

Image Captioning with Synergy-Gated Attention and Recurrent Fusion LSTM

You Yang^{1,2}, Lizhi Chen^{2*}, Longyue Pan², and Juntao Hu²

¹ National Center for Applied Mathematics in Chongqing.

Chongqing, 401331 China

[e-mail: 565357950@qq.com]

² School of Computer and Information Science, Chongqing Normal University.

Chongqing, 401331 China

[e-mail: 2019210516034@stu.cqnu.edu.cn]

*Corresponding author: Lizhi Chen

*Received July 28, 2022; revised September 12, 2022; accepted October 4, 2022;
published October 31, 2022*

Abstract

Long Short-Term Memory (LSTM) combined with attention mechanism is extensively used to generate semantic sentences of images in image captioning models. However, features of salient regions and spatial information are not utilized sufficiently in most related works. Meanwhile, the LSTM also suffers from the problem of underutilized information in a single time step. In the paper, two innovative approaches are proposed to solve these problems. First, the Synergy-Gated Attention (SGA) method is proposed, which can process the spatial features and the salient region features of given images simultaneously. SGA establishes a gated mechanism through the global features to guide the interaction of information between these two features. Then, the Recurrent Fusion LSTM (RF-LSTM) mechanism is proposed, which can predict the next hidden vectors in one time step and improve linguistic coherence by fusing future information. Experimental results on the benchmark dataset of MSCOCO show that compared with the state-of-the-art methods, the proposed method can improve the performance of image captioning model, and achieve competitive performance on multiple evaluation indicators.

Keywords: Image captioning, Synergy-Gated Attention, Recurrent Fusion LSTM, Deep learning.

This work is supported partially by the Chongqing Normal University graduate scientific research and innovation project (Grant No. YKC20038), the 13th five-year plan of Chongqing education science planning (Grant No. 2019-GX-10).

1. Introduction

Image captioning is a task that makes a sentence from reading an image. The sentence should be fluence and hold semantic consistency with image. Such a task is the intersection part of computer vision (CV) and natural language processing (NLP). Benefiting from the rapid development of machine translation [1] and object detection [2] in these two fields, image captioning has attracted more and more attention in recent years. To hold semantic consistency between image vision and sentence text, the image captioning model needs not only to recognize specific objects, but also to capture the relationships between objects.

The Encoder-decoder framework is the leading network of image captions currently. It's essentially a type of deep learning model. Duo to the captions generated by models based on deep learning [3-6] are closer to natural language than the conventional template-based [7] and retrieval-based [8] methods, so it has been widely used. CNN+RNN model is the representative framework instance, which finishes the task through two stages of encoding and decoding. In the first step, CNN [9] encodes the image content and extracts the semantic feature information of an image. In the second stage, RNN [10] decodes image features extracted from the encoder into corresponding captions. Another encoder-decoder instance is the transformer, in which encoding and decoding are all achieved by a specific number of transformer layer.

Before the Up-Down [11] model was proposed, captioning models usually used ResNet [9] and other similar encoders to extract the grid features of images and obtain the spatial semantic information of images. The Up-Down model first used Faster R-CNN [2] as an encoder to remove the salient region features of images for image captioning, and achieved state-of-the-art results at that time. However, most captioning models only use the features extracted by Faster R-CNN, discarding the grid features extracted by ResNet, which leads to the underutilization of spatial feature information. Based on this, a Synergy-Gated Attention (SGA) method is proposed, where the encoder attends to salient and spatial features of the image, and establishes a gated mechanism through the global features of both, to better control the interaction of two kinds of information.

Many current works in image captioning focus on the decoding side, exploring how RNNs decode the image features more efficiently. As a variant of RNNs, the LSTM [12] is widely used in the decoder of image captioning models. It plays a crucial role in processing sequential data, and by introducing the input gate, forget gate, and output gate, LSTM can effectively alleviate the problem of gradient vanishing in RNNs. However, the hidden vector of the LSTM output in one time step usually depends on the output vector of the preceding LSTM, which ignores the visual correlation of the posterior LSTM hidden vectors. To solve the above problem, we design the RF-LSTM to replace the traditional LSTM unit. The RF-LSTM predicts the sequence information of the posterior steps in one time step and uses this information together to guide the output of the current LSTM. In this way, our model can achieve competitive performance.

Experimentally, we explore the effects of SGA and RF-LSTM respectively, and note that both methods show good performance in captioning model. To comparing fairly with other models, we integrate SGA and RF-LSTM, called SGA-RF-LSTM. The overall architecture is shown in Fig. 1. Through quantitative and qualitative analysis, it is demonstrated that the proposed method achieves competitive performance against state-of-the-art methods. Specifically, we obtain 119.1 CIDEr score under cross-entropy loss (XE) and 130.0 CIDEr-D scores with reinforcement learning [13, 14] on MS COCO "Karpathy" offline test split [10]. The main work of this study is as follows:

1) We proposed a Synergy-Gated Attention (SGA), which can attend to salient region features and spatial features simultaneously. The global information of these two features is also used to guide the interaction between the two attended features.

2) We designed a Recurrent Fusion LSTM (RF-LSTM), which can obtain future output information to improve linguistic coherence. This is essentially different from recursing LSTM, which only focuses on previous output information.

3) We proposed an image captioning model combined with SGA and RF-LSTM, which achieve competitive performance compared to the state-of-the-art models, on the MS COCO datasets.

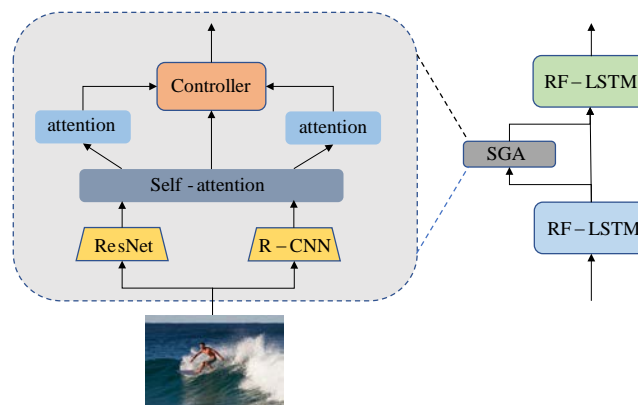


Fig. 1. An overview of the proposed framework.

2. Related work

2.1 Encoder-Decoder based captioning

In recent years, with the significant progress of deep learning, image captioning models have developed rapidly and acquired breakthrough results. Inspired by sequence-to-sequence tasks [15], such as machine translation, the encoder-decoder frameworks were widely used in image captioning models. Vinyals et al. [3] proposed a captioning model, where the encoder was designed by CNN, and the decoder was designed by LSTM. Image features were only utilized at the beginning of LSTM. Xu et al. [16] first introduced the attention mechanism into the image caption model, where attention was used at every moment of the LSTM to focus on the salient position information of images. After that, a series of innovations based on the encoder-decoder framework was proposed to guide the captioning models to generate sentences that meet the description of human language by adding semantic attributes [17, 18]. Moreover, to explore the relationship between visual regions and mine the available semantic information in images, some methods have emerged to build scene graphs [19-21], which enhance the representation of images and quality of captions by constructing visual relationship graphs.

Innovations in the structure of the decoder and LSTM-based refinements play a significant role in image captioning models, and an effective decoder can help the captioning model to generate more accurate descriptions. Ke et al. [22] proposed a reflective decoding network for image captioning, which enhances both the long-sequence dependency and position perception of words in a caption decoder. Li et al. [23] used CNN as a decoder to replace the conventional LSTM, which solved the problems of long-term memory loss and lack of parallel processing in LSTM.

2.2 Attention-based captioning

Currently, attention plays a significant role in the task of image captioning. The attention mechanism will selectively focus on the part of the image, allowing the model to obtain valuable information from the image quickly, which is more in line with human cognitive behavior. Lu et al. [24] proposed an adaptive attention model combined with a visual sentinel, which can adaptively decide whether to focus on visual information or non-visual text, so that meaningful information can be extracted at every time step. Wang et al. [25] proposed that using hierarchical features enables attention to be synchronously calculated on the features of pyramid levels, and multiple multi-modal integration strategies can significantly improve model performance. Huang et al. [26] proposed an attention module that enhances visual attention by further measuring the relevance between the attention result and the query. Based on Transformer [27], Herdade et al. [28] proved the importance of spatial awareness of the model by combining spatial relation information between objects through geometric attention. Given the excellent results obtained by the above methods, it inspires us to use different attention methods and different image features from multiple perspectives.

3. Method

In this paper, a new image captioning model is proposed to explore the diversity of image feature information fusion. Fig. 2 shows this model’s total framework. We first show multi-mode embedding in Section 3.1, then introduce the Synergy-Gated Attention (SGA) and Recurrent Fusion LSTM (RF-LSTM) in Section 3.2 and Section 3.3 respectively, and finally introduce the training implements of the model in Section 3.4.

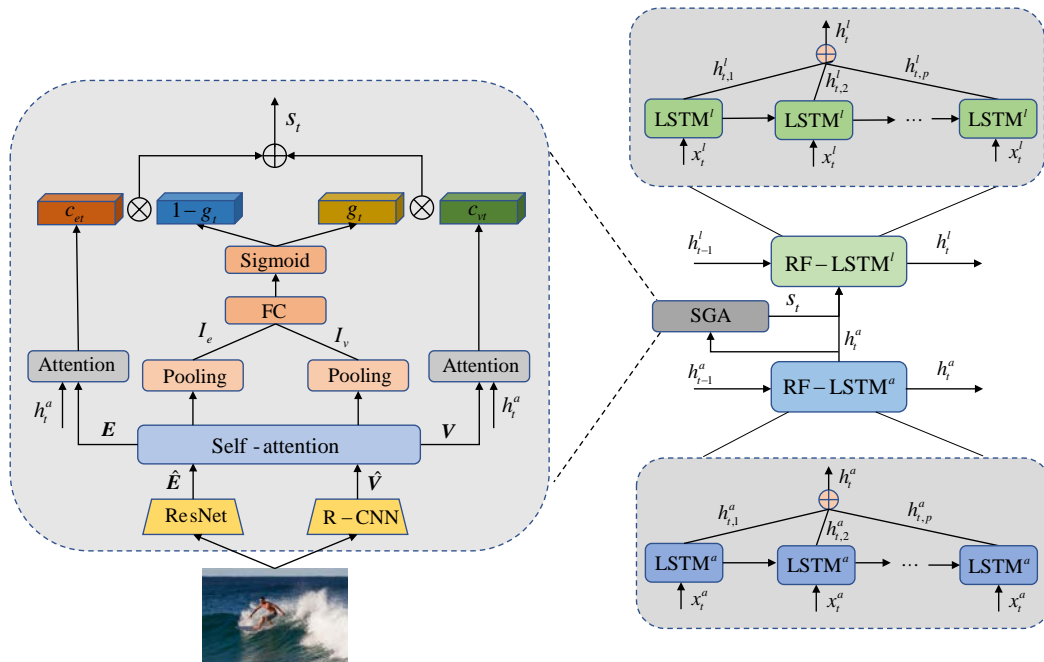


Fig. 2. The framework of the proposed SGA-RF-LSTM.

3.1 Multimodal embedding

We use CNN to extract spatial semantic information and Faster R-CNN to extract salient

region information. Based on the above, we obtain visual embedding $\hat{V} = [\hat{v}_1, \hat{v}_2, \dots, \hat{v}_N]$, $\hat{v}_i \in \mathbb{R}^d$ and space embedding $\hat{E} = [\hat{e}_1, \hat{e}_2, \dots, \hat{e}_M]$, $\hat{e}_i \in \mathbb{R}^d$. Since both the grid features and region features are visual contents and exist in images as a whole, it is necessary to model their interaction. Formally, given \hat{V} and \hat{E} , we use a 3-layer Transformer module $\psi(\cdot; \theta_a)$ to obtain more informative features via a self-attention operation as:

$$\hat{V}, \hat{E} = \psi \left(\begin{bmatrix} \hat{V} \\ \hat{E} \end{bmatrix}; \theta_a \right) \quad (1)$$

where $\hat{E} \in \mathbb{R}^{(w \times h) \times d}$ is the feature map outputted from the last convolutional layer of the ResNet. $M = w \times h$ denotes the number of grids composed of image areas of the same size. $\hat{V} \in \mathbb{R}^{N \times d}$ is the output vector of the Faster R-CNN, which is composed of N d -dimensional image area features v_i . $E = \{e_1, e_2, \dots, e_M\}$, $e_i \in \mathbb{R}^d$, $V = \{v_1, v_2, \dots, v_N\}$, $v_i \in \mathbb{R}^d$.

3.2 Synergy-Gated Attention

Our Synergy-Gated Attention method performs a multimodal task based on the multimodal embedding generation (V, E) in Section 3.1. Both varieties of the information are fused so that the LSTM can employ different regional and spatial information of the image simultaneously to generate the current captions at each moment.

The formula for calculating the attentive weights of E is defined as follows:

$$Z_{et} = W_{eh}^T \tanh(W_e E + (W_{eh} h_t^a) a_e^T) \quad (2)$$

$$\alpha_{et} = \text{soft max}(Z_{et}) \quad (3)$$

$$c_{et} = \sum_{i=1}^M \alpha_{et,i} e_i \quad (4)$$

where W_{eh}^T , W_e and W_{eh} are the matrices for learning spatial attentive weight. $\alpha_{et} = \{\alpha_{et,1}, \alpha_{et,2}, \dots, \alpha_{et,M}\}$ is the related weights of E , which sums to 1. c_{et} is a weighted sum of E , and it indicates the most relevant position of grid region of the images.

The formula of calculating the attentive weights of V is defined as follows:

$$Z_{vt} = W_{vh}^T \tanh(W_v V + (W_{vh} h_t^a) a_v^T) \quad (5)$$

$$\alpha_{vt} = \text{soft max}(Z_{vt}) \quad (6)$$

$$c_{vt} = \sum_{i=1}^N \alpha_{vt,i} v_i \quad (7)$$

where W_{vh}^T , W_v and W_{vh} are the matrices for learning regional attentive weight. $\alpha_{vt} = \{\alpha_{vt,1}, \alpha_{vt,2}, \dots, \alpha_{vt,N}\}$ are the related weights of V , which sums to 1. c_{vt} is a weighted sum of V , and it indicates the most relevant position of salient region of the images.

The utilization of two features may produce semantic noises during the fusion process. To solve the problem, we concatenate the pooling features I_e extracted by the CNN model and the pooling features I_v extracted by Faster R-CNN. Then we send the concatenated matrix to the gate control unit to obtain the gate output value g_t :

$$I_e = \frac{1}{M} \sum_{i=1}^M e_{ti} \quad (8)$$

$$I_v = \frac{1}{N} \sum_{i=1}^N v_{ti} \quad (9)$$

$$g_t = \sigma(W_g [I_e, I_v]) \quad (10)$$

where $I_e \in \mathbb{R}^{d \times 1}$ denotes the mean value of the feature map vector E , and $I_v \in \mathbb{R}^{d \times 1}$ denotes the mean value of vector V of salient region features. $\sigma(\cdot)$ denotes a Sigmoid function.

Through the dual gate control [29], the attentive information of the salient region is guided by the gate output value g_t , and the supplementary value $(1 - g_t)$ guides the spatial semantic information to achieve the final attentive fusion:

$$s_t = g_t \odot c_{vt} + (1 - g_t) \odot c_{et} \quad (11)$$

where \odot indicates the Hadamard product, and $s_t \in \mathbb{R}^{d \times 1}$ represents the output of SGA.

3.3 Recurrent Fusion LSTM

To improve the performance of sequence generation in LSTM, we introduce Recurrent Fusion LSTM (RF-LSTM). As shown in Fig. 2, the structure is an encoder-decoder framework based on double-layer RF-LSTM. The first layer is the attention Recurrent Fusion LSTM, which denotes $RF - LSTM^a$ composed of attention LSTMs that generate attentive weight. The second layer is the language Recurrent Fusion LSTM, which indicates $RF - LSTM^l$ composed of language LSTMs that generate words.

Our RF-LSTM, by recurring multiple LSTMs in one time step, focuses on modeling the same input, and establishes the relationship between the input information. The number of recurrences in the same layer is P , which means that there are P different fusion outputs in each layer.

On the first layer, the hidden state of attention RF-LSTM h_t^a is calculated as follows:

$$h_t^a = RF - LSTM^a(x_t^a, h_{t-1}^a) = \frac{1}{P} \sum_{i=1}^P h_{t,i}^a \quad (12)$$

$$h_{t,i}^a = LSTM^a(x_t^a, h_{t-1,i}^a), i = 1, 2, \dots, P \quad (13)$$

where x_t^a is the input vector of attention RF-LSTM, and h_{t-1}^a is the hidden state of previous time step of attention RF-LSTM. $h_{t,i}^a$ is the i -th output of attention RF-LSTM at time step t .

The input of attention RF-LSTM consists of the word embedding in the current time step and visual vector $I_v + h_{t-1}^a$, where I_v is the pooling feature extracted by Faster R-CNN and h_{t-1}^a is the context vector of the previous time step of language RF-LSTM (h_{t-1}^a is initialized to 0 at the beginning):

$$x_t^a = [Ew_{t-1}, I_v + h_{t-1}^a] \quad (14)$$

where E is the embedding matrix of words, and w_{t-1} is the word generated by language RF-LSTM at the previous time step. We follow the earlier approach in the traditional image captioning baseline model, in which the embedding of each word token depends on its context. Specifically, create a learnable weight of shape (x, y) , where x represents the size of the dictionary and y represents the dimension of the embedding vector, initialized as a random number in the range $(0, 1)$. It is worth noting that the word embedding generation method can also be replaced with the pre-trained language model to extract language features.

On the second layer, the hidden state of language LSTM is calculated as:

$$h_t^l = RF - LSTM^l(x_t^l, h_{t-1}^l) = \frac{1}{P} \sum_{i=1}^P h_{t,i}^l \quad (15)$$

$$h_{t,i}^l = LSTM^l(x_t^l, h_{t-1,i}^l), i = 1, 2, \dots, P \quad (16)$$

where x_t^l is the input of language RF-LSTM, and h_{t-1}^l is the hidden state of previous time step of language RF-LSTM. $h_{t,i}^l$ is the i -th output of language RF-LSTM at time step t .

The input of language RF-LSTM is denoted as x_t^l , which is defined as follows:

$$x_t^l = [s_t, h_t^a] \quad (17)$$

where s_t denotes the output of SGA, and h_t^a is the hidden state of attention RF-LSTM at current time step.

The probability distribution of the output word of the SGA-RF-LSTM model at time step t is denoted as $p(y_t|y_{1:t-1})$:

$$p(y_t|y_{1:t-1}) = \text{softmax}(W_p h_t^l) \quad (18)$$

where h_t^l denotes the hidden state of language RF-LSTM at time step t , and $W_p \in \mathbb{R}^{d \times 1}$ is a parameter matrix.

3.4 Objectives

Given the target ground truth sequence $y_{1:T}^*$ and the captioning model with parameter θ , SGA-RF-LSTM is trained by minimizing cross entropy L_{XE} :

$$L_{XE}(\theta) = -\sum_{t=1}^T \log(p_\theta(y_t^*|y_{1:t-1}^*)) \quad (19)$$

Since reinforcement learning has been used to train captioning models, we follow this training strategy to optimize non-differentiable metrics, and then seek the minimum negative expected score from the initialization of the trained model under cross-entropy:

$$L_R(\theta) = -E_{y_{1:T} \sim p_\theta}[r(y_{1:T})] \quad (20)$$

where r is the CIDEr-D score function. We directly optimize the non-differentiable metrics with Self-Critical Sequence Training (SCST) [14], and the gradient can be approximated:

$$\nabla_\theta L_R(\theta) \approx -(r(y_{1:T}^s) - r(\hat{y}_{1:T})) \nabla_\theta \log p_\theta(y_{1:T}^s) \quad (21)$$

where $y_{1:T}^s$ denotes a result sampled from a probability distribution, and $r(\hat{y}_{1:T})$ is the baseline score of greedily decoding.

4. Experiments

4.1 Datasets

The proposed method is evaluated on the MSCOCO 2014 [30] and Flickr30K dataset [39]. The MSCOCO dataset is the largest offline dataset for the image captioning task, which contains 123,287 images with five different annotations. For offline evaluation, we use the "Karpathy" Data Split [10] where 113,287 images are used for training and 5000 respectively for validation and testing. For the Flickr30K dataset, 29014, 1000 and 1000 images are used to train, validate, and test respectively. To quantitatively evaluate the performance of the method proposed in the paper and compare it with other methods, we use standard automatic evaluation metrics, including BLEU [31], METEOR [32], ROUGE-L [33], CIDEr-D [34] and SPICE [35].

4.2 Implementation Details

We use the pre-trained ResNet-101 to extract the grid features of the images, and the Faster R-CNN [2] to extract the salient region features of images. In training implementation, the dimension of the original encoding feature vector is 2048, and we project it into a new space with a dimension of 1024. The dimension of pooling and attentive layer of SGA is 1024. We follow the training strategy of AoA model [26]. In XE training stage, the batch size is set to 10 with 40 epochs. We initialize the learning rate to $2e-4$, and anneal it by 0.8 every 3 epochs.

The predetermined sampling probability [36] is increased by 0.05 every 5 epochs. In reinforcement learning stage, we initialize the learning rate to 2e-5 to train 20 epochs. When the validation score does not improve on some metrics, we anneal it to 0.5. We employ Adam optimizer in both stages and the beam size is set to 2.

4.3 Performance Comparison

As shown in Table 1, we report the performance of the proposed model on the offline COCO Karpathy test split and compare the performance of our approach with that of several recent image captioning models. The compared models include: SCST [14], which applies advanced attribute features to image captioning tasks; RFNet [4], which fused the encoding features of multiple CNN networks to form a representation of the decoder; Up-Down [11], which employed the Faster R-CNN as the bottom-up mechanism, extracting the salient region features; GCN-LSTM [19], which used Graph Convolutional Networks to explore pair-wise relations between image regions; HAN [25], which proposed the adoption of hierarchical features so that attention could be calculated synchronously on the features of pyramid levels; AoANet [26], which enhanced traditional visual information attention by further measuring the correlation between attention results and queries; SRT [6], which proposed a new recall mechanism consisting of recall unit, semantic guidance and recall words; MT [37], which constructed a fully connected architecture between each encoder layer and decoder layer. We can see that our model has achieved the highest scores compared with other models in most of the metrics. Model it is important to note that the repeat five times experiments, we found five experimental results scores than the baseline model, the results score value fluctuates up and down in the experiment between the average of five times, the probability of 0.5 to the fifth power, The P value is 0.03125, less than significant level, the considerable difference statistically significant.

Table 1. Performance of our method on MS COCO Karpathy’s test split under XE loss and CIDEr reward optimization, where B-1 / B-4 / M / R / C / S means BLEU1/ BLEU4 / METEOR / ROUGE-L / CIDEr / SPICE scores

	Cross-Entropy Loss					CIDEr-D Score Optimization				
	B-1	B-4	M	R	C	B-1	B-4	M	R	C
SCST [14]	-	30.0	25.9	53.4	99.4	-	35.5	27.3	56.8	118.3
RFNet [4]	76.4	35.8	27.4	56.8	112.5	79.1	36.5	27.7	57.3	121.9
Up-Down [11]	77.2	36.2	27.0	56.4	113.5	79.8	36.3	27.7	56.9	120.1
GCN-LSTM [19]	77.3	36.8	27.9	57.0	116.3	80.5	38.2	28.5	58.5	128.3
HAN [25]	77.2	36.2	27.5	56.6	114.8	80.9	37.6	27.8	58.1	121.7
AoANet [26]	77.4	37.2	28.4	57.5	119.8	80.2	38.9	29.2	58.8	129.8
SRT [6]	77.1	36.6	28.0	56.9	116.9	80.3	38.5	28.7	58.4	129.1
MT [37]	-	-	-	-	-	80.8	39.1	29.2	58.6	131.2
SGA-RF-LSTM	77.6	37.4	28.5	57.7	119.1	80.6	39.5	29.4	59.0	130.0

As shown in **Table 2**, we report experimental results on the Flickr30K dataset. Our approach significantly outperforms all the compared methods, indicating that the performance improvement produced by our proposed SGA-RF-LSTM model is equally effective on different datasets.

Table 2. The image captioning results obtained on the Flickr30K Karpathy test split under XE loss

Models	B-1	B-4	M	R	C	S
Adaptive [24]	67.7	25.1	20.4	46.7	53.1	-
GLA [40]	56.8	14.6	16.6	41.9	36.2	-
DHEDN-3 [41]	65.3	23.1	19.2	-	-	-
DSEN+T [42]	-	28.7	21.9	-	68.7	-
SGA-RF-LSTM	68.9	29.5	22.4	49.1	70.2	15.5

For better qualitatively evaluating the generated results, we visualize the evolutions of the contribution of visual features to the model output for AoA [26] and SGA in **Fig. 3**. The assistance of one region related to the output is given by non-linear correlation. Hence, we employ the Integrated Gradients approach [38], which approximates the integral of gradients concerning the given input. In addition, our SGA attaches the attention heat map on the grid, which can cooperate with the salient region information to focus on the corresponding grid spatial features, thereby helping the model to effectively use the image information and generate accurate captions.

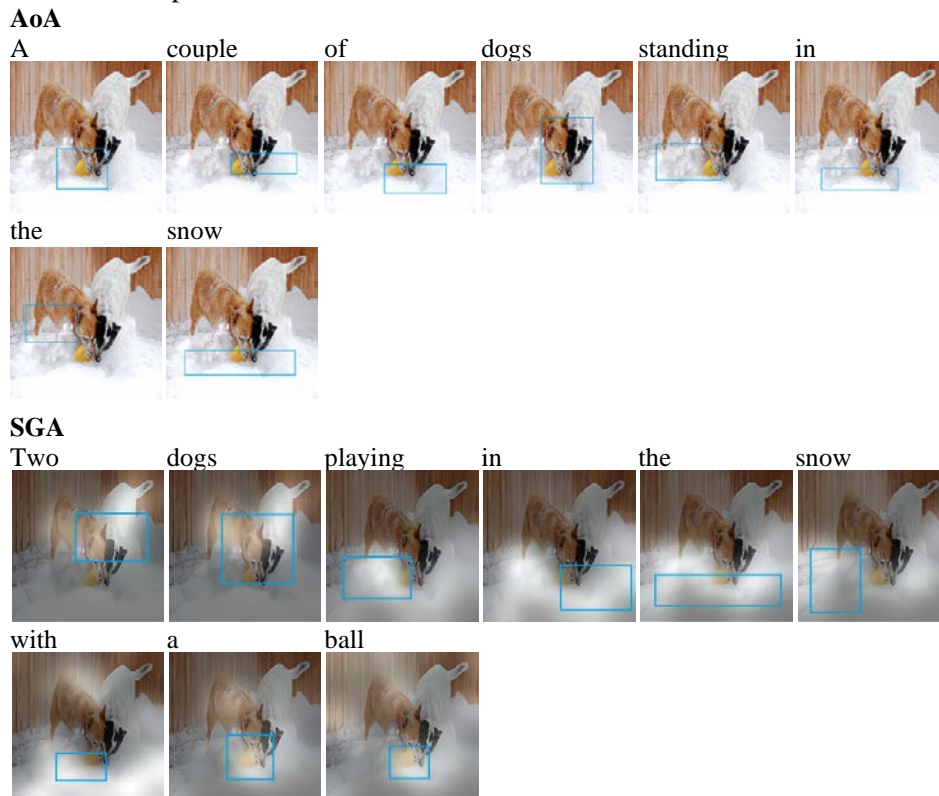





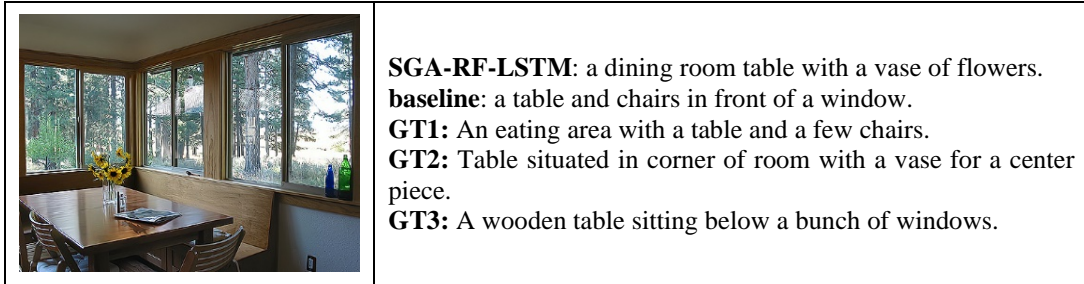
Fig. 3. The visualization and captions generation processes of the AoA model and the SGA model.

4.4 Qualitative Analysis

Table 3 shows several example image captions, which contain images and their caption generated by the proposed SGA-RF-LSTM, Up-Down [11] baseline, and two ground truths (GT) respectively. From these examples, we found that the captions generated by the baseline model are logically correct, but not accurate enough, and even some captions do not match the image content. SGA-RF-LSTM generates more descriptive and precise captions. For example, the baseline model generates "a bathroom with a toilet and a shower". Although the caption is correct, it does not clearly describe the positional relationship between the objects. SGA-RF-LSTM accurately describes the position information of "next to a white toilet". Besides, the baseline model generates "a group of people are skiing on the snow", while SGA-RF-LSTM specifically describes "a man and little girl". SGA-RF-LSTM has such advantages, because it combines the grid features and salient region features of the image simultaneously, and uses the recurrent fusion method in the decoding stage to enhance the output of LSTM.

Table 3. Examples of image captioning results generated by our SGA-RF-LSTM

Image	Captions
	<p>SGA-RF-LSTM: a bathroom with a walk in shower next to a white toilet.</p> <p>baseline: a bathroom with a toilet and a shower.</p> <p>GT1: A bathroom with an enclosed shower next to a sink and a toilet.</p> <p>GT2: A bathroom featuring a walk in shower, mirror, sink and toilet.</p> <p>GT3: There are a toilet, a sink, and a shower stall in a large bathroom.</p>
	<p>SGA-RF-LSTM: a brown dog laying on the ground next to a pool.</p> <p>baseline: a large brown dog laying in a pool.</p> <p>GT1: A dog laying down next to a pool in a backyard.</p> <p>GT2: A golden retriever sleeps at the edge of the pool.</p> <p>GT3: A golden retriever laying down on the side of a pool.</p>
	<p>SGA-RF-LSTM: A man and little girl are on skis in the snow.</p> <p>baseline: A group of people are skiing on the snow.</p> <p>GT1: A man and little girl are on skis in the snow.</p> <p>GT2: A man and a child skiing on a snowy plain.</p> <p>GT3: a person riding ski on a snowy surface.</p>



4.5 Ablation Study

To quantify the impact of the proposed SGA-RF-LSTM on the image captioning model in the sequence generation stage and the attention stage, we performed ablation experiments by comparing the different variants of SGA-RF-LSTM. As shown in **Table 4**, the first two rows represent the baseline model using only grid features and regional features, rows 3 through 11 represent ablation experiments with SGA and RF-LSTM, and the penultimate row represents the performance scores combined with SGA and RF-LSTM. It is worth noting that the last line experiments the results of the generation method based on BERT [43] word embedding based on the model proposed in this paper, and it is found that the model's performance is further improved compared with the previous one.

Table 4. Ablation study about the SGA-RF-LSTM under XE loss

Encoder	Decoder	B-1	B-4	M	R	C	S
ResNet(Grid)	LSTM	75.8	34.8	27.2	56.1	109.9	20.3
Fatser – RCNN(Region)	LSTM	76.0	35.8	27.6	56.5	113.3	20.8
Region + Region + A	LSTM	76.6	35.9	27.7	56.6	113.8	20.8
Grid × Region + A	LSTM	76.6	36.0	27.7	56.7	113.9	20.8
Grid + Region + A	LSTM	76.9	36.0	27.8	56.7	114.9	21.0
Grid + Region + GA(SGA)	LSTM	77.0	36.1	27.9	56.8	115.5	21.1
Region	P – LSTM ^l	76.4	35.8	27.7	56.6	113.3	20.8
Region	PF – LSTM ^l	76.1	35.7	27.8	56.5	114.1	20.9
Region	RF – LSTM ^l	76.4	36.0	27.8	56.7	114.4	20.9
Region	RF – LSTM ^a	76.6	36.1	27.7	56.8	113.7	20.7
Region	RF – LSTM ^{l+a}	76.8	36.2	27.9	56.9	115.4	21.0
SGA – RF – LSTM ^{l+a}		77.6	37.4	28.5	57.7	119.1	21.4
SGA – RF – LSTM ^{l+a} + BERT		77.9	37.5	28.7	57.9	119.7	21.4

4.5.1 Effect of SGA

We set up different schemes to evaluate the effect of the SGA method in the attention stage. First, to explore the impact of paying attention to two salient region features simultaneously in the attention phase, we set "Region+Region+A", where "A" represents the soft attention

processing; Second, to verify the influence of paying attention to the grid features and salient region features on the model simultaneously, we set "Grid \times Region+A" and "Grid+Region+A", where " \times " means the relationships between features established through matrix multiplication, and "+" means relationships between features shown through matrix summation. Finally, to verify the impact of the gating mechanism on the model by using the pooling information of both parts while focusing on the grid features and the salient region features, we set "Grid+Region+GA", where "GA" means adding the gating mechanism. From [Table 4](#), we observe that paying attention to two salient region features simultaneously can improve the performance of the model compared to single attention. However, the improvement of cooperatively attending grid and region feature is more prominent, which demonstrates that the grid space information can be better used as a complement to the salient region information. Furthermore, we find that the gating mechanism between two different features can effectively alleviate semantic noise and guide the interaction of information between them.

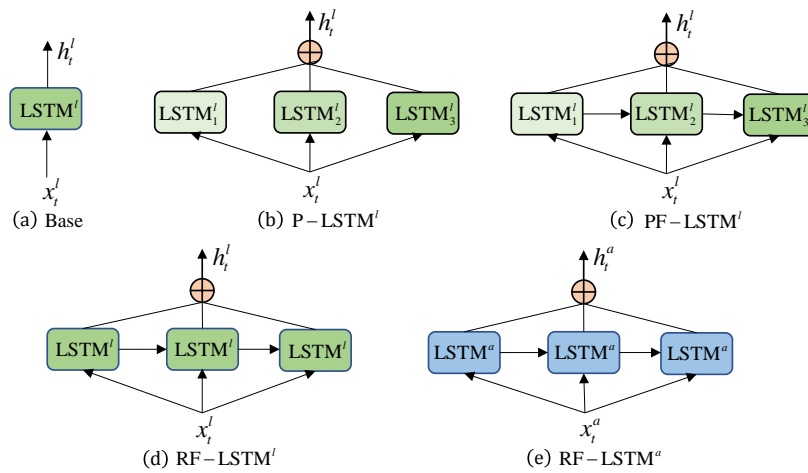


Fig. 4. Different schemes of outputting LSTM hidden state. (a) a base LSTM^l output h_t^l . (b) Pooled and merged output h_t^l by three parallel LSTM^l. (c) Fusion output h_t^l by three different recurrent LSTM^l. (d) Fusion output h_t^l by three identical recurrent LSTM^l. (e) Fusion output h_t^a by three identical recurrent LSTM^a.

4.5.2 Effect of RF-LSTM

As shown in [Fig. 4](#), we design different structures of LSTM and compare the use of different variants to model the vectors of hidden state. From [Table 4](#), we observed that three parallel different LSTMs do not improve the model's performance. Compared with the paralleled LSTM in the same layer, we use the proposed RF-LSTM method. The recurrent fusion of three different LSTMs in the same layer can improve performance slightly. We also find that the same parameters used in the same layer of LSTM can further improve performance. Meanwhile, we also evaluate the use of RF-LSTM in the first layer of the decoder and find that the performance is better than the original LSTM, but the execution is not as good as in the second layer. Eventually, we use the RF-LSTM method on both LSTM layers and find that the model outperforms the other structures to achieve the highest performance.

We combined the Synergy-Gated Attention and Recurrent Fusion LSTM method to form SGA – RF – LSTM^{l+a}. We set the variable on times of LSTM's recurrence to verify the impact of recurrence times of single time step on model performance in [Table 5](#). Generally, more repetitions in a single time step can get more different outputs, and the fusion of these output vectors can improve the performance of the model. We observe that three times of recurrent

fusion can achieve the best performance, which verifies the effectiveness of recurrent fusion LSTM for modeling inputs in one time step.

Table 5. Ablation on the times of LSTM's recurrence under XE loss

Models	B-1	B-4	M	R	C	S
SGA – (RF – LSTM ^{l+a} × 1)	76.8	36.2	28.1	56.9	117.9	21.1
SGA – (RF – LSTM ^{l+a} × 2)	77.1	36.3	28.0	57.0	118.7	21.3
SGA – (RF – LSTM ^{l+a} × 3)	77.6	37.4	28.5	57.7	119.1	21.4
SGA – (RF – LSTM ^{l+a} × 4)	76.9	36.2	28.0	56.9	117.3	21.1
SGA – (RF – LSTM ^{l+a} × 5)	77.3	36.6	28.1	57.1	118.0	21.3

5. Conclusion

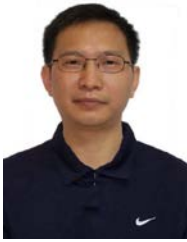
In this paper, we propose a Synergy-Gated Attention (SGA) method, which can attend to the salient region features and grid features simultaneously, so that the image information can be better used at the attention stage. We also propose a gating mechanism by using the global information of the two feature sources, which effectively guides the interaction between the two source information, and alleviates the problem of semantic noise generated during the fusion processing. In addition, we replace the original LSTM with the RF-LSTM. The new architecture not only relies on the previously hidden vector information, but also integrate information from future predictions to guide the current word generation, resulting in better performance than the original LSTM. Extensive experiments show the superiority of the proposed method compared with the state-of-the-art methods on the benchmark dataset.

References

- [1] M. T. Luong, H. Pham, C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412-1421, 2015. [Article \(CrossRef Link\)](#)
- [2] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, June 2017. [Article \(CrossRef Link\)](#)
- [3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156-3164, 2015. [Article \(CrossRef Link\)](#)
- [4] W. Jiang, L. Ma, Y. Jiang, W. Liu and T. Zhang, "Recurrent Fusion Network for Image Captioning," in *Proc. of European Conference on Computer Vision*, vol. 11206, pp. 510-526, October 2018. [Article \(CrossRef Link\)](#)
- [5] X. Chen, L. Ma, W. Jiang, J. Yao and W. Liu, "Regularizing RNNs for Caption Generation by Reconstructing the Past with the Present," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7995-8003, 2018. [Article \(CrossRef Link\)](#)
- [6] L. Wang, Z. Bai, Y. Zhang, and H. Lu, "Show, Recall, and Tell: Image Captioning with Recall Mechanism," in *Proc. of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 07, pp. 12176-12183, 2020. [Article \(CrossRef Link\)](#)
- [7] 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. of European conference on computer vision*, pp. 15-29, 2010. [Article \(CrossRef Link\)](#)

- [8] R. Mason, E. Charniak. "Nonparametric method for data-driven image captioning," in *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, pp. 592-598, 2014. [Article \(CrossRef Link\)](#)
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. [Article \(CrossRef Link\)](#)
- [10] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 664-676, 1 April 2017. [Article \(CrossRef Link\)](#).
- [11] P. Anderson, X. D. 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. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077-6086, 2018. [Article \(CrossRef Link\)](#)
- [12] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997. [Article \(CrossRef Link\)](#).
- [13] Z. Ren, X. Wang, N. Zhang, X. Lv and L. Li, "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1151-1159, 2017. [Article \(CrossRef Link\)](#)
- [14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, "Self-Critical Sequence Training for Image Captioning," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179-1195, 2017. [Article \(CrossRef Link\)](#)
- [15] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, pp. 3104-3112, 2014. [Article \(CrossRef Link\)](#)
- [16] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R Salakhutdinov, R. S. Zemel and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. of International conference on machine learning (ICML)*, vol. 37, pp. 2048-2057, July 2015. [Article \(CrossRef Link\)](#)
- [17] T. Yao, Y. Pan, Y. Li, Z. Qiu and T. Mei, "Boosting Image Captioning with Attributes," in *Proc. of 2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4904-4912, 2017. [Article \(CrossRef Link\)](#)
- [18] N. Li and Z. Chen, "Image Captioning with Visual-Semantic LSTM," in *Proc. of Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 793-799, 2018. [Article \(CrossRef Link\)](#)
- [19] T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring visual relationship for image captioning," in *Proc. of the European conference on computer vision (ECCV)*, pp. 711-727, 2018. [Article \(CrossRef Link\)](#)
- [20] X. Yang, H. Zhang and J. Cai, "Auto-encoding and Distilling Scene Graphs for Image Captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2313-2327, 2022. [Article \(CrossRef Link\)](#)
- [21] Y. Zhong, L. Wang, J. Chen, D. Yu and Y. Li, "Comprehensive image captioning via scene graph decomposition," in *Proc. of European Conference on Computer Vision*, pp. 211-229, 2020. [Article \(CrossRef Link\)](#)
- [22] L. Ke, W. Pei, R. Li, X. Shen and Y. Tai, "Reflective Decoding Network for Image Captioning," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8887-8896, 2019. [Article \(CrossRef Link\)](#)
- [23] R. Li, H. Liang, Y. Shi and F. Feng, "Dual-CNN: a convolutional language decoder for paragraph image captioning," *Neurocomputing*, vol. 396, no. 12, pp. 92-101, 2020. [Article \(CrossRef Link\)](#)
- [24] J. Lu, C. Xiong, D. Parikh and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," in *Proc. of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3242-3250, 2017. [Article \(CrossRef Link\)](#)
- [25] W. Wang, Z. Chen, H. Hu, "Hierarchical attention network for image captioning," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8957-8964, 2019. [Article \(CrossRef Link\)](#)

- [26] L. Huang, W. Wang, J. Chen and X. Wei, "Attention on Attention for Image Captioning," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4633-4642, 2019. [Article \(CrossRef Link\)](#)
- [27] 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. of the 31st International Conference on Neural Information Processing Systems*, pp. 6000-6010, 2017. [Article \(CrossRef Link\)](#)
- [28] S. Herdade, A. Kappeler, K. Boakye, J. Soares, "Image captioning: Transforming objects into words," *Advances in Neural Information Processing Systems*, pp. 11137–11147, 2019.
- [29] G. Li, L. Zhu, P. Liu and Y. Yang, "Entangled Transformer for Image Captioning," in *Proc. of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8927-8936, 2019. [Article \(CrossRef Link\)](#)
- [30] T. Y. Lin, M. Maire, S. Belongie, H. James, P. Pietro, R. Deva and D. Piotr, "Microsoft coco: Common objects in context," in *Proc. of European conference on computer vision*, pp. 740-755, 2014. [Article \(CrossRef Link\)](#)
- [31] K. Papineni, S. Roukos, T. Ward and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318, 2002. [Article \(CrossRef Link\)](#)
- [32] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. of the ninth workshop on statistical machine translation*, pp. 376-380, 2014.
- [33] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. of Association for Computational Linguistics Workshop*, pp. 74-81, 2004.
- [34] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566-4575, 2015. [Article \(CrossRef Link\)](#)
- [35] P. Anderson, B. Fernando, M. Johnson and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. of European conference on computer vision*, Springer, Cham, vol.9909, pp. 382-398, 2016. [Article \(CrossRef Link\)](#)
- [36] S. Bengio, O. Vinyals, N. Jaitly and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. of the 28th International Conference on Neural Information Processing Systems*, Vol. 1, pp. 1171-1179, 2015. [Article \(CrossRef Link\)](#)
- [37] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10575-10584, 2020. [Article \(CrossRef Link\)](#)
- [38] M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks," in *Proc. of 2017 International Conference on Machine Learning (PMLR)*, vol. 70, pp. 3319-3328, 2020. [Article \(CrossRef Link\)](#)
- [39] P. Young, A. Lai, M. Hodosh and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *Proc. of 2014 Transactions of the Association for Computational Linguistics (ACL)*, vol. 2, pp. 67-78, 2014. [Article \(CrossRef Link\)](#)
- [40] L. Li, S. Tang, Y. Zhang, L. Deng, Q. Tian. "GLA: Global-local Attention for Image Description," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 726-737, 2018. [Article \(CrossRef Link\)](#)
- [41] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, "Deep Hierarchical Encoder-Decoder Network for Image Captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2942-2956, 2019. [Article \(CrossRef Link\)](#)
- [42] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, "Dense Semantic Embedding Network for Image Captioning," *Pattern Recognition*, pp. 285-296, vol. 90, 2019. [Article \(CrossRef Link\)](#)
- [43] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Article \(CrossRef Link\)](#)



You Yang received the Ph.D. degree in computer application technology from Beihang University, Beijing, China, in 2010. He is currently an associate professor of National Center for Applied Mathematics in Chongqing. His research interests include computer vision and document image processing.



Lizhi Chen received the B.Eng. degree in software engineering from Linyi University, Shandong, China, in 2019. He is currently a candidate for M.Eng. degree in School of Computer and Information Science, Chongqing Normal University, Chongqing, China. His research interests include image captioning and deep neural network.



Longyue Pan received the B.Eng. degree in Information Management and Information System from Tonghua Normal University, Jilin, China, in 2020. She is currently a candidate for M.Eng. degree in School of Computer and Information Science, Chongqing Normal University, Chongqing, China. Her research interests include image captioning.



Juntao Hu received the B.Eng. degree in electronic information science and technology from Chongqing Normal University, Chongqing, China, in 2018. He is currently pursuing his M.Eng. degree in School of Computer and Information Science, Chongqing Normal University, Chongqing, China. His research interests include image captioning and multimodal representation learning.