LG-MLFormer: Local and Global MLP for Image Captioning

Zetao jiang

School of Computer Science and Information Security, Guilin University of Electronic Technology

Xiuxian Wang

School of Computer Science and Information Security, Guilin University of Electronic Technology

Zhongyi Zhai (✉ zhaizhongyi@guet.edu.cn)

School of Computer Science and Information Security, Guilin University of Electronic Technology

Bo Cheng

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications

Research Article

Keywords: Image captioning, reinforcement learning, artificial intelligence, Transformer, Multi-Layer Perceptrons

Posted Date: October 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2148019/v1

LG-MLFormer: Local and Global MLP for Image Captioning

Zetao Jiang[1†], Xiuxian Wang[1†], Zhongyi Zhai[1*] and Bo Cheng[2]

[1*]School of Computer Science and Information Security, Guilin University of Electronic Technology, Street, Guilin, 541004, China.
[2]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Street, Beijing, 100876, China.

*Corresponding author(s). E-mail(s): zhaizhongyi@guet.edu.cn;
†These authors contributed equally to this work.

Abstract

Self-attention-based image captioning model exists visual features?? spatial information loss problem, introducing relative position encoding can solve the problem to some extent. However, it will bring additional parameters and greater computational complexity. To solve the above problem, we propose a novel local-global MLFormer (LG-MLFormer) with specifically designed encoder module Local-global Multi-Layer Perceptron (LG-MLP). The LG-MLP can capture the latent correlations between different images and its linear stacking calculation mode can reduce computational complexity. It consists of two independent local MLP (LM) modules and a cross-domain global MLP (CDGM) module. The LM specially designs the mapping dimension between linear layers to realize the self-compensation of visual features?? spatial information without introducing relative position encoding. The CDGM module aggregates cross-domain potential correlations between grid-based features and region-based features to realize the complementary advantages of these global and local semantic associations. Experiments on the Karpathy test split and the online test server reveal that our approach provides superior or comparable performance to the state-of-the-art (SOTA). Trained models and code for reproducing the experiments are publicly available at: https://github.com/wxx1921/LGMLFormer-local-and-global-mlp-for-image-captioning.

Keywords: Image captioning, reinforcement learning, artificial intelligence, Transformer, Multi-Layer Perceptrons

1 Introduction

Image captioning is a compound research field combining computer vision and natural language processing. It aims to learn the relationships between visual and textual elements and generates sentences to describe the visual content of an image. Most current methods require an encoder-decoder framework[1–3] to describe image samples. The encoder extracts the visual features of the input image, based on which the decoder generates sentences. High-quality visual features are conductive to capture the semantic relationship. Therefore, designing a proper encoder is crucial for image captioning.

In optimizing the encoder design, convolutional neural network (CNN)-based encoders [4–6] have become the leading choices for image captioning, but their ability to capture long-distance relationship dependencies is poor. Then, attention-based models [7–10] have been widely proposed, as these methods can solve the above problem by utilizing self-attention mechanisms[11] to capture correlations between intra- and inter-modal. However, the above methods are conducted on region features, which will inevitably result in spatial information loss. Considering that, recent proposed transformer-based models introduce spatial location information, as demonstrated by the relationship-sensitive transformer (RSTNet) [12], dual global enhanced transformer (DGET) [13], and dual-level collaborative transformer (DLCT) [15]. Though they achieved significant improvement, they still have the following limitations: (1) The overall structure is still based on self-attention mechanisms, which inevitably ignores the latent correlation between images. (2) Though introducing spatial information has solved the spatial information loss problem to some extent, the structural complexity and computational complexity of the corresponding model will increase a lot. (3) Luo at el.[15] proved that utilizing dual-level self-attention mechanisms to fuse grid-based features and region-based features will improve the performance of image captioning. However, it will further increase computational complexity of the model to $O(2 \times n^2)$.

Considering that, we propose an LG-MLFormer for image captioning, which consists of Local-Global Multi-Layer Perceptrons (LG-MLP) Encoder and Cross-Domain Memory Augment Decoder (CDMAD). LG-MLP Encoder can capture the latent correlations between different images and realize the self-compensation of visual features?? spatial information global information in the encoding stage. The CDMAD can use multi-level features at low computational complexity and explore multimodal prior knowledge of vision and language in the decoding stage. In the encoding stage, due to the parameters of MLP module can be shared by parameter interaction between different linear layers in back propagation algorithm. Therefore, the latent correlation

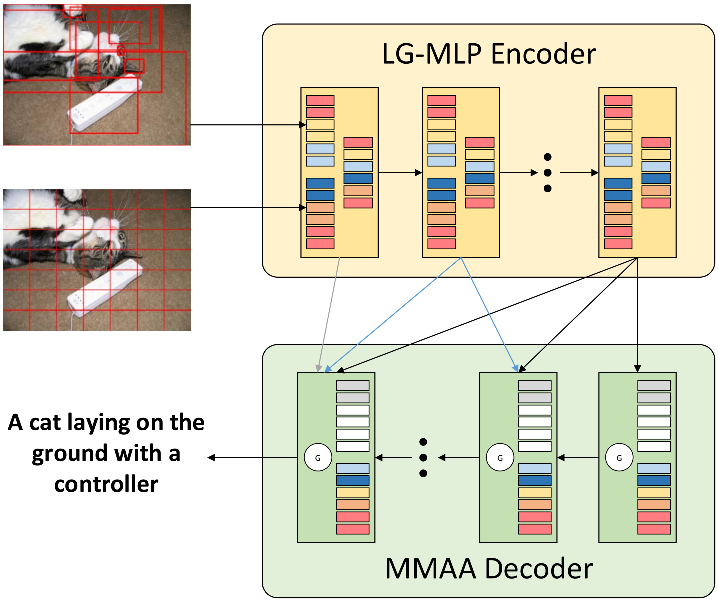


Fig. 1 Our image captioning model encodes global semantic associations exploiting learned a potential correlation between two sources of visual features. Multi-level encodings of image are connected to a text decoder through a CDC schema.

between different images can be captured by the MLP modules. As shown in Figure 1, we first explore the local latent correlation of the two source visual features separately through two independent Local MLP modules (LM). After that, the spatial information of visual features can be imitated by designing mapping dimensions between linear layers. We specifically design the mapping dimensions of the LG-MLP module to eliminate the loss of spatial information that grid and region features without introducing relative position encoding. Finally, we capture global latent semantic correlations between the grid features and the region features though a cross-domain global MLP (CDGM) module. the semantic noise of two sources features can be eliminated by capturing the global latent semantic correlations. Our LG-MLP module realize the complementary advantages of these global and local semantic associations. In additional, the LG-MLP's model structural complexity and computational complexity are reduced by simple linear stacking. In the decoding stage, we propose a vision-language memory-augmented attention (VLMA) module to integrate prior knowledge of vision and language. visual-language memory mechanism is proposed by two learning parameter matrices with different dimensions setting. To reduce the computational complexity meanwhile better exploit the multi-level features, a novel cross-domain connectivity (CDC) schema is proposed by the connection settings with increasing depth. Compare with the quadratic computational complexity of meshed schema, Our CDC requires only computational complexity at addition level and achieve superior performance.

Experimentally, we extensively evaluate different transformer-based models and some advanced proposals and intensively explore the image captioning performance of other advanced models. Our LG-MLFormer, when tested on the COCO benchmark[16], achieves a new state of the art on the "Karpathy"[17] test set.

Our contributions are summarized as follow:

- We propose a Local-Global Multi-Layer Perceptrons (LG-MLP) module to refine the representation of visual features by capturing the potential correlations between different images. At the same time, the mapping dimension between linear layers was designed to realize the self-compensation of visual features?? spatial information without introducing the extra parameter.
- We first attempt to propose a novel encoder-decoder framework, named MLFormer, in which the encoder is based on all-mlp structure, and the decoder is based on self-attention. The MLFormer provide new feasible solutions for image captioning.
- We propose an Visual-Language Memory-augmented attention (VLMA) module to integrate prior knowledge of vision and language for the prediction of sequence words.
- We provides superior results to those of the state-of-the-art (SOTA) models on the COCO benchmark dataset by combining the LG-MLP module and the VLMA module into our LG-MLFormer through the CDC schema.

2 Related works

2.1 All-MLP Models

Convolutional neural network (CNN) have become leading choices, and attention-based networks have become popular in the past few years. However, the CNN has poor ability to capture long-distance dependencies, and self-attention mechanisms ignore the potential correlations between different images. To address these problems, Ilya et al. [18] propose an MLP-Mixer, which uses a simple architecture based exclusively on MLPs to achieve comparable performance to that of the current best CNN and attention models[19–22]. The MLP-Mixer showed that while convolutions and attention are sufficient for good performance, neither is necessary. Meng-hao et al.[23] proposed a novel all-MLP module, named external attention, to replace self-attention mechanisms for visual tasks and overcome the disadvantages of ignoring the external associations among multiple images; Additionally, Hanxiao et al. [24] proposed an all-MLP structure module with a spatial gating unit (called the GMLP), which achieved comparable or superior performance to that of self-attention in natural language understanding and major vision tasks. All-MLP models[18, 24, 25] provide new feasible solutions for optimising the image feature extracted.

Although the all-MLP models have achieved superior performance results in computer vision, hey have the following limitations: Due to the parameter dimension between the linear layers being fixed, so the MLP cannot solve the spatial information loss of visual features. With the aim of filling this gap,we propose a LG-MLP to realize the self-compensation of visual features?? spatial information without introducing extra parameters.

2.2 Image captioning

In recent years, methods based on neural network image captioning (NIC) [5, 6, 26] have been widely proposed. They have achieved better performance than other approaches for image captioning tasks by using the encoder-decoder paradigm [1, 4, 27–29], where a CNN is used as the encoder to extract visual features, and an RNN is utilised as the decoder to generate sentences based on predicted words. This pattern produces a more accurate and vivid caption for the input image than earlier template-based methods [30–32]. To further optimise the extracted visual features, Anderson et al. [3] use an object detector to extract the region-based visual features. After that, Cornia et al. [8]employed memory-augmented self-attention to exploit the self-affinities between the different regional features of a single image and utilise the prior knowledge.

Although region-based visual features have been widely adopted for image captioning, regional features suffer from a lack of contextual information and fine-grained details. After revisiting the visual grid features [33] extracted by a pretrained detector, new models using these grid features have achieved SOTA performance in image captioning tasks. For example, Zhang at el.[12] introduced a genetic algorithm GA module into a fully attentive encoder to overcome the disadvantages of grid feature flattening by integrating spatial information. Luo at el.[15] proposed a locality-constrained cross-attention mechanism to eliminate the semantic noise produced when utilizing grid features and regional features simultaneously. In this paper, we exploited the potential semantic associations which grid-based features and region-based features by all-MLP structure module, and use special parameter designs to effectively solve the problem of geometric information loss caused by input flattening of visual features without relative position information, and get linear computational complexity by using simple linear stacks.

3 Method

In this section, we introduce our Local-Global MLFormer architecture for image captioning, which consists of Local-Global Multi-Layer Perceptrons (LG-MLP) Encoder and Cross-Domain Memory Augment Decoder. In the encoding stage, the LG-MLP can realize the self-compensation of visual features?? spatial information without introducing extra parameters and its explore the potential correlations between different images to extract visual features of richer potential semantic information. In the decoding stage, the

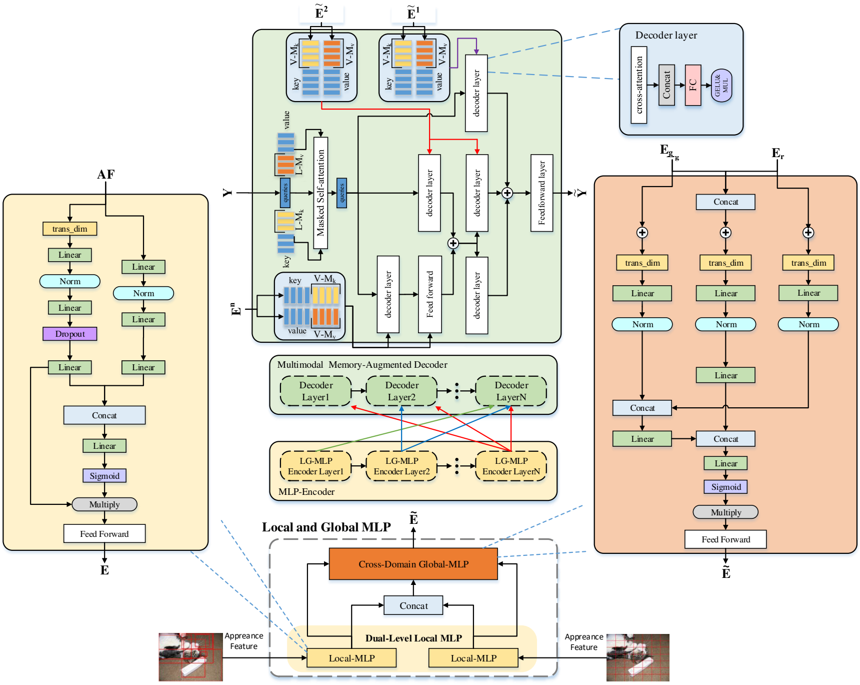


Fig. 2 Architecture of the LG-MLFormer. Our model is composed of a stack of LG-MLP encoding layers, which encodes multi-level global semantic associations with a potential correlation between two sources of visual features, and a stack of multimodal memory-augment decoder layers, in charge of generating word tokens with a visual-language prior knowledge.

Visual-Language Memory-augmented attention (VLMA) module integrates the prior knowledge of vision and language to generate the output caption word by word. The architecture of LG-MLFormer is shown in Figure 2.

3.1 LG-MLP Encoder

Although the self-attention mechanism eliminates the semantic noise between grid features and regional features, the self-attention mechanism has the following defects : (1) self-attention ignores the potential correlations between different images. (2) Self-attention mechanism has quadratic complexity. (3) To capture the spatial geometric information of the grid-based feature, the self-attention needs to introduce relative position encoding led to the model becoming more complex and computationally expensive. To solve the above problems, we propose a LG-MLP module as a visual encoder. Our LG-MLP consist of two independent local MLP modules and a cross-domain global MLP module, which are introduced in 3.11, 3.12.

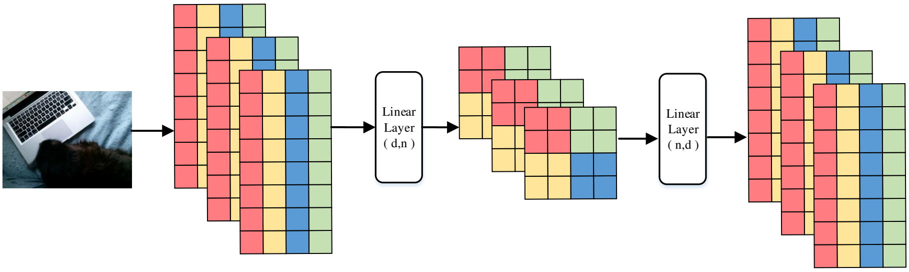


Fig. 3 Operating principle of Local MLP module. Spatial information of images can be simulated by adjusting the shape of feature vectors from one dimension to two dimensions. Our LM module designs the linear mapping dimensions between linear layers according to the number of visual regions for realizing the self- compensation of visual features?? spatial information without introducing extra parameters.

3.1.1 Local MLP

To realize the self-compensation of visual features' spatial information without introducing extra parameters and exploit the potential correlations between different images. As shown in Fig 3, our Local MLP (LM) module specially designs the mapping dimension between linear layers and leverages two linear layers to replace self-attention layers. Its operating principle is as follows.

Given a set of region-based features or grid-based features X extracted from an input image. The LM module captures the potential correlations between different images to acquire visual feature representations of richer potential semantic information. Since the parameter weights of the linear layer can be trained via a backpropagation algorithm for optimization purposes, these weights can be shared globally during the training process. This significantly improves the ability of the resulting model to capture the potential correlations between multiple image samples. At the same time, the model got the linear computational complexity by using simple linear stacks. Therefore, we first introduce two linear layer parameter weight vectors to replace the key and value vectors of the classic self-attention mechanism. After that, due to spatial information of images can be simulated by adjusting the shape of feature vectors from one dimension to two dimensions, so we design the linear mapping dimensions between linear layers according to the number of visual regions for realizing the self-compensation of visual features' spatial information without introducing extra parameters. Finally, we set a gate mechanism to improve the effectiveness of the LM module. The above operation can be defined as:

$$\begin{aligned} FC(X) &= WX + b \\ mlp(X) &= FC(Norm(FC(X))) \end{aligned} \tag{1}$$

$$LM(X) = \sigma(FC([W_1 mlp(X), W_2 mlp(X)])) \otimes W_1 mlp(X) \tag{2}$$

where the [,] indicates concatenation, $\sigma()$ is sigmoid activation function, $\otimes$ is matrix multiplication.

3.1.2 Cross-Domain Global MLP

Recently, the attention-based model has modelled the spatial geometric relationship between the visual region of an image by introducing relative position encoding. The semantic noise between region feature and grid feature is effectively solved cross-attention mechanism. Although the above method has achieved success for image captioning, it also increases unnecessary parameters and computational overhead. Thus, we propose a Cross-Domain Global MLP (CDGM) module to eliminate the semantic noise though exploring the cross-domain potential correlations between grid-based features and region-based features, and it achieves better performance without introducing external parameters. Its operating principle is as follows.

First, We integrate the output of the LM module through operation of concatenation. After that, the original features and integrated features are sent into the same abstract space by linear activation operation. Finally, we use a gate mechanism to weight local and global semantic contributions. The local and global semantic correlations can be aligned by the value of the weight to eliminate semantic noise. The above operation can be defined as:

$$\begin{aligned}\tilde{E_r} &= Norm(FC(E_r))\\ \tilde{E_g} &= Norm(FC(E_g))\\ \tilde{E} &= Norm(FC([E_r, E_g]))\end{aligned} \tag{3}$$

$$CDGM(\tilde{E}, \tilde{E_r}, \tilde{E_g}) = \sigma([FC([\tilde{E_r}, \tilde{E_g}]), FC(\tilde{E})]) \otimes FC([\tilde{E_r}, \tilde{E_g}]) \tag{4}$$

where the E_r and E_g are the two-source features with potential semantic associations extracted by the LM module, and [,] indicates the concatenation, operation $\sigma()$ is sigmoid activation function, $\otimes$ is matrix multiplication.

3.1.3 Encoding layer

We embed the local and global MLP modules into an LG-MLP encoding layer. The output of the encoder is applied to a feedforward position layer, which performs a nonlinear affine transformation on each element of the set. The formula for the feedforward position layer is defined as follows:

$$F(X)_i = V_2\sigma(V_1X_i + b_1) + b_2 \tag{5}$$

where X_i represents the i-th vector of the input set and $F(X)_i$ represents the output of the i-th vector of the output-input set. In addition, $\sigma()$ is the rectified linear unit (ReLU) activation function operation, V_1 and V_2 are learning weight matrices, and b_1 and b_2 are deviation terms.

After that, the LG-MLP and feedforward position layer is encapsulated by residual connection and hierarchical standardization operations to form the overall architecture of the coding layer. The operation formula of the encoder layer is defined as follows:

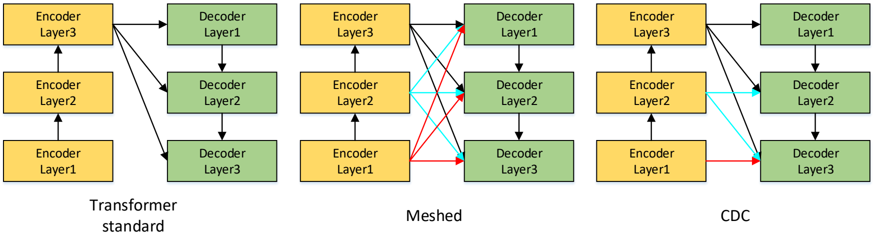


Fig. 4 Connection pattern comparison diagram. Compare with the quadratic computational complexity of meshed schema, Our CDC requires only computational complexity of addition level and achieve superior performance.

$$LGMLP(X_r, X_g) = CDGM(LM(X_r), LM(X_g)) \tag{6}$$

$$\tilde{E} = AddNorm(LGMLP(X_r, X_g)) \tag{7}$$

where AddNorm represents the joint residual connection and hierarchical standardization operation.

Finally, multiple encoder layers are linear stacked so that the input of the i-th encoder layer is the output set calculated by encoder layer i-1. The high-level layer can learn and fine the feature representation of the previous layer. Therefore, the stack of N encoder layers produces multilevel outputs from the outputs of each encoder layer.

3.2 Cross-Domain Memory Augment Decoder

In this paper, the decoder exploits the multilevel visual feature extracted by the encoder to generate a sentence in a word-by-word manner. The decoder adopts a multilayer structure to utilize the high-level and low-level visual features better. In the deep neural network model, the influence of high-level sub-modules on the performance of the model is higher than that of the low-level sub-modules. The information of low-level features is diluted through multi-level feature processing. Therefore, information redundancy be produced by the meshed schema. As shown in Fig 4, to reduce the computational complexity meanwhile better exploit the multi-level features, a novel cross-domain connectivity (CDC) schema is proposed by the connection settings with increasing step by step. Compare with the quadratic computational complexity of meshed schema, Our CDC requires only computational complexity of addition level and achieve superior performance. and its computational complexity formula is as follows:

$$L(N) = \frac{N(N+1)}{2} \tag{8}$$

where N is the number of decoder layers.

3.2.1 Visual-Language Memory Augmented Attention

The defined learning parameter matrix is used as the memory mechanism in the training of the self-attention mechanism. By integrating the memory mechanism into the key and value of the self-attention mechanism, the self-attention operation is used to make it fully learn the visual information in the training set, and the learned prior knowledge is fed back to the query through the dot product operation. the trained memory mechanism can effectively provide prior knowledge for the model. However, previous work often uses a single visual memory mechanism to provide visual prior knowledge for the model, ignoring the use of prior knowledge of text features. In view of the fact that the vector dimension of the text feature is often smaller than that of the visual feature in image captioning task. To better adapt to the input of two different modes of text and image, this paper designs two learning parameter matrices with different dimensions, in which the parameter matrices with small dimension is used for textual memory mechanism, and the parameter matrices with large dimension is used for visual memory mechanism. In the decoding stage, the textual memory mechanism provides prior knowledge for self-attention computations for the model corpus samples, and the visual memory provides prior knowledge to better use visual semantic correlations. Its operating principle is as follows.

First, the textual memory-augmented attention mechanism captures the long-distance relationships between the words in the input text sequence, generates a word vector feature sequence Y, and integrates the prior knowledge derived from the corpus sample. After that, the visual memory-augmented attention mechanism connects the outputs of all encoder layers E and word vector feature sequence Y through the CDC. Unlike the transformer, which uses only the output of the last coding layer, the decoding layer of our method performs cross-attention operations on different encoder layers in a deep incremental mode to exploit multilevel features. The VLMA based on CDC can be defined as:

$$CDC(\tilde{E}, Y) = \sum_{i=1}^{N} \sum_{j=N+1-i}^{N} \alpha_j \otimes VLMA(\tilde{E}_j, Y) \tag{9}$$

where VLMA represents the VLMA module, which is computed by using queries obtained from the output of the visual-language memory-augmented attention module and keys and values from the encoder:

$$\begin{aligned} VLMA(\tilde{E}_j, Y) &= softmax\left(\frac{W_q Y \cdot \tilde{K}_j}{\sqrt{d}}\right) \cdot \tilde{V}_j \\ \tilde{K}_j &= [W_k \tilde{E}_j, \tilde{M}_k^j] \\ \tilde{V}_j &= [W_v \tilde{E}_j, \tilde{M}_v^j] \end{aligned} \tag{10}$$

where M_k and M_v are learnable weight matrices with n rows, and [,] indicates the concatenation operation. Keys and values can be learned by adding two learning memory units, thereby providing prior knowledge for the attention operation. α is a weight matrix. This weight realizes the contribution to each coding layer and the modulation of priorities. These functions are calculated by measuring the correlation between the cross-attention results and input queries. The formula is defined as follows:

$$\alpha_j = gelu(W_j[Y, VLMA(\tilde{E}_j, Y)] + b_j) \tag{11}$$

where GELU is the Gaussian error linear unit activation function, [,] denotes the concatenation operation, W_j is a weight matrix with dimensions of 2d ?? d, and b_j is a learning bias vector.

3.2.2 Decoding layer

Regarding the decoder, during the image captioning generation stage, since the prediction of the t-th word depends entirely on the t-1th predicted word, the self-attention process in the decoder layer requires a mask operation, where Y=t. In addition, the decoding layer encapsulates the memory-augmented attention mechanism and the feedforward position layer via the AddNorm operation (as shown in formulas 6 and 7 above). The operation formula of the decoder layer is defined as follows:

$$\tilde{Y} = AddNorm(CDC(\tilde{E}, AddNorm(VLMA_{mask}))) \tag{12}$$

where Y is the word vector sequence and $VLMA_{mask}$ is the visual-language memory augmented attention operation with a mask.

3.3 Training Details

According to the standard image captioning process, we use the two-stage training mode to train our model, in which the first stage pretrains our model by using the cross-entropy loss function (XE), and the second stage fine-tunes the text generation ability of our model by using a reinforcement learning algorithm.

During the first stage of training, according to the previously provided ground truth words, the model's ability is optimized; the model predicts the next token by utilizing the cross-entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log(p(y_t^* \| y_{1:t-1}^*)) \tag{13}$$

Therefore, the decoder can realize the sequential parallel operation for the input sequence so that the calculation of the whole output sequence can be completed simultaneously. During the second stage of training, we optimize our model so that it can obtain a consensus-based Image Description Evaluation (CIDEr-D) score via self-critical sequence training (SCST)[34]:

$$b = \frac{1}{k}(\sum_{j}^{k} r(y_{1:T}^{j})), \tag{14}$$

$$\nabla_\theta L_R(\theta) = -\frac{1}{k}\sum_{j=1}^{k}(r(y_{1:T}^{j}) - b)\nabla_\theta \log p(y_{1:T}^{j}), \tag{15}$$

where $y_{1:T}^{j}$ is the j-th sampled caption, k is the beam size of the beam search process[3] and r defines the baseline score (the reward is calculated by using the mean of all rewards).

4 Experiment setup

4.1 Dataset

We use the MS-COCO dataset[16] to evaluate our LG-MLFormer model. The dataset contains 123,287 images, which are divided into 82,783 training images, 40,504 validation images, and 40,775 test images in a dense manner. Each image possesses 5 different captions. We employ the split provided by Karpathy et al.[17] , where 5,000 images are used for validation, 5000 images are utilized for testing, and the remaining images are employed for training. We also evaluate the model on the COCO online test server, for which annotations are not made publicly available.

Following the standard evaluation protocol, we use the full set of captioning metrics, including BLEU [35], METEOR [36], ROUGE [37], CIDEr [38], and SPICE [39], to evaluate the achieved image captioning quality. Extensive experiments on the Karpathy test split and on an online test server reveal that our approach provides comparable or superior results to those of the state-of-the-art (SOTA) models.

4.2 Image Features

We used two sets of image features in the experiments.

(a) Grid features extracted from Faster-RCNN with ResNext101, which the grid features are leveraged by average pooling them to a 7 × 7 grid size, the same as the representation used in RSTNet[12].

(b) Region features use Faster-RCNN with ResNext101 to extract 2048-dimensional features after the first fully connected (FC) layer of the detection head, the same as the representation used in M2 Transformer[8]. We used these features as the input of LM module and the concat of the two features as the feature of the whole image.

(c) Cross-modal features, Kuo et al[48]. use the CLIP[49] to retrieve a contextual description, which represents a visual relationship, as a supplemental input for representing missing information, due to all necessary information in the image can not be fully represented by the methods based on the visual region features.

4.3 Implementation Details

To eliminate the semantic noise of two sources of visual features and improve the semantic associations between different features, we propose a novel all-MLP structure consisting of a LM module and a CDGM module as an image feature encoder. Specifically, an input image I is represented as grid and regional features by the Faster-RCNN with ResNext101, in which the grid features are leveraged by average pooling them to a 7×7 grid size, and the region features use the same model to extract 2048-dimensional features after the first fully connected (FC) layer of the detection head. In our approach, we set the d_{model} of each layer to 512, the number of heads to 8, and the numbers of visual and textual memory vectors to 40 and 20, respectively. Then, during CIDEr-D optimization, we use a baseline lambda reinforcement learning rate is 5×10^{-6}, and the lambda reinforcement learning policy is defined as follows:

$$lam\_rl\_lr = \begin{cases} base\_lr, & e <= 28 \\ base\_lr * 0.1, & 28 < e <= 40 \end{cases} \tag{16}$$

where the initial $base\_lr$ is baseline lambda reinforcement learning rate, e is the current epoch number (starting from 19).

5 Result

5.1 Comparison of Performance-Parameters Ratio

To intuitively show the high efficiency of our LG-MLFormer in improving model performance, we fully verify the performance-parameter ratio of the LG-MLFormer and state-of-the-art models. As shown in Figure 5, Our LG-MLFormer achieve a excellent performance with fewer parameters. Based on the above experimental results, the following conclusions can be drawn: 1) Our LG-MLFormer can utilize the external correlations between grid and region features to improve the performance of the model. 2) The ingenious structure design of LG-MLFormer enables it to achieve comparable performance with the large-scale model at a low parameter overhead.

5.2 Ablation Study

To fully validate the impacts of our proposed LG-MLP, CDC, and VLMA on model performance and explore their relationship, we perform ablation experiments to compare the different deformations of LG-MLFormer. Firstly, to demonstrate the advantages of the all-MLP structure over the self-attention architecture, we use the transformer as the baseline model. Secondly, we replace the encoder of the self-attention mechanism with the all-MLP encoder built by the LG-MLP module. After that, encoding and decoding layers are connected in the CDC schema. Finally, we combine the VLMA module and LG-MLP module into our LG-MLFormer through the CDC schema. The above experiment is conducted on grid-based features, region-based features and cross-modal features, and the results are shown in Table1, Table2 and Table3.

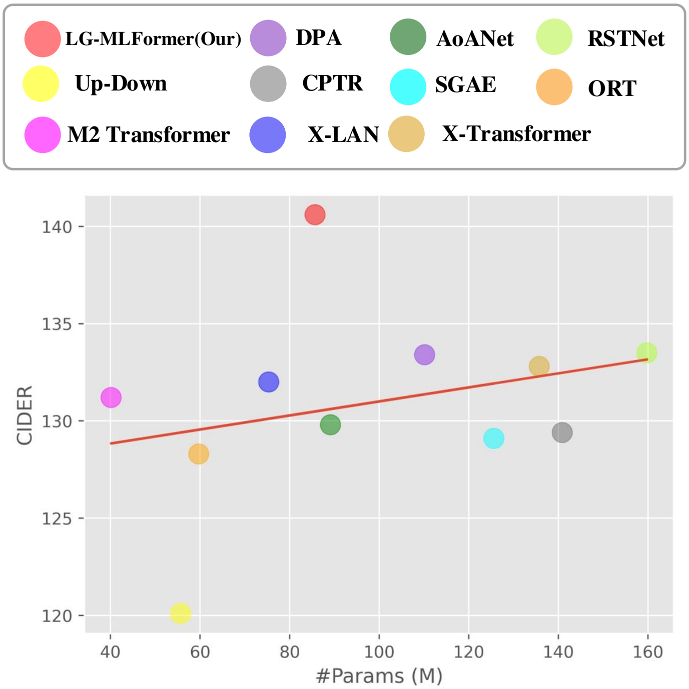


Fig. 5 Comparison of our LG-MLFormer and existing approaches in terms of number of parameters and caption quality.

Table 1 Ablation study on the MSCOCO karpathy test set features extracted from ResNext101 features

LG-MLP	CDC	VLMA	B@1	B@4	M	R	C	S
✗	✗	✗	80.9	38.9	29.0	58.5	131.2	22.7
✓	✗	✗	81.0	39.1	28.9	58.8	132.6	22.8
✓	✓	✗	81.1	39.3	28.8	58.9	132.8	22.8
✓	✓	✓	**81.7**	**39.7**	**29.4**	**59.1**	**133.9**	**23.0**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

In the ResNext101 experiments, the reason for the observed elevation in model performance obtained when utilizing the LG-MLP module is that this module can capture the potential correlations between different images, making it possible to acquire visual features of rich potential semantic associations. Additionally, the LG-MLP module can eliminate the semantic noise by exploring the potential global correlations between grid-based features and region-based features. Finally, we achieve a larger improvement (from 131.2

Table 2 Ablation study on the MSCOCO karpathy test set region-based features extracted from ResNext101 features.

LG-MLP	CDC	VLMA	B@1	B@4	M	R	C	S
✗	✗	✗	79.1	36.2	27.7	56.9	121.8	20.9
✓	✗	✗	80.9	38.9	28.6	58.6	131.0	22.8
✓	✓	✗	80.9	38.9	28.8	58.8	131.4	22.8
✓	✓	✓	**81.0**	**39.1**	**29.1**	**58.8**	**132.1**	**22.9**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

Table 3 Ablation study on the MSCOCO karpathy test set cross-modal features extracted from CLIP.

LG-MLP	CDC	VLMA	B@1	B@4	M	R	C	S
✗	✗	✗	82.2	40.0	29.9	59.4	137.7	23.3
✓	✗	✗	82.8	40.9	29.9	59.8	139.9	23.7
✓	✓	✗	83.0	40.6	30.1	60.2	140.3	24.0
✓	✓	✓	**83.2**	**41.5**	**30.4**	**60.7**	**140.6**	**24.2**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

to 133.9) when incorporating the LG-MLP and the VLMA modules into our model through the CDC schema.

5.3 Offline Evaluation Analysis

We compare the image captioning performance of our LG-MLFormer with that of the SOTA models on the offline COCO Karpathy test split in Table 4. The comparison models include SCST [34], which uses attention over the input grid of features and advances the self-critical training policy. Additionally, we compare the Up-Down method [3] and the region-aware fusion network (RFNet) [40], which both use attention over the observed regional features to optimize the image captioning abilities of the resulting models. A graph convolutional network with long short-term memory (GCN-LSTM) [41], which uses a GCN to utilize the pairwise relationships between image regions for refining the semantic representation of the input image, is also compared. The scene-graph autoencoder (SGAE) [42], which acquires features with rich sematic information by using auto-encoding scene graphs, is tested. ORT [7], which was the first attempt to introduce the transformer architecture into image captioning tasks, exploits the spatial relationships between detections with a region encoder. Furthermore, we compare our approach with the attention-on-attention network (AoANet) [43] and M^2 Transformer [8] approaches, which respectively use attention to optimize visual encodings and memory vectors

Table 4 Performance comparison with the state of the art on the MSCOCO Karpathy test split

Model	B@1	B@4	M	R	C	S
UP-Down[3]	75.0	37.3	28.1	57.9	123.8	21.6
RFNet[40]	79.1	36.5	27.7	57.3	121.9	21.2
GCN-LSTM[41]	80.5	38.2	28.5	58.3	127.6	22.0
SGAE[42]	80.8	38.4	28.4	58.6	127.8	22.1
ORT[7]	80.5	38.6	28.7	58.4	128.3	22.6
Transformer[8]	80.9	38.9	29.0	58.5	131.2	22.7
AoA Transformer[43]	80.8	39.1	29.1	59.1	130.3	22.7
M^2 Transformer[8]	80.8	38.9	29.1	58.5	131.8	22.7
X-Transformer[10]	81.0	39.5	29.1	59.0	130.2	22.8
CoSA-Net[14]	-	39.0	29.0	58.7	129.5	22.5
RSTNet[12]	81.1	39.3	29.3	58.8	133.3	23.0
GET[9]	81.5	39.5	29.3	58.9	131.6	22.8
CTX+M2[48]	81.5	39.7	30.0	59.5	135.9	23.7
BLIP[46]	-	39.7	-	-	133.3	-
LEMON[47]	-	**41.5**	30.8	-	139.1	24.1
CaMEL[45]	82.7	40.9	30.3	60.1	138.9	**24.5**
LG-MLFormer(Ours)	**83.2**	**41.5**	**30.4**	**60.7**	**140.6**	24.2

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

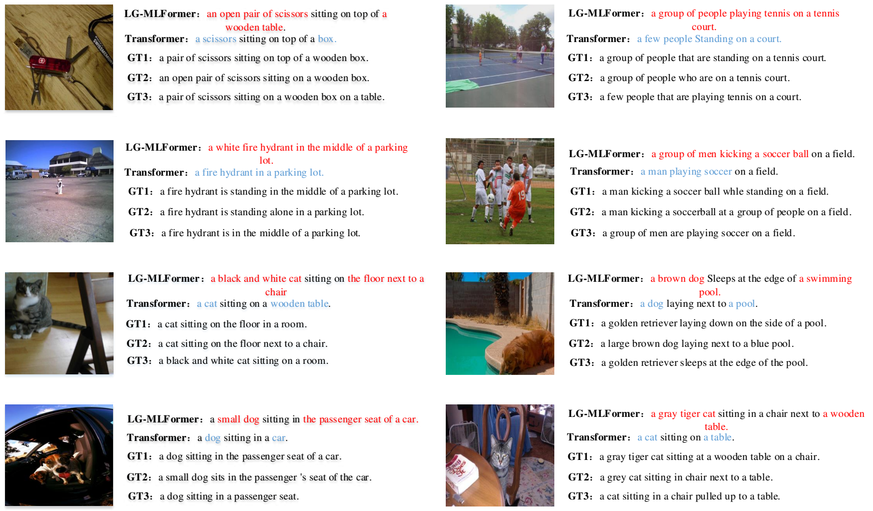
to provide prior knowledge for visual encoding. X-Transformer[10] introduces bilinear pooling into the attention module of a base transformer. we compare our network with the RSTNet [12], which use attention over the grid features with absolute and relative position encodings to overcome the disadvantage regarding the loss of spatial information. Finally, we compare our network with the GET[9], CTX+M2[48], BLIP[46], LEMON[47] and CaMEL[45]. Our approach provides superior results to those of the SOTA models in terms of most evaluation metrics.

5.4 Comparison with advanced baseline

Table 5 Comparing with SOTAs on ResNext101 grid-based and region-based features

Model	B@1	B@4	M	R	C	S
UP-Down[3]	75.0	37.3	28.1	57.9	123.8	21.6
Transformer[8]	80.9	38.9	29.0	58.5	131.2	22.7
AoA Transformer[43]	80.8	39.0	29.1	59.0	130.3	22.7
M^2 Transformer[8]	80.8	38.9	29.1	58.5	131.8	22.7
X-Transformer[10]	81.0	39.6	29.1	59.0	130.2	22.8
RSTNet[12]	81.1	39.3	**29.4**	58.8	133.3	**23.0**
LG-MLFormer(Ours)	**81.7**	**39.7**	**29.4**	**59.1**	**133.9**	**23.0**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores



LG-MLFormer： an open pair of scissors sitting on top of a wooden table.
Transformer： a scissors sitting on top of a box.
GT1： a pair of scissors sitting on top of a wooden box.
GT2： an open pair of scissors sitting on a wooden box.
GT3： a pair of scissors sitting on a wooden box on a table.

LG-MLFormer： a white fire hydrant in the middle of a parking lot.
Transformer： a fire hydrant in a parking lot.
GT1： a fire hydrant is standing in the middle of a parking lot.
GT2： a fire hydrant is standing alone in a parking lot.
GT3： a fire hydrant is in the middle of a parking lot.

LG-MLFormer： a black and white cat sitting on the floor next to a chair
Transformer： a cat sitting on a wooden table.
GT1： a cat sitting on the floor in a room.
GT2： a cat sitting on the floor next to a chair.
GT3： a black and white cat sitting on a room.

LG-MLFormer： a small dog sitting in the passenger seat of a car.
Transformer： a dog sitting in a car.
GT1： a dog sitting in the passenger seat of a car.
GT2： a small dog sits in the passenger 's seat of the car.
GT3： a dog sitting in a passenger seat.

LG-MLFormer： a group of people playing tennis on a tennis court.
Transformer： a few people Standing on a court.
GT1： a group of people that are standing on a tennis court.
GT2： a group of people who are on a tennis court.
GT3： a few people that are playing tennis on a court.

LG-MLFormer： a group of men kicking a soccer ball on a field.
Transformer： a man playing soccer on a field.
GT1： a man kicking a soccer ball whle standing on a field.
GT2： a man kicking a soccerball at a group of people on a field.
GT3： a group of men are playing soccer on a field.

LG-MLFormer： a brown dog Sleeps at the edge of a swimming pool.
Transformer： a dog laying next to a pool.
GT1： a golden retriever laying down on the side of a pool.
GT2： a large brown dog laying next to a blue pool.
GT3： a golden retriever sleeps at the edge of the pool.

LG-MLFormer： a gray tiger cat sitting in a chair next to a wooden table.
Transformer： a cat sitting on a table.
GT1： a gray tiger cat sitting at a wooden table on a chair.
GT2： a grey cat sitting in chair next to a table.
GT3： a cat sitting in a chair pulled up to a table.

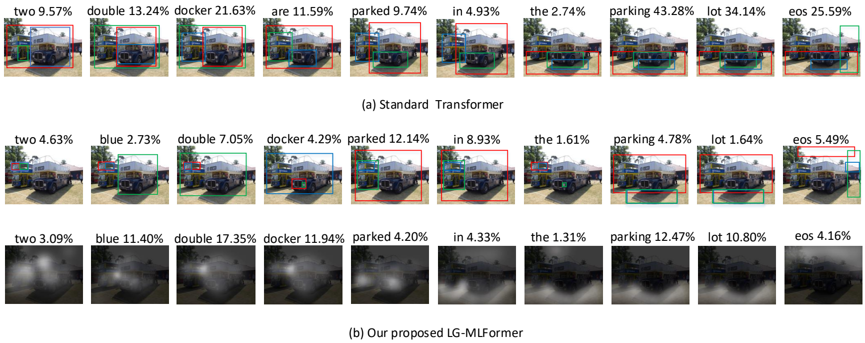
Fig. 6 Examples of image captioning results by original Transformer and our LG-MLFormer with ground truth sentences. Original Transformer and our LG-MLFormer use the same features when generate the captions.

Table 6 Comparing with SOTAs on ResNext101 grid-based and region-based features

Model	B@1	B@4	M	R	C	S
UP-Down[3]	79.8	36.3	27.7	56.9	120.1	21.4
Transformer[8]	79.1	36.2	27.7	56.9	121.8	20.9
ORT[7]	80.5	38.6	28.7	58.4	128.3	22.6
AoA Transformer[43]	80.2	38.9	**29.1**	**58.8**	129.8	22.4
M^2 Transformer[8]	80.8	**39.1**	**29.1**	58.6	131.2	22.6
LG-MLFormer(Ours)	**81.0**	**39.1**	**29.1**	**58.8**	**132.1**	**22.9**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

To prove that the excellent performance of our model does not depend entirely on the visual features extracted by ResNext101, we also conduct experiments to compare our proposed LG-MLFormer with the SOTAs on the two sources features extracted by ResNext101 in Table 5 and 6. In the experimental stage, we set the d_{model} parameter to 512 in all models, and choose the 50 epoch results of self-critical training for displaying. the above experiments result demonstrate that our LG-MLFormer can acquire a superior result when comparing with other SOTA methods under the same visual features and architecture configuration.



(a) Standard Transformer

(b) Our proposed LG-MLFormer

Fig. 7 visualization of attention states for region-based Transformer (a) and our DLCT (b).For each generated word, we show top-3 attended regions (red, blue, green respectively) and the attention heatmap on grids (only available in LG-MLFormer) with the highest attention weight in the title.

5.5 Online Evaluation Analysis

Finally, we also report the performance of our method on the online COCO test server. The model we use is an ensemble of 4 LG-MLFormer models trained on the Karpathy training split. Table 7 report the results of our LG-MLFormer in comparison with those of other high-performance models on the leaderboard. The official testing images are divided into c5 and c40, among which c5 is an image with 5 reference captions and c40 is an image with 40 reference captions. Extensive experiments on the online test server reveal that our MLFormer architecture provides superior results to those of the state-of-the-art (SOTA) models.

5.6 Qualitative results and visualization

Fig 6 displays some results generated by our LG-MLFormer and the original transformer. In general, compared with those output by the original transformer, the captions generated by our LG-MLFormer are more accurate and comprehensive.

To fully and qualitatively evaluate the visual representation of encoding, we visualize the contribution of each visual feature to the model output, as shown in Fig 7. The principle is that we averaged self-attention weights of 8 heads in the last Encoder-Decoder multi-head attention layer. In addition, compared with the standard Transformer model, our LG-MLFormer can focus on the corresponding grid when generating the word "blue". When generating the word "parking", the attention heatmap on grids provides a more fine-grained semantic segmentation of parking, which shows the advantages of our LG-MLFormer.

Table 7 Leaderboard of various methods on the online MS-COCO test server.where B@N are short for BLEU@N scores

	B@1		B@2		B@3		B@4		METEOR		ROUGE-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [34]	78.1	93.7	61.9	86.0	47.9	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [3]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [40]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	37.1	122.9	125.1
GCN-LSTM [41]	80.8	**95.9**	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 [42]	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
ETA [44]	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
AoANet [43]	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
CoSA-Net[14]	81.0	95.4	65.7	89.9	51.0	81.2	39.0	70.9	28.8	38.3	58.8	74.1	126.2	128.5
DGET [13]	80.8	95.1	65.6	89.5	51.2	80.8	39.2	70.4	28.9	38.2	58.8	74.4	126.3	129.2
LG-MLFormer(Ours)	**81.6**	**95.9**	**66.0**	**90.5**	**51.4**	**82.2**	**39.4**	**72.1**	**29.1**	**38.6**	58.8	74.2	**128.3**	**131.1**

Note: B@1, B@4, M, R, C, and S are short for BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr, SPICE scores

6 Conclusion

In this paper, we present LG-MLFormer, the first attempt to build an all-MLP model to construct an MLFormer architecture model for image captioning. We propose an LG-MLP module consisting of two independent local MLP (LM) modules and a cross-domain global MLP (CDGM) to improve the ability to capture the potential correlations between different images and realize the complementary advantages of the global and local semantic associations between grid-based features and region-based features. On the one hand, our LG-MLFormer specially designs the mapping dimension between linear layers to realize the self-compensation of visual features?? spatial information without introducing relative position encoding. On the other hand, our LG-MLFormer eliminates the semantic noise between two source features by capturing the potential semantic correlations between grid-based features and region-based features. In addition, our LG-MLFormer introduces two types of memory vectors with visual and language modes to utilize multimodal prior knowledge to generate sentence outputs. Its linear stacking calculation mode can reduce computational complexity. Extensive experiments on the MS-COCO dataset and the online test server reveal that our approach provides superior or comparable results to the SOTA models.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants (62172118, 61876049) and Nature Science key Foundation of Guangxi (2021GXNSFDA196002); in part by the Guangxi Key Laboratory of Image and Graphic Intelligent Processing under Grants (GIIP2006, GIIP2007, GIIP2008); and in part by the Innovation Project of Guangxi Graduate Education under Grants (YCB2021070 , YCBZ2018052, YCSW2022269).

Conflict of Interest. All authors disclosed no relevant relationships.

Data Availability Statement. The datasets generated during and/or analysed during the current study are available in the MSCOCO repository, https://cocodataset.org/.

References

[1] Vinyals Oriol, Toshev Alexander , Bengio Samy ,Erhan Dumitru. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Transaction on pattern analysis and machine intelligence. 39(4):652−663, 2016.

[2] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 375−383, 2017.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6077−6086, 2018.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[5] Zhe Gan, Chuang Gan,Xiaodong He, et al.Semantic compositional networks for visual captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[7] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In Advances in Neural Information Processing Systems pages 11137−11147, 2019.

[8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10578−10587, 2020.

[9] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network. In Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[10] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10971−10980, 2020.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998−6008, 2017.

[12] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 15465−15474, 2021.

[13] Tiantao Xian and Zhixin Li and Canlong Zhang and Huifang Ma. ADual Global Enhanced Transformer for image captioning. Neural Networks, 148:129−141, 2022.

[14] Jing Wang, Yehao Li, Yingwei Pan, et al. Contextual and Selective Attention Networks for Image Captioning. Sci China Inf Sci, 2022.

[15] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-Level Collaborative Transformer for Image Captioning. arXiv preprint arXiv:2101.06462, 2021.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll??ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740−755, 2014.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[18] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, et al. MLP-Mixer: An all-MLP Architecture for Vision. arXiv preprint arXiv:2105.01601, 2021.

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, et al. Big Transfer (BiT): General Visual Representation Learning. arXiv preprint arXiv:1912.11370, 2019.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, 2021.

[21] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. High-Performance Large-Scale Image Recognition Without Normalization. arXiv preprint arXiv:2102.06171, 2021.

[22] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling Local Self-Attention for Parameter Efficient Visual Backbones. arXiv preprint arXiv:2103.12731, 2021.

[23] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. Beyond Self-attention: External Attention using Two Linear Layers for Visual Tasks. arXiv preprint arXiv:2105.02358, 2021.

[24] Hanxiao Liu, Zihang Dai, David R. So, and Quoc V. Len. Pay Attention to MLPs. arXiv preprint arXiv:2105.08050, 2021.

[25] Xiaohan Ding, Chunlong Xia, Xiangyu Zhang, Xiaojie Chu, Jungong Han, and Guiguang Ding. RepMLP: Re-parameterizing Convolutions into Fully-connected Layers for Image Recognition. arXiv preprint arXiv:2105.01883, 2021.

[26] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing RNNs for Caption Generation by Reconstructing The Past with The Present. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[27] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156−3164, 2015.

[28] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014.

[29] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,pages: 8307−8316 2019.

[30] Girish Kulkarni, et al. Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12):2891−2903, 2013.

[31] Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, et al. Common subspace for model and similarity: Phrase learning for caption generation from images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2668−2676, 2015.

[32] Margaret Mitchell, et al. Midge: Generating image descriptions from computer vision detections. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 747−756, 2012.

[33] Huaizu Jiang, et al. In Defense of Grid Features for Visual Question Answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10267−10276, 2020.

[34] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002.

[36] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005.

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL Workshop, volume 8, 2004.

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[39] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In Proceedings of the European Conference on Computer Vision, 2016.

[40] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In Proceedings of the European Conference on Computer Vision, pages 499−515, 2018.

[41] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In Proceedings of the European Conference on Computer Vision, pages 684−699, 2018.

[42] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10685−10694, 2019.

[43] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In Proceedings of the International Conference on Computer Vision, 2019.

[44] Guang Li,et al. Entangled transformer for image captioning. In Proceedings of the International Conference on Computer Vision, pages 8928−8937, 2019.

[45] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. CaMEL: Mean Teacher Learning for

Image Captioning. In International Conference on Pattern Recognition, 2022.

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086, 2022.

[47] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17980−17989, 2022.

[48] Chia-Wen Kuo,and Zsolt Kira. Beyond a Pre-Trained Object Detector: Cross-Modal Textual and Visual Context for Image Captioning. In CVPR, 2022.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In ICML, 2021.