Article

Fusion Transformer for Image Captioning

Shujing Guo, xxx

[1] School of Software Engineering, Dalian University of Technology, Dalian 116120, China; ShujingGuo1016@gmail.com; xxx
* Correspondence: xxx

Abstract: Image captioning aims to automatically describe the visual content of a given image with fluent and reasonable sentences, which combines both computer vision (CV) and natural language processing (NLP). By leveraging region features, grid features or relative direction between objects, transformer-based models have achieved impressive and promising performance. However, existing approaches for image captioning only integrate some of the three types of information mentioned above and thus are arduous to obtain satisfactory results. In this paper, we propose a novel Fusion Transformer (FT) network to fuse both region and grid visual features considering directional relationships between objects. Such approach enhances the orientation perception between visual features and can also capture both the high-level and fine-grained details in the image. Specifically, in encoder, a modified multi-head attention is proposed to integrate the relative direction information between objects based on the original attention mechanism. It plays an important role in mining intrinsic spatial and contextual relationships between visual features with the fusion of relative direction encoding. We express both kinds of visual features as region-level and grid-level visual modalities, and word representations as the language modality. To use the complementary advantages of both region and grid features, we apply a Fusion Attention (FA) module to integrate these two types of visual features with word representations. This module performs attention over the target visual modality (grid or region) by the guidance of previous modality (region or grid). In FA, a Language Guidance Block (LGB) is employed to perform preliminary attention by infusing processed word features with each kind of previous visual modality, in order to make multiple modalities fully integrated. Further, the representations of the preliminary attention affect the target visual modality to obtain the integrated information. Moreover, to control the flow of integrated information, we apply a Fusion Gate Operation (FGO) module to do further fusion for visual and language modalities. The extensive experimental results on MS-COCO dataset show that our proposed Fusion Transformer performs competitively on various evaluation metrics, especially the CIDEr score reaches 133.4%. (And the CIDEr score reaches 134.7% in the further improvement of Fusion Transformer.)

Keywords: image captioning; Fusion Transformer; relative direction; region features; grid features

Citation: Guo, S.;xxx Title. *Mathematics* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Image captioning is a vital task at the intersection of computer vision and natural language processing, which generates a descriptive statement automatically for an input image by precisely understanding the scene meaning. More than simply recognizing the entity objects, image captioning has to master the spatial relations between objects. To generate reasonable and fluent sentences that match visual semantics, image captioning also has to bridge the gap between visual modality and language modality.

Inspired by the Seq2Seq [1] type tasks in machine translation, Vinyals *et al.* [2] transferred the basic encoder-decoder architecture [3] to image captioning simply. Subsequently, most image captioning methods follow this classical framework. In previous image captioning methods [2,4,5], the encoder generally utilizes a Convolutional Neural Network (CNN) to extract global features of images, while the decoder applies Recurrent Neural Network (RNN) to generate captions corresponding to the input images. Recently, to enhance the accuracy and consistency of encoder-decoder based frameworks for image captioning, the attention mechanism is exploited to assign different weights to different part of the inputs. Among these attention-based encoder-decoder architectures, the proposal of the Transformer [6] represents a milestone especially for its multi-head attention. More than relating different positions of a single sequence, multi-head attention in Transformer is also adept at capturing relationships between different modalities.

Most existing Transformer-based architectures utilize three sources of features to enhance captioning performance, which refers to region features, grid features and spatial relations between objects. Since region-based features extracted by Faster R-CNN [7] are first utilized in image captioning [4], a flurry of methods [8–10] adopt region features as input for providing high-level individual object information. Region features are accomplished in recognizing salient regions, which obviously indicates their lack of obtaining fine-grained detailed information in images. However, grid features like vanilla grid convolutional feature maps [11,12] are widely utilized for their contextual detailed information. Since a novel method in [13] is proposed to extract grid features for image captioning, a few models [14] are established to advance the performance by feeding grid features as input. Recently DLCT [16] is proposed to realize the integration of both region and grid features by leveraging a dual-way encoder. Naturally, in this paper, we also apply an encoder utilizing both visual features as input according to their complementary nature. Though transformer-based methods achieve outstanding performances in capturing relations between objects, the spatial relations (e.g., relative positional relations, absolute positional relations and relative directional relations) are still indispensable for image captioning. Inspired by the previous methods adding only positional information [8,17] or only directional information [18], we propose a novel method containing both positional and directional information to describe more accurate relations between objects.

In this paper, we propose a novel modified multi-head attention in encoder to fuse both region and grid information with additional relative directional encoding. To explore relations between region features, grid features and word representations, we introduce a Fusion Attention (FA) module to extend the initial cross-attention in the decoder part of Transformer. Further, to address the noises caused by direct fusion of two visual representations (region and grid), Language Guidance Block (LGB) in FA is proposed to integrate representations concluding two visual information indirectly. Concretely, LGB utilizes word attention from the decoder as a guidance to attend to the prior region features (or grid features). Then the result of LGB considered as a previous representation attends to the target grid features (or region features) and subsequently the value of target attention is obtained. It should be noticed that the value of another target modality attention is calculated the same way as the grid one. Ultimately, we employ a Fusion Gate Operation (FGO) module to further integrate two kinds of target attention computed by FA. This realizes the fusion of multiple types of information in an interlaced way involving region information, grid information and language information. As described, the final output of FGO will be propagated to the next Feed Forward Network (FFN) layer in decoder.

To integrate multiple types of information, DLCT [16] computes representations of multimodal information separately in encoder and concatenates them before feeding to decoder. Different from DLCT, we fuse the two types of visual features (region and grid) guided by text information in the decoder part, rather than computing the cross-attention value of region and grid information in encoder part. Moreover, we integrate various information in an interlaced way instead of mixing two single modal representations directly. And the worthy merits of FA in Fusion Transformer are that the word representations have chances to guide each kind of visual information directly and the previous fusion attention with one kind of visual modality can affect the target visual modality. Compared with another method in [17], we have low input requirements for encoder, which is a more practical way instead of extracting semantic attributes from the attribute detector as additional inputs.

The contributions of this paper can be summarized as follows:

1. A modified Multi-Head Self-Attention is proposed to capture relationship between visual features with directional relations. It extends the original positional encoding vector by fusing the relative directional encoding between objects.
2. A Fusion Attention (FA) is introduced to explore the relations between region features, grid features and word embeddings in an interlaced way. FA also extends the initial cross-attention in transformer-decoder.
3. A Fusion Gate Operation (FGO) module is infused to control the further propagation of region attention and grid attention.

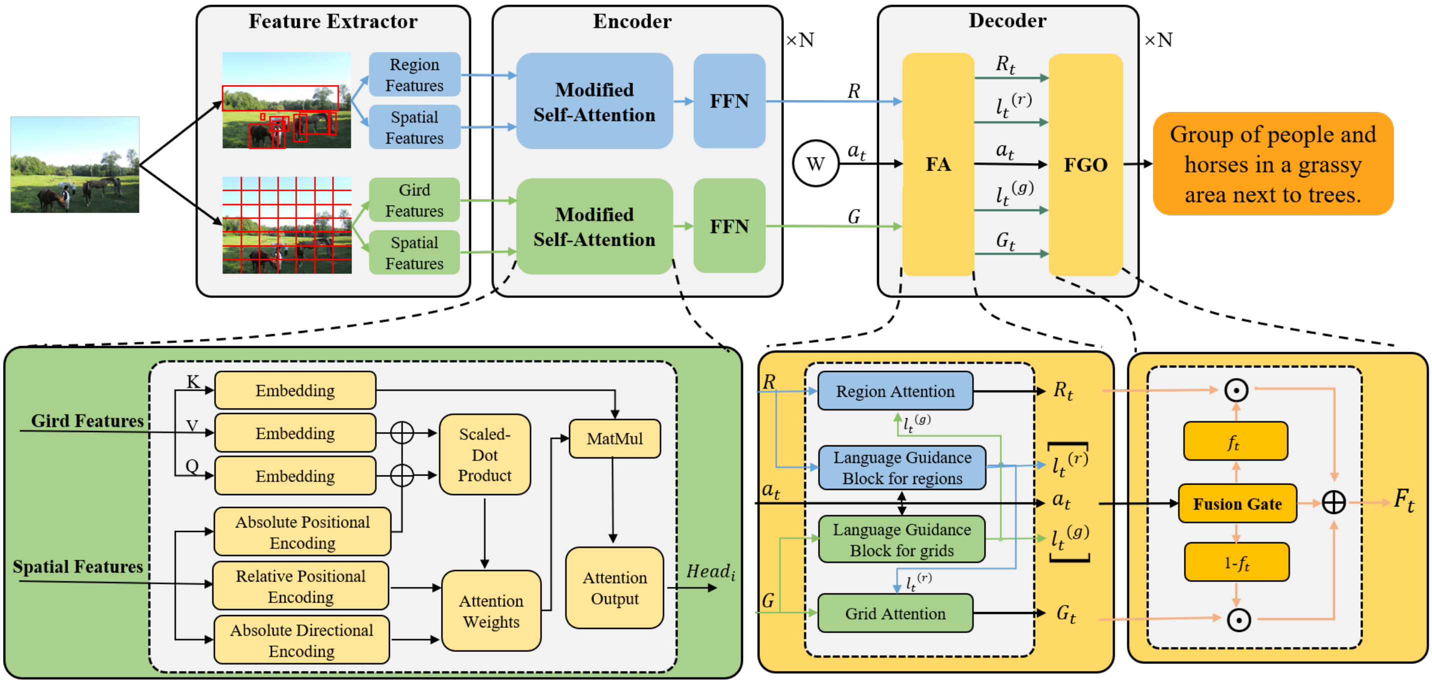


Figure 1. The overview of our proposed Fusion Transformer (FT) architecture. We propose a novel Fusion Transformer which creatively fuses region information, grid information and word representations with additional directional encoding. Notice that the word "W" in the figure refers to word representations.

2. Related Work

2.1. Encoder-decoder in Image Captioning

As the pioneering work of the encoder-decoder structure in image captioning, the Neural Image Caption Generator [2] replaces the RNN with a CNN-based InceptionNet to extract image features, and then applies LSTM-based sentence generator as the decoder to generate captions. Most methods [5,9,19] in image captioning are typically based on the encoder-decoder framework. Earlier works [5,20] make full use of grid-based features which contain contexual information. Recently, Jiang *et al.* [13] propose a novel method which extracts fixed-size patches as grid features from Faster R-CNN [7]. Since then, this method has been widely utilized in recent models. However, grid-based features have drawbacks in grasping salient characteristics of high-level objects. The proposal of region features by Anderson *et al.* [4] addresses the limitation of grid features. Recent approaches [8–10,21] in image captioning demonstrate the superiority of region features as well. In spite of these advances, few works study the complementary merits of encoding multimodal visual features including grid visual features and region visual features.

2.2. Attention Mechanism

The concept of attention mechanism is first proposed to mimic the human pattern of visual attention. Bahdanau *et al.* [22] first propose the attention mechanism formally. Later, Xu *et al.* [5] introduce the Soft Attention and Hard Attention in image captioning tasks for the first time, making the model more efficient by only focusing on the source domain information that is most relevant to the target task each time. Ahmed *et al.* [23] propose a novel network structure called Transformer which contains the multi-head attention mechanism. Earlier methods [4,5,20,24,25] employ attention mechanism to explore the relations between monomodal information. Recently, few methods [16,26,27] are emerging to combine multiple types of information. In this paper, we apply the Transformer structure to integrate multiple types of information by the Fusion Attention in decoder part. Rather than performing attentions over the few modalities separately and later integrating the individual attention representations in [16], our FA succeeds in grasping fusion nature between each kind of visual information and word representations by the guidance of language information.

2.3. Positional and Directional Information

Although the migration of the original Transformer from machine translation to image captioning can gain huge improvement, the initial positional encoding does not apply to 2-D images. The relations between objects in 2-D space cannot be accurately calculated only by absolute positions like the sequence of token representations in sentences. This obvious defect motivates recent approaches [8,9,16,18] to fuse positional information or directional information, attempting to achieve comprehensively better performances. Herdade *et al.* [8] propose an Object Relation Transformer which modifies the attention weights by fusing relative geometric information between two features. Guo *et al.* [9] introduce a Geometry-aware Self-attention which adds a bias calculating the attention over two objects with their relative position information. Luo *et al.* [16] integrate the absolute position information for the first time. Song *et al.* [18] devise a high-level relative directional category to judge the directional relations between objects, which aims to fuse directional information. However, these methods failed to fuse directional information with positional information in a simple vector representation.

3. Proposed Methodology

We propose our Fusion Transformer which is illustrated in Figure 1. We first introduce the standard Transformer structure for image captioning in section 3.1. Then in section 3.2, we devise a modified multi-head attention to fuse region and grid features with directional information. In addition, the two visual features fed to encoder are processed into highly abstract representations in section 3.3. In section 3.4, a Fusion Attention is devised to integrate two visual representations and mono language information in an interlaced way. Finally, we continue to integrate the fusion information, in order to generate the comprehensive captions in section 3.5.

3.1. Standard Transformer Architecture

The Transformer architecture in [6] eschews recurrence and relies entirely on attention mechanisms to draw global dependencies between the input and the output. Transformer is composed of a stack of N identical layers for both encoder and decoder. Each encoder layer consists of a multi-head attention sub-layer followed by a Feed Forward Network (FFN) sub-layer, while each decoder layer consists of two multi-head attention sub-layers also followed by a FFN. Moreover, residual connections and layer normalization are employed around all sub-layers to reduce the vanishing gradient phenomenon. In image captioning with Transformer, region-based features extracted by object detector are fed to encoder as input and the decoder generates caption words by utilizing previous words and the decoder generates caption words by utilizing previous words and the intermediate visual representations from encoder.

Given N input image feature vectors $x_1, x_2, ..., x_N$, where x_i denotes the i-th feature vector. These image features are first input into an embedding layer of the encoder and converted to vectors of dimension d_{model}. The embedded feature vectors are subsequently fed to encoder. Each encoder layer contains a multi-head attention sub-layer which has h identical heads. Each head of multi-head attention is a scaled dot-product self-attention which calculates queries $\boldsymbol{Q} \in \mathbb{R}^{N\times d_k}$, keys $\boldsymbol{K} \in \mathbb{R}^{N\times d_k}$ and values $\boldsymbol{V} \in \mathbb{R}^{N\times d_k}$ as follows:

$$\boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times d_{model}}$ denotes the matrix which is stacked by input image feature vectors $x_1...x_N$ and $\boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{W}_{\boldsymbol{V}}$ are all $d_{model} \times d_k$ dimensional learned projection matrices. Then, the similarity score matrix $\boldsymbol{E} \in \mathbb{R}^{N\times N}$ between any keys $\boldsymbol{K}$ and all queries $\boldsymbol{Q}$ is computed as:

$$\boldsymbol{E} = \frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_k}}, \tag{2}$$

where the element E_{mn} represents the similarity between the m-th image feature x_m and the n-th image feature x_n $(m, n = 1...N)$. The output of the scaled dot-product attention head is a weighted sum of values, which is formulated as:

$$head(\boldsymbol{X}) = attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_k}})\boldsymbol{V}. \tag{3}$$

Equations 1 to 3 are calculated for a single head of the multi-head self-attention mechanism. Furthermore, the output of multi-head attention is formulated as:

$$multi{-}head(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = Concat(head_1, ..., head_h)\boldsymbol{W}_{\boldsymbol{O}}, \tag{4}$$

where $head_i$ denotes the i-th $(i = 1, 2, ..., h)$ head of the multi-head attention mechanism. For convenience, here we denote the result of $\boldsymbol{multi - head(Q, K, V)}$ as $\boldsymbol{X'}$.

Finally, a position-wise FFN is connected with the multi-head attention, which contains two linear transformations with a ReLU activation in between [6]:

$$FFN(\boldsymbol{X'}) = ReLU(\boldsymbol{X'}W_1 + b_1)W_2 + b_2. \tag{5}$$

Further, the output of visual representations from encoder will be fed to a multi-head attention sub-layer in decoder for producing the next word of a caption.

3.2. Positional and Directional information Integrating

The initial positional encoding for machine translation in Transformer retains the relative position information of tokens, which is not applicable to images. Previous methods either exploit positional information or modify the attention mechanism only in a directional manner. Nevertheless, we devise a relative spatial encoding method which combines relative positional information with relative directional encoding.

3.2.1. Relative Directional Encoding

The relative direction between objects can crucially guide the model to generate descriptive sentences which are more consistent with human orientation cognition. Therefore, a set of standard direction vectors are artificially settled to do similarity calculation with the real visual orientation vectors obtained from region pairs or grid pairs.

Concretely, on planar 2-D coordinates, we first define a system of $4d$ standard direction vectors shown in Figure 2, where d denotes the number of divided sub-quadrant regions in each quadrant. Each vector with an arrow pointing from the center of the coordinate represents a normal direction.

In this system, every standard direction vector $u_i \in \mathbb{R}^2$ can be formulated as:

$$u_i = (cos(\frac{i\pi}{2d}), sin(\frac{i\pi}{2d})), \tag{6}$$

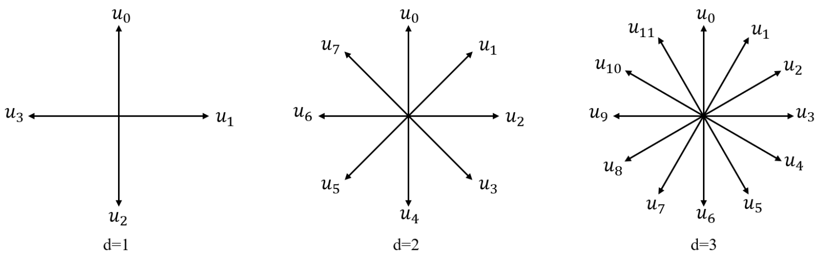


Figure 2. Examples of standard direction vectors for $d = 1, 2, 3$. Notice that if $k = 0$, it means without using relative directional information in the model.

where there are $4d$ normal direction vectors $u_0, u_1, ..., u_{4d-1}$. Notice that the embedding of u_i is a randomly initialized parameter which can be optimized during training like the word embedding in natural language processing.

Given N image features as well and their corresponding bounding boxes (or grid boxes) are represented as B_i $(i = 1, 2, ..., N)$. For regions, the relative direction between the m-th region box B_m and the n-th region box B_n can be measured as v_{mn} by the center coordinates of their bounding boxes, and the euclidean distance [28] between them is calculated as:

$$v_{mn} = \left(\frac{x_m - x_n}{\sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}}, \frac{y_m - y_n}{\sqrt{(x_m - x_n)^2 + (y_m - y_n)^2}}\right), \tag{7}$$

where (x_m, y_m) and (x_n, y_n) denote the center coordinates of the region boxes B_m and B_n, respectively. Finally, the relative directional scalar s_{mn} is represented by the cosine similarity of v_{mn} and u_i, which is as follows:

$$s_{mn} = \arg\max_i \frac{v_{mn} \cdot u_i}{\|v_{mn}\| \cdot \|u_i\|}. \tag{8}$$

For grids, (x_i, y_i) can also be noted as the center coordinates for the i-th $(i = 1, 2, ..., N)$ grid box.

3.2.2. Relative spatial Encoding

Among region features, the relative locations and directions for bounding boxes play important roles in grasping geometric relations for objects. As defined in section 3.2.1, there are N image features as well and their corresponding bounding boxes (or grid boxes) are represented as B_i $(i = 1, 2, ..., N)$.

We first calculate a vector Ω which represents geometric relationship between the m-th region box B_m and the n-th region box B_n:

$$\Omega(m, n) = \left(log\left(\frac{|x_m - x_n|}{w_m}\right), log\left(\frac{|y_m - y_n|}{h_m}\right), log\left(\frac{w_n}{w_m}\right), log\left(\frac{h_n}{h_m}\right)\right), \tag{9}$$

where (x_i, y_i) stands for the center coordinates of the bounding box B_i. w_i, h_i denote width and height for B_i. Then, Equation 9 is extended with the modification of Equation 8, which is as follows:

$$\Omega'(m, n) = (\Omega(m, n), log(s_{mn})), \tag{10}$$

where Ω' denotes the relative spatial encoding with directional information. $\Omega'(m, n)$ is embedded in a high-dimensional embedding and then mapped to a scalar:

$$\Omega'(m, n) = ReLU(Emb(\Omega')\boldsymbol{W_G}), \tag{11}$$

where $Emb(.)$ introduced in [6] calculates a high-dimensional embedding and $\boldsymbol{W_G}$ is a learned matrix.

For grids, the i-th grid box B_i's center coordinates can also be represented as (x_i, y_i) and its width and height are denoted as w_i, h_i.

3.2.3. Absolute Positional Encoding

The absolute positional information determines the unique coordinate positions for each region box and grid identified in an image.

Given N region features whose bounding boxes are denoted as $B_1, B_2, ..., B_i, ..., B_N$. For region boxes, the absolute position of the bounding box B_i is represented as $(x_{min}, y_{min}, x_{max}, y_{max})$, where (x_{min}, y_{min}) is the top-left corner of the box B_i and (x_{max}, y_{max}) denotes the bottom-right corner. Then, the absolute positional encoding for regions (we abbreviate it to RAPE) is calculated as:

$$RAPE = B_i \boldsymbol{W_{emb}}, \tag{12}$$

where $\boldsymbol{W_{emb}}$ is an embedding matrix.

However, the absolute positional encoding for grids (we abbreviate it to GAPE) is introduced as a vector concatenating two 1-D embeddings:

$$GAPE = [PE_r; PE_c], \tag{13}$$

where r, c denote the row index and column index of the grid in the feature map. Positional encoding PE_r and PE_c refer to:

$$\begin{aligned} PE(pos, 2i) &= sin(pos/10000^{2i/(d_{model}/2)}), \\ PE(pos, 2i+1) &= cos(pos/10000^{2i/(d_{model}/2)}), \end{aligned} \tag{14}$$

where pos denotes the row index or the column index of a grid box, and i denotes the dimension of a grid. For example, there is a grid box whose row index and column index are 3 and 5, and we first calculate $PE_r = PE_3$ over its dimension (i ranges from 0 to $d_{model}/2$). Then we compute $PE_c = PE_5$ over its dimension, where i ranges from 0 to $(d_{model}/2) - 1$. Finally, $[PE_r; PE_c] = [PE_3; PE_5]$ is represented as the absolute positional encoding for this grid box. Moreover, PE_r, $PE_c \in \mathbb{R}^{d_{model}/2}$.

3.3. *Modified Multi-head Attention*

Despite feeding to encoder in one manner, we apply an encoder feeding both region features and grid features separately. Notice that the Multi-Head Self-Attention in our FT is modified with our devised relative spatial encoding and absolute positional encoding illustrated in 3.2.2, 3.2.3. The similarity score $\boldsymbol{E}$ in Equation 2 is modified to $\boldsymbol{E'}$ as follows:

$$\boldsymbol{E'} = \frac{(\boldsymbol{Q} + APE_q)(\boldsymbol{K} + APE_k)^T}{\sqrt{d_k}} + log(\Omega'(m, n)), \tag{15}$$

where APE_q represents the absolute positional encoding of queries and APE_k denotes the absolute positional encoding of keys. Overall, the modified multi-head attention is calculated as follows:

$$Modified\ multi{-}head(Q, K, V) = Concat(head_1, ..., head_h)\boldsymbol{W^O}, \tag{16}$$

where

$$\begin{aligned} head_i &= attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, APE_q, APE_k, \Omega') \\ &= softmax(\boldsymbol{E'})\boldsymbol{V} \\ &= softmax(\frac{(\boldsymbol{Q} + APE_q)(\boldsymbol{K} + APE_k)^T}{\sqrt{d_k}} + log(\Omega'(m, n)))\boldsymbol{V}. \end{aligned} \tag{17}$$

And $attention(.)$ in Equation 17 denotes the modified self-attention.

3.4. Fusion Attention

The intermediate representations of both visual features from the encoder are fed to decoder for fusing multiple types of information containing two types of visual information with relative directional encoding and word representations. Previous approaches typically perform attention over few modalities separately in encoder and then integrate the individual attention representations before feeding to decoder, which fail to measure intrinsic relations of various types of information. To overcome this deficiency, we apply a Fusion Attention (FA) shown in Figure 3 concluding a Language Guidance Block (LGB) which enables the word representations to guide each kind of visual information directly.

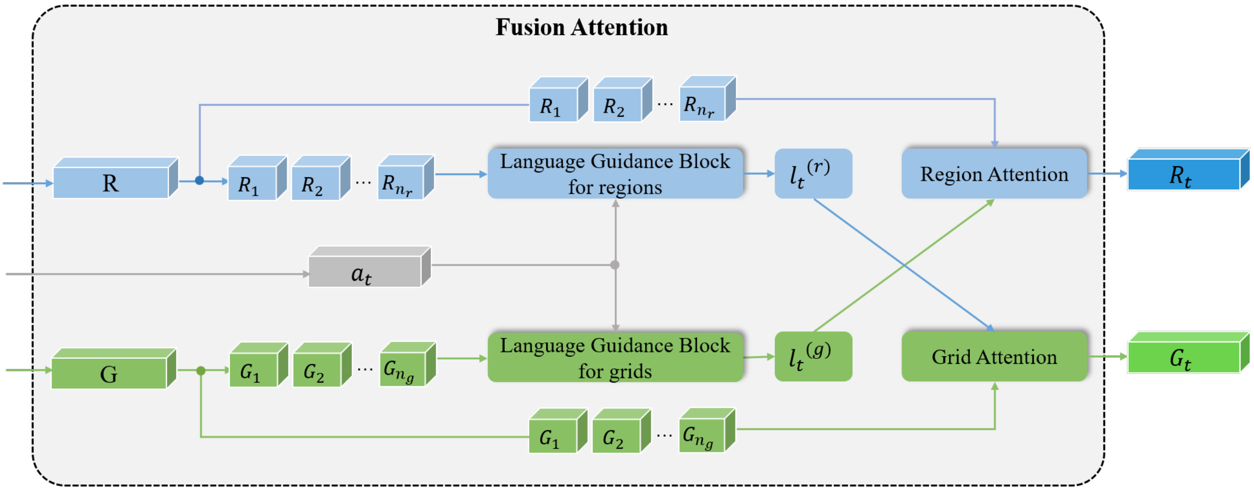


Figure 3. The illustration of Fusion Attention (FA). The Language Guidance Block (LGB) is devised to enable the text information a_t to fuse with each of the visual representations directly and respectively. Then the output of LGB can be regarded as the previous fusion attention with single visual modality which can affect the target visual modality.

We denote the processed region features from encoder as $\boldsymbol{R}$ and the number of them are represented as n_r. Further, the processed grid features from encoder are identified as $\boldsymbol{G}$ and the number of them are denoted as n_g. Subsequently, region representations $R_1, R_2, ..., R_{n_r}$ and grid representations $G_1, G_2, ..., G_{n_g}$ are fed to decoder.

Here we start from LGB which enables the output of normal multi-head attention $\boldsymbol{a_t}$ from decoder (calculated by Equation 4, where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ refer to language information here) to guide one of the injected visual representations (region or grid) from the encoder. This manner can be mimicked as the multi-head attention mechanism in Equation 4 :

$$\boldsymbol{l_t}^{(r)} = multi{-}head(\boldsymbol{a_t}, \boldsymbol{R}, \boldsymbol{R}), \tag{18}$$

where $\boldsymbol{R}$ is considered as the preliminary injected region modality and $\boldsymbol{l_t}^{(r)}$ denotes the generation of LGB for regions.

Furthermore, we exploit the fusion region guidance $\boldsymbol{l_t}^{(r)}$ to attend the target modality (grid or region) to generate the final fusion representation for visual features:

$$\boldsymbol{G_t} = multi{-}head(\boldsymbol{l_t}^{(r)}, \boldsymbol{G}, \boldsymbol{G}), \tag{19}$$

where $\boldsymbol{G}$ is considered as the target grid modality and $\boldsymbol{G_t}$ denotes the final output of FA for the target grid modality affected by the preliminary region modality. In addition, grid features injected as preliminary modality are processed the same way as above and the final output of FA for region modality is represented as $\boldsymbol{R_t}$.

3.5. Fusion Gate Operation

To further integrate the fusion representation $\boldsymbol{G_t}$ (in Equation 19) under the guidance of region modality and the $\boldsymbol{R_t}$ (in 3.4) under the guidance of grid modality, we employ the Fusion Gate Operation (FGO) to retain truly meaningful information in memory cells, to discard invalid information, and then to remember newly emerging states. Therefore, this property is commonly used in neural networks, like LSTM [15], GRU [3], etc., and it can address the long-term dependencies and gradient explosion (or gradient vanishing).

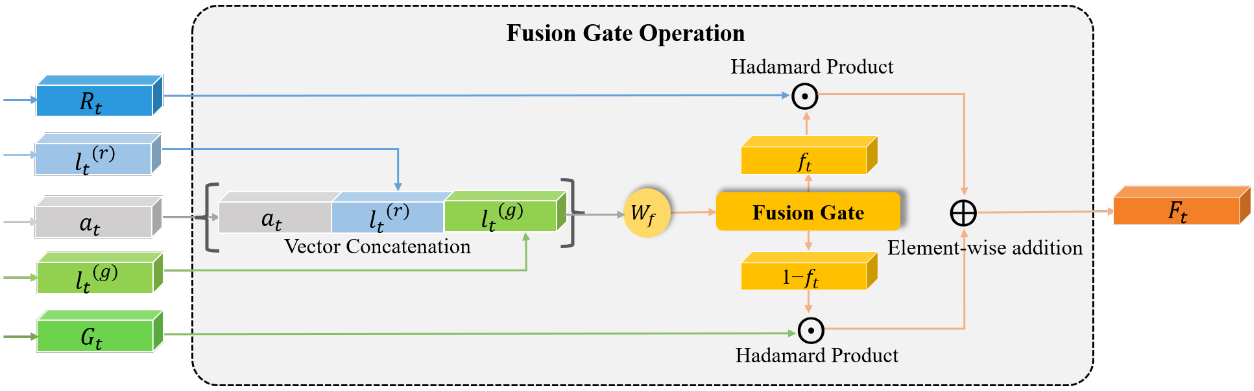


Figure 4. Framework of the Fusion Gate Operation (FGO). FGO is applied to further fuse the output fusion information from Fusion Attentiion, and the contextual information F_t is finally generated.

We first calculate a fusion gate f_t to fuse the output of normal Multi-Head Self-Attention $\boldsymbol{a_t}$ (also calculated by Equation 4, here $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ refer to the word representations) with the output of LGB for regions and grids:

$$f_t = sigmoid(\boldsymbol{W_f} \cdot [\boldsymbol{a_t}, \boldsymbol{l_t}^{(r)}, \boldsymbol{l_t}^{(g)}]), \tag{20}$$

where $f_t \in \mathbb{R}^{d_{model} \times 1}$ and $\boldsymbol{W_f} \in \mathbb{R}^{d_{model} \times 3d_{model}}$.

Then, a dual-way gate is applied to control the bilateral pathway, which is calculated as follows:

$$\boldsymbol{F_t} = f(\boldsymbol{R_t}) \odot f_t + f(\boldsymbol{G_t}) \odot (1 - f_t), \tag{21}$$

where f_t controls one output flow of FA for the target region modality and $(1 - f_t)$ controls the other output flow of FA. In addition, $\odot$ refers to hadamard product, $f(.)$ refers to the identity function and $\boldsymbol{F_t}$ denotes the final output of FGO. Finally, $F_t \in \mathbb{R}^{d_{model} \times 1}$ will be fed to the next FFN sub-layer in decoder.

4. Experiments

In this section, we assess the validity of FT by comparing to the state-of-the-art models on evaluation metrics. We also adopt extensive experiments to respectively demonstrate the effectiveness of fusing two kinds of visual features and integrating directional information.

4.1. Experimental Settings

4.1.1. Datasets

We utilize the typical MS-COCO 2014 dataset [29] to evaluate the performance of our proposed method. MS-COCO dataset contains 123,287 images annotated with 5 different ground truth captions for each. In this paper, the extensively available Karpathy split [30] is leveraged for offline evaluation, where 113,287 images are used for training, 5,000 images for validation and the remaining 5,000 for training. We follow standard practice and only perform minimal text pre-processing [4], which converts all sentences to lower case and tokenizes on white space.

4.1.2. Evaluation Metrics

To evaluate the quality of image captions with the Fusion Transformer, a set of standard evaluation metrics are widely employed, including BLEU [31], METEOR [32], ROUGE [33] and CIDEr [34]. BLEU and ROUGE are based on N-gram matching. BLEU mainly measures the accuracy but cannot evaluate the completeness of the generated sentences. ROUGE calculates the co-occurrence probability of N-gram in both reference descriptions and generated descriptions.

4.1.3. Implementation details

In our implementation, we utilize the pre-trained Faster R-CNN provided by [13] to extract region features and grid features. Jiang *et al.* [13] build a detector by employing the dilated C_5 backbone and 1×1 RoIPool followed by two FC layers. It modifies the original feature extractor by removing the dilated C_5 layer and applies a conventional ResNet [12] C_5 layer to extract grid features. For region features, the same model is utilized for extracting region representations after the first FC layer.

We adapt ResNeXt-101 as the backbone network of visual representations. We average-pool the grid features to 7×7 grid size and extract 2048-d region features by Faster R-CNN. Subsequently, we set d_{model} to 512. Different from the standard Transformer framework, we set the numbers of both encoder and decoder to 3 instead of 6. The number of multi-head is 8. In the cross-entropy pre-training stage, our implementation details follow [16]. In reinforcement learning stage, we optimize our model with CIDEr reward with the learning rate of 5×10^{-6} and the batch size is set to 100. Furthermore, Adam optimizer is used in both cross-entropy pre-training and reinforcement learning stage. The beam size is set to 5. And the value of d is set to 1 in the relative directional encoding part.

Table 1. Performance comparisons on the MS-COCO Karpathy offline test split. Notice that all values are reported as percentage (%).

Model	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE**	**CIDEr**
SCST [36]	-	34.2	26.7	57.7	114.0
Up-Down [4]	79.8	36.3	27.2	56.9	120.1
HAN [37]	80.9	37.6	27.8	58.1	121.7
GCN-LSTM [26]	80.5	38.2	28.5	58.5	128.3
SGAE [38]	80.8	38.4	28.4	58.6	127.8
ORT [8]	80.5	38.6	28.7	58.4	127.8
SRT [39]	80.3	38.5	28.7	58.4	129.1
AoA [40]	80.2	38.9	29.2	58.8	129.8
HIP [41]	-	39.1	28.9	59.2	130.6
M2 [19]	80.8	39.1	29.2	58.6	131.2
X-Transformer [42]	80.9	39.7	29.5	59.1	132.8
DRT [18]	81.7	**40.4**	29.5	59.3	133.2
ETA [17]	81.5	39.9	28.9	59.0	127.6
RSTNet [14]	81.1	39.3	29.4	58.5	133.3
region + grid (Ours)	81.5	39.9	**29.6**	59.0	**133.4**
segmentation + grid (Improvement)	**81.9**	40.0	**29.6**	**59.4**	**134.7**

4.2. Quantitative Analysis

4.2.1. Comparison with State-of-the-Art Models

Table 1 indicates the performance of the state-of-the-art models and Fusion Transformer on the offline test split. Specifically, we compare with the following models: SCST [36], Up-Down [4], HAN [37], GCN-LSTM [26], SGAE [38], ORT [8], SRT [39], AoA [40], HIP [41], M2 [19], X-Transformer [42], DRT [18] and ETA [17]. Especially, Up-Down is the first to introduce region features into image captioning. SGAE introduces scene graphs. ORT pioneers the fusion of relative positional information between objects. M2 designs a mesh-like structure to exploit both low-level and high-level contributions from encoder. X-Transformer proposes the X-Linear Attention Networks (X-LAN) that novelly integrates X-Linear attention block(s) into image encoder and sentence decoder of image captioning model to leverage higher order intra- and inter-modal interactions [42]. DRT proposes the direction matrix which consists of relative directional information between objects. And ETA utilizes both region and grid features as input without considering absolute positional information and relative directional information. However, our proposed FT model not only utilizes two kinds of visual features but also fuses additional relative directional encoding.

CIDEr is specially utilized for evaluating the quality of generated captions while BLEU and ROUGE are commonly used for text translation. Thus, CIDEr is regarded as the most representative metric in image captioning. As shown in Table 1, our Fusion Transformer model surpasses all other approaches in terms of METEOR and CIDEr. It also achieves comparable scores in BLEU and ROUGE compared to the ETA model. Overall, our method outperforms the competitor ETA in all metrics. Focusing on the BLEU and ROUGE metrics, our FT performs slightly worse than DRT.

Notice that the CIDEr score of our Fusion Transformer reaches 133.4%, which advances ETA by 5.8%, DRT by 0.2% and RSTNet by 0.1%. On the one hand, the significant boost of performance compared to ETA demonstrates the advantages of creatively fusing the vital relative directional encoding with general relative positional encoding. On the other hand, the improvement compared to DRT also substantiates the complementary of two kinds of visual features and the importance of positional relations.

4.2.2. Ablation Study

We conduct mainly two ablation studies to quantify the significant designs in our Fusion Transformer. The first study is designed to evaluate the importance of feeding two kinds of visual features. The other one is proposed to assess the effectiveness of fusing additional relative directional encoding information.

Table 2. Performance comparison of different feature settings. First three lines of this table are based on vanilla Transformer.

Model	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
Grid (G)	81.2	39.0	29.0	58.6	131.2
Region (R)	80.1	39.0	28.9	58.6	130.1
G + R	80.9	38.9	29.2	58.6	131.6
Ours (G + R)	**81.5**	**39.9**	**29.6**	**59.0**	**133.4**

As shown in Table 2, we conduct several experiments on our features utilizing the vanilla Transformer. For standard Transformer model exploiting only one kind of visual feature (region or grid), the result of fusing only grid features consistently exhibits better performance than the one fusing only region features. In sum, the CIDEr score of feeding grid features in Transformer reaches 131.2%, which advances the one feeding region features by 1.1%. This indicates the better performance and the worthy effect of grid features for standard Transformer. Further, compared to the results of every single feature, the concatenation of both region and grid features comprehensively achieves better score especially in CIDEr. Moreover, our Fusion Transformer with both visual features reaches obviously higher score in CIDEr which refers to 133.4%.

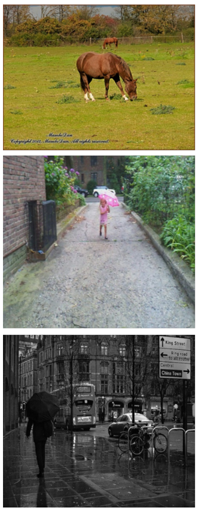
Table 3. Performance with / without relative directional information fused into modified Self-Attention, where dir means directional information.

Model	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
G + R	80.9	38.9	29.2	58.6	131.6
Ours (w/o dir)	81.4	39.8	29.5	59.1	132.9
Ours	**81.5**	**39.9**	**29.6**	**59.0**	**133.4**

For the relative directional encoding, we also conduct several experiments in Table 3 to demonstrate the effectiveness of our modified multi-head self-attention integrating relative directional encoding with conventional positional encoding. As shown in Table 3, three alternatives are considered : 1. vanilla Transformer using both region and grid features as input to encoder without relative directional encoding; 2. Fusion Transformer without fusing relative directional encoding; 3. Fusion Transformer with relative directional information.

Compared to our Fusion Transformer without fusing relative directional encoding, the whole Fusion Transformer model obviously improves the quality of generated captions which boosts the CIDEr score from 132.9% to 133.4% and performs better in all metrics. This shows the significance of relative directional encoding fused to attention mechanism.

Above all, the highlight of this paper contains double visual features input (region and grid features) and modified self-attention with relative directional encoding. All above improve the comprehensive performance of our Fusion Transformer model.



GT: A horse is standing in a green field.
Transformer: A brown horse grazing in a field.
Ours: Two horses grazing in a field of grass. # CIDEr : 1.86

GT: A little girl walking down a driveway carrying a pink umbrella.
Transformer: A little girl holding an umbrella on a sidewalk.
Ours: A little girl walking down a sidewalk with a pink umbrella. # CIDEr : 2.25

GT : A person walking in the rain on the sidewalk.
Transformer : A person walking down a city street with an umbrella.
Ours : A person walking in the rain with an umbrella on a street. # CIDEr : 2.96

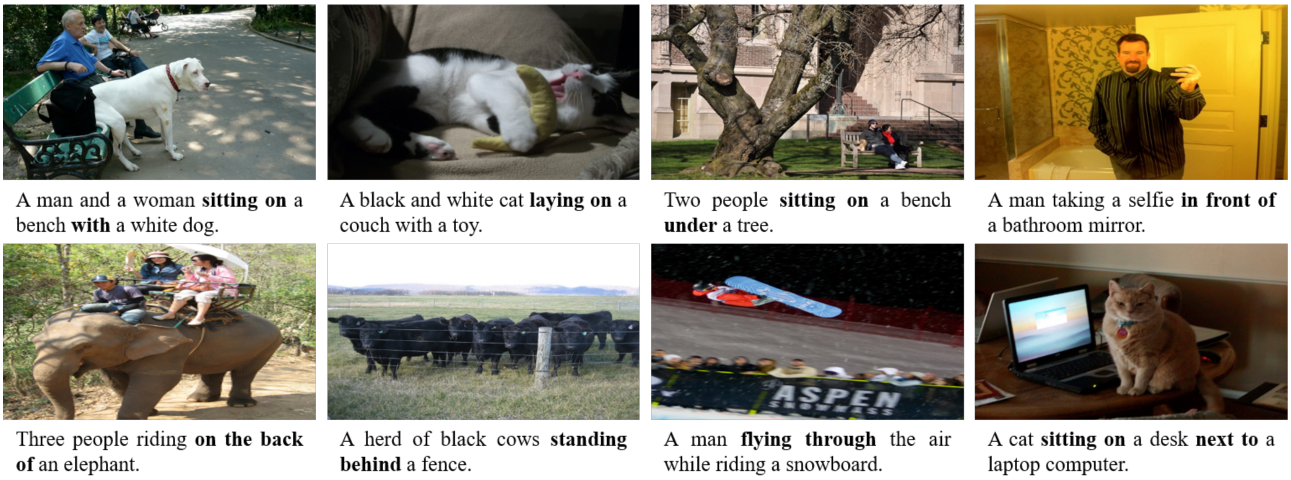
Figure 5. Examples of image captioning results by vanilla Transformer and our proposed FT, coupled with ground truth sentences and the corresponding CIDEr scores. The underlined words or phrases show the detailed information grasped from raw images.

4.3. Case Study

To validate the benefits of our proposed Fusion Transformer, we conduct qualitative analysis with visualization examples over the image / caption pairs which are shown in Figure 5. It illustrates few image captions by vanilla Transformer and our method.

As indicated by these examples, our Fusion Transformer can grasp more detailed and contextual information to generate more accurate captions than standard Transformer, which attributes to the complementary of region and grid features. For example, in the first example of Figure 5, our model can identify the real entire number of horses even though there is another horse standing far away from the most visualized horse in the center of the image. And for the second example in Figure 5, "pink" refers to the fine-grained color which can be difficult to identify for Transformer, while ours enable to deal with this detailed problem. Overall, our method can generate more fine-grained information contained in the entire image thanks to the fusion of two kinds of visual information.

With more reasonable directional information, our method performs better than baseline Transformer, which can measure spatial information, e.g., relative direction utilizing more accurate orientation prepositions. As illustrated in Figure 6, we can see that our novel modified self-attention fusing positional encoding with relative directional encoding proposed in this paper assists our model to enhance the orientation perception capturing the representative spatial relationship between objects which contains "on", "under", "in front of", "behind", "through", "on the back of", "over", "next to", etc.



A man and a woman **sitting on** a bench **with** a white dog.

A black and white cat **laying on** a couch with a toy.

Two people **sitting on** a bench **under** a tree.

A man taking a selfie **in front of** a bathroom mirror.

Three people riding **on the back of** an elephant.

A herd of black cows **standing behind** a fence.

A man **flying through** the air while riding a snowboard.

A cat **sitting on** a desk **next to** a laptop computer.

Figure 6. Performances of our FT model fused with relative directional encoding. As shown in this figure, the bold phrases or words represent the improvement of our model in directional relations' capturing.

For example, in the last image of Figure 6, the relative directional information can guide the model to generate "A cat sitting on a desk next to a laptop computer" instead of "A cat sitting next to a desk on a laptop computer" or other generations. Compared to ours, the vanilla Transformer regards the region features as a bag of tokens. It exploits the multi-head attention to grasp the appearance relations between objects. However, the vanilla Transformer ignores the spatial information like relative and absolute positional information, especially for the relative directional information. Hence, the standard Transformer has less sensitive orientation awareness than our Fusion Transformer.

5. Conclusions

In this paper, we propose a novel Fusion Transformer (FT). It creatively fuse two visual features and additional directional encoding with text information, while other methods fuse at most three of the above four types of information. First, we apply an encoder to grasp fine-grained contextual information of images by exploiting the complementary advantages of region and grid representations. Meanwhile a novel modified Multi-head Self-Attention is devised to explore broader spatial information containing positional information and relative directional information between objects. This modified attention enables the model to achieve better orientation perception. Fusion Attention (FA) with a Language Guidance Block (LGB) is applied to enable the word features to guide each kind of visual information directly. This makes the previous fusion attention with single visual information affect the target visual information. Further, Fusion Gate Operation (FGO) module is employed to do further integration. Experiment results demonstrate the superiority of our approach reaching 133.4% in CIDEr on offline test, which outperforms all other compared approaches. The Extensive ablation studies prove the significance of relative direction information. And the effect of inputting both region and grid features is also demonstrated in extensive experiments.

Author Contributions: Conceptualization, Hong Yu and Han Liu; Data curation, Shujing Guo; Formal analysis, Shujing Guo; Funding acquisition, Hong Yu and Han Liu; Investigation, Hong Yu; Methodology, Hong Yu and Shujing Guo; Project administration, Hong Yu and Han Liu; Resources, Yuanqiu Liu; Software, Yuanqiu Liu; Supervision, Han Liu; Validation, You Li; Visualization, Yuanqiu Liu; Writing – original draft, Shujing Guo; Writing – review and editing, Hong Yu and Han Liu. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China, No. 62106035 and 61876028, and Fundamental Research Funds for the Central Universities, No. DUT20RC(3)040, DUT20RC(3)066 and DUT20GF106.

Data Availability Statement: All data, models or code generated or used during the study are available from the authors upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS): Annual Conference on Neural Information Processing Systems, 2014, pp. 3104–3112.
2. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156–3164.
3. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734.
4. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 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 (CVPR), 2018, pp. 6077–6086.
5. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning, 2015, pp. 2048–2057.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 91–99.
8. Herdade, S.; Kappeler, A.; Boakye, K.; Soares, J. Image Captioning: Transforming Objects into Words. In Proceedings of the Advances in Neural Information Processing Systems, 2019, pp. 11135–11145.
9. Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; Lu, H. Normalized and Geometry-Aware Self-Attention Network for Image Captioning. In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10324–10333.
10. He, S.; Liao, W.; Tavakoli, H.R.; Yang, M.; Rosenhahn, B.; Pugeault, N. Image Captioning Through Image Transformer. In Proceedings of the Asian Conference on Computer Vision (ACCV), 2021, pp. 153–169.
11. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations (ICLR), 2015.
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
13. Jiang, H.; Misra, I.; Rohrbach, M.; Learned-Miller, E.G.; Chen, X. In Defense of Grid Features for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10264–10273.
14. Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; Ji, R. RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15465–15474.
15. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* 1997, 9, 1735–1780.
16. Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.; Ji, R. Dual-level Collaborative Transformer for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021, pp. 2286–2293.
17. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled Transformer for Image Captioning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 8927–8936.
18. Song, Z.; Zhou, X.; Dong, L.; Tan, J.; Guo, L. Direction Relation Transformer for Image Captioning. In Proceedings of the ACM International Conference on Multimedia (ACM), 2021, pp. 5056–5064.
19. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10575–10584.

20. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. 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 (CVPR), 2017, pp. 3242–3250.
21. Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; Ji, R. Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2021, pp. 1655–1663.
22. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations(ICLR), 2015.
23. Ahmed, K.; Keskar, N.S.; Socher, R. Weighted Transformer Network for Machine Translation. *CoRR* **2017**, *abs/1711.02132*.
24. Fang, H.; Gupta, S.; Iandola, F.N.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR, 2015, pp. 1473–1482.
25. Wu, Y.; Zhu, L.; Jiang, L.; Yang, Y. Decoupled Novel Object Captioner. In Proceedings of the ACM International Conference on Multimedia (ACM), 2018, pp. 1029–1037.
26. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring Visual Relationship for Image Captioning. In Proceedings of the Computer Vision European Conference (ECCV), 2018, pp. 711–727.
27. Li, N.; Chen, Z. Image Cationing with Visual-Semantic LSTM. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI, 2018, pp. 793–799.
28. Van Der Heijden, F.; Duin, R.P.; De Ridder, D.; Tax, D.M. *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*; John Wiley & Sons, 2005.
29. Lin, T.; Maire, M.; Belongie, S.J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision European Conference (ECCV), 2014, pp. 740–755.
30. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 664–676.
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Annual Meeting of the Association for Computational Linguistics(ACL), 2002, pp. 311–318.
32. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation, 2005, pp. 65–72.
33. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.
34. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4566–4575.
35. Jones, K.S. Index term weighting. *Information storage and retrieval* 1973, 9, 619–633.
36. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-Critical Sequence Training for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1179–1195.
37. Wang, W.; Chen, Z.; Hu, H. Hierarchical Attention Network for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2019, pp. 8957–8964.
38. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10685–10694.
39. Wang, L.; Bai, Z.; Zhang, Y.; Lu, H. Show, Recall, and Tell: Image Captioning with Recall Mechanism. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), 2020, pp. 12176–12183.
40. Huang, L.; Wang, W.; Chen, J.; Wei, X. Attention on Attention for Image Captioning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 4633–4642.
41. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Hierarchy Parsing for Image Captioning. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 2621–2629.
42. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-Linear Attention Networks for Image Captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10968–10977.