-LAN + MGAs: a blue and white train at a train station.

FIGURE 5. Examples of captions generated by Up-Down, AoANet, X-LAN and these models with MGAs respectively.

features used for Up-Down are not. Moreover, it is worth noting that additional exponential linear units are adopted in the encoder of X-LAN, but not in MGAs. Even so, the MGAs still brings further performance boost over these two strong baselines. The performance improvement on these three models indicates the advantages of constraining attention weight and considering the intra-object attention distribution.

Figure 5 illustrates some qualitative comparisons of image captioning results from Up-Down, AoANet, X-LAN and these models with MGAs. Compared with Up-Down, the Up-Down+MGAs can generate more accurate descriptions like *"trash can"* and *"two sinks"* instead of wrong object or quantities like *"toilet"* and *"a sink"* as shown in the first example. And in the second and third instances, the captions provided by Up-Down all lack vital objects as *"man"* and *"boat"*. The sentences produced by the AoANet are relevant and exact to image content. However, the AoANet+MGAs further generates more specific and detailed descriptions. For instance, AoANet+MGAs infers that the scene in the image is *"a public restroom"* rather than *"a bathroom"* in the first example; the *"train"* is *"blue and yellow"* in the fourth instance. As for X-LAN and X-LAN+MGAs, the generated sentences of these two models are similar, and both are accurate for the image content.

V. CONCLUSION

In this paper, we propose a Multi-Gate Attention (MGA) block, which modifies and extends the conventional self-attention by following components: an Attention Weight Gate (AWG) module that constrains the attention mechanism to focus on the most relevant information; a Self-Gated (SG) module that explicitly calculates the intra-object attention distribution of individual object feature vector; and a modified transformer, namely pre-layernorm transformer. By stacking multiple MGA blocks in the image encoder and applying AWG module in the language decoder, we devise the Multi-Gate Attention Network (MGAN) for image captioning. Extensive comparative experiments and ablation studies on MS COCO demonstrate the effectiveness of the MGAN as well as each component. Furthermore, the results on other models combined with MGA indicate the generalizability of our proposal.

REFERENCES

[1] K. Xu, 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. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[2] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 375–383.
[3] 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. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[4] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4634–4643.
[5] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4651–4659.
[6] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 21–29.
[7] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5659–5667.
[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[9] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8928–8937.
[10] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10971–10980.
[11] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10327–10336.
[12] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.
[13] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3156–3164.
[14] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.
[15] 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. Eur. Conf. Comput. Vis.* Springer, 2010, pp. 15–29.
[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "BabyTalk: Understanding and generating simple image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2891–2903, Dec. 2013.

[17] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, and H. Daumé, III, "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2012, pp. 747–756.
[18] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2012, pp. 359–368.
[19] A. Gupta, Y. Verma, and C. Jawahar, "Choosing linguistics over vision to describe images," in *Proc. 26th AAAI Conf. Artif. Intell.*, 2012, pp. 606–612.
[20] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "TreeTalk: Composition and compression of trees for image descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 351–362, Dec. 2014.
[21] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4894–4902.
[22] M. Chen, G. Ding, S. Zhao, H. Chen, Q. Liu, and J. Han, "Reference based LSTM for image captioning," in *Proc. AAAI*, 2017, pp. 3981–3987.
[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7008–7024.
[24] J. Wang, J. Tang, and J. Luo, "Multimodal attention with image text spatial relationship for OCR-based image captioning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4337–4345.
[25] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, "Multistage attention network for image inpainting," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107448.
[26] Q. Guo, X. Qiu, P. Liu, X. Xue, and Z. Zhang, "Multi-scale self-attention for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 7847–7854.
[27] F. Luo, H. Huang, Y. Duan, J. Liu, and Y. Liao, "Local geometric structure feature for dimensionality reduction of hyperspectral imagery," *Remote Sens.*, vol. 9, no. 8, p. 790, Aug. 2017.
[28] G. Shi, H. Huang, and L. Wang, "Unsupervised dimensionality reduction for hyperspectral imagery via local geometric structure feature learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1425–1429, Aug. 2020.
[29] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*. [Online]. Available: http://arxiv.org/abs/1901.02860
[30] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," 2018, *arXiv:1806.00187*. [Online]. Available: http://arxiv.org/abs/1806.00187
[31] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," 2020, *arXiv:2002.04745*. [Online]. Available: https://arxiv.org/abs/2002.04745
[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
[34] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3128–3137.
[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[36] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.
[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[39] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correalation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[40] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[41] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4566–4575.
[42] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 382–398.
[43] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 499–515.
[44] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of caption," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7454–7464.
[45] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 211–229.

WEITAO JIANG received the B.E. degree in electronic information science and technology from Dalian Maritime University, Dalian, China, in 2019. He is currently pursuing the M.E. degree in information and communication engineering with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, China. His current research interests include computer vision and image captioning.

XIYING LI (Member, IEEE) received the Ph.D. degree in optical engineering from the Beijing Institute of Technology, in 2002. She was a Visiting Scholar with Washington University. She is currently an Associate Professor with the School of Intelligent Systems Engineering, Sun Yat-sen University. Her main research interests include vehicle recognition and tracking, traffic video analysis and understanding, pedestrian detection and recognition, and video big data.

HAIFENG HU (Member, IEEE) received the Ph.D. degree from Sun Yat-sen University, in 2004. He is currently a Professor with the School of Electronics and Information Technology, Sun Yat-sen University. Since 2000, he has been published more than 120 articles. His research interests include computer vision, pattern recognition, image processing, and neural computation.

QIANG LU received the master's degree in agricultural mechanization from South China Agricultural University, in 2018. He is currently a Research Assistant with the School of Intelligent Systems Engineering, Sun Yat-sen University. His research interests include object detection and recognition, image understanding, and traffic behavior analysis.

BOHONG LIU received the B.E. degree in communication engineering from Sun Yat-sen University, Guangzhou, China, in 2019, where he is currently pursuing the M.E. degree in electronics and communication engineering with the School of Electronics and Information Technology. His research interests include computer vision and person re-identification.