Image Captioning Based on An Improved Transformer with IoU Position Encoding

Yazhou Li†, Yihui Shi†, Yun Liu, Ruifan Li‡ and Zhanyu Ma
† Both authors contributed equally to this research. ‡ Corresponding author.
School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
E-mail: {yhshi, yazhouli, yunliu, rfli, mazhanyu}@bupt.edu.cn

***Abstract*—The task of image captioning aims to automatically generate descriptive sentences for a given image. Most existing works use recurrent neural network as language decoder. In this paper, we use a transformer structure to generate descriptive captions. When applied in the task of image captioning, the transformer network exists two problems. The first is the disappearance of the query vector information in stacking network. The second is the lacking of spatial information between objects in the decoding process. To solve these problems, we propose an improved *T*ransformer with *I*oU *P*osition encoding model, i.e., TIP. We improve the transformer from two aspects. First, we propose an intra-modal attention mechanism to alleviate the problem of vanishing query vectors. Second, we propose an Intersection-over-Union (IoU) spatial position encoding method to enhance the semantic information of images. Extensive experiments on MS-COCO datasets demonstrate the effectiveness of our model.**

I. Introduction

Image captioning [1], [2] is a challenging task, which aims to describe image content in language sentences. It involves two different fields: computer vision and natural language processing. Image captioning not only needs to identify the objects in the image and the relationship between the objects, but also needs to describe them in coherent language sentences. Inspired by machine translation [3], most advanced models adopt the encoder-decoder framework, which use convolutional neural networks(CNN) [4] to extract image feature information, recurrent neural networks (RNN) [2] to generate captions.

To enable the model to dynamically focus on different regions of the image, Xu et al. [5] add an attention module to the encoder-decoder framework. Their method focuses on visual information when generating each word, but it needs to pay more attention to non-visual information when generating non-visual words. In order to solve this problem, Lu et al. [6] propose an adaptive attention mechanism, which can make the model pay attention to visual or non-visual information. Anderson et al. [4] propose a combined top-down and bottom-up attention model approach and apply it to image captioning. Bottom-up attention module is used to extract the region of interest in the image and obtain the feature of the object. The advantage of this method is that we can set a threshold to select the number of regions of interest. To enrich the features of the image, You et al. [7] propose to use a target detector to detect all objects, and combine the objects' attributes with the features of the image as the input of the decoder. Simao et al. [8] propose the object relation model. Their model encodes position and relationships between detected objects in images. Existing approaches generate sentences with low diversity and do not consider the content of interest. Chen et al. [9] propose to use scene graphs to generate image captions in a fine-grained way.

The aforementioned models use RNN structure as decoder. However, the sequential structure of RNN leads to limited long-term memory capacity in the decoding process. Recently, the transformer structure has verified its effectiveness in sequential tasks, and self-attention networks have been widely used in multi-modal tasks. Zhang et al. [10] introduce an adaptive attention mechanism based on the transformer, which makes decoder to determine where and when to use image region information. To enable the model use multi-level features, Marcella et al. [11] introduce a meshed-memory transformer. Their model learns a multi-level representation of the relationships between image regions,and uses a mesh-like connectivity at decoding stage to exploit low-level and high-level features. In order to solve the problem that the embedding of a word only uses itself, Yu et al. [12] propose a transformer model based on knowledge graph. This model can use not only the word itself, but also the information of the word's neighbors when decoding. The typical self-attention has difficulty in solving the problem of semantic gap between vision and language. Li et al. [13] introduce an entangled transformer structure. The entangled attention module enables the model to solve the problem of semantic gap.

Although the model structure based on self-attention networks has achieved state-of-the-art performance in image captioning, it still has two problems that need to be solved. First, the query vector in the self-attention mechanism is prone to cause the query vector to disappear in the stacked network. In the network structure, the query vector of the current layer is obtained from the output of the previous layer. It is difficult for the query vector of the next layer to use the query information of the current layer. Second, the image captioning models ignore the spatial information between objects in the image. The above two reasons result in generated sentences fail to accurately describe the image's content.

In this paper, we propose an improved Transformer with IoU Position encoding model, called TIP, based on the classical

transformer. Specifically, in order to solve the problem of the disappearance of the query vector in the stacked network, we designed an intra-modal self-attention mechanism. Meanwhile, in order to solve the problem of lack spatial information between objects when decoding, we propose a spatial position encoding method based on IoU. Since IoU captures the relationships between objects, the generated captions can clearly express the semantic information of the images.

Our contributions are summarized as follows. 1) We propose an intra-modal attention module that can alleviate the problem of vanishing query vectors when the network is propagating forward. 2) We propose a spatial position encoding method based on IoU which enriches image features and enhances the model's understanding of the spatial information between objects. 3) We conduct experiments to compare image captioning models, and verify the effectiveness of our proposed TIP model through quantitative analysis and qualitative analysis.

II. Related Works

In recent years, inspired by the encoder-decoder framework of machine translation task [14], a series of deep learning image captioning models have been proposed. The Neural Image Captioing (NIC) model [2] uses the convolutional neural network to extract the feature information of the image, and uses long-short term memory (LSTM) to translate the features into the corresponding sentence. However, NIC model uses image features only in the initial moment. To enhance the guidance of image features for generating sentences, Jia et al. [15] propose to input the image features at each moment of decoding.

In order to better utilize the features of the image, Karpathy et al. [16] propose to extract the features of different regions of the image using R-CNN network [17]. The model uses the feature information of different regions to generate the corresponding sentences. To enable the model to dynamically focus on different regions in the image at the moment of decoding, Xu et al. [5] propose the attention mechanism. The proposed attention module effectively improves the performance of the image captioning model. However,this attention mechanism only focus on different visual information and ignore model language information. Lu et al. [6] propose an adaptive attention mechanism. The adaptive attention mechanism can decide whether to focus more on visual or language information at the current moment. All above models use an encoder based on RNN structure. However, the inherent sequential structure makes the model has limited long-time memory capability.

In order to alleviate the long-time dependency problem, we adopt the transformer [18] structure as the decoder. Since the network is stacked by multiple layers, it is difficult to propagate the query information to the next layer during the network forward propagation. To make our TIP model better applicable to image captioning, we propose an intra-modal attention module. Moreover, the spatial information of the image should be used along with the feature information in the decoding process. Hu et al. [19] propose to use the coordinates and size of the objects as auxiliary information. However, they only utilize the coordinate information of the image and do not utilize the relationship between the objects in the image. In order to make our TIP model better utilize the spatial relationship of objects, we propose an IoU spatial position encoding module.

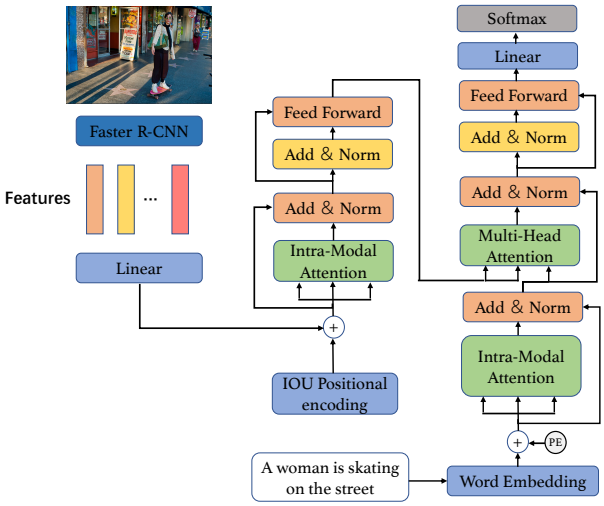

Fig. 1. The framework of our proposed TIP model. The TIP model is comprised of three primary components: image encoder, IoU position module, and caption decoder.

III. Method

In this section, we first describe the structure of our proposed TIP model. Then we introduce the intra-modal attention module. Finally, we introduce the spatial information encoding module.

A. Transformer model for image captioning

The overall model structure is shown in Fig. 1. We adopt the encoder-decoder framework. Here, Fast R-CNN is used as the encoder which encodes the image into visual features. The decoder is mainly used for the generation of word sequences. Based on the extracted visual features, the corresponding captions are generated. The decoder uses a self-attention network, which can solve the problem of long-term dependence due to the sequential structure. To be brief, the decoder consists of identical blocks. Each block includes multi-head self attention and feed-forward network and each block employs a residual connection [20], followed by layer normalization [21].

B. Intra-modal attention

We propose an intra-modal attention module to alleviate the problem of vanishing query vectors. It enables the model to better preserve image and text information. Intra-modal attention mechanism is shown in Fig. 2. We first calculate the similarity using the query vector and the key vector to get the weight values. The weight is transformed into a value between 0 and 1 by a softmax function. The calculation formula is as follows,

$$a_{i,j} = f_{sim}(q_i, k_j) \tag{1}$$

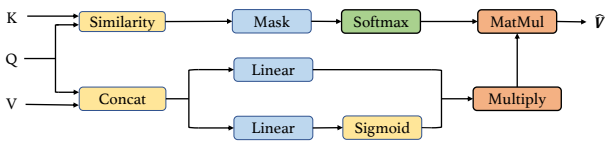

Fig. 2. The structure of intra-modal attention.

where q_i denotes the query vector of the ith word, and k_j denotes the key vector of jth word, and $a_{i,j}$ denotes the similarity result of the ith and jth words.

$$\alpha_{i,j} = \frac{e^{a_{i,j}}}{\sum_j e^{a_{i,j}}} \tag{2}$$

The query vector is concatenated with the value vector, and the V_c vector is obtained after the fully connected layer. The calculation formula is as follows,

$$V_c = W_q^c q + W_v^c v + b_c \tag{3}$$

where W_q^c and W_v^c are the parameters to be learned. The parameter b_c represents the bias. The vector g is obtained from the concatenation by a sigmoid activation function, i.e.,

$$g = \sigma(W_q^s q + W_v^s v + b_s) \tag{4}$$

The parameters W_q^s and W_v^s are learned and the symbol σ indicates the sigmoid activation function. The b_s parameter represents the bias. And the role of the gate is to selectively remember the information of the V_c vector as follows,

$$\hat{v} = g \odot V_c \tag{5}$$

where the symbol $\odot$ represents the element-wise product. The attended $\hat{V}$ is calculated as follows,

$$\hat{V} = \textstyle\sum_i \alpha_i \hat{v}. \tag{6}$$

Note that the intra-modal attention module adopts a different structure from the original self-attention. To preserve the information within the query vector, we use a concatenation method to fuse the query vector with the value vector. Meanwhile, we use the gate mechanism to remove useless information.

C. Position encoding

We enhance the visual information by adding spatial features of objects. The region proposal network extracts the information of each object. The coordination of the ith box can be represented by $\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2})\}$. During decoding, we hope that feature information can play a key role and spatial information can guide sentence generation. The fusion of spatial and visual features makes the decoder understand the image content better. However, the data distribution of the feature and the box coordinate information is not consistent.

In order to better fuse the features of the image and the spatial information of objects, we try several spatial location encoding methods. We compare the direct use of coordinate information with use IoU of box with other boxes. Finally, the following approach of visual and spatial feature fusion achieved the best results. We first compute the IoU between each object bounding box and other object bounding boxes in the image. Then we use a fully connected layer to map it to a dimension consistent with the object feature map. The output is summed with the visual features. We use the obtained result as the input of the decoder.

TABLE I
PERFORMANCE COMPARISON USING CROSS-ENTROPY.

Model	B-1	B-4	M	R	C	S
Review Net [29]	72.1	31.3	25.6	53.3	96.5	-
Adaptive [6]	74.2	33.2	26.6	-	108.5	-
PG-BCMR [30]	75.4	33.2	25.7	55.0	101.3	-
SCST [31]	-	30.0	25.9	53.4	99.4	-
LSTM-A [32]	75.4	35.2	26.9	55.8	108.8	20.0
Up-Down [4]	77.2	36.2	27.0	56.4	113.5	20.3
RF-Net [33]	76.4	35.8	27.4	56.8	112.5	20.5
TKG [12]	75.6	34.3	27.7	56.3	112.8	20.9
TIP	75.5	35.7	27.9	56.5	113.9	20.8
TIP (w/o intra)	75.7	34.9	27.7	56.1	112.3	20.9

IV. EXPERIMENTAL SETTINGS

A. Datasets and metrics

To verify the effectiveness of our proposed TIP model, we conduct experiments on the MS-COCO [22] dataset. The MS-COCO dataset contains 82,783 training images, 40,504 validation images and 40,775 test images. To better compare with other baseline methods, we use the same dataset division method as in [2]. The offline dataset contains 113,287 images with five annotated sentences for each image. Both the validation set and the testing set contain 5000 different images.

We use CIDEr (Consensud-based Image Description Evaluation) [23], BLEU (Bi-Lingual Evaluation Understudy) [24], METEOR (Metric for Evaluation of Translation withe Explicit Ordering) [25], ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [26] and SPICE (Semantic propositional image caption evaluation) [27] to evaluate our proposed method.

B. Implementation details

The number of regions of interest detected by the region detector is set to 36. We set the dimension of the region features to 2048. The dimension of the word embedding vector is set to 512. The number of encoder and decoder layers of the network structure is set to 6, and the number of multi-head attention mechanisms is set to 8. We set the maximum length of the sentence to 16. We set the batch size in the training process to 10. We update the parameters of the network using the Adam method [28]. Adam's momentum and weight decay are 0.8 and 0.999, respectively. The initial learning rate of model is set to 4×10^{-4}, with 2000 warm-up steps. We train 15 epochs using the cross-entropy loss.

V. RESULTS AND ANALYSIS

A. Result Comparison

Table I shows the metric scores of different models. TIP achieves the highest scores in the CIDEr and METEOR.

TABLE II
PERFORMANCE COMPARISON USING REINFORCEMENT LEARNING.

Model	B-1	B-4	M	R	C	S
LSTM-A [32]	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down [4]	79.8	36.3	27.7	56.9	120.1	21.4
RF-Net [33]	79.1	36.5	27.7	57.3	121.9	21.2
ICS [12]	80.2	38.0	28.6	58.4	128.6	22.1
TIP	78.4	37.9	28.0	58.3	123.1	21.8

CIDEr calculates weights according to the importance of the words and measures the similarity of the generated sentences to the annotated sentences. The CIDEr is also used specifically to evaluate image captioning. TIP model achieves a high score in CIDEr, indicating that the model has better performance. The METEOR evaluation method is highly correlated with manual evaluation. A high score in this metric indicates that the generated sentences are more readable. The sentences generated by TIP model are not the highest in BLEU. BLEU calculates the similarity of annotated sentences to generated sentences. It does not consider the importance of different words. Therefore, the low score of this metric indicates that the generated sentences do not match well with the annotated sentences. It does not indicate that the sentences are not readable.

The model without intra-modal attention module decreased by 1.6 on CIDEr, 0.8 on BLEU-4, and 0.6 on ROUGE. Intra-modal attention module is designed to alleviate the problem of disappearing query information. The improvement on metrics verify the effectiveness of intra-modal attention module.

Reinforcement learning solves the problem of image captioning exposure bias. As shown in Table II, the TIP model metric score improve by 9.2 on CIDEr over the model without reinforcement learning. The model we proposed score higher on BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr and SPICE than the LSTM-A, Up-Down, and RF-Net structures.

B. On Spatial Encoding

Table III shows the results of different position encoding methods. Coord model indicates that we use a fully connected layer to map coordinates to certain dimensions. The output is added to the feature information. As we can see from the Table III that the Coord model decreases by 0.2 on BLEU-2, BLEU-3, BLEU-4 and SPICE compared with baseline model. The model decreased by 2.5 on CIDEr. We infer that the inconsistent distribution of coordinate information and feature map leads to a decrease in the metrics.

Coord(hw) method adds the length and width of the object box to the Coord method. In table III, Coord(hw) improves by 0.6 on CIDEr and 0.2 on SPICE, and decreases on BLEU instead. This indicates that the length and width information of the box does not improve the model performance. Coord-Norm(hw) model adds normalization to the box coordinates based on Coord(hw). In Table III that the results improve by 0.4 on BLEU-1, 0.2 on BLEU-2, BLEU-3, BLEU-4 and ROUGE, and 0.9 on CIDEr compared with the Coord(hw) model. Normalization solves the problem of inconsistent distribution of coordinates and image features to some extent. Therefore, CoordNorm(hw) model improves the performance.

TABLE III
PERFORMANCE COMPARISON ON SPATIAL ENCODING METHODS.

Model	B-1	B-2	B-3	B-4	R	S	C
Baseline	75.1	58.8	45.0	34.3	55.8	20.4	110.7
Coord	75.1	58.6	44.8	34.1	55.7	20.2	108.2
Coord(hw)	74.8	58.6	44.7	34.0	55.5	20.4	108.8
CoordNorm(hw)	75.2	58.8	44.9	34.2	55.7	20.4	109.7
IoUc	75.2	58.8	44.9	34.2	55.6	20.5	110.0
IoU+	75.5	59.3	45.6	34.9	55.8	20.7	110.6

TABLE IV
PERFORMANCE COMPARISON USING DIFFERENT BEAM SIZE.

Beam	B-1	B-2	B-3	B-4	R	S	C
1	74.9	58.5	44.1	32.8	55.3	20.3	108.8
2	75.8	59.7	45.8	34.9	55.9	20.7	112.1
3	75.8	59.8	46.0	35.3	55.9	20.7	112.0
4	75.6	58.5	45.8	35.3	55.9	20.6	110.8

The IoUc model first calculates the IoU information between the current object and other objects in the image. Then the IoU information is fused with the feature information by concatenating. Finally we use a fully connected layer to make the concatenated dimension reduced. In table III, the IoUc model improves 0.1 on BLEU-1, BLEU-3, and BLEU-4, 0.2 on BLEU-2, 0.3 on SPICE, and 1.8 on CIDEr compared to the Coord model. The performance of the IoU model achieves an improvement compared to the Coord model. We infer that the performance improvement of the model is due to two factors. The first is that the IoU captures the relationship between objects in the image. The second is that the spatial information distribution of the IoUc model is closer to the image feature information. It enables the image feature information to be better fused with spatial information.

IoU+ model first calculates IoU between the current object and other objects in the image. Then the model use a fully connected layer to keep the IoU position encoding dimension consistent with the image feature dimension. Finally, the output of the fully connected layer is summed with image feature values. In table III, IoU+ model improves 0.3 on BLEU-1, 0.5 on BLEU-2, 0.7 on BLEU-3, BLEU-4, 0.2 on ROUGE, SPICE, and 0.6 on CIDEr compared to the IoUc model. The IoU+ model uses the spatial information of IoU compared with the IoUc model. But the only difference is the way of using features. IoU+ model uses the IoU features mapped by the fully connected layer and adds them to image features. This position encoding is more similar to the text encoding of the transformer. Thus, the position encoding method of IoU+ has a higher score.

C. On Beam Size

Table IV shows the results of using different sizes of beam search. When the size of beam search is 1, the search space for word sampling is too small. A large amount of decoding

Standard: **A man** riding a snowboard down a snow covered slope.
Ours: **Two people** riding snowboards on snowy slope.
GT1. Two men riding snowboards in snow down a slope.
GT2. Two people are snowboarding down a hill fast.
GT3. two men are riding snowboards in a snow slope.

Standard: **a couple of boats** that are sitting in the water .
Ours: **a boat** sitting on top of a lake next to a shore .
GT1. A boat full of Asian people is sailed through reeds and bushes .
GT2. A large boat filled with people navigate through a still lake .
GT3: Around 15 people are on a boat on a river going somewhere .

Standard: a black cow standing in a field of tall grass.
P: A cow standing in a grassy field next to **a house**.
GT1. Two cows grazing on grass in a field by a house.
GT2. A white faced cow stands in the tall grass.
GT3. a cow stands in the grassy area of a yard.

B: A man riding a skateboard down a sidewalk.
P: A man riding a skateboard next to **another man**.
GT1. Two men on skateboards on the pavement
GT2.Two young skateboarders are skating near each other.
GT3. There are two men riding on their skateboards.

Fig. 3. Qualitative results of our TIP model.

information is lost and many better sentences are missed. When the size of beam search is increased to 2 and 3, the search space of words increases. The metric scores the highest and the model is more likely to reach the optimal performance. And when the size of beam search increases to 4, the model tends to generate shorter sentences. The high similarity between the generated words and the lack of diversity lead to some slight decrease in the metrics. Also, as the beam increases, the memory usage of the algorithm increases and the speed of generated sentences is slower.

D. On Number of Network Layers

We conduct an experiment on the number of network layers in the transformer. In Table V, EN means the number of encoder layers and DN means the number of decoder layers. We try four combinations of the number of encoders and decoders. We can see that as the network becomes deeper, the metric score for the image captioning is higher. It is also found that the increase in the number of encoder layers does not have a particularly large improvement in the performance of the image captioning. The increase in the number of layers in the decoder part results in larger improvement in the metrics of BLEU-1, BLEU-2, BLEU-3, BLEU-4, and CIDEr. Therefore, under the condition that the number of model parameters is guaranteed, increasing decoder layer numbers can effectively improve the performance of the model.

E. Qualitative Results and Visualization

We randomly select some images on the dataset for qualitative analysis. The results are shown in Fig. 3. In the first image, the standard model generates the result of a person skiing on the snow. Our proposed model can clearly identify there are two persons in the image. The reason is that the feature maps of the two persons extracted by the neural network are relatively similar. Thus, the standard model considers the similar feature maps as one object and easily forgets the other

TABLE V
PERFORMANCE COMPARISON USING DIFFERENT NUMBER OF LAYERS.

EN	DN	B-1	B-2	B-3	B-4	R	S	C
6	6	76.0	56.0	46.1	35.2	56.0	21.0	112.9
8	6	75.8	59.4	45.6	34.7	55.9	20.8	112.8
6	8	76.1	59.7	45.9	34.9	56.2	21.1	112.9
8	8	76.1	59.7	45.8	34.9	56.0	21.2	113.2

object. Our IoU position encoding method contains the spatial information of the object. Since the IoU value of two persons in the image is 0, it can be clearly identified as two persons. Therefore, the generated sentences are closer to the annotated sentences.

Similarly, it can be seen from the third image, which expresses that a black cow is standing on the grass in front of the house. The standard model does not recognize the house. Our model identifies that two objects do not intersect according to the IoU information of the house and the cow. Therefore, our model can identify both the cow and the house. It can be seen that the sentences generated by our TIP model have richer semantic information. From the above qualitative analysis, we can see that the model with IoU position encoding can identify the relationship between objects and generated sentences are closer to the annotated sentences.

VI. CONCLUSION

In this paper, we propose an improved transformer with IoU position encoding model, i.e., TIP. By introducing an intra-modal attention module, the TIP model saves useful query information to the higher layer. Moreover, We propose an IoU position encoding method that fuses visual features and spatial features. Finally, we conduct extensive experiments with qualitative and quantitative analysis to verify the effectiveness of the intra-modal attention module and IoU position encoding method.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303300 and Subject II under Grant 2019YFF0303302.

REFERENCES

[1] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *ICML*. JMLR, 2014, pp. 595–603.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[3] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014, p. 3104–3112.

[4] 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 *CVPR*, 2018, pp. 6077–6086.

[5] 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 *ICML*, 2015, pp. 2048–2057.

[6] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017, pp. 375–383.

[7] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016, pp. 4651–4659.

[8] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NIPS*, vol. 32, 2019.

[9] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *CVPR*, 2020, pp. 9962–9971.

[10] W. Zhang, W. Nie, X. Li, and Y. Yu, "Image caption generation with adaptive transformer," in *2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 2019, pp. 521–526.

[11] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *CVPR*, 2020, pp. 10 578–10 587.

[12] Y. Zhang, X. Shi, S. Mi, and X. Yang, "Image captioning with transformer and knowledge graph," *Pattern Recognition Letters*, vol. 143, pp. 43–49, 2021.

[13] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *ICCV*, October 2019.

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[15] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *ICCV*, 2015.

[16] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, vol. 30, 2017.

[19] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *CVPR*, 2018, pp. 3588–3597.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[21] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," in *NIPS deep learning symposium*, 2016.

[22] 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 *ECCV*. Springer, 2014, pp. 740–755.

[23] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[25] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[26] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[27] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ICCV*, 2016, pp. 382–398.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *NIPS*, vol. 29, 2016.

[30] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *ICCV*, 2017, pp. 873–881.

[31] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017, pp. 7008–7024.

[32] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017, pp. 4894–4902.

[33] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *ICCV*, 2018, pp. 499–515.