Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network

Jiayi Ji[1], **Yunpeng Luo**[1], **Xiaoshuai Sun**[1,2*], **Fuhai Chen**[1], **Gen Luo**[1], **Yongjian Wu**[3], **Yue Gao**[4], **Rongrong Ji**[1,2]

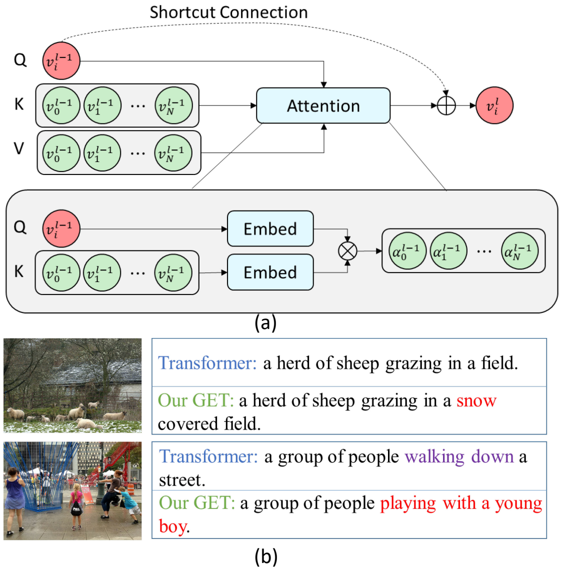
[1] Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University
[2] Institute of Artificial Intelligence, Xiamen University
[3] Tencent Youtu Lab
[4] Tsinghua University
jjyxmu@gmail.com,lyricpoem1997@gmail.com,xssun@xmu.edu.cn,chenfuhai3c@163.com,
luogen@stu.xmu.edu.cn, littlekenwu@tencent.com, gaoyue@tsinghua.edu.cn, rrji@xmu.edu.cn

Abstract

Transformer-based architectures have shown great success in image captioning, where object regions are encoded and then attended into the vectorial representations to guide the caption decoding. However, such vectorial representations only contain region-level information without considering the global information reflecting the entire image, which fails to expand the capability of complex multi-modal reasoning in image captioning. In this paper, we introduce a Global Enhanced Transformer (termed GET) to enable the extraction of a more comprehensive global representation, and then adaptively guide the decoder to generate high-quality captions. In GET, a Global Enhanced Encoder is designed for the embedding of the global feature, and a Global Adaptive Decoder are designed for the guidance of the caption generation. The former models intra- and inter-layer global representation by taking advantage of the proposed Global Enhanced Attention and a layer-wise fusion module. The latter contains a Global Adaptive Controller that can adaptively fuse the global information into the decoder to guide the caption generation. Extensive experiments on MS COCO dataset demonstrate the superiority of our GET over many state-of-the-arts.

Introduction

Image captioning aims to describe the semantic content of an image via neural language, which has recently attracted extensive research attention. Inspired by the sequence-to-sequence model for machine translation, most captioning models (Vinyals et al. 2016; Xu et al. 2015; Anderson et al. 2018; Huang et al. 2019) mainly adopt a encoder-decoder framework, where an encoder network encodes the input image into a vectorial feature, and a decoder network takes the vectorial feature as input and generates the output caption. Such an encoder-decoder framework is recently well promoted with the development of the Transformer (Vaswani et al. 2017), where the self-attention is efficiently utilized to capture the correlations among the regions and words (Liu et al. 2019; Huang et al. 2019; Li et al. 2019a; Herdade et al. 2019; Cornia et al. 2020).


Shortcut Connection
Q
K
V
v_i^{l-1}
v_0^{l-1} v_1^{l-1} ··· v_N^{l-1}
Attention
v_i^l
Embed
Embed
α_0^{l-1} α_1^{l-1} ··· α_N^{l-1}
(a)
(b)

Transformer: a herd of sheep grazing in a field.

Our GET: a herd of sheep grazing in a snow covered field.

Transformer: a group of people walking down a street.

Our GET: a group of people playing with a young boy.

Figure 1: (a) The self-attention mechanism in the l-th layer of a standard Transformer. The vectorial representation v_i^l is region-biased, which only focuses on the region-level information (Devlin et al. 2018; Song et al. 2020; Weng et al. 2020). (b) Two key issues of the traditional Transformer-based captioning model that we try to address: object missing (top: missing "snow") and false prediction (bottom: predicting "playing with a boy" as "waking down").

In the Transformer architecture, a set of image regions are encoded and attended into vectorial representations, as shown in Fig. 1 (a). These representations are then fused into the decoder to generate the corresponding captions. However, as demonstrated by earlier works (Devlin et al. 2018; Song et al. 2020; Weng et al. 2020), even though the vectorial representations of these regions are hierarchically calculated by being attended to all regions in the image, they still ignore the image-level characteristics and are thereby less effective for the decoder (Weng et al. 2020; Anderson et al. 2018). It causes the problem of object missing when gener-

*Corresponding Author

ating descriptions, which is attributed to the limit number of categories in object detectors. As shown in the top of Fig. 1 (b), an important concept, *i.e.* "snow", is not presented. Besides, it is more error-prone by focusing on local information while ignoring global guidance, as shown in the bottom of Fig. 1 (b), which is attributed to treating each object in isolation, to lead to a relationship bias.

To improve the caption quality, a natural way is to capture and leverage global representation to guide the selection of attractive objects and their relationships, which is however nontrivial due to two challenges. First, directly extracting a global representation from an image by techniques like pooling might introduce strong contextual noises, which severely cause semantic ambiguity and damage the representation accuracy. Such damage can be even accumulated for multi-step self-attention in Transformers. Second, the extracted global representation can not be directly used by the Transformer decoder since the need for global guidance varies during the generation of captions.

To solve the above problems, we propose a new Transformer architecture, *i.e.*, Global Enhanced Transformer (termed GET) as shown in Fig. 2. GET captures the global feature via Global Enhanced Attention and utilizes the global feature to guide the caption generation via Gated Adaptive Controller. In GET, we first design a Global Enhanced Encoder to extract intra- and inter-layer global representations. Specifically, we adopt Global Enhanced Attention to aggregate local information from each layer to form intra-layer global representation. After that, the global features are sequentially aggregated among layers via recurrent neural networks, which discard useless information from the previous layers. Then we adaptively fuse the distilled global representation into the decoder via a Global Adaptive Controller module, which can be implemented by two alternative gating modules to control the fusion, *i.e.,* Gate Adaptive Controller and Multi-Head Adaptive Controller. As the local vectorial representations may be insufficiently comprehensive in detail, GET explores the global parts of images to supplement the local vectorial representation, which could be more comprehensive and instructive for caption generation.

To sum up, our major contributions are itemized below:

- We address the issue of object missing and relationship bias by leveraging global representation to provide more comprehensive visual information and play the role of connecting various local parts, which is fundamental in image captioning task.
- We devise a unique encoder, termed Global Enhanced Encoder, which enables the Transformer framework to model intra- and inter-layer global information simultaneously, and propose a novel gating mechanism named Gated Adaptive Controller to provide an adaptive and sophisticated control for the fusion of global information.
- Through extensive experiments, we demonstrate that our Global Enhanced Transformer (GET) model can achieve new state-of-the-art performance on MS COCO dataset.

Related Work

Image Captioning. Inspired by the encoder-decoder architectures in machine translation (Bahdanau, Cho, and Bengio 2014; Sutskever, Vinyals, and Le 2014), most existing image captioning approaches typically adopt the CNN-RNN framework (Vinyals et al. 2016; Karpathy and Fei-Fei 2015), where a convolution neural network (CNN) (He et al. 2016; Lin et al. 2020) is used to encode a given image, which is followed by a recurrent neural network (RNN) (Hochreiter and Schmidhuber 1997) to decode the CNN output into a sentence. Recently, a variety of advanced models (Yao et al. 2018; Yang et al. 2019; Anderson et al. 2018; Lu et al. 2017) have been proposed with attention (Xu et al. 2015) and RL-based training objectives (Rennie et al. 2017).

Transformer-based Image Captioning. Some recent approaches have explored the use of the Transformer model (Vaswani et al. 2017) in Vision-Language tasks. (Huang et al. 2019) introduced a Transformer-like encoder to encode the regions into the hidden states, which was paired with an LSTM decoder. Recently, (Zhu et al. 2018; Herdade et al. 2019; Pan et al. 2020; Guo et al. 2020; Li et al. 2019b; Cornia et al. 2020) proposed to replace conventional RNN with the Transformer architecture, achieving new state-of-the-art performance. On the same line, (Li et al. 2019a; Liu et al. 2019, 2020) used the Transformer to integrates both visual information and additional semantic concepts given by an external tagger. However, leveraging global information in the Transformer for the image captioning task has never been explicitly explored, which motivates our work in this paper.

Preliminaries

The Transformer-based models formulate the calculation of the t-th hidden state of decoder as

$$h_t = Decoder(Encoder(I), w_1, \cdots, w_{t-1}), \quad (1)$$

where w_i represents the feature embedding of the i-th word. The Transformer contains an encoder which consists of a stack of self-attention and feed-forward layers, and a decoder which uses self-attention on textual words and cross-attention over the vectorial representations from the encoder to generate the caption word by word.

We first present a basic form of attention, called "Scaled Dot-Product Attention" , which is first proposed as a core component in Transformer (Vaswani et al. 2017). All intra-modality and cross-modality interactions between word and image-level features are modeled via this basic form of attention. The attention module operates on some queries Q, keys K and values V and generates weighted average vectors $\hat{V}$, which can be formulated as:

$$\hat{V} = \text{ Attention } (Q, K, V) = \operatorname{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \quad (2)$$

where Q is a matrix of n_q query vectors, K and V both contain n_k keys and values, all with the same dimensionality, and d is a scaling factor.

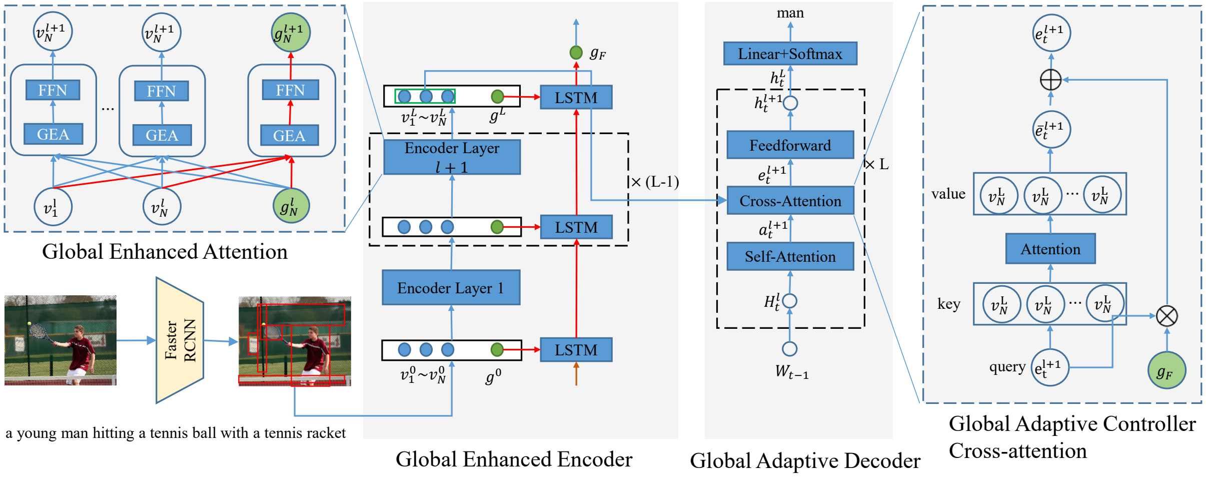


Figure 2: Overview of our Global Enhanced Transformer Networks (GET) for image captioning. A set of regions are first fed into a global enhanced encoder to extract intra- and inter-layer global information and region-level representation, which are then adaptively fused into the decoder to generate captions. Notice that the Residual Connections, Layer Normalizations, and Embedding Layers are omitted.

To extend the capacity of exploring subspaces, Transformer employs an effective module called multi-head attention, which is defined as

$$MultiHead(Q,K,V) = Concat\,(H_1,\ldots,H_h)\,W^O, \quad (3)$$

$$H_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right), \quad (4)$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{\frac{d}{h}\times d}$ are the independent head projection matrices, $i = 1, 2, \cdots, h$, and W^O denotes the linear transformation.

Our Method

In this section, we devise our Global Enhanced Transformer (GET) for image captioning. As shown in Fig. 2, the overall architecture follows the encoder-decoder paradigm. First, a global-enhanced encoder maps the original inputs into highly abstract local representations and extracts the intra- and inter-layer global representation. Then the decoder adaptively incorporates the multimodal information simultaneously through the proposed global adaptive controller to generate the caption word by word.

Global-enhanced Encoder

The image is represented as a group of visual features $V = \{v_1, v_2, \cdots, v_N\}$ extracted from a pre-trained object detector as (Ren et al. 2015), where N is the number of visual regions. Specifically, the detector is a Faster-RCNN model pre-trained on the Visual Genome dataset (Krishna et al. 2017). We can represent the images as:

$$g = \frac{1}{N}\sum_{i=1}^{N} v_i. \quad (5)$$

Each encoder is a stack of L identical layers, of which each one contains a novel structure, *i.e.*, the global-enhanced self-attention (GEA). To adapt the feature dimensionality to the encoder, the visual features V is first fed into a fully-connected layer, then we get projected features $V^0 = \{v_1^0, v_2^0, \cdots, v_N^0\}$ and g^0.

Global-enhanced attention. The early methods only feed regions to the encoder to extract the vectorial representation. As shown in (Devlin et al. 2018; Song et al. 2020; Weng et al. 2020), even though the vectorial representation of each region is hierarchically calculated by attending to all regions in the image, these vectorial representations only contain local features which focus on the region-level information. To capture a comprehensive global representation, both region features V and global feature g are fed into the multi-head self-attention module in each layer. By this way, the local information can be aggregated to form the global representation, through which we can capture the intra-layer global information. Specifically, the output of the l-th $(0 \leq l{<}L)$ layer $O^l \in R^{d\times(n+1)}$ is fed into the multi-head self-attention module in the $(l+1)$-th layer, which is then followed by a residual connection and a layer-normalization:

$$\begin{aligned}\overline{V}^{l+1} &= GEA(O^l)\\ &= MultiHead(O^l, O^l, O^l),\end{aligned} \quad (6)$$

$$V^{l+1} = LayerNorm(O^l + \overline{V}^{l+1}), \quad (7)$$

where $O^0 = (V^0; g^0)$, and the residual connections help avoid the vanishing gradient problem the training phase. Then a final feed-forward neural network is adopted for additional processing of the outputs, which is also followed by a residual connection and a layer normalization step:

$$O^{l+1} = LayerNorm(V^{l+1} + FFN(V^{l+1})), \quad (8)$$

As illustrated in (Dou et al. 2018; Wang et al. 2020c), the representations in different layers have different meanings.

Thus we integrate the global representation from different layers to fuse all the low- and high-level information. Note that such a fusion can also help ease the information flow in the stack (Wang et al. 2020c). A straightforward way is pooling (*e.g.*, average pooling), which however loses layer information. In contrast, we adopt LSTM network (Hochreiter and Schmidhuber 1997) for layer-wise fusion and achieve the final global representation g_F:

$$h_i = LSTM(g^i, h_{i-1}), g_F = h_L, \tag{9}$$

where the LSTM control the model to forget useless information from previous layers via the forgetting gate, which aggregates the global representation from the first layer to L-th layer to obtain inter-layer information.

Global Adaptive Decoder

In the decoding phase, the global representation was adaptively fused into the decoder to guide caption generation. Similar to the encoder, the decoder consists of N identical layers. We start with the basic layer of the global adaptive decoder, which contains a global adaptive controller (GAC) to decide how much the global contextual information should be considered.

Based on the local representation V^L and global representation g_F, the decoder generates captions for the image word-by-word. Suppose the decoder is generating the t-th word in the target sentence. We denote $w_t \in \mathbb{R}^{d\times 1}$as the vector representation of the t-th word, which is the sum of word embedding and positional encoding. Therefore, the input matrix representation for time step t is:

$$W_{t-1} = (w_0, w_1, \cdots, w_{t-1}), \tag{10}$$

where w_0 represents the start of sentence.

For the $(l + 1)$-th layer, the inputs $H_t^l = \{h_1^l, h_2^l, \cdots, h_t^l\} \in \mathbb{R}^{d\times t}$ are fed into a multi-head self-attention module:

$$\overline{a}_t^{l+1} = MutiHead(h_t^l, H_t^l, H_t^l). \tag{11}$$

Note that W_{t-1} are the inputs of the first layer and $h_t^0 = w_{t-1}$. Then there is a residual connection around them, which is followed by a layer-normalization step:

$$a_t^{l+1} = LayerNorm(h_t^l + \overline{a}_t^{l+1}). \tag{12}$$

Subsequently, the output a_t^{l+1} is passed into the other multi-head cross-attention, *e.g.*, GAC to incorporate with features V and g_F, which is followed by a residual connection and a layer-normalization:

$$\overline{e}_t^{l+1} = GAC(a_t^{l+1}, V^L, g_F) \tag{13}$$

$$e_t^{l+1} = LayerNorm(a_t^{l+1} + \overline{e}_t^{l+1}), \tag{14}$$

where e_t^{l+1} contains multi-model information, which is adaptively refined by the global representation to model a more comprehensive and suitable representation. The detail of GAC is described in the next subsection. Then we feed it into a feed-forward neural network (FFN), which is followed by a residual connection and a layer-normalization to obtain the output:

$$h_t^{l+1} = LayerNorm(e_t^{l+1} + FFN(e_t^{l+1})). \tag{15}$$

Finally, the output of layer N is fed into the classifier over vocabulary to predict the next word. Let the predicted caption be $Y_t = \{y_0, y_1, \cdots, y_t\}$, where $y_i \in V$, and V is the vocabulary of the captions. Then the conditional probability distribution of words at time t is $p(y_t|Y_{t-1})$, which can be calculated by:

$$p(y_t|Y_{t-1}) = softmax(W_y h_t^L), \tag{16}$$

where $W_y \in \mathbb{R}^{|V|\times d}$, and $|V|$ is the number of words in the vocabulary.

Global Adaptive Controller Cross-attention

In the generation process, we design two alternative functions for the global adaptive controller to fuse the global information into decoder according to the contextual signals, *i.e.*, Gate Adaptive Controller (GAC) and Multi-Head Adaptive Controller (MAC).

Gate Adaptive Controller Self-Attention. The demand for global information for each target word is different. Motivated by (Lu et al. 2017), we propose a context gating mechanism to control the importance of global information. The context gate is determined by the query a_t^{l+1} and the global representation g_L:

$$\alpha = sigmoid\big((a_t^{l+1})^{\mathrm{T}} g_L\big). \tag{17}$$

We then adaptively fuse the global representation to refine the output from multi-head self-attention as below:

$$\hat{e}_t^{l+1} = MultiHead(a_t^{l+1}, V^L, V^L), \tag{18}$$

$$\overline{e}_t^{l+1} = \hat{e}_t^{l+1} + \alpha * g_L. \tag{19}$$

Multi-Head Adaptive Controller Self-Attention. A more sophisticated method is to use the multi-head attention for fusion, which naturally fuses the local represention and the global representation by taking a weighted sum of region vectors V^L and global vector g_L. We set $V_g = (V^L; g_F) \in \mathbb{R}^{(N+1)\times d}$

$$\overline{e}_t^{l+1} = MultiHead(a_t^{l+1}, V_g, V_g) \tag{20}$$

Noticeably, attentive weights depend solely on the pairwise similarities between visual vectors (*e.g.,* region vectors and global vectors) and the query vector. In such a way, the output can capture suitable global information to refine the original local representation. Besides, the multi-head mechanism allows the model to jointly attend to information from different representation subspaces.

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:

$$p_l(Y) = \prod_{t=0}^{T} p(y_t|Y_{t-1}). \tag{21}$$

The training process consists of two phases: pre-training by supervised learning and fine-tuning by reinforcement

	B-1	B-4	M	R	C	S
SCST	-	34.2	26.7	55.7	114.0	-
Up-Down	79.8	36.3	27.7	56.9	120.1	21.4
RFNet	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM	80.5	38.2	28.5	58.3	127.6	22.0
Up-Down+HIP	-	38.2	28.4	58.3	127.6	22.0
SGAE	80.8	38.4	28.4	58.6	127.8	22.1
ETA	**81.5**	39.3	28.8	58.9	126.6	22.7
SRT	80.3	38.5	28.7	58.4	129.1	22.4
AoANet	80.2	38.9	29.2	58.8	129.8	22.4
ORT	80.5	38.6	28.7	58.4	128.3	22.6
MMT	80.8	39.1	29.2	58.6	131.2	22.6
POS-SCAN	80.2	38.0	28.5	-	125.9	22.2
CBT	-	39.0	29.1	**59.2**	128.1	**22.9**
Ours(w/ GAC)	80.8	38.8	29.0	58.6	130.5	22.4
Ours(w/ MAC)	**81.5**	**39.5**	**29.3**	58.9	**131.6**	22.8

Table 1: Comparison with the state of the art on the "Karpathy" test split, in single-model setting. All values are reported as percentage (%).

Model	B-1	B-4	M	R	C	S
Ensemble/Fusion of 2 models						
GCN-LSTM	80.9	38.3	28.6	58.5	128.7	22.1
SGAE	81.0	39.0	28.4	58.9	129.1	22.2
ETA	81.5	39.9	28.9	59.0	127.6	22.6
GCN-LSTM+HIP	-	39.1	28.9	59.2	130.6	22.3
MMT	81.6	39.8	29.5	59.2	133.2	23.1
Ours	**81.9**	**40.3**	**29.6**	**59.4**	**133.5**	**23.3**
Ensemble/Fusion of 4 models						
SCST	-	35.4	27.1	56.6	117.5	-
RFNet	80.4	37.9	28.3	58.3	125.7	21.7
AoANet	81.6	40.2	29.3	59.4	132.0	22.8
MMT	82.0	40.5	29.7	59.5	134.5	23.5
Ours	**82.1**	**40.6**	**29.8**	**59.6**	**135.1**	**23.8**

Table 2: Comparison with the state of the art on the "Karpathy" test split, using ensemble technique, where B-N, M, R, C and S are short for BLEU-N, METEOR, ROUGE-L, CIDEr and SPICE scores. All values are reported as percentage (%).

learning. Let θ be the parameters of the model. In pretraining, given a target ground truth sequence $Y^* = \{y_0^*, \cdots, y_T^*\}$, the objective is to minimize the cross-entropy loss (XE):

$$L(\theta) = -\sum_{t=0}^{T} log\big(p(y_t^* | Y_{t-1}^*)\big). \tag{22}$$

At the fine-tuning stage, we employ a variant of the self-critical sequence training approach (Rennie et al. 2017) on sequences sampled using beam search to directly optimize the metric, following previous works (Rennie et al. 2017; Anderson et al. 2018). The final gradient for one sample is calculated as:

$$\nabla_\theta L(\theta) = -\frac{1}{k}\sum_{i=1}^{k}\Big(\Big(r\big(\boldsymbol{Y}^i\big) - b\Big)\nabla_\theta \log p\big(\boldsymbol{Y}^i\big)\Big) \tag{23}$$

where $r(\cdot)$ can be any evaluation score metric, and we use the CIDEr-D score as a reward. $Y^i = \{y_0^i, \cdots, y_T^i\}$ is the i-th sentence in the beam, and $b = \Big(\sum_i r\big(Y^i\big)\Big)/k$ is the baseline, computed as the mean of the rewards obtained by the sampled sequences.

Experiments

Dataset and Implementation Details

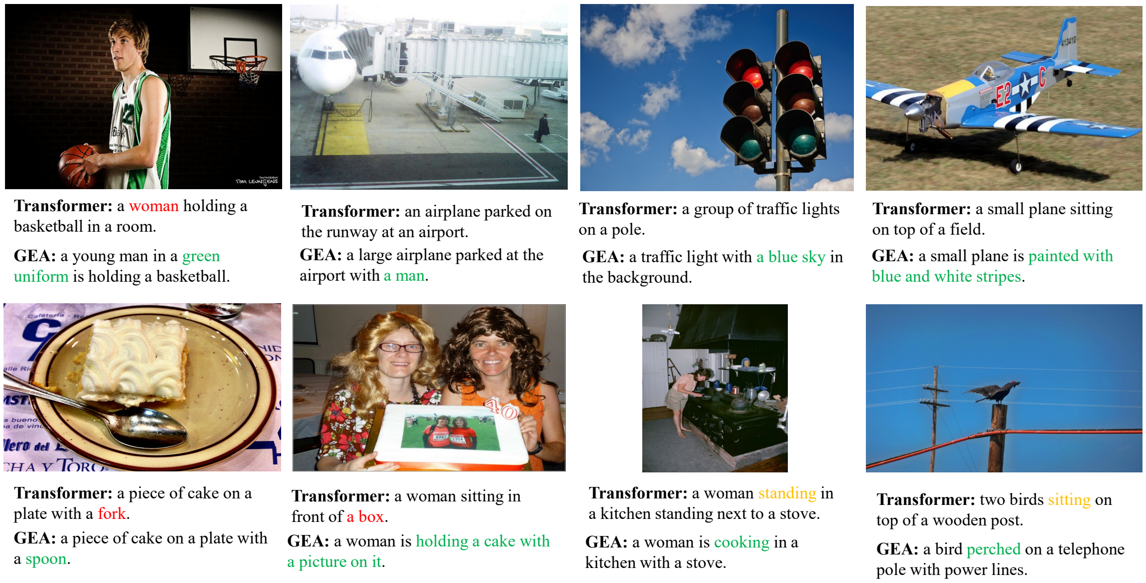
All the experiments are conducted on the most popular benchmark dataset of image captioning, i.e., MS COCO (Lin et al. 2014). The whole MSCOCO dataset contains 123,287 images, which includes 82,783 training images, 40,504 validation images, and 40,775 testing images. Each image is equipped with five 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 (Karpathy and Fei-Fei 2015) 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.

We use Faster R-CNN (Ren et al. 2015) with ResNet-101 (He et al. 2016) finetuned on the Visual Genome dataset (Krishna et al. 2017) to represent image regions. In our model, we set the dimensionality d of each layer to 512, and the number of heads to 8. We employ dropout with a keep probability of 0.9 after each attention and feed-forward layer. Pre-training with XE is done following the learning rate scheduling strategy with a warmup equal to 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 (Kingma and Ba 2014), a batch size of 50, and a beam size equal to 5. At the inference stage, we adopt the beam search strategy and set the beam size as 3. Five evaluation metrics, *i.e.*, BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE-L (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016), are simultaneously utilized to evaluate our model.

Performance Comparison

Offline Evaluation. Tab. 1 and Tab. 2 show the performance comparisons between the state-of-the-art models and our proposed approach on the offline COCO Karpathy test split. We show the performances for both the single model version and the ensemble version. The baseline models we compared include SCST (Rennie et al. 2017), LSTM-A (Yao et al. 2017), Up-Down (Anderson et al. 2018), RFNet (Ke et al. 2019), GCN-LSTM (Yao et al. 2018), SGAE (Yang et al. 2019), AoANet (Huang et al. 2019) ORT (Herdade et al. 2019), ETA (Li et al. 2019a), MMT (Cornia et al. 2020), SRT (Wang et al. 2020b), POS-SCAN (Zhou et al. 2020) and CBT (Wang et al. 2020a). We present the results of the proposed GET with two different global adaptive controllers (*e.g.,* GAC and MAC). For clarity, the symbol "ours" only represents the latter one in the following section.

Single model. In Tab. 1, we report the performance of our method in comparison with the aforementioned state-of-the-art methods, using captions predicted from a single model and optimization on the CIDEr score. Our method



Transformer: a woman holding a basketball in a room.

GEA: a young man in a green uniform is holding a basketball.

Transformer: an airplane parked on the runway at an airport.

GEA: a large airplane parked at the airport with a man.

Transformer: a group of traffic lights on a pole.

GEA: a traffic light with a blue sky in the background.

Transformer: a small plane sitting on top of a field.

GEA: a small plane is painted with blue and white stripes.

Transformer: a piece of cake on a plate with a fork.

GEA: a piece of cake on a plate with a spoon.

Transformer: a woman sitting in front of a box.

GEA: a woman is holding a cake with a picture on it.

Transformer: a woman standing in a kitchen standing next to a stove.

GEA: a woman is cooking in a kitchen with a stove.

Transformer: two birds sitting on top of a wooden post.

GEA: a bird perched on a telephone pole with power lines.

Figure 3: Examples of captions generated by our approach and the standard Transformer model. Some detailed and accurate words are marked in green, the wrong words are marked in red, and the inaccurate words are marked in yellow. Our method yields more detailed and accurate descriptions.

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	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	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	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
RF-Net	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	-	-	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	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	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	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
MMT	**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
Ours	**81.6**	**96.1**	**66.5**	**90.9**	**51.9**	**82.8**	**39.7**	**72.9**	**29.4**	38.8	59.1	74.4	**130.3**	**132.5**

Table 3: MS COCO Online Evaluation. All values are reported as percentage (%), with the highest value of each entry highlighted in boldface.

surpasses all other approaches in terms of BLEU-4, METEOR and CIDEr, and achieves competitive performance on SPICE and ROUGE-L compared to the SOTA. In particular, it advances the current state of the art on CIDEr by 0.4%.

Ensemble model. Following the common practice (Rennie et al. 2017; Huang et al. 2019) of building an ensemble of models, we also report the performances of our approach when averaging the output probability distributions of multiple and independently trained instances of our model. In Tab. 2, we use ensembles of two and four models, trained from different random seeds. Noticeably, when using four models, our approach achieves the best performance according to all metrics, with an increase of 0.6 CIDEr points with respect to the current state of the art (Cornia et al. 2020).

Online Evaluation. Finally, we also report the performance of our method on the online COCO test server. In this case, we use the ensemble of four models previously described, trained on the "Karpathy" training split. Results are reported in Tab. 3, in comparison with the top-performing approaches on the leaderboard. For fairness of comparison, they also used an ensemble configuration. As can be seen, our method surpasses the current state of the art on most of the metrics, achieving an improvement of 1.0 CIDEr points with respect to the best performer.

Qualitative Analysis. Fig. 2 shows several image captioning results of the plain Transformer and our GET. Generally, compared with the captions of the plain Transformer which are somewhat relevant to image content and logically correct, our GET produces more accurate and descriptive sentences by exploiting intra- and inter-modal interactions. For example, our GET generates the phrase of "a green uniform" and "a man", while they are missing from the plain Transformer. Besides, our GET generates more precise phrases, such as "holding a cake with a picture on it" and "cooking". These also confirm the advantage of capturing and leveraging the intra- and inter-layer global represen-

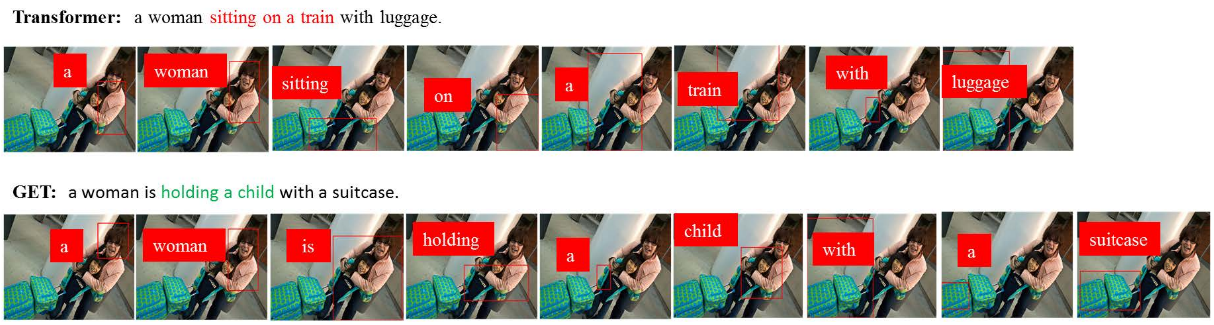

Transformer: a woman sitting on a train with luggage.
a
woman
sitting
on
a
train
with
luggage
GET: a woman is holding a child with a suitcase.
a
woman
is
holding
a
child
with
a
suitcase

Figure 4: The visualization of attended image regions along with the caption generation process for plain Transformer and the proposed GET. At the decoding step for each word, we outline the image region with the maximum output attribution in red.

Layer	BLUE-4	METEOR	ROUGE-L	CIDEr
2	38.2	28.9	58.3	129.7
3	**39.5**	**29.2**	**58.9**	**131.6**
4	39.2	29.2	58.6	130.7
5	39.0	28.9	58.5	130.3
6	39.0	29.0	58.5	130.3

Table 4: Ablation on the number of encoding and decoding layers. All values are reported as percentage (%).

tation in the Transformer architectures.

Experimental Analysis

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

Firstly, we investigate the impact of the number of the encoding and decoding layers on captioning performance for the GET. As shown in Tab. 4, varying the number of layers, we observe a slight decrease in performance when increasing the number of layers. Following this finding, all subsequent experiments uses three layers.

Then, we investigate the impact of all the proposed modules in both encoder and decoder. We choose the plain Transformer as the baseline, which is shown in the third line in Tab. 5. Then we extend the baseline model by adopting the GEA module, which slightly improves the performance. The results indicate that the GEA module can also improve the region level presentation via aggregate information from global representation. Then we investigate the impact of different global representations. As shown in the 5-th line and 6-th line, the performance improvements validate the effectiveness of GEA to obtain better presentation than the original presentation g_0 via aggregating the intra-layer information. Then we exploit different strategies to fuse the inter-layer information, and the LSTM network obtains the best performance, which basically validates the effectiveness of such layer-wise global representation. Both the GAC and MAC gain expected performance, which further indicates the effectiveness of our intra- and inter-layer global representation. And MAC is the better one, which shows that the Multi-Head mechanism works better at feature fusion for its ability of complex relationship modeling.

encoder		decoder	B-4	M	R	C
intra-layer	inter-layer					
-	-	-	37.9	28.0	57.9	128.1
GEA	-	-	38.1	28.1	58.2	128.3
g_0	-	MAC	38.2	28.3	58.0	128.6
GEA	-	MAC	38.4	28.3	58.2	128.9
GEA	average	MAC	38.5	28.7	58.1	129.4
GEA	attention	MAC	38.7	29.0	58.2	129.8
GEA	LSTM	MAC	**39.5**	**29.2**	**58.9**	**131.6**
GEA	LSTM	GAC	38.8	29.0	58.6	130.5

Table 5: Ablation on different variants of the Transformer. All values are reported as percentage (%).

Attention Visualization. In order to better qualitatively evaluate the generated results with GET, we visualize the evolutions of the contribution of detected regions to the model output along with the caption generation processes for plain Transformer and the proposed GET in Fig. 4. The contribution of one region with respect to the output is given by complex non-linear dependencies, which cannot be extracted easily. Therefore, we employ the Integrated Gradients approach (Sundararajan, Taly, and Yan 2017), which approximates the integral of gradients with respect to the given input via a summation. Results presented in Fig. 4 show that our approach can help to ground the correct image regions to words by exploring the proposed global representation.

Conclusion

In this paper, we present Global Enhanced Transformer (GET) for image captioning. GET addresses the problem of traditional Transformer-based architectures on the ignorance of global contextual information that limits the capability of reasoning in image captioning. Our model incorporates the Global Enhanced Encoder which captures both intra- and inter-layer global representation to provide more comprehensive visual information and play the role of connecting various local parts, and the Global Adaptive Decoder which adaptively fuses the global information into the decoder to guide caption generation. We show the superior performance of the proposed GET both quantitatively and qualitatively on the MS COCO datasets.

Acknowledgments

This work is supported by the National Science Fund for Distinguished Young (No.62025603), the National Natural Science Foundation of China (No.U1705262, No. 62072386, No. 62072387, No. 62072389, No. 62002305, No.61772443, No.61802324 and No.61702136) and and Guangdong Basic and Applied Basic Research Foundation (No.2019B1515120049).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv* .

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL (Workshops)*.

Cornia, M.; Stefanini, M.; Baraldi, L.; and Cucchiara, R. 2020. Meshed-Memory Transformer for Image Captioning. In *CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dou, Z.-Y.; Tu, Z.; Wang, X.; Shi, S.; and Zhang, T. 2018. Exploiting deep representations for neural machine translation. *arXiv preprint arXiv:1810.10181* .

Guo, L.; Liu, J.; Zhu, X.; Yao, P.; Lu, S.; and Lu, H. 2020. Normalized and Geometry-Aware Self-Attention Network for Image Captioning. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image Captioning: Transforming Objects into Words. In *NeurIPS*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* .

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on Attention for Image Captioning. In *ICCV*.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Ke, L.; Pei, W.; Li, R.; Shen, X.; and Tai, Y.-W. 2019. Reflective Decoding Network for Image Captioning. In *ICCV*.

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

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* .

Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019a. Entangled Transformer for Image Captioning. In *ICCV*.

Li, J.; Yao, P.; Guo, L.; and Zhang, W. 2019b. Boosted transformer for image captioning. *Applied Sciences* .

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL (Workshops)*.

Lin, M.; Ji, R.; Wang, Y.; Zhang, Y.; Zhang, B.; Tian, Y.; and Shao, L. 2020. HRank: Filter Pruning Using High-Rank Feature Map. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, F.; Liu, Y.; Ren, X.; He, X.; and Sun, X. 2019. Aligning visual regions and textual concepts for semantic-grounded image representations. In *NeurIPS*.

Liu, F.; Ren, X.; Liu, Y.; Lei, K.; and Sun, X. 2020. Exploring and distilling cross-modal information for image captioning. *arXiv preprint arXiv:2002.12585* .

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-Linear Attention Networks for Image Captioning. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*.

Song, K.; Wang, K.; Yu, H.; Zhang, Y.; Huang, Z.; Luo, W.; Duan, X.; and Zhang, M. 2020. Alignment-Enhanced Transformer for Constraining NMT with Pre-Specified Translations. AAAI.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *ICML*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE PAMI* .

Wang, J.; Xu, W.; Wang, Q.; and Chan, A. B. 2020a. Compare and Reweight: Distinctive Image Captioning Using Similar Images Sets. *arXiv preprint arXiv:2007.06877* .

Wang, L.; Bai, Z.; Zhang, Y.; and Lu, H. 2020b. Show, Recall, and Tell: Image Captioning with Recall Mechanism. In *AAAI*.

Wang, Q.; Li, F.; Xiao, T.; Li, Y.; Li, Y.; and Zhu, J. 2020c. Multi-layer representation fusion for neural machine translation. *arXiv preprint arXiv:2002.06714* .

Weng, R.; Wei, H.; Huang, S.; Yu, H.; Bing, L.; Luo, W.; and Chen, J. 2020. GRET: Global Representation Enhanced Transformer. In *AAAI*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Autoencoding scene graphs for image captioning. In *CVPR*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *ECCV*.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017. Boosting image captioning with attributes. In *ICCV*.

Zhou, Y.; Wang, M.; Liu, D.; Hu, Z.; and Zhang, H. 2020. More Grounded Image Captioning by Distilling Image-Text Matching Model. In *CVPR*.

Zhu, X.; Li, L.; Liu, J.; Peng, H.; and Niu, X. 2018. Captioning transformer with stacked attention modules. *Applied Sciences* .