



End-to-end Image Captioning via Visual Region Aggregation and Dual-level Collaboration

Jingkuan Song (宋井宽), Pengpeng Zeng (曾鹏鹏), Jiayang Gu (顾嘉扬),
Jinkuan Zhu (朱晋宽), Lianli Gao (高联丽)

(School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Corresponding author: Jingkuan Song, jingkuan.song@gmail.com

Abstract To date, Transformer-based pre-trained models have demonstrated powerful capabilities of modality representation, leading to a shift towards a fully end-to-end paradigm for multimodal downstream tasks such as image captioning, and enabling better performance and faster inference. However, the grid features extracted with the pre-trained model lack regional visual information, which leads to inaccurate descriptions of the object content by the model. Thus, the applicability of using pre-trained models for image captioning remains largely unexplored. Toward this goal, this paper proposes a novel end-to-end image captioning method based on Visual Region Aggregation and Dual-level Collaboration (VRADC). Specifically, to learn regional visual information, this paper designs a visual region aggregation that aggregates grid features with similar semantics to obtain a compact visual region representation. Next, dual-level collaboration uses the cross-attention mechanism to learn more representative semantic information from the two visual features, which in turn generates more fine-grained descriptions. Experimental results on the MSCOCO and Flickr30k datasets show that the proposed method, VRADC, can significantly improve the quality of image captioning, and achieves state-of-the-art performance.

Keywords image captioning; end-to-end training; pre-trained model; visual region aggregation; dual-level collaboration

Citation Song JK, Zeng PP, Gu JY, Zhu JK, Gao LL. End-to-end image captioning via visual region aggregation and dual-level collaboration, *International Journal of Software and Informatics*, 2023, 13(2): 221–241. <http://www.ijsi.org/1673-7288/316.htm>

With the popularization of intelligent terminals and multimedia devices, massive, multimedia data such as images and videos are uploaded to sharing websites or platforms such as Instagram, Twitter, TikTok, and YouTube everyday. Since these data contain rich and valuable information, how to enable computers to automatically understand and analyze the content of multimedia data and serve users with a better interactive experience has become a research hotspot in computer science. As a basic task in this field, image captioning^[1–3] aims to automatically generate a natural language sentence consistent with the content of a

This is the English version of the Chinese article “基于视觉区域聚合与双向协作的端到端图像描述生成. 软件学报, 2023, 34(5): 2152–2169. DOI: 10.13328/j.cnki.jos.006773”

Funding items: National Key Research and Development Program of China (2022YFC2009900, 2022YFC2009903); National Natural Science Foundation of China (62122018, 62020106008, 61772116, 61872064)

Received 2022-04-18; Revised 2022-05-29, 2022-08-03; Accepted 2022-08-24; IJSI published online 2023-06-29

given image. This task needs to be accomplished in a variety of scenarios such as content-based image retrieval and recommendation^[4], intelligent blind guide system^[5], and human-machine interaction^[6, 7]. Although the current computer vision and natural language processing technologies have achieved great progress in visual content understanding and language semantic analysis, image captioning is challenging because it needs to both understand the rich visual information and generate complex semantic description. It can be regarded as the touchstone to judge whether AI has completed the leap from perception to cognition.

The available image captioning methods are mostly based on the encoder-decoder framework of the neural machine translation^[8]. The encoder uses visual feature extractors such as Convolutional Neural Networks (CNNs)^[9] and object detection networks^[10] to analyze and extract the visual information of the images. The decoder uses language generators such as Recurrent Neural Networks (RNNs)^[11] and Transformer^[12] to generate natural language descriptions. The mainstream image captioning methods are two-stage models, as shown in Figure 1(a)(i). Specifically, the pre-trained visual feature extractor is used to obtain offline visual features, with which the proposed network frameworks such as RSTNet^[3] and DLCT^[13] are trained. Significant progress has been achieved in the above methods. However, due to the differences in data domain and task form between the pre-trained visual feature extractor and the downstream image captioning task, how to improve the model performance becomes a bottleneck. Moreover, extracting offline visual features is time-consuming, which makes it inapplicable to real-time image captioning scenarios.

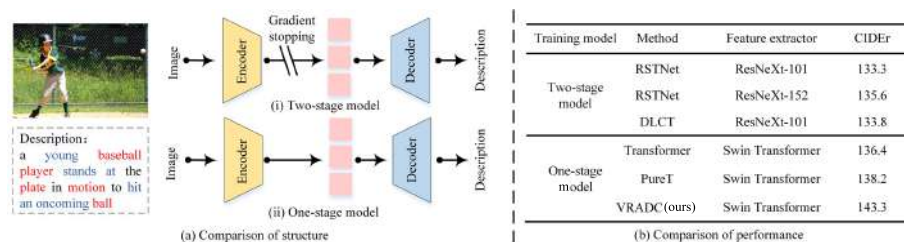


Figure 1 Comparison of structure and performance between two-stage model and one-stage model in image captioning

Transformer-based pre-trained models such as vision Transformer^[14] and swin Transformer^[15] have recently achieved great progress in visual tasks such as image classification and object detection. The fully end-to-end image captioning model based on Transformer, i.e., the one-stage model shown in Figure 1(a)(ii), has more powerful capabilities and thus has become a new research hotspot. This model can ensure the architecture unity between the encoder and the decoder and optimize the parameters of them at the same time. As shown in Figure 1(b), the one-stage model demonstrates improved performance compared with the two-stage model, with an average increase of 3.7 points in CIDEr^[16]. Although the Transformer-based fully end-to-end model shows great potential, how to align the strong visual features obtained from the pre-trained model with the language words remains to be explored. Fang *et al.*^[17] assisted text generation by extracting entity concepts or attribute representations from a given image. Wang *et al.*^[18] added pre-fusion between the generated word embedding and the global image features to increase the interaction between multimodal features, so as to improve the inference from image to caption.

Currently, visual features can be classified into two types: grid features and region features. The former can be extracted from the evenly divided grids with the same size of an image through a visual feature extractor, covering the information of the entire image. The latter refers to the representative region feature of objects in an image imported into an object detection model.

The vision Transformer-based end-to-end image captioning method extracts grid features and not the region object information in the image. The semantic information involved in the complete description of the content of an image is complex. As shown in Figure 1, the semantic information of this description is complex, and the detailed words such as “young”, “stands at”, and “hit an oncoming” need the context information of the whole image for inference, which is the advantage of grid features. However, the nouns such as “baseball player”, “plate”, “motion”, and “ball” in the sentence often correspond to the object information in the image, which is the advantage of region features. One solution is to integrate the end-to-end object detection technology into the model, but the complex model design and the high computing cost make it inconducive to the training and test of the end-to-end model.

To solve the above problems, this paper proposes a novel end-to-end image captioning method based on Visual Region Aggregation and Dual-level Collaboration (VRADC). Figure 2 illustrates the architecture of this model. First, vision Transformer is used to extract visual grid features from the original image as the initial features. To learn the compact visual region representation, this paper proposes a visual region aggregation module, which uses learnable semantic clustering to aggregate the features with the same semantic information in the grid features. Each cluster center implicitly represents a representative visual region feature. In addition, this paper proposes a dual-level collaboration module to learn the relationship between grid features and region features. Specifically, under the guidance of region features, the cross-attention mechanism is used to obtain the semantic features with higher-quality visual information, which will improve the image captioning quality.

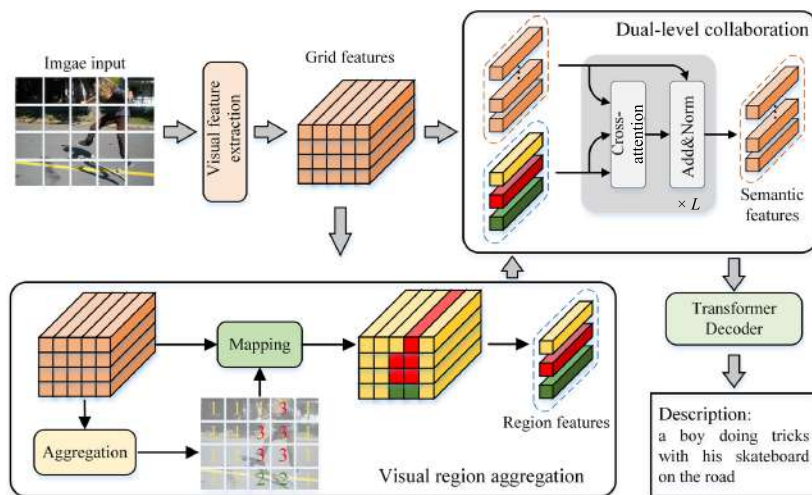


Figure 2 Schematic diagram of VRADC

This paper makes the following contributions.

- (1) This paper is the first work to consider the region object information in end-to-end image captioning. Without explicit supervision, this paper proposes a visual region aggregation module, which implicitly learns the distinctive and representative region features from the grid features.
- (2) This paper proposes a dual-level collaboration module, which integrates the learned region features into the grid features to obtain more representative semantic features for generating more fine-grained descriptions.
- (3) The VRADC method proposed in this paper has been verified by a large number of

experiments with two public datasets: MSCOCO and Flickr30k. The experimental results show that our method achieves state-of-the-art performance and is superior to the mainstream methods, and even better than some large-scale pre-trained visual and language models.

Section 1 of this paper introduces the relevant methods and research status of image captioning, pre-trained vision models, and pre-trained vision-language models. Section 2 details the image captioning based on the VRADC method. Section 3 introduces the experimental settings (including datasets and training details) and the quantitative and qualitative experimental analysis to prove the effectiveness of the method proposed in this paper. Section 4 makes a summary and prospects the future research directions.

1 Related Work

1.1 Image captioning

Image captioning is a research hotspot with emerging research achievements in recent years. The available image captioning methods are mostly based on the encoder-decoder framework, where the encoder extracts image features and the decoder uses the extracted image features to generate descriptions. According to the structure of the image captioning model, this section introduces two classes of image captioning methods: the RNN-based encoder-decoder framework and the Transformer-based encoder-decoder framework.

RNN-based encoder-decoder framework. The RNN-based encoder-decoder framework has significantly improved the quality of machine translation, and it is thus used for image captioning. The initial works^[19–21] used CNN to extract the visual information from images and RNN to generate the final description. However, these methods mainly use global visual information and lose local features in the images. As a result, the decoder fails to accurately analyze the details of the image. People focus on the regions of interest when describing an image. Inspired by this phenomenon, researchers apply the attention mechanism to the RNN-based encoder-decoder framework. Xu *et al.*^[22] proposed a LSTM-based attention mechanism module, which calculated the attention weight of LSTM's current hidden features to each region in the image and dynamically selected the region related to the current word as the visual input of the decoder, so that the generated text can accurately focus on the significant object and details. Chen *et al.*^[23] proposed a hierarchical attention-based image captioning method, which used hierarchical attention to adaptively select the convolutional feature map generated by CNN to guide word generation. Furthermore, Chen *et al.*^[20] proposed a RNN-based attribute attention module to help the model predict the attribute information in the image, and used the predicted attribute information to assist the captioning, thus improving the prediction accuracy of the model for high-frequency attribute words. Later, Anderson *et al.*^[11] used Faster R-CNN^[10] as the encoder and replaced the region features extracted by CNN with the features extracted by Faster R-CNN. Moreover, they proposed that bottom-up and top-down attention focused on regional objects rather than grid features. Yang *et al.*^[24] mapped the fine-grained semantic scene map into the RNN-based encoder-decoder, aiming to improve the quality of image captioning by using the inductive bias of language.

Transformer-based encoder-decoder framework. As major breakthroughs have been made for Transformer in the field of natural language recently, the Transformer-based image captioning methods have sprung out. The core idea of this framework is to use the self-attention or cross-attention mechanism in Transformer to strengthen the visual information encoding and the information interaction between vision and language. In order to make the encoder and the decoder of Transformer more adaptive to image captioning, researchers have proposed a variety of schemes for improvement. Pan *et al.*^[25] proposed a bilinear attention module, which used spatial and channel bilinear attention distribution to capture the interactive information between

object features. By embedding the relative position information into the multi-head attention mechanism, Chen *et al.*^[26] improved the direction perception between visual features. Cornia *et al.*^[27] stored the relationship between object features at different levels as prior knowledge, and then used the prior knowledge to assist in the generation of semantic descriptions at different levels. Later, Jiang *et al.*^[28] reported that grid features were superior to region features in terms of performance and time cost after re-examination. Zhang *et al.*^[3] proposed a grid enhancement module based on grid visual features to consider the relative relationship between grid locations. The proposed adaptive attention module would adaptively measure the contribution of visual cues and language cues to predict the final word based on the Transformer decoder.

However, the above methods belong to two-stage models: extraction of the visual features and training of the designed model. Recently, the Transformer-based pre-trained model demonstrates good performance in visual classification, and even surpasses the CNN-based pre-trained model. Therefore, the Transformer-based image captioning method, a one-stage method, is proposed. This method is applicable in real-time scenes and can generate accurate image descriptions. Wang *et al.*^[18] first applied swin Transformer to this field and increased the interaction between the two modes through the pre-fusion of global visual features and current predicted words. Fang *et al.*^[17] used vision Transformer as an encoder and introduced an additional module to predict high-frequency concept words, thereby improving the perception of concept-related grid features by the encoder. They then incorporated the extracted concept word features into the grid features to further improve the visual features. These methods prove that the grid features encoded by Transformer contain rich semantic information and are suitable for the downstream visual language tasks such as image captioning. In addition, the Transformer-based encoder-decoder network realizes the unification of the network structure.

1.2 Pre-trained vision model

The development of visual language tasks is inseparable from the pre-trained vision models, which provide semantically rich visual features for the downstream tasks of visual language. In the early visual language works, CNNs such as ResNet^[29] and VGG^[30] are used as visual feature extractors. Generally, these CNNs are pre-trained by classification loss with image classification datasets such as ImageNet. However, because the classification dataset has only a rough category label for each image, the image feature obtained by such training does not contain rich information. The emergence of object detection networks represented by Faster R-CNN^[10] and YOLO^[31] solves this problem. Different from the classification task, the pre-trained task and the pre-trained dataset of object detection networks are more suitable for the downstream task of visual language. VisualGenome^[32] is used for training since it uses massive annotation data with more fine-grained information including attributes and objects, which is exactly the focus of image captioning. Moreover, the resolution of the image processed by the object detector is higher than that by CNNs (such as 448×448 vs. 256×256), and thus the region features of the object obtained have higher quality than the grid features extracted by CNNs. The recently emerged Transformer^[14, 15] encodes the image into a series of visual tokens. Compared with CNNs, Transformer has a larger field of perception and no inductive bias, thus achieving state-of-the-art performance in large-scale multi-label classification visual datasets such as ImageNet-21k. In addition, the Transformer-based structure is more suitable for combining with the downstream tasks of visual language. Therefore, using the Transformer-based pre-trained model as the image feature extractor has become the common practice of current visual language tasks.

1.3 Pre-trained vision-language model

Inspired by the great success of the pre-trained method in the field of vision or language, large-scale pre-trained vision-language models have gradually become a research hotspot. Generally, such models are also based on Transformer. For this purpose, a series of agent tasks such as visual masking, text masking, and image-text matching are designed to train the model. The datasets used for pre-training are usually the datasets of large-scale image-text pairs, such as Conceptual Captions^[33] and HowTo100M^[34]. The existing pre-trained vision-language models can be classified into single-stream models and dual-stream models according to the fusion of visual and text information. The single-stream models represented by VisualBERT^[35], VideoBERT^[36], VinVL^[37], and MDETR^[38] assemble text and visual features together and then pass them into a Transformer to complete the fusion of features. The dual-stream models represented by LXMERT^[39], ALIGN^[40], ALBEF^[41], and Frozen^[42] pass the text and visual features into the Transformer of each stream and then use the cross-attention module of each stream to fuse the features.

In addition, the pre-trained vision-language models can be classified into encoder models and encoder-decoder models according to the architecture. Most of the existing pre-trained vision-language models are encoder models, such as ALIGN, VinVL, and ALBEF. They directly feed the cross-modal representation into the output layer to generate the final output. The encoder-decoder models represented by MDETR and UniVL import the cross-modal representation to the encoder and then transmit it to the decoder for decoding.

2 Methods

This paper proposes an end-to-end image captioning method, VRADC, which is a Transformer-based encoder-decoder model. Specifically, for a given image, the pre-trained swin Transformer is used as the visual feature extractor to extract the grid features, which are expressed as $G = \{g_1, g_2, \dots, g_M\}$, where M represents the number of grids in the image and g_m represents the feature of the grid m . To learn the regional information in the image, this paper designs a visual region aggregation module, which aggregates the semantically similar grid features into compact region features expressed as $R = \{r_1, r_2, \dots, r_N\}$, where N represents the number of aggregation regions and r_n represents the feature of the region n . Then, a dual-level collaboration module is introduced to learn the relationship between grid features and regional features and generate more semantic and complete visual semantic features expressed as $V = \{v_1, v_2, \dots, v_M\}$. Finally, V is imported into the Transformer decoder to generate the final image description $Y = \{y_1, y_2, \dots, y_T\}$, where T is the longest length of the generated description. Figure 2 illustrates the overall framework of VRADC. This section details the specific implementation of VRADC.

2.1 Visual feature extraction

The existing algorithms have proved that representative visual features are more conducive to image captioning. Transformer-based models and their variants such as BERT^[43] and GPT^[44] have shown improved performance in natural language processing. Some pioneering work has introduced them into computer vision tasks. Transformer-based models such as vision Transformer and swin Transformer surpass CNN-based models in terms of performance and representation capacity. In this paper, swin Transformer with powerful performance is used as a visual feature extractor, which can establish a Transformer-based model for end-to-end image captioning from the raw image.

In this module, the image is first divided into B disjoint patches recorded as $I_p \in R^{p \times p \times 3}$, where $\{H, W\}$ and $\{P, P\}$ represent the sizes of the image and the patch, respectively. $N =$

$(H * W)/P^2$ is the number of patches and also the effective sequence length input into swin Transformer. The patches are then tiled and input into a trainable embedding layer to obtain the patch embedding vector. To preserve the location information, we use the position embedding into the patch embedding vector. Next, the patch embedding vector goes through four encoding stages. Each stage contains a patch feature fusion layer and multiple adjacent swin Transformer core units, which are used to obtain hierarchical visual representation. The patch feature fusion layer reduces the total size of local features to a quarter of the original size by assembling the region features of 2×2 patches. Each core unit is composed of a multi-head self-attention module based on a shift window, a multi-layer perceptron module, a GELU nonlinear layer, and a normalization module. Multiple core units can introduce cross-window connections while maintaining the effective computation of the originally non-overlapping windows, thereby improving the representation of the whole model. Finally, this paper uses the features output in the last stage of the swin Transformer as grid features, which are represented as G and input into the next module.

2.2 Visual region aggregation

As mentioned above, the grid features extracted by the swin Transformer lose the regional information, which leads to the inaccurate description of some object information. One feasible solution is to integrate the end-to-end object detection into the model to make up for its short comings. However, the object detection is not conducive to the end-to-end model training and testing due to the complex process and high cost of computation. If grid features with the same semantics can be implicitly selected and aggregated into multiple subspaces to obtain pseudo-region features, the operation will become more flexible. Therefore, this paper designs a visual region aggregation module, which aggregates grid features into multiple cluster centers without explicit supervision. Each cluster center integrates similar semantic information in grid features, representing a possible representative region.

In this module, the distance between the grid feature and the cluster center was calculated to obtain the compact region feature R . First, N learnable cluster centers are set and expressed as $C = \{c_1, c_2, \dots, c_N\}$. Then, the correlation between the grid feature g_m and the cluster center c_n is calculated as

$$r_{m,n} = \frac{\exp(g_m c_n^T + b_n)}{\sum_{k=1}^N \exp(g_m c_k^T + b_k)} \quad (1)$$

where b_n and b_k are learnable parameters. After weighting all grid features, the region feature r_n is obtained as

$$r_n = \text{Norm} \left(\sum_{m=1}^M r_{m,n} (g_m - \tilde{c}_n) \right) \quad (2)$$

where Norm is the L_2 norm normalization operation and $\tilde{c}_n$ is the learnable parameter with the same size as c_n . The final feature R is defined as the region feature.

2.3 Dual-level collaboration

This paper uses the dual-level collaboration module to establish the relationship between grid features and region features, so as to obtain richer visual representation. This module mainly adopts the cross-attention mechanism. As shown in Figure 2, first, different fully connected layers are used to map the grid feature G to query Q_G and the region feature R to the key K_R and value V_R . The similarity between G and R is calculated by the dot-product attention

function and expressed as a matrix.

$$S = \text{Similarity}(Q_G, K_R) = \text{Softmax} \left(\frac{Q_G K_R^T}{\sqrt{d}} \right) \quad (3)$$

where $S(m, n)$ represents the importance of the region feature r_n to the grid feature g_m . Considering the semantically related region information, this paper employs the Multi-Head Attention (MHA) mechanism to recalculate the similarity S according to Eq. (3).

In the multi-head attention mechanism, Q_G , K_R , and V_R are mapped to multiple different subspaces, and then the similarity S^h of each subspace is calculated according to Eq. (3). The outputs of all the heads are connected together to give a feature $\bar{P}$ through the nonlinear layer:

$$\begin{aligned} P^h &= S^h V_R^h, \quad h = 1, 2, \dots, H \\ \bar{P} &= [P^1; P^2; \dots; P^H] W^R \end{aligned} \quad (4)$$

where H is the number of subspaces. Finally, the feature is normalized by layer normalization and added to an original grid feature to produce a region-related grid feature $\tilde{G}$:

$$\tilde{G} = G + \text{LayerNorm}(\bar{P}) \quad (5)$$

A more refined visual feature can be obtained by stacking MHA blocks in L layers. The output of the last MHA block is taken as the final visual semantic feature V .

The visual semantic feature is input into the Transformer decoder to generate the final description Y .

$$Y = \text{TransformerDecoder}(V) \quad (6)$$

2.4 Model training

To train the model proposed in this paper, a two-stage training strategy the same as that of mainstream algorithms^[1, 3] is adopted. In the supervised learning stage (stage 1), the cross-entropy loss is used to learn the difference between the generated word and the reference word. In the reinforcement learning stage (stage 2), the CIDEr of sentences is used as the reward of reinforcement learning to optimize the generation performance of the model.

In stage 1, the cross-entropy loss of the model is minimized by giving the sentence $Y^* = \{y_1^*, y_2^*, \dots, y_T^*\}$ of the reference description of the image:

$$L_{CE} = - \sum_{t=1}^T \log(p_\theta(y_t^* | y_{1:t-1}^*)) \quad (7)$$

where θ is the parameter of the model. In the training of this stage, the words predicted by the model in the current time step largely depend on the words generated previously. The inaccurate word generation by the model in the previous time step will lead to error accumulation, further lowering the accuracy of the description generated by the model. In addition, the training loss in stage 1 is not correlated with the evaluation criteria. That is, the training and test objectives of the model are inconsistent.

To solve the above problems, in the stage 2, the idea of SCST^[45] is adopted and reinforcement learning is introduced into the model training process to optimize CIDEr as a reward, thus minimizing the negative reward expectations.

$$L_{RL} = -E_{y_{1:T} p_\theta} [r(y_{1:T})] \quad (8)$$

where $r(\cdot)$ is the CIDEr-D of the description $y_{1:T}$. The Monte Carlo method is used to approximate the gradient $\nabla_{\theta} L_{RL}$.

$$\begin{aligned}\nabla_{\theta} L_{RL} &\approx -\frac{1}{k} \sum_{i=1}^k \left(r(y_{1:T}^i) - b \right) \nabla_{\theta} \log p_{\theta}(y_{1:T}^i) \\ b &= \frac{1}{k} \left(\sum_{i=1}^k r(y_i) \right)\end{aligned}\quad (9)$$

where k refers to the number of sampled sequences; $y_{1:T}^i$ refers to the description sequence obtained by the sampling i ; and b represents the average reward value obtained with the sampled sequences. Through reinforcement learning, the generation of description can be optimized directly from the criteria, avoiding the local optimum caused by only one cross-entropy function.

3 Experiments

To evaluate the performance of the proposed VRADC method, this paper carried out a large number of experiments with two public datasets: MSCOCO^[46] and Flickr30k^[47]. The experimental results were quantitatively and qualitatively analyzed. Furthermore, VRADC is compared with the current mainstream algorithms for image captioning. This section first introduces the experimental dataset, evaluation criteria, and implementation details, and then analyzes the experimental results.

3.1 Datasets and evaluation criteria

3.1.1 Datasets

Like the mainstream methods, this paper conducts experiments on the MSCOCO and Flickr30k datasets to evaluate the performance of the proposed VRADC model.

MSCOCO is a large English dataset commonly used in image captioning, with a total of 164,062 images collected from online websites. To facilitate offline training and testing, 123,287 images are labeled with sentence descriptions, and each image has at least 5 reference descriptions. In this paper, the MSCOCO dataset is divided into a training set, a validation set, and a test set with 113,287, 5,000, and 5,000 images, respectively, according to the division method provided by Karpathy^[48]. In addition, MSCOCO provides an online test platform, which has 40,775 images with unavailable descriptions for further assessing the model performance.

The Flickr30k dataset is composed of 31,783 images from the Flickr website. The images cover a variety of human activities, events, scenes, and other visual content, and each image is labeled with five English sentences generated by network users. For fairness, this paper employs the division method provided by Young *et al.*^[47], using 29,783 images for training, 1,000 images for validation, and 1,000 images for test.

3.1.2 Evaluation criteria

To ensure the fairness of evaluation, this paper uses five evaluation indicators including BLEU@N^[49], METEOR^[50], ROUGE-L^[51], CIDEr^[16], and SPICE^[52] commonly used in image captioning to measure the quality of the description generated by the model. Among the indicators, BLEU@N is often used to evaluate the accuracy of the machine-generated description by comparing the n -gram coincidence degree between the reference description and the generated description. A higher coincidence degree indicates more accurate generated description. METEOR, an improved version of BLEU, is used to calculate the recall rate and accuracy rate of unigram weighting to evaluate the matching of sequence, synonym, root, affix, and interpretation between sentences. ROUGE-L is often used to evaluate the quality of text

abstract, using the longest common substring between sentences to calculate the accuracy rate and recall rate. CIDEr performs TF-IDF weighting on each n -gram between sentences and calculates the cosine similarity between TF-IDF weight vectors to measure the consistency between the reference description and the generated description, thereby evaluating the consistency and richness of image captioning. SPICE, a semantic analysis indicator based on the scene graph structure, converts the reference description and the generated description into a sentence scene graph and then analyzes the objects, attributes, and relationship between sentences. The code developed in this paper is available on GitHub (<https://github.com/jkdxg/VRADC>).

3.2 Experimental settings

Swin Transformer is used as the backbone of the encoder for visual feature extraction. Because of the different configurations in model size and image resolution, this paper selects two model sizes of Swin-B and Swin-L and two image resolutions of 224×224 and 384×384 . The visual features output by swin Transformer will be mapped to 512 dimensions. For each description, this paper first uses a tokenizer to delete punctuation marks and then divide the sentence into words. Then, all the words are converted into lower case letters. Finally, 10,201 words are selected to establish as the final word vocabulary after removal of the words that appear less than 5 times. Both the dual-level collaboration module and the Transformer-based decoder have 512 feature dimensions. Each layer has 8 self-attention headers, with 2,048 dimensions in the fully connected layer. The dual-level collaboration and decoder have 5 and 3 layers, respectively.

In model training, the Adam optimizer is used to optimize the parameters, which is completed into two stages: supervised learning and reinforcement learning. In the supervised learning stage, the number of total batches and iterations is set as 50 and 20, respectively. For the network parameters extracted from the visual features and other parameters, the initial learning rates are set as 4×10^{-5} and 4×10^{-4} , respectively. After 10 iterations, the learning rate decays once every 3 iterations, with a decay rate of 0.8. In the reinforcement learning stage, the number of total batches and iterations is set as 10 and 30, respectively, and the model learning rate is fixed at 2×10^{-5} . After each round of training, the performance of the model was evaluated with the validation set. Finally, the model with the highest CIDEr is used for test with the validation set. In the testing stage, the Beam Search^[53] method is used for modeling, with the Beam size set as 5.

3.3 Quantitative analysis of experimental results

3.3.1 Framework exploration and comparison

The core of VRADC is to obtain visual information with richer semantics for high-quality image captioning. This section introduces the comparative experiments for proving the powerful performance of VRADC, mainly focusing on the following questions.

- RQ1: In the visual region aggregation, does the different number of cluster centers affect the experimental results?
- RQ2: In the dual-level collaboration, does the cross-attention of different layers affect the experimental results?
- RQ3: Does different visual feature extractor skeleton affect the experimental results?
- RQ4: In the dual-level collaboration, do different attention mechanisms affect the experimental results?
- RQ5: Does the adoption of a fully end-to-end training strategy affect the experimental results?

RQ1: In the visual region aggregation, does the different number of cluster centers affect the experimental results?

To answer this question, this study sets a different number of cluster centers ($N = 5, 7, 9, 11, 13$), 3 layers in the dual-level collaboration, Swin-B as the visual feature extractor, and the image resolution of 224×224 . The experimental results are shown in Figure 3.

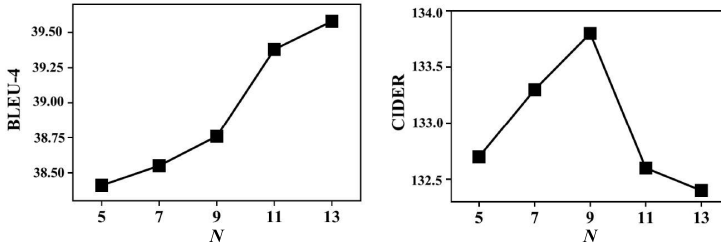


Figure 3 Effects of different cluster center numbers in visual region aggregation on model performance

It should be noted that the visual region aggregation module is designed to obtain representative regional information. As shown in Figure 3, the performance is the best at the number of cluster centers being 9. Too many cluster centers may make it hard to find representative regions, thus affecting the performance of the model. Too few cluster centers lead to the loss of massive weak semantic information, also resulting in poor performance of the model. Therefore, an appropriate number of cluster centers is the key for the model to obtain the optimal results. The final model is set with 9 cluster centers.

RQ2: In the dual-level collaboration, does the cross-attention of different layers affect the experimental results?

With 9 cluster centers, the model is set with different cross-attention layers ($L = 1, 3, 5, 7, 9$), Swin-B as the visual feature extractor, and the image resolution of 224×224 . The experimental results are shown in Figure 4, which demonstrate that different cross-attention layers have small impact on the performance of the model, and too many layers mean the introduction of more parameters. Finally, the dual-level collaboration is set with 5 layers.

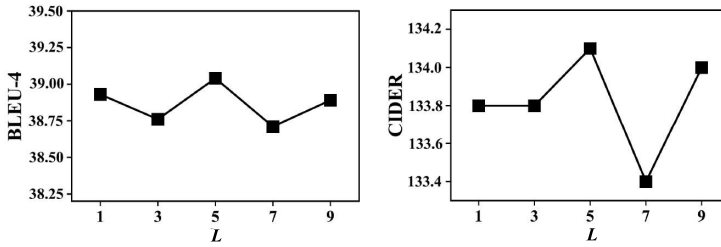


Figure 4 Effects of different cross-attention layers in dual-level collaboration on model performance

RQ3: Does different visual feature extractor skeleton affect the experimental results?

To answer this question, the study adopts different pre-training settings in swin Transformer to initialize the visual feature extractor from two aspects: pre-trained model size (Swin-B and Swin-L) and image resolution (224×224 and 384×384). The experimental results are shown in Table 1. In addition to the image captioning results, Table 1 lists the original performance and parameters of pre-trained models in image classification.

The following conclusions can be drawn from Table 1. First, the performance of the pre-trained model is crucial to the method proposed in this paper. When Swin-L is used as a visual feature extractor, the model demonstrates improved performance since a larger model has better performance in image classification and modality representation. Second, with the same

Table 1 Effects of different visual feature extractors on model performance

Pre-trained model	Image resolution	Image captioning					Image classification		
		B1	B4	M	R	C	Acc@1 (%)	Acc@5 (%)	Params (M)
Swin-B	224 × 224	81.02	39.04	29.16	58.55	134.1	85.2	97.5	88
Swin-B	384 × 384	82.75	41.30	30.28	59.82	141.5	86.4	98.0	88
Swin-L	224 × 224	82.48	40.64	29.73	59.43	139.5	86.3	97.9	197
Swin-L	384 × 384	83.09	41.33	30.56	60.34	143.3	87.3	98.2	197

pre-trained model, the increase in image resolution from 224×224 to 384×384 improves the model performance (CIDEr) from 134.1 to 141.5 for Swin-B and from 139.5 to 143.3 for Swin-L.

RQ4: In the dual-level collaboration, do different attention mechanisms affect the experimental results?

To evaluate the performance of different attention mechanisms in dual-level collaboration, the ablation experiment was carried out. As shown in Table 2, there are two attention mechanisms: parallel attention and cross-attention. The former refers to inputting grid features and learned region features into the self-attention module to obtain representative features. The cross-attention mechanism adopted in this paper integrates one feature into another feature to obtain representative feature. The cross-attention mechanism includes two ways: (1) cross-attention (region): integrating grid features into the learned region features to obtain representative visual features; (2) cross-attention (grid): integrating the learned region features into the grid features to obtain the representative visual features. The results in Table 2 indicate that the cross-attention (grid) mechanism outperforms the other two methods in terms of CIDEr. Therefore, the cross-attention (grid) mechanism is adopted in this study.

Table 2 Effects of different attention mechanisms on model performance

Attention mechanism	B1	B4	M	R	C
Parallel attention	81.33	39.00	29.18	58.66	133.9
Cross-attention (region)	81.07	38.53	28.83	58.19	131.9
Cross-attention (grid)	81.02	39.04	29.16	58.55	134.1

RQ5: Does the adoption of a fully end-to-end training strategy affect the experimental results?

One of the core viewpoints of this paper is that the fully end-to-end training strategy based on Transformer is conducive to image captioning. To verify this viewpoint, this paper designs an ablation experiment to train the proposed model with different parameter settings of swin Transformer. The experimental results are shown in Table 3. The end-to-end training strategy outperforms the non-end-to-end training strategy in other criteria except B1. In particular, the CIDEr is increased from 130.3 to 134.1, indicating the effectiveness of the end-to-end training strategy.

Table 3 Effects of end-to-end training strategy on model performance

Training strategy	B1	B4	M	R	C
Non-end-to-end (loading pre-training parameters)	81.19	38.46	28.38	57.81	130.3
End-to-end (loading pre-trained parameters)	81.02	39.04	29.16	58.55	134.1
End-to-end (not loading pre-trained parameters)	64.7	21.89	20.24	46.19	65.9

In addition, Table 3 suggests that the end-to-end training strategy with loading of the pre-trained model parameters significantly influences the experimental results. The pre-training of swin Transformer with ImageNet, a large image dataset, improves the image representation. However, if the pre-trained parameters are not loaded, the model is trained from scratch. Due to the large number of parameters and the lack of data, the model can be hardly fitted, which further results in poor performance.

3.3.2 Performance comparison on the MSCOCO dataset

This section compares the proposed method VRADC with the mainstream algorithms on the MSCOCO dataset to verify the effectiveness of the proposed method. The mainstream algorithms can be roughly classified into three classes. The first class is about two-stage models, which use offline visual feature training models and generate descriptions, including Adaptive^[54], SCST^[45], Up-down^[1], VRCDA^[55], GCN-LSTM^[56], AoANet^[57], M²Transformer^[27], GET^[58], X-Transformer^[25], DRT^[26], RSTNet^[3], and DLCT^[13]. The second class includes one-stage models. Such models input the original image and optimize the feature extractor and decoder, including ViTCAP^[17] and PureT^[18]. The third class is for pre-trained vision-language models, which use massive data to learn the general multimodal representation, including Oscar^[59], VinVL^[37], and SimVLM^[60]. VRADC proposed in this paper belongs to the second class, being a one-stage model. To fairly compare the above models, this paper carries out offline testing and online testing.

(1) **Offline testing.** Table 4 illustrates the performance of VRADC and other algorithms in offline testing. This table only shows the comparison results of single models. Firstly, one-stage models, especially VRADC, have better performance than two-stage models. Compared with RSTNet (ResNeXt152), VRADC (Swin-L 384) improves CIDEr by 7.7 points. Secondly, VRADC surpasses one-stage models in all the evaluation criteria. Thirdly, compared with the pre-trained vision-language models using massive multimodal data, VRADC still achieves excellent performance. Specifically, VRADC improves CIDEr by 0.8 point compared with SimVLM. In general, the above experimental results fully demonstrate the high potential of VRADC.

Table 4 Comparison of performance between VRADC and mainstream algorithms in offline testing

Model	Method	Visual feature extractor	B1	B4	M	R	C	S
Two-stage model	Adaptive	ResNet101	74.2	33.2	26.6	—	108.5	—
	SCST	ResNet101	—	34.2	26.7	55.7	114.0	—
	Up-down	F-RCNN101	79.8	36.3	27.7	56.9	120.1	21.4
	VRCDA	F-RCNN101	80.6	37.9	28.4	58.2	123.7	21.8
	GCN-LSTM	F-RCNN101	80.9	38.3	28.6	58.5	128.7	22.1
	AoANet	F-RCNN101	80.2	38.9	29.2	58.8	129.8	22.4
	M ² Transformer	F-RCNN101	80.8	39.1	29.2	58.6	131.2	22.6
	GET	F-RCNN101	81.5	39.5	29.3	58.9	131.6	22.8
	X-Transformer	F-RCNN101	80.9	39.7	29.5	59.1	132.8	23.4
	DRT	F-RCNN101	81.7	40.4	29.5	59.3	133.2	23.3
	RSTNet	ResNeXt101	81.1	39.3	29.4	58.8	133.3	23.0
	DLCT	ResNeXt101	81.4	39.8	29.4	59.1	133.8	23.0
	RSTNet	ResNeXt152	81.8	40.1	29.8	59.5	135.6	23.3
One-stage model	ViTCAP	ViT-B 384	—	40.3	29.4	59.5	133.6	23.3
	PureT	Swin-L 384	82.1	40.9	30.2	60.1	138.2	24.2
	VRADC (this study)	Swin-B 224	81.0	39.0	29.2	58.5	134.1	23.0
	VRADC (this study)	ViT-B 384	81.5	39.7	29.6	59.0	135.3	23.3
	VRADC (this study)	Swin-L 384	83.09	41.3	30.6	60.3	143.3	24.3
Pre-trained vision-language model	Oscar	ResNeXt152	—	41.7	30.6	—	140.0	24.5
	VinVL	ResNeXt152	—	41.0	31.1	—	140.9	25.2
	SimVLM	ViT-L	—	40.3	33.4	—	142.6	24.7

(2) **Online testing.** To further validate the effectiveness of VRADC, we submitted this method to the official online test server for comparison with mainstream algorithms. Similar to other models, four VRADC models are trained with the same parameters, and then the descriptions generated by the four models are integrated to generate the final description. This paper tested two configurations: Swin-B 224 and Swin-L 384. The experimental results are shown in Table 5. Compared with all the other models, VRADC with the Swin-L 384

Table 5 Comparison of performance between VRADC and mainstream algorithms in online testing

Model	B1		B2		B3		B4		M		R		C	
	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	116.7
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
GCN-LSTM	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	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
M2Transformer	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
GET	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
X-Transformer	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
DRT	82.7	96.5	67.7	91.5	53.1	83.4	40.9	73.6	29.6	39.0	59.8	75.0	132.2	133.9
RSTNet	81.7	96.2	66.5	90.0	51.8	82.7	39.7	72.5	29.3	38.7	59.2	74.2	130.1	132.4
(ResNext101)														
RSTNet	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
(ResNext152)														
DLCT	82.0	96.2	66.9	91.0	52.3	83.0	40.2	73.2	29.5	39.1	59.4	74.8	131.0	133.4
(ResNext1101)														
DLCT	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.4	135.4
(ResNext152)														
PureT (Swin-L 384)	82.8	96.5	68.1	91.8	53.6	83.9	41.4	74.1	30.1	39.9	60.4	75.9	136.0	138.3
VinVL*	81.9	96.9	66.9	92.4	52.6	84.7	40.4	74.9	30.6	40.8	60.4	76.8	134.7	138.7
VRADC	81.9	96.1	66.5	90.7	51.7	82.1	39.4	71.8	29.0	38.3	58.7	73.4	130.4	132.7
(Swin-B 224)														
VRADC	83.6	97.5	68.9	93.3	54.3	85.7	42.0	76.2	30.7	40.6	60.6	76.3	140.6	143.4
(Swin-L 384)														

configuration still maintains the best performance in all indicators. In particular, VRADC (Swin-L 384) has the c5 and c40 of CIDEr 4.6 and 5.1 points higher than PureT (Swin-L 384). In addition, the VRADC model with the Swin-B 224 configuration in this paper has also achieved comparable performance with most mainstream models.

3) Comparison of parameter number, computational complexity, and inference speed. In addition to the evaluation of the quality of sentences generated from image captioning, this paper compares the computational complexity (FLOPS), inference speed, and model parameters between VRADC and other models. VRADC and VITCAP are one-stage models and directly generate the final image description from the original image. RSTNet and DLCT are two-stage models which perform visual feature extraction and model prediction in image captioning. For fair comparison, the calculation, inference speed, and model parameters of the two stages are summed up for the two-stage models, and all the above four models are tested in the same environment. The experimental results are shown in Figure 5.

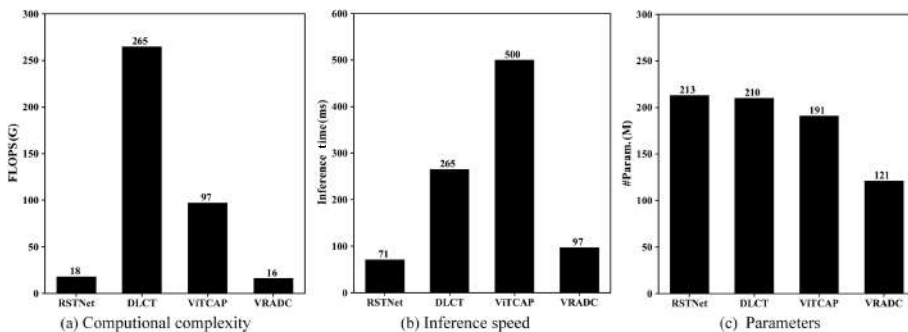


Figure 5 Comparison between VRADC and mainstream algorithms in terms of computational complexity, inference speed, and parameters

The following conclusions can be drawn from Figure 5. (1) VRADC has the lowest FLOPs (Figure 5(a)), demonstrating low computational complexity; (2) VRADC ranked the second in inference speed (Figure 5(b)), with the inference time 26 ms longer than RSTNet but much shorter than the other two models. RSTNet is a two-stage model which needs data access, conversion and other operations and thus has additional time consumption, while VRADC proposed in this paper does not have the above problems; (3) VRADC has the fewest parameters (Figure 5(c)), which means that it can be deployed on devices with a small storage space. In general, VRADC proposed in this paper can well balance the computational complexity, inference speed, and parameter number, demonstrating great advantages.

3.4 Performance comparison on the Flickr30k dataset

This section compares the performance of VRADC with the mainstream algorithms on the Flickr30k dataset. The mainstream algorithms to be compared include Adaptive^[54], Soft-Attention^[22], Hard-Attention^[22], SCA-CNN^[23], Semantic-Attention^[61], NBT^[62], A_R_L^[63], IVAIC^[64], and VRCDA^[55]. As shown in Table 6, VRADC achieves excellent performance compared with other mainstream algorithms in all the criteria. The experimental results further prove the effectiveness of VRADC.

Table 6 Comparison of performance between VRADC and other mainstream algorithms on the Flickr30k dataset

Approach	B1	B4	M	R	C
Soft-Attention	66.7	19.1	18.5	–	–
Hard-Attention	66.9	19.9	18.5	–	–
Adaptive	67.7	25.1	20.4	–	53.1
SCA-CNN	66.2	22.3	19.5	–	–
Semantic-Attention	64.7	23.0	18.9	–	–
NBT	69.0	27.1	21.7	–	57.5
A_R_L	69.8	27.7	21.5	48.5	57.4
IVAIC	70.8	30.6	22.5	49.8	63.0
VRCDA	73.2	30.6	22.7	50.6	66.0
VRADC	74.9	32.6	33.4	51.9	74.9

3.5 Qualitative analysis of experimental results

To qualitatively evaluate the visual representation obtained, this paper visualizes the importance of visual features to the final output words, with the results shown in Figure 6. Each example includes a raw image, as well as words and attention maps generated by a reference model and VRADC at each time step. The cross-attention weight of the last layer of the Transformer decoder is presented in this paper.

As shown in Figure 6, the reference model and VRADC can focus on the corresponding regions when generating words. VRADC outperforms the reference model in terms of focus region, object content, details, and context coherence. In the first example in Figure 6, the description generated by VRADC is “a man wearing a hat sitting at a table with beer”, which can accurately describe the environment where the human stays. However, the reference model generates the description “a man wearing a green hat and a green tie”, which is only a rough description of human appearance. VRADC can accurately focus on the corresponding region of “hat”. In the second example in Figure 6, the description generated by VRADC is in line with human expression habits and smoother than that generated by the reference model.

To intuitively prove that VRADC can generate more fine-grained descriptions, Figure 7 illustrates some descriptions generated by VRADC and the corresponding reference descriptions. VRADC performs well in the captioning in terms of included details, content and other aspects, generating descriptions even better than the reference. In addition, Figure 7 shows the heatmaps

of the region information learned by the visual region aggregation. In each heatmap, different colors represent different indicator values and region information. As shown in Figure 7, the visual region aggregation module not only focuses on the visual region of the specific foreground but also retains the discriminative background information in the image, which further proves that VRADC can well learn the region information in the study space.

In addition, Figure 8 illustrates the descriptions generated by VRADC and Transformer. The following conclusions can be summarized.

- (1) VRADC can more accurately describe the object information such as the wood cutting board in the first image and the two bowls in the second image than Transformer.
- (2) The sentences generated by VRADC are more consistent with human’s description

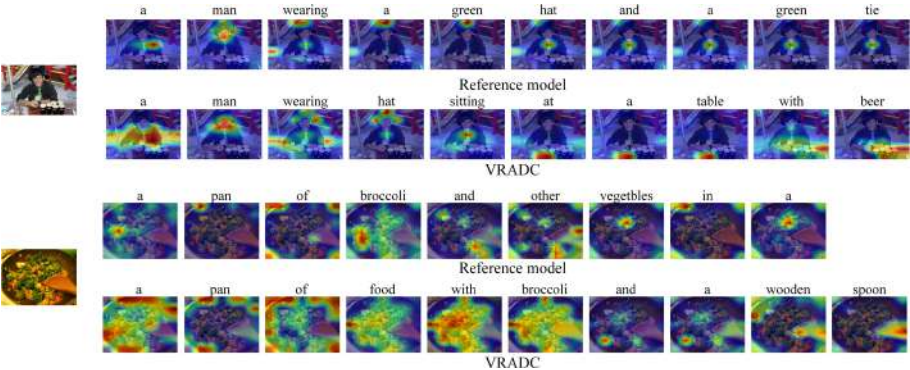


Figure 6 Visualization of the visual attention

Image				
Heatmap				
Reference description	A young girl inhales with the intent of blowing out a candle	A very cute brown dog with a disc in its mouth	A cat looking at his reflection in the mirror	A cat that is laying on a computer keyboard
VRADC	Two women sitting at a table with a bowl of food	A dog running with a purple frisbee in its mouths	A siamese cat looking at its reflection in a mirror	A cat laying on top of a computer keyboard

Figure 7 Descriptions generated by VRADC and reference descriptions

Image				
Reference description	A square pizza on a wooden cutting board	Two bowls filled with broccoli soup on top of a table	A man on a horse in a field with cows and a dog	A young boy in a sweatshirt and baseball cap by a bus
Transformer	A pizza sitting in a box on a table	A bowl of soup and a plate of broccoli on a table	A group of cows grazing in a field of grass	A young boy holding a baseball bat
VRADC	A pizza sitting on top of a wooden cutting board	Two bowls of soup and broccoli on a wooden table	A person riding a horse in a field with cows	A young boy wearing a hat standing in front of a bus

Figure 8 Descriptions generated by VRADC and Transformer

habits, such as “A person riding a horse in a field with cows” in the third image, which is more coherent, being smoother.

(3) VRADC generates more diverse sentences, as manifested by description of the fourth image that a young boy “wearing a hat” standing in front of a bus.

4 Summary and Prospect

This paper presented the novel Transformer-based image captioning method VRADC. This method includes a simple and effective visual region aggregation module, which implicitly learns representative regional information from grid features without any explicit supervision. Then, to integrate regional information into grid features, this paper proposed a dual-level collaboration module to obtain more semantic information, which can guide the model to generate a high-quality description consistent with the image content. Further, experiments were carried out on the MSCOCO and Flickr30k datasets, and the performance of VRADC was compared with that of the mainstream algorithms. The experimental results showed that VRADC achieves state-of-the-art performance in a variety of criteria and test environments. It even surpasses some large pre-trained vision-language models. Quantitative and qualitative experimental analysis also validated the effectiveness of VRADC.

In the future, a more effective visual region aggregation module will be designed to learn the region information in images. Moreover, the dual-level collaboration module will be optimized to improve the feature fusion method for generating a more fine-grained image description. In addition, some pre-trained language models or knowledge maps can be introduced to enhance the interpretability of the model, and the lightweight design of the end-to-end model to facilitate model deployment to mobile intelligent terminals.

References

- [1] Anderson P, He XD, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]
- [2] Gao LL, Li XP, Song JK, Shen HT. Hierarchical LSTMs with adaptive attention for visual captioning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2020, 42(5): 1112–1131. [doi: 10.1109/TPAMI.2019.2894139]
- [3] Zhang XY, Sun XS, Luo YP, Ji JY, Zhou YY, Wu YJ, Huang FY, Ji RR. RSTNet: Captioning with adaptive attention on visual and non-visual words. *Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021. 15460–15469. [doi: 10.1109/CVPR46437.2021.01521]
- [4] Cui H, Zhu L, Li JJ, Yang Y, Nie LQ. Scalable deep hashing for large-scale social image retrieval. *IEEE Trans. on Image Processing*, 2019, 29: 1271–1284. [doi: 10.1109/TIP.2019.2940693]
- [5] Wu SM, Wieland J, Farivar O, Schiller J. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. *Proc. of the 2017 ACM Conf. on Computer Supported Cooperative Work and Social Computing*. Portland: ACM, 2017. 1180–1192. [doi: 10.1145/2998181.2998364]
- [6] Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JMF, Parikh D, Batra D. Visual dialog. *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 1080–1089. [doi: 10.1109/CVPR.2017.121]
- [7] Jain U, Schwing AG, Lazebnik S. Two can play this game: Visual dialog with discriminative question generation and answering. *Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018. 5754–5763. [doi: 10.1109/CVPR.2018.00603]
- [8] Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *Proc. of the 27th Int'l Conf. on Neural Information Processing Systems*. Montreal: MIT Press, 2014. 3104–3112.

- [9] Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. Proc. of the 2017 Int'l Conf. on Engineering and Technology. Antalya: IEEE, 2017. 1–6. [doi: 10.1109/ICEngTechnol.2017.8308186]
- [10] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: NIPS, 2015. 91–99.
- [11] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]
- [12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 5998–6008.
- [13] Luo YP, Ji JY, Sun XS, Cao LJ, Wu YJ, Huang FY, Lin CW, Ji RR. Dual-level collaborative transformer for image captioning. Proc. of the 35th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2021. 2286–2293. [doi: 10.1609/aaai.v35i3.16328]
- [14] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai XH, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.
- [15] Liu Z, Lin YT, Cao Y, Hu H, Wei YX, Zhang Z, Lin S, Guo BN. Swin Transformer: Hierarchical vision transformer using shifted windows. Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 9992–10002. [doi: 10.1109/ICCV48922.2021.00986]
- [16] Fang ZY, Wang JF, Hu XW, Liang L, Gan Z, Wang LJ, Yang YZ, Liu ZC. Injecting semantic concepts into end-to-end image captioning. Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 17988–17998. [doi: 10.1109/CVPR52688.2022.01748]
- [17] Vedantam R, Zitnick CL, Parikh D. CIDEr: Consensus-based image description evaluation. Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 4566–4575. [doi: 10.1109/CVPR.2015.7299087]
- [18] Wang YY, Xu JG, Sun YF. End-to-end transformer based model for image captioning. Proc. of the 36th AAAI Conf. on Artificial Intelligence. Palo Alto: AAAI Press, 2022. 2585–2594. [doi: 10.1609/aaai.v36i3.20160]
- [19] Xue ZY, Guo PY, Zhu XB, Zhang NG. Image description method based on generative adversarial networks. Ruan Jian Xue Bao/Journal of Software, 2018, 29(S2): 30–43 (in Chinese with English abstract). <http://www.jos.org.cn/jos/article/abstract/18015?st=search>
- [20] Chen H, Ding GG, Lin ZJ, Zhao SC, Han JG. Show, observe and tell: Attribute-driven attention model for image captioning. Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI.org, 2018. 606–612. [doi: 10.24963/ijcai.2018/84]
- [21] Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(4): 652–663. [doi: 10.1109/TPAMI.2016.2587640]
- [22] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel RS, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. Proc. of the 32nd Int'l Conf. on Machine Learning. Lille: JMLR.org, 2015. 2048–2057. [doi: 10.5555/3045118.3045336]
- [23] Chen L, Zhang HW, Xiao J, Nie LQ, Shao J, Liu W, Chua TS. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6298–9306. [doi: 10.1109/CVPR.2017.667]
- [24] Yang X, Tang KH, Zhang HW, Cai JF. Auto-encoding scene graphs for image captioning. Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 10677–10686. [doi: 10.1109/CVPR.2019.01094]
- [25] Pan YW, Yao T, Li YH, Mei T. X-linear attention networks for image captioning. Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10968–10977. [doi: 10.1109/CVPR42600.2020.01098]
- [26] Song ZL, Zhou XF, Dong LH, Tan JL, Guo L. Direction relation transformer for image captioning. Proc. of the 29th ACM Int'l Conf. on Multimedia. Virtual Event: ACM, 2021. 5056–5064. [doi:

- 10.1145/3474085.3475607]
- [27] Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. Proc. of the 2020 IEEE Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10575–10584. [doi: 10.1109/CVPR42600.2020.01059]
 - [28] Jiang HZ, Misra I, Rohrbach M, Learned-Miller E, Chen XL. In defense of grid features for visual question answering. Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10264–10273. [doi: 10.1109/CVPR42600.2020.01028]
 - [29] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]
 - [30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.
 - [31] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: 10.1109/CVPR.2016.91]
 - [32] Krishna R, Zhu YK, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA, Bernstein MS, Li FF. Visual genome: Connecting language and vision using crowd sourced dense image annotations. Int'l Journal of Computer Vision, 2017, 123(1): 32–73. [doi: 10.1007/s11263-016-0981-7]
 - [33] Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: ACL, 2018. 2556–2565. [doi: 10.18653/v1/P18-1238]
 - [34] Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 2630–2640. [doi: 10.1109/ICCV.2019.00272]
 - [35] Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW. VisualBERT: A simple and performant baseline for vision and language. arXiv:1908.03557, 2019.
 - [36] Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: A joint model for video and language representation learning. Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 7464–7473. [doi: 10.1109/ICCV.2019.00756]
 - [37] Zhang PC, Li XJ, Hu XW, Yang JW, Zhang L, Wang LJ, Choi YJ, Gao JF. VinVL: Revisiting visual representations in vision-language models. Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 5575–5584. [doi: 10.1109/CVPR46437.2021.00553]
 - [38] Kamath A, Singh M, LeCun Y, Synnaeve G, Misra I, Carion N. MDETR-modulated detection for end-to-end multi-modal understanding. Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1760–1770. [doi: 10.1109/ICCV48922.2021.00180]
 - [39] Tan H, Bansal M. LXMERT: Learning cross-modality encoder representations from transformers. Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 5100–5111. [doi: 10.18653/v1/D19-1514]
 - [40] Jia C, Yang YF, Xia Y, Chen YT, Parekh Z, Pham H, Le QV, Sung YH, Li Z, Duerig T. Scaling up visual and vision-language representation learning with noisy text supervision. Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 4904–4916.
 - [41] Li JN, Selvaraju R, Gotmare A, Joty S, Xiong CM, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 2021, 34: 9694–9705.
 - [42] Bain M, Nagrani A, Varol G, Zisserman A. Frozen in time: A joint video and image encoder for end-to-end retrieval. Proc. of the 2021 IEEE Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 1728–1738. [doi: 10.1109/ICCV48922.2021.00175]
 - [43] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. of the 2019 Annual Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis:

- Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]
- [44] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020. 1877–1901.
- [45] Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 1179–1195. [doi: 10.1109/CVPR.2017.131]
- [46] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. *Proc. of the 13th European Conf. on Computer Vision*. Zurich: Springer, 2014. 740–755. [doi: 10.1007/978-3-319-10602-1_48]
- [47] Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. of the Association for Computational Linguistics*, 2014, 2: 67–78. [doi: 10.1162/tacl_a_00166]
- [48] Karpathy A, Li FF. Deep visual-semantic alignments for generating image descriptions. *Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Boston: IEEE, 2015. 3128–3137. [doi: 10.1109/CVPR.2015.7298932]
- [49] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: A method for automatic evaluation of machine translation. *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: Association for Computational Linguistics, 2002. 311–318. [doi: 10.3115/1073083.1073135]
- [50] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. *Proc. of the 9th Workshop on Statistical Machine Translation*. Baltimore: Association for Computational Linguistics, 2014. 376–380. [doi: 10.3115/v1/W14-3348]
- [51] Lin CY. ROUGE: A package for automatic evaluation of summaries. *Proc. of the 2004 Text Summarization Branches Out*. Barcelona: Association for Computational Linguistics, 2004. 74–81.
- [52] Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic propositional image caption evaluation. *Proc. of the 14th European Conf. on Computer Vision*. Amsterdam: Springer, 2016. 382–398. [doi: 10.1007/978-3-319-46454-1_24]
- [53] Wang PD, Ng HT. A beam-search decoder for normalization of social media text with application to machine translation. *Proc. of the 2013 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta: Association for Computational Linguistics, 2013. 471–481.
- [54] Lu JS, Xiong CM, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017. 3242–3250. [doi: 10.1109/CVPR.2017.345]
- [55] Liu MF, Shi Q, Nie LQ. Image captioning based on visual relevance and context dual attention. *Ruan Jian Xue Bao/Journal of Software*, 2022, 33(9): 3210–3222 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/6623.htm> [doi: 10.13328/j.cnki.jos.006623]
- [56] Yao T, Pan YW, Li YH, Mei T. Exploring visual relationship for image captioning. *Proc. of the 15th European Conf. on Computer Vision*. Munich: Springer, 2018. 711–727. [doi: 10.1007/978-3-030-01264-9_42]
- [57] Huang L, Wang WM, Chen J, Wei XY. Attention on attention for image captioning. *Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision*. Seoul: IEEE, 2019. 4633–4642. [doi: 10.1109/ICCV.2019.00473]
- [58] Ji JY, Luo YP, Sun XS, Chen FH, Luo G, Wu YJ, Gao Y, Ji RR. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. *Proc. of the 35th AAAI Conf. on Artificial Intelligence*. Palo Alto: AAAI Press, 2021. 1655–1663. [doi: 10.1609/aaai.v35i2.16258]
- [59] Li XJ, Yin X, Li CY, Zhang PC, Hu, XW, Zhang L, Wang LJ, Hu HD, Dong L, Wei FR, Choi YJ, Gao JF. OSCAR: Object-semantics aligned pre-training for vision-language tasks. *Proc. of the 16th European*

- Conf. on Computer Vision. Glasgow: Springer, 2020.121–137. [doi: 10.1007/978-3-030-58577-8_8]
- [60] Wang ZR, Yu JH, Yu AW, Dai ZH, Tsvetkov YL, Cao Y. SimVLM: Simple visual language model pretraining with weak supervision. arXiv:2108.10904, 2021.
- [61] You QZ, Jin HL, Wang ZW, Feng C, Luo JB. Image captioning with semantic attention. Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 4651–4659. [doi: 10.1109/CVPR.2016.503]
- [62] Lu JS, Yang JW, Batra D, Parikh D. Neural baby talk. Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7219–7228. [doi: 10.1109/CVPR.2018.00754]
- [63] Wang JB, Wang W, Wang L, Wang ZY, Feng DD, Tan TN. Learning visual relationship and context-aware attention for image captioning. Pattern Recognition, 2020, 98: 107075. [doi: 10.1016/j.patcog.2019.107075]
- [64] Li ZX, Wei HY, Huang FC, Zhang CL, Ma HF, Shi ZZ. Combine visual features and scene semantics for image captioning. Chinese Journal of Computers, 2020, 43(9): 1624–1640 (in Chinese with English abstract). [doi: 10.11897/SP.J.1016.2020.01624]



Jinkuan Song, Ph.D. His research interests include deep learning and information retrieval.



Jinkuan Zhu, master's degree candidate. His research interests include information retrieval and video understanding.



Pengpeng Zeng, Ph.D. candidate. His research interests include multimodal fusion and computer vision.



Lianli Gao, Ph.D. Her research interests include computer vision, deep learning, semantic network, and knowledge inference.



Jiayang Gu, master's degree candidate. His research interests include deep learning and multimodal infusion.