Multi-Branch Distance-Sensitive Self-Attention Network for Image Captioning

Jiayi Ji†, Xiaoyang Huang†, Xiaoshuai Sun*,Yiyi Zhou, Gen Luo, Liujuan Cao, Jianzhuang Liu, *Senior Member, IEEE*, Ling Shao, *Fellow, IEEE*, Rongrong Ji, *Senior Member, IEEE*

***Abstract*—Self-attention (SA) based networks have achieved great success in image captioning, constantly dominating the leaderboards of online benchmarks. However, existing SA networks still suffer from distance insensitivity and low-rank bottleneck. In this paper, we aim to optimize SA in terms of two aspects, thereby addressing the above issues. First, we introduce a Distance-sensitive Self-Attention (DSA), which considers the raw geometric distances between query-key pairs in the 2D images during SA modeling. Second, we present a simple yet effective approach, named Multi-branch Self-Attention (MSA) to compensate for the low-rank bottleneck. MSA treats a multi-head self-attention layer as a branch and duplicates it multiple times to increase the expressive power of SA. To validate the effectiveness of the two designs, we apply them to the standard self-attention network, and conduct extensive experiments on the highly competitive MS-COCO dataset. We achieve new state-of-the-art performance on both the local and online test sets, *i.e.*, 135.1% CIDEr on the Karpathy split and 135.4% CIDEr on the official online split. The source codes and trained models for all our experiments are publicly available at https://github.com/Young499/image-captioning-MDSANet.**

***Index Terms*—Image Captioning, Multi-Branch Techniques, Distance-Sensitive Positional Embedding**

I. Introduction

IMAGE captioning aims to automatically translate an image into a natural language description [1]–[4], which is of great importance in enabling computers to understand images. In addition to recognizing the objects in an image, an image captioning model should also be able to analyze their state and relationships, and then translate this information into a natural language sentence [5]–[7]. Most successful image captioning approaches adopt the encoder-decoder framework [8]–[11], which is inspired by the sequence-to-sequence model for machine translation [12], [13]. The encoder transforms an image into intermediate representations, and is immediately followed by a decoder that generates a descriptive sentence. Recently, the Transformer [14] and its variants have gained widespread popularity in image captioning, and dominated the leaderboards as the de facto standard [10], [11]. Despite the great success of these self-attention (SA) networks, there are two key issues left unexplored.

Firstly, existing SA networks still fail to model the geometric distances between visual instances, which is critical for visual content understanding. In a standard SA module, information on the spatial orders and distances of objects is typically ignored. Some recent works have tried to address the loss of spatial information by introducing position embedding methods, *i.e.*, absolute positional encoding [15] and relative positional encoding [16], [17]. However, these embeddings are usually unable to maintain precise information of geometric object distances in the image, which are insufficient for modeling the relations of objects and reasoning complex scenes.

Secondly, existing SA networks also suffer from a low-rank bottleneck [18]. Specifically, to extend the capacity of exploring subspaces, the self-attention has multiple heads, *i.e.*, it is a multi-head self-attention (MHA). To ensure that the number of parameters in the attention layer stays unchanged irrespective of the amount of heads, dimensionality reduction is used in the input projections. Thus, increasing the number of heads reduces the head size. This in turn limits the expressivity of individual heads, leading an information bottleneck [18], [19]. An intuitive solution to this problem is to increase the embedding size of multiple heads. However, this will greatly increase the number of parameters and make it difficult to optimize the deep network.

To address the above two issues, we introduce two effective designs, namely *Distance-sensitive Self-Attention* (DSA) and *Multi-branch Self-Attention* (MSA). For the issue of geometric distance modeling, DSA uses the distance information to guide the self-attention modeling. Specifically, in DSA, the spatial distance of each query-key pair is transformed into a scalar by a dynamic sigmoid function, whose value can directly reflect the geometric distance between the key and the query. Notably, this sigmoid function can dynamically determine the proper range of the scalar. Then, the scalar is used to adjust the corresponding raw attention weights. To tackle the information bottleneck, MSA is employed to duplicate the MHA several times, averaging the outputs. With such a simple operation, the shortcomings of self-attention on small subspaces can be compensated for by iterative modeling, thereby improving the

*Corresponding author. †Equal contribution.

J. Ji, R. Ji, X. Sun (Corresponding Author), Y. Zhou, L. Cao, X. Huang and G. Luo are with the Media Analytics and Computing Laboratory, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: xssun@xmu.edu.cn).

R. Ji and X. Sun are also with Institute of Artificial Intelligence, Xiamen University and Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Xiamen University, 361005, China.

J. Liu is with e Noah's Ark Lab, Huawei Technologies Co. Ltd., Shenzhen 518129, China.

L. Shao is with Terminus Group, China.

R. Ji is also with Peng Cheng Lab, Shenzhen, China.

Manuscript received September 17, 2021; revised December 25, 2021 and March 18, 2022; accepted April 05, 2022. This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, and No. 62002305), Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049), and the Natural Science Foundation of Fujian Province of China (No.2021J01002).

expressive power of SA networks.

These two designs are further combined as an enhanced SA module. We build a new image captioning model called the Multi-branch Distance-sensitive Self-Attention Network (MD-SAN) by integrating the new SA module into a classical Transformer structure [14]. Extensive experiments on MS-COCO are conducted to validate the effectiveness of MD-SAN, as well as the two newly proposed designs. Notably, our MD-SAN establishes a new state-of-the-art on the MS-COCO evaluation sever, improving the best result in terms of CIDEr from 134.0% to 135.4% on the official online test split. To examine the generalization ability of DSA and MSA, we also apply them to the Transformer networks of two other multimodal tasks, *i.e.*, Visual Question Answering (VQA) [20] and Visual Grounding (VG) [21], and conduct experiments on the VQA-v2 [22], RefCOCO, RefCOCO+, and RefCOCOg [23] datasets. When integrated into the strong Transformer-based baselines, our method can consistently increase their accuracies on various tasks with negligible extra computational cost.

To summarize, the contributions of this paper are three-fold:

- We propose a distance-sensitive self-attention approach which explicitly models the real distances between objects in an image to improve scene understanding.
- We introduce a multi-branch self-attention to break the low-rank bottleneck and increase the expressive power of the multi-head SA.
- By combining the DSA and MSA and applying them to the self-attention network, we establish a new state-of-the-art on the MS-COCO image captioning benchmark. Further experiments on visual question answering and visual grounding tasks verify the generalization of our method.

II. Related Work

A. Image Captioning

Inspired by the encoder-decoder framework in machine translation [12], [13], most existing image captioning approaches adopt the CNN-RNN architectures [1], [2]. Recently, a variety of improved models [24]–[30] have been proposed using attention [31] and Reinforcement Learning (RL) based training objectives [32]. Xu *et al.* [31] introduced soft and hard attention mechanisms to automatically focus on salient objects when generating each word by mimicking the human visual system. Lu *et al.* [33] proposed an adaptive attention mechanism with a visual sentinel determining whether to attend to the image. Anderson *et al.* [8] introduced an object detector to identify salient image regions (objects) and extract a feature vector for each object, which is then fed into the decoder for caption generation. Rennie *et al.* [32] explored reinforcement learning with a self-critical reward for model training. At the same time, some GNN-based methods [34] have been introduced to the image captioning task. Yao *et al.* [24] used a graph convolutional neural network to integrate semantic and spatial relationships between objects, aiming to further improve the encoding of objects and their relationships. Yang *et al.* [25] utilized graph convolution to incorporate scene graph into the encoder-decoder image captioning framework. Despite their wide adoption, RNN-based models suffer from limited representation power and a sequential nature.

B. Transformers in Image Captioning

Recently, Vaswani *et al.* [14] showed that solely using the Transformer model can achieve state-of-the-art results for machine translation. Other recent approaches have explored the use of Transformers in vision-language tasks. For instance, Huang *et al.* [9] introduced a Transformer-like encoder to encode regions into hidden states, which was paired with an LSTM decoder. Futher, Herdade *et al.* [10], [11], [16], [17], [26] proposed to replace conventional RNNs with the Transformer architecture, achieving new state-of-the-art performances. Along the same line, Li *et al.* [35], [36] used a Transformer to integrate both visual information and additional semantic concepts given by an external tagger. Herdade *et al.* [16], [17] incorporated geometry relationships between region features into the Transformer architecture by leveraging relative positional encoding for captioning. Luo *et al.* [15] adopted both relative positional encoding and absolute positional encoding to enhance the visual feature representations. However, none of these models explicitly model real distances, which is important for understanding the relations between objects and reasoning visual scenes.

C. Position Representation

There are mainly two classes of methods for position representation: absolute position encoding and relative position encoding. The absolute ones [14], [15] encode the absolute position of the input tokens from 1 to the maximum sequence length and each position has a corresponding embedding. The relative ones [15]–[17] transform the relative position between input tokens into the high-dimensional vectors and learns the pairwise relationship between tokens. Both methods map the absolute position or relative position to the high-dimensional space through the learnable embedding vectors. The process is black-boxed, and the distance information is not interpretable. Thus, both of them may not be optimal for modeling distance information because they usually cannot keep the precise information of token distances. A core innovation of our paper is that we directly use the distance information as a priori information and map it into a real number via a monotonic function to revise the attention weights. Then the model independently chooses whether to focus on short-range features or long-range features.

D. Multi-Branch Techniques

Each block of the multi-branch architecture of a neural network consists of more than one parallel component. Typical structures in computer vision include the inception architectures [37], ResNet [38], ResNeXt [39], and DenseNet [40]. In natural language process, bi-directional LSTM [41] models can be regarded as a two-branch architecture, also benefitting from the multi-branch technique.

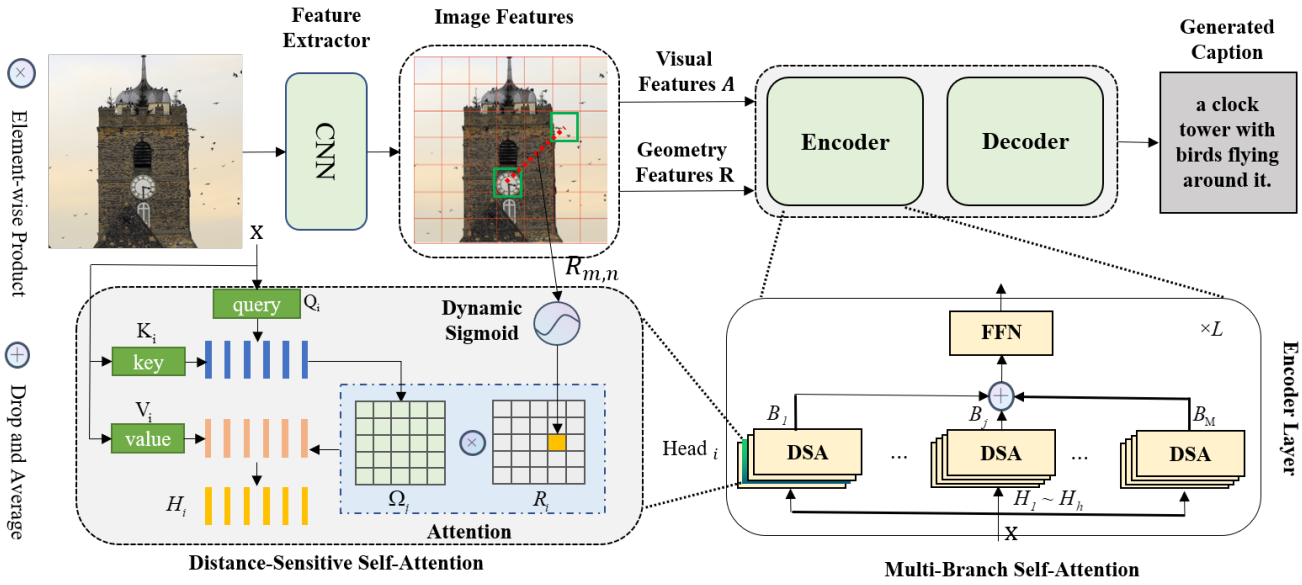


Fig. 1. Overview of our proposed Multi-branch Distance-sensitive Self-Attention Network (MD-SAN) for image captioning. A set of visual features A and geometry features R are first fed into the encoder, which are then adaptively fused into the decoder to generate a caption. For the first layer, the input $X = A$. Note that the residual connections, layer normalizations, and embedding layers are omitted for simplicity.

III. Preliminaries

The Transformer model consists of an encoder and a decoder, both of which have a stack of layers. We first present a core component of the Transformer, called the multi-head self-attention, which has h attention heads with independent parameters. For the i-th attention head, the independent head projection matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times \frac{d}{h}}$ are used to transform the input $X \in \mathbb{R}^{N \times d}$ into queries Q_i, keys K_i and values V_i, which are defined as:

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V. \tag{1}$$

The attention weight matrix Ω_i between any queries and keys for the i-th head are computed as

$$E_i = \frac{Q_i K_i^T}{\sqrt{d}}, \Omega_i = \text{softmax}\left(E_i\right), \tag{2}$$

where $\sqrt{d}$ is a scaling factor. The output of the head is then calculated as

$$H_i = \text{ Attention } \left(Q_i, K_i, V_i\right) = \Omega_i V_i. \tag{3}$$

Next the outputs of the h attention heads are concatenated together and the final output is a linear projection of the concatenated representations, which is represented as:

$$O' = MultiHead(Q, K, V) = Concat\left(H_1, \ldots, H_h\right) W^O. \tag{4}$$

This is then followed by a residual connection and a layer normalization:

$$O = LayerNorm(X + O'). \tag{5}$$

Note that the residual connection helps avoid the vanishing gradient problem in the training phase. Then, a final feed-forward network (FFN) is adopted for additional processing of the outputs, which is also followed by another residual connection and layer normalization [14].

IV. Our Method

Given an image I, the captioning model needs to generate a word sequence $Y_T = \{y_0, \cdots, y_T\}$, $y_t \in D$, where D is the vocabulary dictionary and T is the sequence length. The image is represented as a group of visual features $A = \{a_1, a_2, \cdots, a_N\}$ extracted from a pre-trained object detector, where N is the number of visual features, and $a_i \in \mathbb{R}^d$. As shown in Fig. 1, we adopt a variant of the Transformer architecture for image captioning. In particular, we propose a novel MD-SAN based encoder to extract accurate internal relationships between features and increase the expressiveness. The decoder then uses the generated features from the last encoder layer as input to generate the caption, which is identical to the original Transformer implementation [14].

A. Distance-Sensitive Self-Attention (DSA)

In this section, we first introduce our proposed distance-sensitive self-attention, which can effectively leverage the real distance information between objects in an image to enhance the internal relation modeling.

Herdade *et al.* [16], [17] proposed to embed the relative positions to high-dimensional representations, which are then projected to scalar scores to directly modify the attention weight matrix Ω_i in Eq. 2. Such a process is an uncontrollable blind box, and thus cannot capture the precise distances. We incorporate the real distance information by directly modifying the attention weight matrix Ω_i in Eq. 2. Thus, we map the real distances into the re-scaled coefficients via a monotonic function that is suitable for adjusting the self-attention weights. The coefficients can reflect the real distances. For the input $X \in \mathbb{R}^{N \times d}$, we denote the geometry features of the m-th and n-th vectors as (x_m, y_m) and (x_n, y_n), respectively. Note that the geometry features are center coordinates for region features

[8]; for grid features [42], they are the 2D indexes. We use the Manhattan distance to compute the relative distance $R_{m,n}$ between the m-th and n-th vectors, i.e.,

$$R_{m,n} = |x_n - x_m| + |y_n - y_m|. \tag{6}$$

With Eq. 6, we obtain the relative distance matrices $R \in \mathbb{R}^{N\times N}$. Then we exploit a monotonic function to project them into the re-scaled coefficients, the scale of which needs to be tunable. Thus, we adopt the dynamic sigmoid function [43] as follows:

$$R_i = \sigma_d(R) = \frac{1+\exp(v_i)}{1+\exp(v_i - w_i R)}, \tag{7}$$

where w_i and v_i are two learnable parameters. The range of the dynamic sigmoid depends on v_i. Finally, R_i is used to adjust the raw attention weights as follows:

$$E_i = ReLU(\frac{Q_i K_i^T}{\sqrt{d}}) * R_i, \Omega_i = \text{softmax}\left(E_i\right), \tag{8}$$

where $*$ represents an element-wise product, and the subscript i denotes the i-th head. Besides, we notice that the raw attention weight matrix has both positive and negative values, which may make the influence of distance oscillate, resulting in invalidity. To continuously reflect the influence of distance information, following Wang *et al.* [43], we add the ReLU activation function to the attention weights to keep them non-negative.

B. Multi-Branch Self-Attention (MSA)

As demonstrated by earlier works [18], [19], multi-head self-attention suffers from a low-rank bottleneck, resulting from the decrease in head size. A natural way to address this is to increase the embedding size of the different heads. However, this will greatly increase the number of parameters and make it difficult to optimize the deep network.

In this paper, we present a simple yet effective approach, *i.e.*, the multi-branch self-attention, to break the bottleneck. The multi-branch technique is one of the keys to the success of deep neural models and has been well studied in computer vision [38]–[40] and natural language processing [41]. Following this premise, we treat a multi-head attention layer as a branch and duplicate it multiple times [44], as follows:

$$MSA(Q,K,V) = \frac{1}{M}\sum_{j=1}^{M} MultiHead^{(j)}(Q^{(j)}, K^{(j)}, V^{(j)}), \tag{9}$$

where M is the number of branches, and the superscript j denotes the j-th branch. Motivated by [44], [45], we leverage the drop branch technique during training to avoid co-adaptation among different branches. Thus, each branch has a certain probability of being randomly erased. Eq. 9 is redefined as follows:

$$\beta^{(j)} = \frac{\mathbb{I}\{U_j \geq \rho\}}{1-\rho}, \tag{10}$$

$$MSA(Q,K,V) = \frac{1}{M}\sum_{j=1}^{M} \beta^{(j)} MultiHead^{(j)}(Q^{(j)}, K^{(j)}, V^{(j)}), \tag{11}$$

where $\mathbb{I}$ is the indicator function, U_j is a uniformly sampled value from $[0,1]$ and the $\rho \in [0,1]$ is drop rate. To make the expectation of $\beta^{(j)}$ equal to 1, we use $1-\rho$ to revise the sampled results. Note that $\beta^{(j)} = 1$ during the inference phase.

The effect of MSA. We theoretically analyze how MSA solves the low-rank bottleneck. We make the following derivation from Eq. 9:

$$\begin{aligned} MSA(Q,K,V) &= \frac{1}{M}\sum_{j=1}^{M} MultiHead^{(j)}(Q^{(j)}, K^{(j)}, V^{(j)}) \\ &= \frac{1}{M}\sum_{j=1}^{M} Concat(H_1^{(j)}, \cdots, H_h^{(j)}) W^{O^{(j)}} \\ &= \frac{1}{M}\sum_{j=1}^{M} [H_1^{(j)}; \cdots; H_h^{(j)}][(W_1^{O^{(j)}})^T; \cdots; (W_h^{O^{(j)}})^T]^T \\ &= \frac{1}{M}\sum_{j=1}^{M} (H_1^{(j)} W_1^{O^{(j)}} + \cdots + H_h^{(j)} W_h^{O^{(j)}}) \\ &= \frac{1}{M}\sum_{j=1}^{M} H_1^{(j)} W_1^{O^{(j)}} + \cdots + \frac{1}{M}\sum_{j=1}^{M} H_h^{(j)} W_h^{O^{(j)}}, \end{aligned} \tag{12}$$

where the second equation is the multi-branch version of Eq. 4. The third equation is the block representation of matrices, where $H_i^{(j)} \in \mathbb{R}^{N\times \frac{d}{h}}$, $W_i^{O^{(j)}} \in \mathbb{R}^{\frac{d}{h}\times d}$. Finally, each term in the last equation is the sum of the projections of a head. These summation terms can be seen as simple ensembles of relatively shallow layers [46], which can greatly improve the representation ability of each head, thereby effectively solving the low-rank bottleneck. Thus the essence of the multi-branch mechanism is to strengthen the representation of each head.

C. Applying DSA and MSA to MHA

We combine both DSA and MSA to replace the standard multi-head self-attention. In the encoder, we first use DSA to replace the original SA, and then we replace the single-branch MHA with our proposed multi-branch distance-sensitive self-attention. Note that the decoder is identical to the original Transformer implementation [14].

D. Training

For a given caption $Y_T = \{y_0, \cdots, y_T\}$, the distribution is calculated as the product of the conditional distributions at all time steps:

$$p_l(Y) = \prod_{t=0}^{T} p(y_t | Y_{t-1}). \tag{13}$$

The training process consists of two phases, pre-training by supervised learning and fine-tuning by reinforcement learning. Let θ be the parameters of the model. During pre-training, given a target ground truth sequence $Y^* = \{y_0^*, \cdots, y_T^*\}$, the objective is to minimize the cross-entropy (CE) loss:

$$L_{CE}(\theta) = -\sum_{t=0}^{T} log\big(p(y_t^* | Y_{t-1}^*)\big). \tag{14}$$

In the fine-tuning stage, we employ a variant of the self-critical sequence training approach [32] on sequences sampled using beam search to directly optimize the metric, following previous works [8], [32]. The objective is to minimize the negative expected relative score:

$$L_R(\theta) = -E_{y\sim p_l}\big[r(Y^s)\big], \tag{15}$$

where $Y^s = \{y_0^s, \cdots, y_T^s\}$ is a sequence sampled through the beam search, and $r(\cdot)$ can be any reward metric. We use the CIDEr-D score as the reward. The final gradient of $L_R(\theta)$ for one sample is calculated as:

$$\nabla_\theta L_R(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\left(\left(r\left(Y^i\right) - b\right)\nabla_\theta \log p\left(Y^i\right)\right), \quad (16)$$

where $Y^i = \{y_0^i, \cdots, y_T^i\}$ is the i-th sentence in the beam, and $b = \left(\sum_i r\left(Y^i\right)\right)/k$ is the baseline, computed as the mean of the rewards obtained by the k sampled sequences.

V. Experiments

A. Dataset and Implementation Details

All experiments are conducted on the most popular benchmark for image captioning, MS COCO [47]. The official dataset of the MSCOCO contains 123,287 images, which includes 82,783 training images, 40,504 validation images, and 40,775 testing images. We use both offline and online evaluation to verify our proposed model. Each image is equipped with five ground-truth sentences. The online evaluation is done on the MS COCO test split, for which ground-truth annotations are not publicly available. In offline testing, we use the Karpathy splits [2] that have been used extensively for reporting results in previous works. This split contains 113,287 training images, and 5K images respectively for validation and testing.

To better verify the validity of MD-SAN, two kinds of features are used to conduct extensive experiments: (1) **Grid features**. To represent them, we follow [15], [42] and use Faster R-CNN [48] with ResNeXt-101 [39] to extract grid features. (2) **Region features**. We use the region features provided by Bottom-Up [8] for training.

In our model, we set the dimensionality d of each layer to 512, the number of heads to 8 in each branch, and the number of branchs M to 3, respectively. We employ dropout with a keep probability of 0.9 after each attention and feed-forward layer. Pre-training with cross entropy is done following the learning rate scheduling strategy with a warmup of 10, 000 iterations. Then, during CIDEr-D optimization, we use a fixed learning rate of 5×10^{-6}. We train all models using the Adam optimizer [49], a batch size of 50, and a beam size of 5. During the inference stage, we adopt the beam search strategy and set the beam size to 3.

B. Metrics

Five evaluation metrics, BLEU [50], METEOR [51], ROUGE-L [52], CIDEr [53], and SPICE [54], are simultaneously utilized to evaluate our model. The most classic metrics for image captioning are based on n-gram similarity of reference sentences and generated descriptions, *i.e.*, BLEU [50], METEOR [51], ROUGE-L [52], and CIDEr [53]. BLEU is a modified form of precision to compare the generated captions against multiple reference sentences, the number of which is between 0 and 1. METEOR is based on the harmonic mean of unigram precision and recall of exact, stem, synonym, and paraphrase matches between sentences. ROUGE-L is a set of automated evaluation criteria designed to evaluate text summarization algorithms. CIDEr represents the sentences in the form of a Term Frequency Inverse Document Frequency (TF-IDF) vector, and then calculates the cosine similarity of the reference description to the description generated by the model. Besides, SPICE [54] is a semantic evaluation metric for image caption that measures how image captions effectively recover objects, attributes, and relationships between them, which is better able to capture human judgments about the model's subtitles.

TABLE I
COMPARISON WITH THE STATE OF THE ARTS ON THE "KARPATHY" TEST SPLIT, IN THE SINGLE-MODEL SETTING. B-N, M, R, C AND S ARE SHORT FOR BLEU-N, METEOR, ROUGE-L, CIDER AND SPICE SCORES, RESPECTIVELY, ALL VALUES OF WHICH ARE REPORTED AS PERCENTAGES (%). PARAMS REPRESENTS THE PARAMETER SIZE OF THE MODEL.

Model	B-1	B-4	M	R	C	S	Params
	Region features (ResNet-101)						
SCST [32]	-	34.2	26.7	55.7	114.0	-	-
Up-Down [8]	79.8	36.3	27.7	56.9	120.1	21.4	-
RFNet [55]	79.1	36.5	27.7	57.3	121.9	21.2	-
GCN-LSTM [24]	80.5	38.2	28.5	58.3	127.6	22.0	-
SGAE [25]	80.8	38.4	28.4	58.6	127.8	22.1	-
ETA [35]	**81.5**	39.3	28.8	58.9	126.6	22.7	-
AoANet [9]	80.2	38.9	29.2	58.8	129.8	22.4	-
ORT [16]	80.5	38.6	28.7	58.4	128.3	22.6	-
Transformer [14]	80.4	38.3	29.0	58.2	129.5	22.5	33.6M
MMTransformer [11]	80.8	39.1	29.2	58.6	131.2	22.6	38.3M
XTransformer [10]	80.9	**39.7**	**29.5**	59.1	132.8	**23.4**	138.3M
Ours	81.1	39.5	29.4	**59.2**	**133.5**	23.1	39.9M
	Grid features (ResNext-101)						
Up-Down [8]	75.0	37.3	28.1	57.9	123.8	21.6	-
AoANet [9]	80.8	39.1	29.1	**59.1**	130.3	22.7	-
Transformer [14]	80.9	38.9	29.0	58.5	131.2	22.7	33.6M
MMTransformer [11]	80.8	38.9	29.1	58.5	131.8	22.7	38.3M
XTransformer [10]	81.0	39.7	29.4	58.9	132.5	23.1	138.3M
RSTNet [56]	81.1	39.3	29.4	58.8	133.3	23.0	-
Ours	**81.5**	**39.8**	**29.6**	**59.1**	**135.1**	**23.2**	39.9M

TABLE II
COMPARISON WITH THE STATE OF THE ARTS ON THE "KARPATHY" TEST SPLIT, USING THE ENSEMBLE TECHNIQUE, WHERE B-N, M, R, C AND S ARE SHORT FOR BLEU-N, METEOR, ROUGE-L, CIDER AND SPICE SCORES, RESPECTIVELY. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

Model	B-1	B-4	M	R	C	S
	Ensemble/Fusion of 2 models					
GCN-LSTM [24]	80.9	38.3	28.6	58.5	128.7	22.1
SGAE [25]	81.0	39.0	28.4	58.9	129.1	22.2
ETA [35]	81.5	39.9	28.9	59.0	127.6	22.6
GCN-LSTM+HIP [57]	-	39.1	28.9	59.2	130.6	22.3
MMTransformer [11]	81.6	39.8	29.5	59.2	133.2	23.1
Ours	**82.2**	**40.5**	**29.7**	**59.6**	**136.9**	**23.3**
	Ensemble/Fusion of 4 models					
SCST [32]	-	35.4	27.1	56.6	117.5	-
RFNet [55]	80.4	37.9	28.3	58.3	125.7	21.7
AoANet [9]	81.6	40.2	29.3	59.4	132.0	22.8
MMTransformer [11]	82.0	40.5	29.7	59.5	134.5	23.5
Ours	**82.3**	**40.9**	**29.8**	**59.7**	**137.8**	**23.7**

C. Baselines

The baseline models we compare include SCST [32], LSTM-A [58], Up-Down [8], RFNet [55], GCN-LSTM [24], SGAE [25], ETA [35], AoANet [9], ORT [16], MMTransformer [11], XTransformer [10], and RSTNet [56].

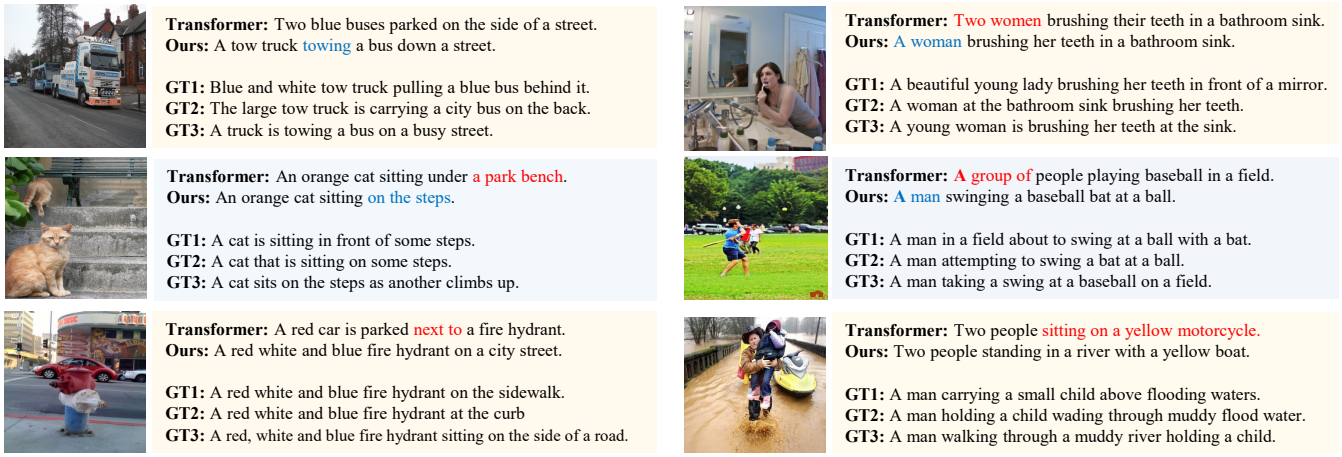


Fig. 2. Examples of captions generated by our approach and the standard Transformer model. Some accurate words are marked in blue, while wrong and inaccurate words are marked in red. Our method yields more accurate descriptions.

TABLE III
MSCOCO ONLINE EVALUATION. ALL VALUES ARE REPORTED AS PERCENTAGES (%), WITH THE HIGHEST VALUE OF EACH ENTRY HIGHLIGHTED IN BOLDFACE.

Model	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr-D	
Metric	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [32]	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.0
LSTM-A [58]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [8]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [55]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
GCN-LSTM [24]	-	-	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [25]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet [9]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
ETA [35]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
MMTransformer [11]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
XTransformer(ResNet-101) [10]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
XTransformer(SENet-154) [10]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	**75.0**	131.1	133.5
RSTNet(ResNext-101) [56]	81.7	96.2	66.5	90.0	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
RSTNet(ResNext-152) [56]	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
Ours(ResNeXt-101)	82.0	96.0	66.7	90.8	52.1	82.8	39.9	72.8	29.5	38.9	59.3	74.4	131.3	133.5
Ours(ResNeXt-152)	**82.4**	**96.5**	**67.4**	**91.6**	**52.8**	**83.6**	**40.7**	**73.7**	**29.8**	**39.4**	**59.8**	**75.0**	**133.4**	**135.4**

D. Performance Comparison

Offline Evaluation Tab. I and Tab. II provide performance comparisons between the state-of-the-art models and our proposed approach on the offline COCO Karpathy test split, for both the single-model version and ensemble version, respectively.

Single Model In Tab. I, we report the performance of our method in comparison with the aforementioned state-of-the-art models on grid-based features (ResNeXt-101 [39]), using captions predicted from a single model and optimization of the CIDEr score. To further validate the effectiveness of our method, following previous work [8], we also conduct experiments on region-based features (ResNet-101 [38]). Our method achieves competitive results in both cases. On region-based features, our method surpasses all other approaches in terms of ROUGE-L and CIDEr, while being competitive in BLEU-1, BLEU-4, METEOR and SPICE. On grid-based features, our approach achieves the best performance in all metrics. In particular, it advances the current state of the art in CIDEr by 1.8%.

Ensemble Model Following the common practice [9], [32] of building an ensemble of models, we also report the performances of our approach when averaging the output probability distributions of multiple independently trained versions of our model. In Tab. II, we use ensembles of two and four models, trained from different random seeds. Notably, for both ensembles, our approach achieves the best performance in all metrics, with increases of 3.7 and 3.3 CIDEr points over the current state-of-the-art MMTransformer [11], respectively.

Online Evaluation Finally, we also report the performance of our method on the online COCO test server. We use the ensemble of four models previously described, trained on the Karpathy training split. Results are reported in Tab. III, in comparison with the top-performing approaches on the leaderboard. For fair comparison, they also use an ensemble configuration. As can be seen, our method (with ResNeXt-152 [39]) surpasses the current state of the arts (*e.g.*, RSTNet with ResNeXt-152 [39]) in almost all metrics, achieving an improvement of 1.4 CIDEr points over the previous best performer.

Qualitative Analysis Fig. 2 shows several image captioning results of the standard Transformer and the proposed MD-

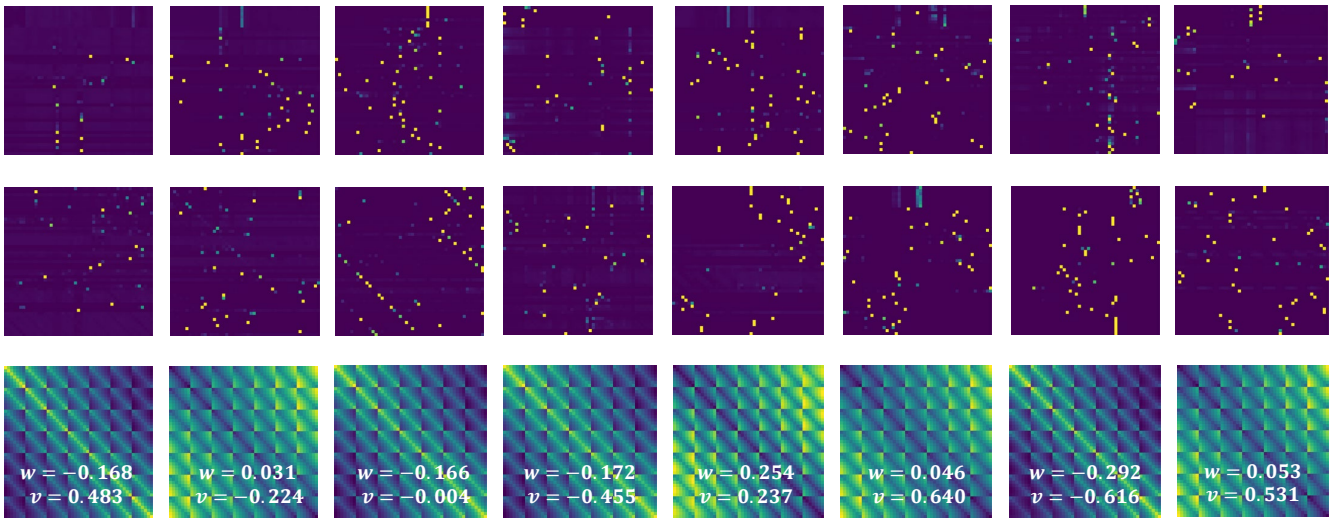
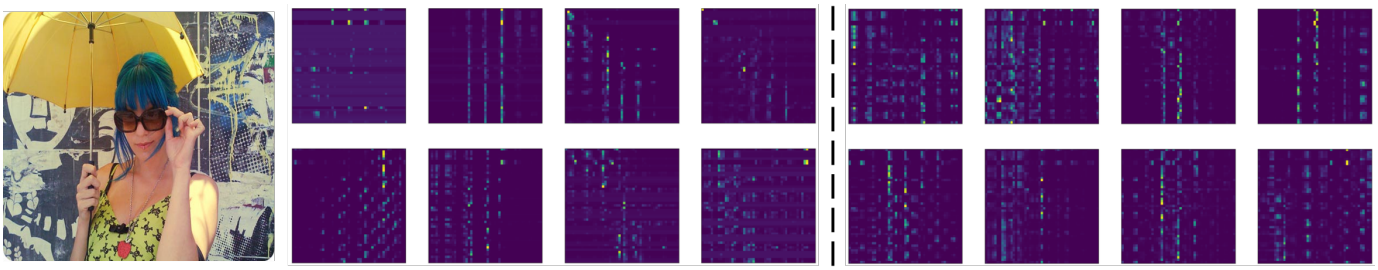


Fig. 3. The first, second and third rows represent the corresponding weight matrices of eight heads for absolute position embedding, relative position embedding and our proposed DSA, respectively. We split the single-head input into 49 grid representations. The x-axis and y-axis both represent the grid index. The colored box at position (i,j) presents the refined weight conditioned on the distance between the i-th and j-th grid. The brighter the color, the higher the value.



Transformer+DSA: A woman with blue hair holding an umbrella. **Transformer:** A woman with a blue umbrella.

Fig. 4. Attention weights and captions generated by Transformer+DSA and the standard Transformer model. The highlighted parts of the 1st, 3rd, 4th and 7th heatmaps are distributed diagonally, corresponding to the heads of $w < 0$ in Fig. 3. This shows that the real distance information is indeed injected into the attention weights of each head.

SAN. Generally, compared with the captions generated by the standard Transformer, which are only somewhat relevant to the image content though logically correct, our approach produces more accurate sentences by exploiting the real distances and the multi-branch technique. For example, MD-SAN generates the accurate phrases of "towing a bus" and "on the steps", which cannot be obtained by the standard Transformer. Besides, our approach generates more precise phrases with counting, such as "a woman" and "a man". This may be because the introduction of real distances improves the capacity of complex multi-modal reasoning, which is important for image captioning with human language. However, the proposed MD-SAN can only elaborate the content of the picture and still lacks the ability of common sense reasoning. For example, our model can only generate the description "two people standing in a river with a yellow boat". Even though this is more accurate than the standard Transformer, it still cannot describe "flooding waters", which is easy for humans.

E. Experimental Analysis

Ablation Study To validate the effectiveness of our proposed modules, we conduct ablation studies by comparing different variants of the proposed MD-SAN.

Firstly, we investigate the impact of the number of the encoding and decoding layers on captioning performance. As shown in Tab. IV, changing the number of layers, we observe a slight overall decrease in performance when the number of layers is larger than 3. Following this finding, all experiments use three layers.

Then, we investigate the impact of the number of the branches M and the drop rate ρ in MSA. As shown in Tab. V and Tab. VI, changing the number of branches, we observe a significant increase on the CIDEr score when using three branches and $\rho = 0.4$. Following this finding, all subsequent experiments use $M = 3$ and $\rho = 0.4$.

Analysis of DSA First, we choose the standard Transformer as our baseline, which is shown in the first row of Tab. VII. We then extend the baseline model to include our DSA module, which improves the performance. This verifies the importance of our DSA. Then we compare it with the Transformer equipped with absolute position embedding (denoted as AbsPE) [32], [59] and relative position embedding (RelPE) [16], respectively. Our module again achieves significant performance gains, proving the importance of capturing real

Fig. 5. Examples of captions generated by our proposed DSA and the standard Transformer model. Our method yields more accurate descriptions.

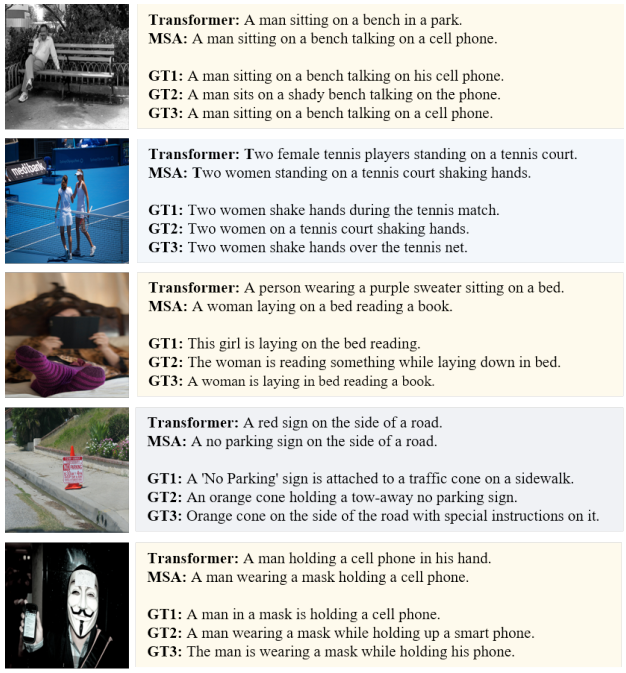


Fig. 6. Examples of captions generated by our proposed MSA and the standard Transformer model. Our method yields more accurate descriptions.

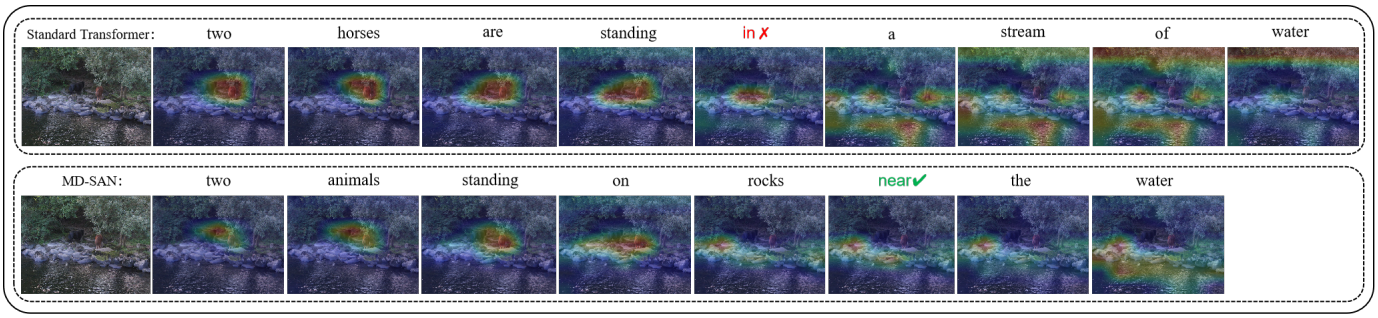


Fig. 7. Examples of attention visualization and captions generated by our approach and the standard Transformer model. Some accurate words are marked in green, and the wrong and inaccurate words are marked in red. Our method yields more accurate descriptions.

TABLE IV
ABLATION ON DIFFERENT NUMBERS OF LAYERS IN THE TRANSFORMER MODEL. ALL VALUES ARE REPORTED AS PERCENTAGE (%).

Both the encoder and decoder have the same number of layers.

Layer	BLEU-4	METEOR	ROUGE-L	CIDEr-D
$l = 2$	39.9	29.5	59.2	134.9
$l = 3$	39.8	**29.6**	59.1	**135.1**
$l = 4$	**40.0**	29.5	59.1	134.6
$l = 5$	39.6	**29.6**	**59.3**	134.6
$l = 6$	39.8	29.5	59.0	134.6

TABLE V
ABLATION ON DIFFERENT BRANCHES IN MSA WITH $p = 0.4$. ALL VALUES ARE REPORTED AS PERCENTAGE (%).

MSA	BLEU-4	METEOR	ROUGE-L	CIDEr-D	Params
$M = 1$	38.9	29.0	58.5	131.2	33.6M
$M = 2$	39.3	29.3	58.9	133.3	36.7M
$M = 3$	**39.5**	**29.4**	58.9	**134.0**	39.9M
$M = 4$	**39.5**	**29.4**	**59.1**	133.6	43.0M

distances effectively. In addition, we explore the effects of other different distances on our DSA, including Euclidean distance and Chebyshev distance (denoted as EucliDSA and ChebyDSA, respectively). We find that the adoption of these distances barely affects the performance of the model, so we choose the Manhattan distance for the least amount of calculation. Also, we explored the effect of different scaling functions in DSA for Eq. 7, including piecewise functions and a naïve scalar [60]: (1) Transformer+DSA-LPF adopts linear piecewise functions as the scaling function, *i.e.*,

$$R_i = \begin{cases} 0, & w_i R \leq 0, \\ w_i R, & 0 < w_i R \leq 2, \\ 2, & w_i R > 2, \end{cases} \tag{17}$$

where w_i is a learnable parameter; (2) Transformer+DSA-QPF

TABLE VI
ABLATION ON DIFFERENT DROP RATES IN MSA WITH $M = 3$. ALL VALUES ARE REPORTED AS PERCENTAGE (%).

MSA	BLEU-4	METEOR	ROUGE-L	CIDEr-D
$\rho = 0.0$	39.0	29.2	58.6	131.6
$\rho = 0.1$	39.4	**29.4**	**59.0**	132.9
$\rho = 0.2$	39.2	**29.4**	58.9	133.2
$\rho = 0.3$	39.4	29.3	58.9	133.6
$\rho = 0.4$	**39.5**	**29.4**	58.9	**134.0**
$\rho = 0.5$	39.4	**29.4**	**59.0**	132.8

adopts quadratic piecewise functions, *i.e.*,

$$R_i = \begin{cases} 0, & w_i R^2 \leq 0, \\ w_i R^2, & 0 < w_i R^2 \leq 2, \\ 2, & w_i R^2 > 2; \end{cases} \tag{18}$$

(3) Transformer+DSA-NS adopts a naïve scalar [60] as the scaling function. As shown in Tab. VIII, these scaling functions also lead to good performance gains, which also demonstrates the effectiveness of introducing distance information.

In order to further demonstrate the role of DSA, we first provide several image captioning results of the standard Transformer and the proposed DSA. As shown in Fig. 5, compared with the captions generated by the standard Transformer, our proposed DSA can describe more accurate spatial location or spatial-related relationship information. For example, our proposed DSA accurately identified the passenger seat of the car, which was identified as the back seat by the baseline method.

To further explore the ability of absolute position embedding, relative position embedding and our proposed DSA to model the real distance, we propose a new matrix Z_i, which is given by:

$$Z_i = \text{softmax}(\frac{E_i}{E_i^0}), \tag{19}$$

where E_i denotes the similarity matrix of various Transformer variants before softmax function, which can be referred to Eq. 8, and E_i^0 represents the counterpart after removing the positional embedding. We then show the matrices of Z_i of eight heads for absolute position embedding, relative position embedding and our proposed DSA in Fig. 3. We can see that the matrices of absolute position embedding and relative position embedding are haphazard and almost unregulated. In terms of DSA, different heads respond differently to distance information. When $w{>}0$, they tend to capture the long-distance relationships; when $w{<}0$, they tend to capture the short-distance relationships. $|w|$ reflects the sensitivity to the distance. Fig. 4 shows that real distance information is indeed injected into the attention weights of each head, helping to yield more accurate captions.

Like the locality principle of CNN, our DSA imposes the strong assumption that either the closer the grid is the more important it is, or the more distant it is the more important it is. This assumption significantly narrows the search space, and allows our DSA to introduce far fewer parameters than existing positional embedding methods. From another perspective, we know that self-attention is good at capturing global dependencies but local relationships can easily be ignored. While our DSA just focuses on the modeling of local relations, which greatly makes up for the deficiency of transformer and makes the model more powerful. Then we also add the local-window designs for comparison, including [61] and [62], denoted as Transformer+Swin and Transformer+Halo, respectively. As shown in Tab. IX, we can find that our DSA achieves the best performance, which is due to the fact that our DSA can adaptively introduce local information without the limitation of a fixed window size.

TABLE VII
ABLATION STUDY ON DIFFERENT POSITION EMBEDDING VARIANTS OF THE STANDARD TRANSFORMER. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

	B-1	B-4	M	R	C	S
Transformer	80.9	38.9	29.0	58.5	131.2	22.7
Transformer+AbsPE	81.0	39.2	29.2	58.7	132.5	22.8
Transformer+RelPE	**81.1**	39.2	29.3	58.8	132.9	**23.1**
Transformer+ChebyDSA	81.0	39.2	29.3	58.9	133.1	23.0
Transformer+EucliDSA	**81.1**	39.2	29.3	**59.0**	133.2	23.0
Transformer+DSA	**81.1**	**39.3**	**29.4**	58.9	**133.3**	23.0

TABLE VIII
ABLATION STUDY OF DIFFERENT SCALING FUNCTIONS.

	B-1	B-4	M	R	C	S
Transformer	80.9	38.9	29.0	58.5	131.2	22.7
Transformer+DSA-LPF	81.1	39.1	29.3	58.7	133.0	22.9
Transformer+DSA-QPF	**81.3**	**39.3**	29.3	**58.9**	133.2	22.8
Transformer+DSA-NS	81.1	39.1	29.2	58.6	132.8	22.9
Transformer+DSA	81.1	**39.3**	**29.4**	**58.9**	**133.3**	**23.0**

Analysis of MSA As can be seen from Tab. X, the Transformer with our MSA module outperforms the standard Transformer in all metrics, in particular with an increase of 2.8 CIDEr points. This validates the effectiveness of MSA in breaking the low-rank bottleneck. Next, we further introduce four strong baselines: Transformer+Talking [19], which includes linear projections across the attention-heads dimension before and after the softmax operation, Transformer+Mh-Heads, which is also the standard transformer where the number of attention heads and the dimensionality of each head is the same as MSA (with 3*8 attention heads), Transformer+ExtendQK [18], which sets the head size of an at-

TABLE IX
COMPARISONS WITH THE LOCAL-WINDOW DESIGNS.

	B-1	B-4	M	R	C	S
Transformer	80.9	38.9	29.0	58.5	131.2	22.7
Transformer+Swin	81.0	39.2	29.3	58.7	132.5	22.8
Transformer+Halo	80.9	**39.4**	29.3	58.8	132.8	22.9
Transformer+DSA	**81.1**	39.3	**29.4**	**58.9**	**133.3**	**23.0**

TABLE X
ABLATION STUDY ON DIFFERENT VARIANTS OF THE PLAIN TRANSFORMER. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

	B-1	B-4	M	R	C	S	Params
Transformer	80.9	38.9	29.0	58.5	131.2	22.7	33.6M
Transformer+Talking	**81.1**	39.3	29.2	**58.9**	132.4	22.7	33.6M
Transformer+Mh-Heads	81.0	39.2	29.3	58.8	132.2	22.9	39.9M
Transformer+ExtendQK	**81.1**	39.3	**29.4**	**58.9**	133.1	22.9	44.6M
Transformer-Large	81.0	39.1	29.2	58.8	132.0	22.8	39.9M
Transformer+MSA	**81.1**	**39.5**	**29.4**	**58.9**	**134.0**	**23.1**	39.9M

TABLE XI
ABLATION STUDY OF DIFFERENT DROP BRANCH RATE FOR TRANSFORMER+MH-HEADS AND TRANSFORMER+EXTENDQK.

	ρ	B-1	B-4	M	R	C	S
Transformer+Mh-Heads	0.0	81.0	39.2	29.3	58.8	132.2	22.9
	0.2	81.0	39.4	29.3	58.8	132.9	22.8
	0.3	**81.2**	39.3	29.3	58.8	133.1	22.9
	0.4	81.1	39.4	**29.4**	**58.9**	133.3	23.0
Transformer+ExtendQK	0.0	81.1	39.3	**29.4**	**58.9**	133.1	22.9
	0.2	81.0	39.4	29.2	58.7	133.4	22.8
	0.3	81.0	39.4	29.3	58.8	133.3	22.9
	0.4	81.1	39.2	29.2	58.8	133.0	22.8
Transformer+MSA	0.4	81.1	**39.5**	**29.4**	**58.9**	**134.0**	**23.1**

tention unit to the input sequence length (*i.e.*, 512), and Transformer-Large, which is the standard Transformer with 8 heads, and has as many parameters as the proposed MSA-based Transformer (8 heads and the dimensionality of each head is 192). Our MSA-based Transformer is significantly better than these four strong baselines in almost all metrics. Then, we apply the drop branch technique to Transformer+Mh-Heads and Transformer+ExtendQK, respectively. As shown in Tab. XI, we can see that the performance does increase a bit, which shows that the adoption of drop branch technique can improve the performance of the model after increasing the discriminability of each subspace feature. However, the above improvements are still much less than that made by the proposed MSA. First, this multi-branch structure can greatly enhance the capacity of the model [37], trading structural complexity for expressive power. Secondly, our MSA inherits the property of global dependency in multi-head self-attention. The effectiveness of self-attention mainly lies in its multi-head attention (MHA), which captures feature dependencies in different truncated feature spaces. Our MSA design can be regarded as step forward than MHA to generate more diversity subspaces with greater granularity on branches, encouraging each branch to learn discriminative representation.

Meanwhile, we further explored the role of MSA in different layers, as shown in Tab. XII. We can see that adding MSA to almost all layers results in some performance improvement, especially in the third layer. One speculation is that the third layer is responsible for the extraction of higher-level semantics, and enhancing the expressiveness of this layer can improve the performance of the model in a very direct way.

In order to further demonstrate the role of MSA, we provide several image captioning results of the standard Transformer and the proposed MSA, as shown in Fig. 6. Generally, compared with the captions generated by the standard Transformer, the latter produces more accurate sentences by exploiting the multi-branch technique.

Combination of DSA and MSA From Tab. XIII, we can see that both DSA and MSA outperform the standard Transformer across all metrics. Moreover, MD-SAN further outperforms DSA and MSA, which demonstrates that they are compatible with each other. Specifically, our MSA further enhances the global dependency modeling capability of MHA; while DSA introduces the modeling capability of local dependencies along with positional information. Both automatically find a reasonable balance through the learning of parameters w and v.

TABLE XII
ABLATION STUDY ON DIFFERENT VARIANTS OF THE MD-SAN TO FURTHER EXPLORE THE THE ROLE OF MSA. WE TAKE TRANSFORMER+DSA AS THE BASELINE, AND TRY TO ADD MSA IN DIFFERENT LAYERS SEPARATELY. THE NUMBER AFTER THE SYMBOL "MSA-" INDICATES THE LAYER TO WHICH THE MSA IS ADDED. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

	B-1	B-4	M	R	C	S	Params
Transformer+DSA	81.1	39.3	29.4	58.9	133.3	23.0	33.6M
+MSA-1	81.2	39.5	29.3	58.9	133.4	23.0	35.7M
+MSA-2	81.0	39.3	29.4	59.0	133.7	23.1	35.7M
+MSA-3	81.4	39.6	29.3	58.9	133.5	22.9	35.7M
+MSA-1,2	80.8	39.1	29.5	58.9	133.4	23.0	37.8M
+MSA-1,3	**81.5**	39.7	29.5	59.0	134.8	23.0	37.8M
+MSA-2,3	81.4	39.6	29.4	59.0	134.2	23.0	37.8M
+MSA	**81.5**	**39.8**	**29.6**	**59.1**	**135.1**	**23.2**	39.9M

TABLE XIII
ABLATION STUDY ON DIFFERENT VARIANTS OF THE PLAIN TRANSFORMER. ALL VALUES ARE REPORTED AS PERCENTAGES (%).

	B-1	B-4	M	R	C	S
Transformer	80.9	38.9	29.0	58.5	131.2	22.7
Transformer+DSA	81.1	39.3	29.4	58.9	133.3	23.0
Transformer+MSA	81.1	39.5	29.4	58.9	134.0	23.1
Transformer+DSA+MSA	**81.5**	**39.8**	**29.6**	**59.1**	**135.1**	**23.2**

Attention Visualization In order to better qualitatively evaluate the generated results, in Fig. 7 we visualize the attention heatmap of each word of a caption generated by the standard Transformer and the proposed MD-SAN. The contribution of one grid with respect to the output is given by averaging the attention weights of the eight heads in the last layer of the decoder. Our MD-SAN can accurately identify the animals close to the water rather than in the water. Results presented in Fig. 7 show that our approach can help ground correct image grids to words and generate more accurate captions.

VI. EXTENSION TO OTHER TASKS

We further investigate the effectiveness and generality of our method on the VQA and VG tasks. Since VQA and VG are both multi-modal classification problems, we use MCAN [44] as the baseline model, which uses an SA-based network to simultaneously encode image and sentence information. To apply our method to VQA and VG, we replace all the SA modules in MCAN with our DSA and MSA modules.

TABLE XIV
COMPARISON OF VQA ACCURACIES ON THE VQA-V2 DATASET WITH STATE-OF-THE-ART SINGLE-MODEL METHODS.

Method	Overall (%)	Yes/No (%)	Number (%)	Other (%)	Params
	ResNet-101				
BUTD [63]	63.84	81.40	43.81	55.78	-
MFH [64]	66.18	84.07	46.55	57.78	-
BAN-4 [65]	65.86	83.53	46.36	57.56	-
BAN-8 [65]	66.00	83.61	47.04	57.62	-
MCAN [20]	67.14	84.86	49.30	58.39	57.81M
Ours	**67.43**	**85.00**	**49.81**	**58.73**	70.42M
	ResNext-101				
MCAN [20]	67.44	**85.15**	49.83	58.63	57.81M
Ours	**67.72**	85.08	**51.03**	**58.93**	70.42M

TABLE XV
COMPARISON OF ACCURACIES (WITH IOU>0.5) ON REFCOCO, REFCOCO+ AND REFCOCOG WITH STATE-OF-THE-ART METHODS. ALL METHODS USE DETECTED OBJECTS TO EXTRACT VISUAL FEATURES.

	RefCOCO			RefCOCO+			RefCOCOg	
Method	TestA	TestB	Val	TestA	TestB	Val	Test	Val
CMN [66]	71.0	65.8	-	54.3	47.8	-	-	-
VC [21]	73.3	67.4	-	58.4	53.2	-	-	-
Spe.+Lis.+Rein.+MMI [67]	73.7	65.0	69.5	60.7	48.8	55.7	59.6	60.0
Spe.+**Lis.**+Rein.+MMI [67]	73.1	64.9	69.0	60.0	49.6	54.9	59.2	59.3
MAttNet [23]	81.1	70.0	76.7	71.6	56.0	65.3	67.3	66.6
MCAN [20]	83.7	73.2	79.6	77.3	60.4	70.5	73.2	72.8
Ours	**84.6**	**74.7**	**80.6**	**77.4**	**61.7**	**70.7**	**73.7**	**73.0**

A. Visual Question Answering

We conduct experiments on the most commonly used dataset for the VQA task [22]. It contains human annotated question-answer pairs for COCO images, with three questions per image (Yes/No, Number, and Other) and ten answers per question. We first strictly follow MCAN [20] when implementing our models. Specifically, images are represented with region features extracted from the Faster R-CNN object detector and the input questions are transformed with the GloVe word embeddings and an LSTM network. To further validate the effectiveness, we also conduct experiments on grid features.

Tab. XIV shows the overall accuracies of our method and the current state-of-the-art models on the offline test splits. The proposed MD-SAN boosts the accuracy of MCAN from 67.44% to 67.72%.

B. Visual Grounding

We use the same settings for the three visual grounding datasets [68]. For the textual queries, the maximum length is set to 14. For the images, we adopt a pre-trained object detector to extract the visual features, *e.g.*, a Faster R-CNN model trained on Visual Genome. During the training data preparation for the detector, we exclude all the common images that exist in the training, validation and test sets of RefCOCO, RefCOCO+, and RefCOCOg to avoid data leakage. We detect 100 objects for each image.

In Tab. XV, we report the comparative results on RefCOCO, RefCOCO+, and RefCOCOg. We employ the commonly used accuracy metric [69], where a prediction is considered correct if the predicted bounding box overlap with the groundtruth IoU is larger than 0.5. Equipped with the visual features (i.e., the Faster-RCNN model pre-trained on Visual Genome), our proposed MD-SAN obtains further improvement in overall accuracies across all datasets.

VII. CONCLUSION

In this paper, we present the Multi-branch Distance-sensitive Self-Attention Network (MD-SAN) for image captioning, which addresses the distance insensitivity problem and low-rank bottleneck of the standard self-attention. We propose two improvements to the self-attention mechanism, these include the Distance-sensitive Self-Attention (DSA) to explicitly consider the real distances between objects in 2D images and improve image understanding, and the Multi-branch Self-Attention (MSA) to improve the capacity of the model and increase the expressive power at negligible extra computational cost. Extensive experiments on the MS-COCO image captioning dataset validate the effectiveness of DSA, MSA, and their combination. Our method is also extendable to the VQA and VG tasks, improving state-of-the-art models.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[3] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 32–44, 2018.

[4] W. Zhao, X. Wu, and J. Luo, "Cross-domain image captioning via cross-modal retrieval and model adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1180–1192, 2020.

[5] S. Ye, J. Han, and N. Liu, "Attentive linear transformation for image captioning," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5514–5524, 2018.

[6] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[7] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 4, pp. 999–1012, 2018.

[8] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[9] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[10] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.

[11] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[12] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2286–2293.

[16] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: transforming objects into words," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 11 137–11 147.

[17] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, and H. Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 327–10 336.

[18] S. Bhojanapalli, C. Yun, A. S. Rawat, S. Reddi, and S. Kumar, "Low-rank bottleneck in multi-head attention models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 864–873.

[19] N. Shazeer, Z. Lan, Y. Cheng, N. Ding, and L. Hou, "Talking-heads attention," *arXiv preprint arXiv:2003.02436*, 2020.

[20] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.

[21] H. Zhang, Y. Niu, and S.-F. Chang, "Grounding referring expressions in images by variational context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4158–4166.

[22] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[23] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

[24] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.

[25] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[26] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-quality image captioning with fine-grained and semantic-guided visual attention," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1681–1693, 2018.

[27] J. Hou, X. Wu, X. Zhang, Y. Qi, Y. Jia, and J. Luo, "Joint commonsense and relation reasoning for image and video captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 973–10 980.

[28] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin, "Fine-grained image captioning with global-local discriminative objective," *IEEE Transactions on Multimedia*, 2020.

[29] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 4, pp. 1047–1061, 2018.

[30] J. H. Tan, C. S. Chan, and J. H. Chuah, "Comic: Toward a compact image captioning model with attention," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2686–2696, 2019.

[31] 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 *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[32] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[33] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[34] Z. Zheng, W. Wang, S. Qi, and S.-C. Zhu, "Reasoning visual dialogs with structural and partial observations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6669–6678.

[35] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8928–8937.

[36] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," in *NeurIPS*, 2019.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[41] J. Zhou, Y. Cao, X. Wang, P. Li, and W. Xu, "Deep recurrent models with fast-forward connections for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 371–383, 2016.

[42] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, "In defense of grid features for visual question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 267–10 276.

[43] C. Wu, F. Wu, and Y. Huang, "Da-transformer: Distance-aware transformer," *arXiv preprint arXiv:2010.06925*, 2020.

[44] Y. Fan, S. Xie, Y. Xia, L. Wu, T. Qin, X.-Y. Li, and T.-Y. Liu, "Multi-branch attentive transformer," *arXiv preprint arXiv:2006.10270*, 2020.

[45] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[46] A. Veit, M. J. Wilber, and S. Belongie, "Residual networks behave like ensembles of relatively shallow networks," *Advances in neural information processing systems*, vol. 29, pp. 550–558, 2016.

[47] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[49] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[51] 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.

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

[53] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[54] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[55] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8888–8897.

[56] X. Zhang, X. Sun, Y. Luo, J. Ji, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Rstnet: Captioning with adaptive attention on visual and non-visual words," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 465–15 474.

[57] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2621–2629.

[58] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4894–4902.

[59] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

[60] O. Press, N. A. Smith, and M. Lewis, "Train short, test long: Attention with linear biases enables input length extrapolation," *arXiv preprint arXiv:2108.12409*, 2021.

[61] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[62] A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.

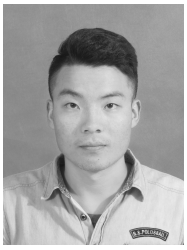
[63] D. Teney, P. Anderson, X. He, and A. Van Den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4223–4232.

[64] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12, pp. 5947–5959, 2018.

[65] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," in *NeurIPS*, 2018.

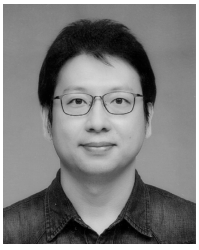
[66] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1115–1124.

[67] L. Yu, H. Tan, M. Bansal, and T. L. Berg, "A joint speaker-listener-reinforcer model for referring expressions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7282–7290.
[68] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, "Deep multimodal neural architecture search," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3743–3752.
[69] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.



Jiayi Ji is currently pursuing the Ph.D. degree with Xiamen University, Xiamen, China. He has published over five papers in international conferences, including the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Neural Information Processing Systems (NeuriPS), AAAI, the ACM Multimedia (MM), and so on. His research interests include MultiModal Machine Learning and Image Captioning.

Xiaoyang Huang received the bachelor's degree in computer science and technology from Xiamen University. He is currently pursuing the master's degree with the MAC Laboratory, Xiamen University. His research interests include MultiModal Machine Learning and Image Captioning.



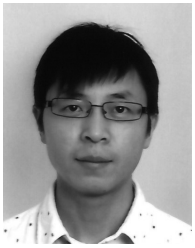
Xiaoshuai Sun (Senior Member, IEEE) received the B.S. degree in computer science from Harbin Engineering University, Harbin, China, in 2007, and the M.S. and Ph.D. degrees in computer science and technology from the Harbin Institute of Technology, Harbin, in 2009 and 2015, respectively. He was a Postdoctoral Research Fellow with the University of Queensland from 2015 to 2016. He served as a Lecturer with the Harbin Institute of Technology from 2016 to 2018. He is currently an Associate Professor with Xiamen University, China. He was a recipient of the Microsoft Research Asia Fellowship in 2011.

Yiyi Zhou received his Ph.D. degree supervised by Prof. Rongrong Ji from Xiamen University, China, in 2019. He is a Post-doctoral Research Fellow of the School of Informatics and a member of Media Analytics and Computing (MAC) lab of Xiamen University, China.

Gen Luo is currently pursuing the master's degree in Xiamen University. His research interests include vision-and-language interactions.

Liujuan Cao received the B.S., M.S., and Ph.D degrees from the School of Computer Science and Technology, Harbin Engineering University. She is currently an associate professor at Xiamen University. Her research interest mainly focuses on computer vision and pattern recognition. She has authored over 40 papers in top and major tired journals and conferences, including CVPR, TIP, etc. She is the Financial Chair of the IEEE MMSP 2015, the Workshop Chair of the ACM ICIMCS 2016, and the Local Chair of the Visual and Learning Seminar 2017.

Jianzhuang Liu (Senior Member, IEEE) received the Ph.D. degree in computer vision from The Chinese University of Hong Kong, Hong Kong, in 1997. From 1998 to 2000, he was a Research Fellow with Nanyang Technological University, Singapore. From 2000 to 2012, he was a Postdoctoral Fellow, an Assistant Professor, and an Adjunct Associate Professor with The Chinese University of Hong Kong. In 2011, he joined the Shenzhen Institute of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China, as a Professor. He is currently a Principal Researcher with Huawei Technologies Company Limited, Shenzhen, China. He has authored more than 150 papers. His research interests include computer vision, image processing, deep learning, and graphics.



Ling Shao is the Chief Scientist of Terminus Group and the President of Terminus International. He was the founding CEO and Chief Scientist of the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. His research interests include computer vision, deep learning, medical imaging and vision and language. He is a fellow of the IEEE, the IAPR, the BCS and the IET.

Rongrong Ji is a Nanqiang Distinguished Professor at Xiamen University, the Deputy Director of the Office of Science and Technology at Xiamen University, and the Director of Media Analytics and Computing Lab. He was awarded as the National Science Foundation for Excellent Young Scholars (2014), the National Ten Thousand Plan for Young Top Talents (2017), and the National Science Fundation for Distinguished Young Scholars (2020). His research falls in the field of computer vision, multimedia analysis, and machine learning. He has published 50+ papers in ACM/IEEE Transactions, including TPAMI and IJCV, and 100+ full papers on top-tier conferences, such as CVPR and NeurIPS. His publications have got over 10K citations in Google Scholar. He was the recipient of the Best Paper Award of ACM Multimedia 2011. He has served as Area Chairs in top-tier conferences such as CVPR and ACM Multimedia. He is also an Advisory Member for Artificial Intelligence Construction in the Electronic Information Education Commitee of the National Ministry of Education.