Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Fast RF-UIC: A fast unsupervised image captioning model☆

Rui Yang [a], Xiayu Cui [a], Qinzhi Qin [c], Zhenrong Deng [a,b,*], Rushi Lan [b], Xiaonan Luo [a]

[a] *Guangxi Key Laboratory of Images and Graphics Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China*
[b] *Nanning Research Institute, Guilin University of Electronic Technology, Guilin 541004, China*
[c] *Guangxi Construction Industry Mostly Lease Co., Nanning 510600, China*

ARTICLE INFO

Keywords:
Image captioning
Fast RF-UIC
Unsupervised
Pre-trained model

ABSTRACT

For visually impaired individuals, image captioning is a crucial task that utilizes deep learning models to recognize an image and generate a descriptive sentence, enabling them to understand the content of the image through words. However, the existing image captioning model needs a lot of manual annotation. Fortunately, the emergence of unsupervised methods provides a new approach to image captioning. Our proposed model Fast RF-UIC achieves unsupervised functionality through the designed Pre-trainer. Compared with the existing pre-trained model, the Pre-trainer has a faster and shorter training cycle. The R2-Inception-V4 model is designed as an encoder that fuse the Res2Net structure with Inception-V4 to obtain more image features. Bi-FGRU is designed as the decoder, which the FReLU activation function is used to improve the character representation ability from two-dimensional space. Furthermore, we expanded the corpus used in existing unsupervised image captioning and included additional captions for common objects, effectively enhancing the model's generalization ability. Through experiments, Fast RF-UIC achieved higher scores than existing unsupervised image captioning methods on several text evaluation metrics such as BLUE, ROUGE, and CIDEr.

1. Introduction

The visually impaired community constitutes a significant proportion of the disabled population, with nearly 20 million visually impaired individuals in China alone. Visual impairment hinders their ability to perceive the world through vision, resulting in significant inconvenience in daily life and work. The image captioning model describes the content of the image with language, so that the visually impaired can experience the beauty of the world through words, and greatly enhance the independence and quality of life of the visually impaired community.

The image captioning task combines the technologies of computer vision and natural language processing, requiring computers to possess the ability to process image tasks and natural language tasks and solve the interaction problems between the two technologies [1]. Currently, the applications of image captioning are becoming increasingly extensive. For instance, image captioning algorithms can assist in transforming images into information which can be read by visually impaired individuals, thereby increasing their access to external information. In visually assistive systems, image captioning algorithm can scan road conditions and display them to the user in real-time in the form of text information, promptly reminding the user of road safety. Image captioning has been studied extensively using supervised methods, dating back to the earliest "encoding-decoding" models. Supervised methods have the advantage of being able to handle large amounts of training data, but require extensive manual annotation and can be resource-intensive. In contrast, unsupervised methods can reduce the cost of dataset annotation and it is beneficial to the learning of large amounts of unlabeled data. With the rise of unsupervised methods, researchers such as Feng et al. [2] have proposed a breakthrough unsupervised image captioning method that do not require a large number of paired image-sentence dataset. Specifically, as deep learning networks become increasingly complex, supervised learning models are more suitable for handling large amounts of training data. Although the increasing of training data improves the performance of the model, it makes labeling task more challenging as well. Consequently, in terms of future development prospects, unsupervised methods may be better suited to the direction of this field. Our model can convert visual information into textual descriptions, allowing visually impaired individuals to perceive and understand images they cannot see. This technology empowers them to access visual content on the internet, social media platforms, educational materials, or even personal photos. And with the help of our model, visually impaired individuals can

☆ This paper was recommended for publication by Prof Guangtao Zhai.

* Corresponding author at: Guangxi Key Laboratory of Images and Graphics Intelligent Processing, Guilin University of Electronic Technology, Guilin 541004, China.

E-mail address: 799349175@qq.com (Z. Deng).

https://doi.org/10.1016/j.displa.2023.102490
Received 10 April 2023; Received in revised form 8 June 2023; Accepted 27 June 2023
Available online 3 July 2023

navigate their surroundings more independently. For instance, by using a smartphone equipped with a camera, they can capture an image of their surroundings, and the image captioning model can provide a verbal description of the scene, enabling them to identify objects, people, or hazards.

This paper mainly consists of five parts, among which section one and section two are the introduction and related work, section three is the Fast RF-UIC model designed in this paper, and the Fast RF-UIC model includes encoder: R2-Inception-V4, decoder: Bi-FGRU and Pre-trainer. Section four is the experimental analysis, and section five is the conclusion. In this paper, our contributions include the following aspects: We propose a three-part unsupervised image captioning framework, including the image feature extraction part R2-Inception-V4, the generation part Bi-FGRU, and the Pre-trainer. We propose a simpler and more effective pre-trained model for generating the “image-overview” pseudo-data set required for model initialization, which reduces the overall training cycle of the model and improves the performance of the generated text. At the same time, the original corpus is expanded to make the model have better generalization.

2. Related work

The field of image captioning has gone through three stages from its birth to the present, based on template [3–6], retrieval [7–9], and deep learning methods [10–13], respectively. With the development of artificial intelligence and the first two methods are limited by grammar and training corpus respectively, image captioning models based on deep learning becomes the mainstream method. Next, we will focus on the development of image captioning based on deep learning.

Vinyals [14] was the first to use deep learning for image captioning. This paper proposed an “encoding-decoding” model called NIC model based on the idea of machine translation, in which the encoding part uses a convolutional neural network (CNN) to extract image features. The decoding part uses Recurrent Neural Network (RNN) to decode the features extracted from the encoding part to generate a caption. Xu [15] and others found that the NIC model did not take into account the spatial characteristics of the image, and combined with biological vision, we humans understand the image for a certain area in the image. Based on this, the paper proposes an attention mechanism, which will generate caption words corresponding to different fields in the image. Lu [16] proposed the sentinel mechanism based on the attention mechanism to calculate the correlation between each word and the image. Xian et al. [17] incorporate global information in the encoding and decoding stages. The above articles are all aimed at the improvement of the decoding part of the model. Chen [18] proposed spatial and multi-channel attention mechanisms to improve the CNN model of the coding part. With the development of target detection technology, Anderson [19] and Lu [20] used different target detection models, and achieved better results by combining the target and template sentences. Wei et al. [21] allows the model to explore various visual feature extraction paths, resulting in robust visual representations. Hua [22] embeds a feature reorganization layer into a deep neural network and simultaneously optimizes classification task and correspondence task via alternate optimization. Xian et al. [23] use the semantic features of the core objects detected from the image to guide the visual features, which incorporate information on the spatial position relationships between the objects. Zhao et al. [24] propose a multi-level cross-modal alignment (MCA) module to align the image scene graph with the sentence scene graph at different level. Feng et al. [25] propose LRBNet in which the captions and questions are embedded in a uniform space with the same dictionary. And Jiang et al. [26] exploit the human captioning attention encoding rich information that human beings perceive during captioning. Unsupervised method was a breakthrough in the field of image captioning, and with the recent rise of unsupervised methods, some people have begun to study unsupervised image captioning in the field of image captioning.

Szegedy et al. [27] first proposed Inception-V1 for feature extraction, known as the GoogleNet network. A Inception structure is proposed which uses more $1 * 1$ convolution kernels and greatly reduces the number of parameters in the computation. Ioffe et al. [28] added Batch Normalization (BN) to Inception-V1 in order to prevent gradient explosion, that is, feature normalization. At the same time, the $3 * 3$ in the high-level Inception structure is decomposed into $1 * 3$ and $3 * 1$. Convolution kernels are concatenated to enhance information representation, forming an Inception-V2/V3 network. Szegedy et al. [29] combined ResNet to design Inception-v4 and Inception-ResNet-v1, Inception-ResNet-v2 networks on this basis. The performance of Inception-ResNet-v2 is higher than the previous two models. And this model is used for image feature extraction. The object detection model evolved from the RCNN model proposed by Girshick et al. [30]. In this paper, several steps of candidate region selection, feature extraction, classification and boundary regression are proposed for object detection. Due to the fact that RCNN takes up a lot of disk space, changing the image size results in missing information, and there are a lot of repeated calculations. The team later proposed the Fast RCNN model [31]. Ren [32] proposed the Faster R-CNN model on this basis, using the RPN network [33] for candidate region selection, and integration in feature extraction, which greatly improved the speed of model calculations. This model is also a more mature and widely adaptable model for object detection.

The unsupervised image captioning model (UIC model) was first proposed by Feng et al. [2] at the CVPR meeting in 2019. The model only needs an image set, a corpus, and an object detection model. The corpus teaches the model to generate fluent sentences to describe an image, the object detector guides the model to identify the key targets in the image, so that the generated fluent sentences associated with the target, finally, the image and caption are projected into a common latent space to reconstruct each other to improve the generalization ability of the model. Li [34] proposed selective-supervised contrastive learning which extends supervised contrastive learning (Sup-CL) and is powerful in representation learning. The unsupervised image captioning model is a big breakthrough in the field of image captioning. On the basis of this model, we have improved and innovated some of the remaining problems, and formed our Fast Reinforcement Learning-based unsupervised Image captioning model.

3. Fast RF-UIC

In our proposed Fast RF-UIC model, R stands for the designed R2-Inception-V4 encoder, F stands for the designed FGRU structure in decoder, and UIC stands for Unsupervised Image Captioning. As shown in Fig. 1, our Fast RF-UIC model mainly consists by Pre-trainer, encoder and decoder. The encoder is used to obtain the image features, and the decoder generates the description according to the image features. The function of the Pre-trainer is to pre-train the decoder. When an image is input, the feature of the image target is extracted through the encoder, and then the extracted image features are passed to the decoder for natural description generation, and the words corresponding to the image features are matched by loading the pre-trained model. In this section, we introduced the details from the above three departments.

3.1. Encoder: R2-Inception-V4

The function of the encoder is to extract image features and encode the input image into a vector containing image features. With reference to the idea of machine translation, we design the R2-Inception-V4 model based on Inception-V4 network. The R2-Inception-V4 model is designed to overcome the limitations of previous models in efficiently extracting image features and encoding them into a vector. As Fig. 2 We replace the reduction part of the original model with Res2Net, making the model more efficient in processing. Inception-v4 model solves the problem of model over fitting as the number of network layers

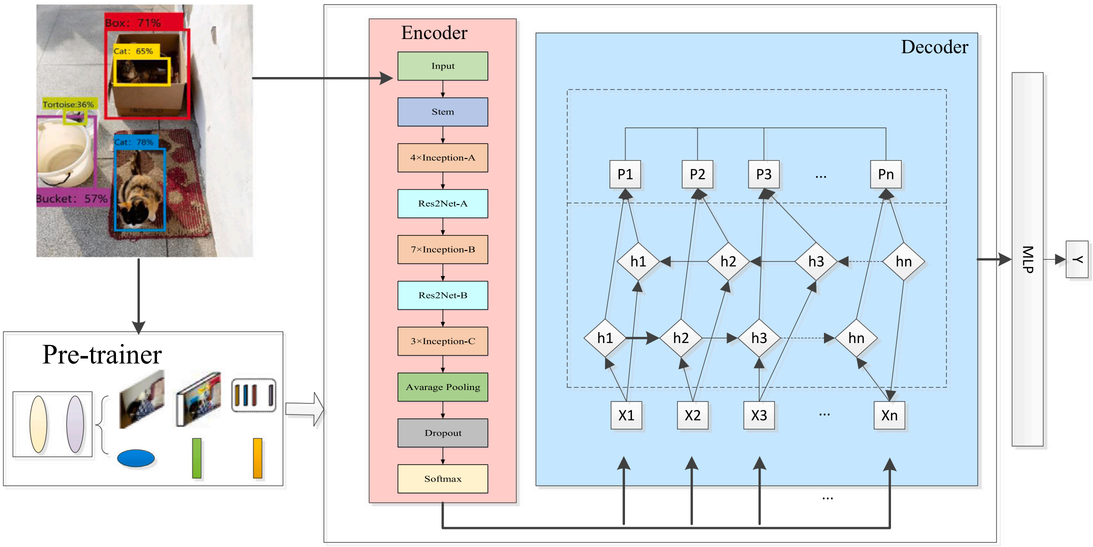


Fig. 1. Fast unsupervised residual image captioning. First input the image into Encoder for feature extraction, and use Pre-trainer to train the Decoder module. Finally, the MLP layer outputs the description.

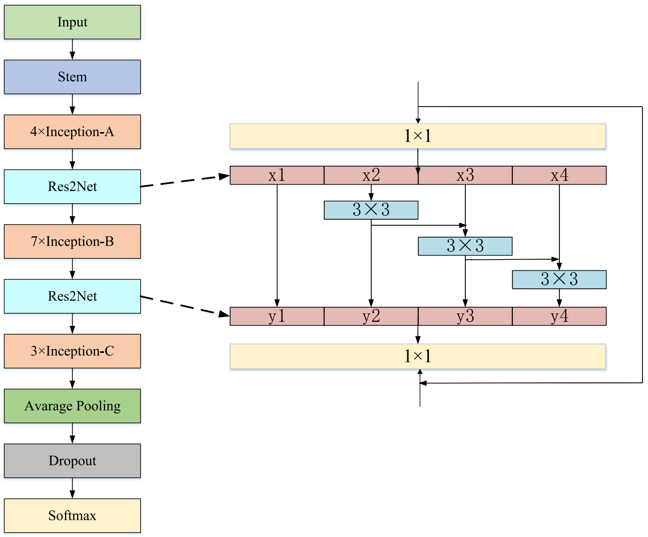


Fig. 2. R2-inception-V4. An image is input into the Stem layer, four Inception-A, one Res2Net, seven Inception-B, one Res2Net and three Inception-C. Finally, the output is obtained through the Average Pooling, Dropout, and Softmax layers. The Res2Net structure is shown on the right.

increases. Inception-V4 introduces a sparse connection structure and uses convolutional kernels of different sizes to obtain sparse features.

The Stem layer plays a crucial role in reducing computation for the subsequent inception modules by quickly reducing the resolution of the feature map. The Reduction module uses Res2Net for processing, and the core of Res2Net feature extraction model is shown on the right of Fig. 2.

Compared with the basic bottleneck, Res2Net uses a smaller convolutional network group to replace the 3 ∗ 3 convolutional layer in the bottleneck, which improves the performance of feature extraction without increasing the computational load, and the output formula of Res2Net is as follows:

$$y_i = \begin{cases} x_i & i = 1 \\ k_i\left(x_i + y_{i-1}\right) & 1 < i \le s \end{cases} \tag{1}$$

where the feature map is divided into s ($s = 4$ in the figure) block after the output of the 1 ∗ 1 convolutional layer, the larger the s, the stronger the performance of Res2Net, but it also increases the computational burden of the computer. x_i means that each subset of the feature graph, $i = \{1, 2, \dots, s\}$ each x_i has a corresponding convolutional layer of 3 ∗ 3, represented by k_i, $i = \{1, 2, \dots, s\}$. At the same time, in order to reduce the corresponding number of parameters when increasing the value of s, Res2Net eliminates the 3 ∗ 3 convolutional layer network of x_1, realizes the reuse of feature subsets, and y_i is the final output. In this way, Res2Net can obtain more than the original number of sensory fields, for example, y_2 can get 3 ∗ 3 sensory fields, y_3 can get 5 ∗ 5 sensory fields, y_4 will also get larger sizes such as 7 ∗ 7 sensory fields. Finally, the s outputs are merged and then passed through an 1 ∗ 1 convolutional layer, which can make convolution process feature maps more efficiently.

The purpose of the Inception network is to simultaneously expand the depth and width of the convolutional neural network, and enhance the sparsity of the network. Inception uses multiple convolution kernels of different scales in the width dimension for multiscale feature extraction. These convolution kernels of different sizes can also be seen as sparse representations of each other's convolution kernels. For example, a 1 ∗ 1 size convolution kernel can be seen as a sparse representation of a 3 ∗ 3 size convolution kernel. Because a 1 ∗ 1 convolution kernel can be considered as a special representation of a 3 ∗ 3 convolution kernel with a position parameter of 0, except for the central position.

Inception-A consists of 4 infrastructures as shown in Fig. 3: average pooling layer, 1 ∗ 1 convolutional layer, 3 ∗ 3 convolutional layer, and 5 ∗ 5 convolutional layer. A 1 ∗ 1 convolutional layer control feature map dimension is added after the average pooling layer, and a 1 ∗ 1 convolutional layer control feature map dimension is used before the 3 ∗ 3 and 5 ∗ 5 convolutional layers, and the 5 ∗ 5 convolution kernel is replaced with two 3 ∗ 3 convolution kernels.

The Inception-B module consists of four infrastructures as shown in Fig. 4: average pooling layer, 1 ∗ 1 convolutional layer, 7 ∗ 7 convolutional layer, and 15 ∗ 15 convolutional layer. A 1 ∗ 1 convolutional layer control feature map dimension is added after the average pooling layer, a 1 ∗ 1 convolutional layer control feature map dimension is used before the 7 ∗ 7 and 15 ∗ 15 convolutional layers, the 7 ∗ 7 convolution kernel is split with asymmetric convolution, and the 15 ∗ 15 convolution kernel is replaced by two 7 ∗ 7 convolutions first, and then split with asymmetric convolution.

Inception-C consists of 4 infrastructures as shown in Fig. 5: average pooling layer, 1 ∗ 1 convolutional layer, 3 ∗ 3 convolutional layer, and 5 ∗ 5 convolutional layer. After the average pooling layer, a 1 ∗ 1

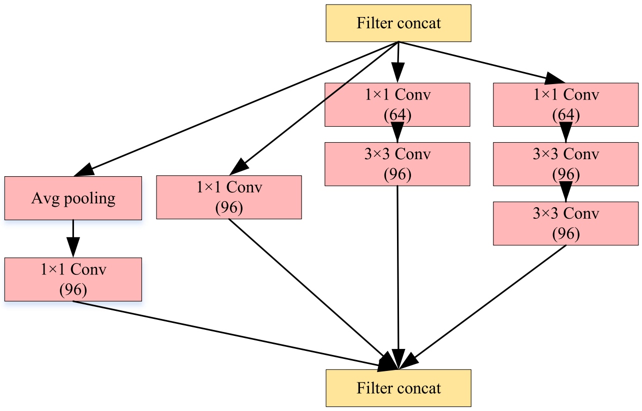


Fig. 3. Inception-A structure.

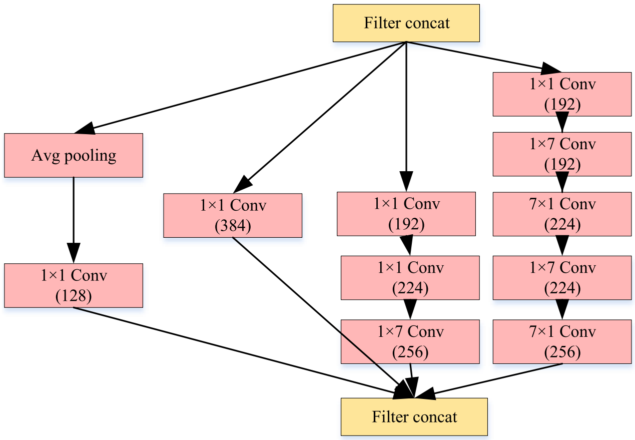


Fig. 4. Inception-B structure.

convolutional layer control feature map dimension is added, and a 1 ∗ 1 convolutional layer is used to control the feature map dimension before the 3 ∗ 3 and 5 ∗ 5 convolutional layers, and the 3 ∗ 3 convolution kernel is processed with two 1 ∗ 3, 3 ∗ 1 asymmetric convolution at the same time and the result of connecting two asymmetric convolutions, the 5 ∗ 5 convolution kernel is replaced by two 3 ∗ 3 convolution kernels, the first 3 ∗ 3 convolution kernel is split with asymmetric convolution, and the latter 3 ∗ 3 convolution kernel is split with two 1 ∗ 3, A 3 ∗ 1 asymmetric convolution processes and joins the results of two asymmetric convolution at the same time.

3.2. Decoder: Bi-FGRU

The text word vector data X_w obtained from the Decoder are input into Bi-FGRU network. GRU is a deep learning network model improved by LSTM. x_t in GRU represents input data, h_t represents GRU unit output, $\widetilde{h}_t$ represents candidate hidden state, r_t represents reset gate, and z_t represents update gate, FReLU (Funnel ReLU) is a visual parameter activation function that extends ReLU/PReLU to pixel level modeling capabilities by using spatial conditions/2D funnel conditions in the activation function (which only incurs negligible computational overhead and is very easy to implement) to extract a fine layout of objects. FReLU has two core components: funnel condition+ pixel wise modeling capability. For funnel conditions, FReLU also samples the same *max*() as a simple nonlinear function. However, for the conditional part, FReLU expands it to 2D conditions that depend on the spatial context of each pixel. The following FReLU activation function formula:

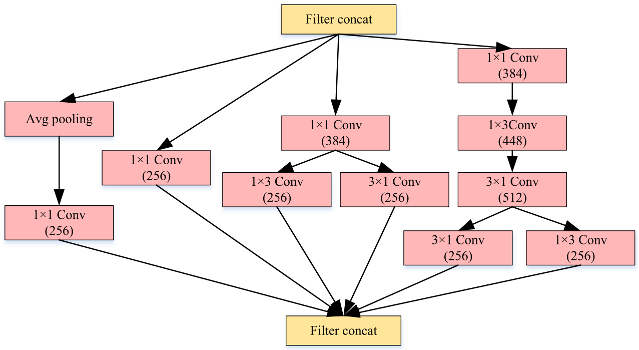


Fig. 5. Inception-C structure.

$$f\left(x_{c,i,j}\right) = \max(x_{c,i,j}, T\left(x_{c,i,j}\right)) \tag{2}$$

$$T\left(x_{c,i,j}\right) = x_{c,i,j}{}^{w} \cdot p^{w}{}_{c} \tag{3}$$

$T(x_{c,i,j})$ is the defined funnel condition that represents the window on the cth channel centered on the 2D position (i, j), and $P_c{}^{w}$ represents the parameters shared by this window in the same channel. The FReLU activation function has pixel level modeling capabilities: the definition of funnel conditions (actually a Depth Seperate convolution operation) allows the network to generate spatial conditions for each pixel in nonlinear activation. Because the function $max(x)$ provides a choice for each pixel, you can choose whether to view the spatial context (as can be seen from the formula here). For example, there are n FReLU layers $\{F1, F2, F3 \ldots Fn\}$ in the network, and each FReLU layer has a k ∗ k parameter window. For convenience, we do not consider the convolutional layer first. The receptive field set of $F1$ in the first layer is $\{1, 1+r\}$ where $r = k - 1$; At the F_n layer, the collection of receptive fields becomes $\{1, 1+r, \ldots, 1+nr\}$, which provides more choices for each pixel. If n is large enough, it can approximate any layout. Using the characteristics of this spatial condition $T(x)$ and max functions, FReLU can provide pixel level modeling or spatial layout capabilities, which can naturally and easily extract the spatial structure of an object as shown in Fig. 6. The update gate in the GRU model is used to determine whether the state information of the hidden layer at the previous moment has affected the current layer. Resetting the gate can delete invalid information from the previous moment. The two gates can store the state information of the hidden layer for a long time, and have a better ability to capture long-distance dependencies. The calculation from hidden state h_{t-1} to h_t is controlled jointly by r_t and z_t. The calculation of each door unit is as follows:

$$\begin{aligned} r_t &= \sigma\left(W_r x_t + U_r h_{t-1}\right), \\ z_t &= \sigma\left(W_z x_t + U_z h_{t-1}\right), \\ \widetilde{h}_t &= f\left(W x_t + U\left(r_t \cdot h_{t-1}\right)\right), \\ \widetilde{h}_t &= f\left(W x_t + U\left(r_t \cdot h_{t-1}\right)\right) \end{aligned} \tag{4}$$

where W_r, W_z, W represent the weight matrices of x_t at reset gate r_t, update gate z_t, candidate hidden state h_{t-1}. In the same way, U_r, U_z, U are the weight matrices of h_{t-1} at reset gate r_t, update gate z_t, candidate hidden state h_{t-1}, and σ is the sigmod activation function, and f is the FReLU activation function.

Bi-FGRU is an improvement based on Bi-GRU (Bidirectional Gate Controlled Cyclic Network). The output of Bi-FGRU is affected by bidirectional effects, avoiding the issue of the impact of one-way GRU text words on the overall logical state. The input of the CrossAttention

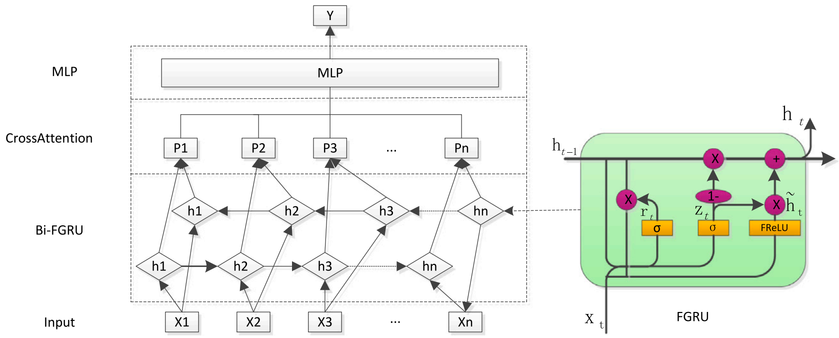


Fig. 6. Bi-FGRU. The input passes through a bidirectional Bi-FGRU layer and is calculated on CrossAttention. Finally, the description is output through the MLP layer.

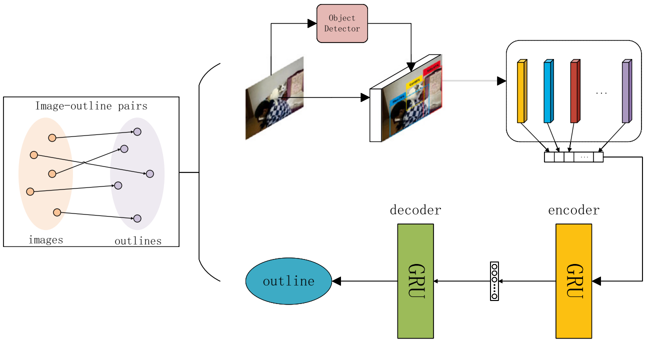


Fig. 7. Pre-trainer.

attention mechanism is the output vector of the previous layer of Bi-FGRU network. CrossAttention highlights local important information by allocating sufficient attention to key information, thereby improving the quality of hidden layer feature extraction. The feature vectors outputted by the CrossAttention mechanism are used as input vectors for the MLP input layer. MLP is a forward structured artificial neural network that can contain many independent divine light elements in each layer. These neurons located in the same layer do not have any connection with each other, but each neuron located in the upper and lower layers has a relative connection. That is, each neuron in the lower layer and each neuron in the upper layer will learn a weight value to express the strength of the connection between the upper and lower layers of neurons. The strength of this connection may be a unique feature belonging to a certain category for emotion classification tasks. For the input feature vectors of CrossAttention, the activation function sigmod is introduced to increase the smoothness of the multi-layer perceptron network, enabling it to segment nonlinear and separable data points.

3.3. Pre-trainer

The function of the Pre-trainer is to pre-train the decoder. In the absence of a image sentence dataset, we need a pre-trained model to generate pseudo captions corresponding to the input images, and use these image-pseudo caption data sets to train the decoder as shown in Fig. 7.

Considering the long training period of the pre-trained model used by the existing unsupervised image captioning methods, we propose a simpler and more effective pre-trained model. As shown in Fig. 2, the model consists of two single-layer GRUs, and uses sentences in the corpus and corresponding keywords for training, so as to generate a sentence similar to the corpus after the key target of the image is given:

$$\begin{aligned} x^{In} &= W_e C_t, t \in \{0 \dots n\}, \\ f_{\hat{S}} &= GRU^{en}\left(x^{In}\right), \\ \hat{S} &\sim GRU^{de}\left(f_{\hat{S}}\right), \end{aligned} \tag{5}$$

given a sentence $\hat{S}$ and its feature $C_t = \left\{C_1 \dots C_n\right\}$ in a corpus, the function of two single-layer GRU is to encode and decode respectively. The input x^{In} of the GRU of the encoding part is the product of the sentence feature C_t and the embedding matrix W_e to obtain the feature vector matrix $f_{\hat{S}}$ of a sentence, then we use this matrix as the input of the GRU in the decoding part, and finally we expect the result of passing through the generator GRU_{de} to be tending to the original sentence $\hat{S}$, $\sim$ stands for progressive equality. With such a feature-sentence pre-trained model and our R2-Inception-V4 to extract image features, we can create an image-sentence pseudo data set. In fact, experiments have proved that our pre-trained model with fewer layers, whose GRU has a shorter training period than LSTM, slightly improve the performance.

4. Experiments

In this section, we will experimentally verify our proposed model from the setting of experimental parameters and the comparison of model indicators. We will compare the improved pre-trained model with existing unsupervised methods through the same data set, and use BLEU [35], METEOR [36], ROUGE [37], CIDEr [38] and other indicators to evaluate and analyze the overall model. Finally, we will show the corresponding caption generated by our model after a given image.

4.1. Experiment settings

In our experiment, we used MSCOCO [39] as the input image set, to ensure robust evaluation and validation of our models, we carefully partitioned the MSCOCO dataset into distinct subsets. For the purpose of verification, we curated a subset of 5000 images, meticulously chosen to represent a balanced distribution of objects and scenes. This verification set served as a means to assess the generalizability and accuracy of our trained models on previously unseen data. For the final evaluation and testing phase, we reserved another 5000 images from the MSCOCO dataset, forming the test set. These images were carefully selected to encompass a wide range of object categories and challenging scenarios, providing a rigorous benchmark for assessing the performance of our models.

We extended the corpus utilized in our experiment by augmenting the corpus produced by Feng [2] and others. The original corpus was created by extracting relevant captions from the image website, Shutterstock, using various keywords such as “cat” and “dog”. The original corpus contained 78 target features and over 2.3 million captions. To expand the corpus, we added more descriptive sentences from other websites such as ‘Pixabay’ and ‘unsplash’. The resulting corpus

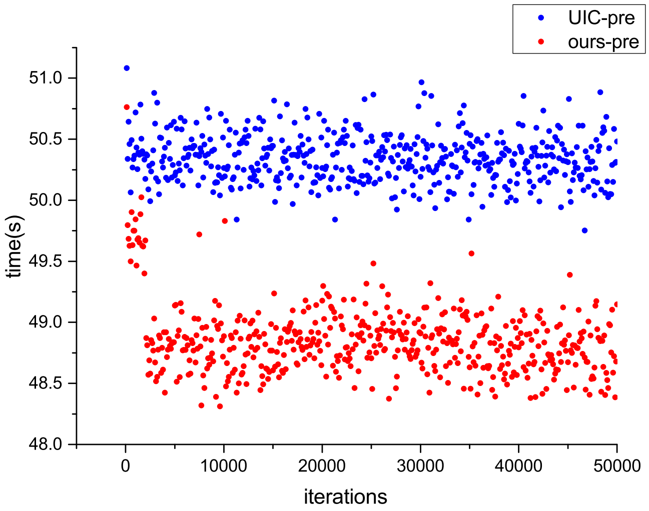


Fig. 8. Comparison of training cycles of pre-trained models.

contains nearly twice as many captions as the original corpus, and the increased number of captions greatly enhances the model's generalization ability. This further highlights the superiority of unsupervised image captioning over traditional methods. We are not constrained by existing datasets and can more flexibly incorporate different features to enhance the generalization capabilities of our model when applying it to different fields.

We fix the hidden layer dimension of GRU at 512. In addition, after a lot of parameters tuning and comparison experiments, we finally set the hyperparameters such as λ_c, λ_{im} and λ_{sen} to 10, 0.2 and 1, the value of γ^s is 0.9. At the same time, we use the stochastic optimization method proposed by Diederik [40] to train our model with a learning rate of 0.001.

4.2. Model training cycle comparison and analysis

In Section 3.3, we present a straightforward and efficient pre-trained model. Its purpose is to create an image-pseudo-caption data set for pre-trained unsupervised models. By comparing the training cycles of previous pre-trained models, we demonstrate the simplicity and effectiveness of our model. To minimize any deviation, we record the time used every 100 steps during pre-trained for each of the 50,000 steps of training conducted under the same experimental conditions. Through the scatter plot, we can visually see the advantages of our model.

In Fig. 8, the horizontal and vertical coordinates represent the number of training steps and time (unit: s) respectively. The red dot indicates the original pre-trained model, while the blue dot represents our pre-trained model. Although there are some data points deviating from the recorded time in the experiment due to various unstable factors, it is apparent from Fig. 8 that our pre-trained model has a shorter average training time per 100 steps than the original pre-trained model (48.7 s vs. 50.3 s). Importantly, the reduction in training time does not compromise model performance. We also compare the scores of pseudo captions generated by different pre-trained models, where indicators C, M, and B2-B4 refer to CIDEr, METEOR, and BLEU2-BLEU4, respectively. Our pre-trained model (Ours-pre) outperforms the original pre-trained model (UIC-pre) in all indicators, as shown in Table 1.

As shown in Table 1, compared with the pre-trained model in the UIC model, the pre-trained model designed in this paper has an average score of about 2 percentage points higher in the five evaluation indexes. The comprehensive model training rate experiment shows that our pre-trained model is superior to existing unsupervised image captioning pre-trained models, both in terms of the training period and model performance.

Table 1
The scores of two pre-trained models.

Method	C	M	B2	B3	B4
UIC-pre [2]	24.8	12.7	5.4	10.9	22.9
Ours-pre	**26.4**	**13.3**	**5.8**	**11.7**	**24.3**

Table 2
The performance distinctions on MSCOCO split test set.

Method	B1	B2	B3	B4	M	R	C	S
UIC w/o init	35.6	18.2	8.7	4.2	10.8	25.2	21.8	6.5
Ours w/o init	37.3	19.3	8.6	4.6	11.6	26.3	20.9	6.9
CLIPRe [41]	39.5	–	–	6.3	14.0	34.5	31.9	8.6
UIC [2]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	**43.3**	**24.2**	**11.7**	**6.1**	**12.5**	**30.2**	**25.9**	**8.4**

4.3. Experimental results and analysis

4.3.1. MSCOCO split test set

We use the controlled variable method, using the same data set, that is, the test set separated from the MSCOCO2014 data set, and compare the generated captions in the following indicators.

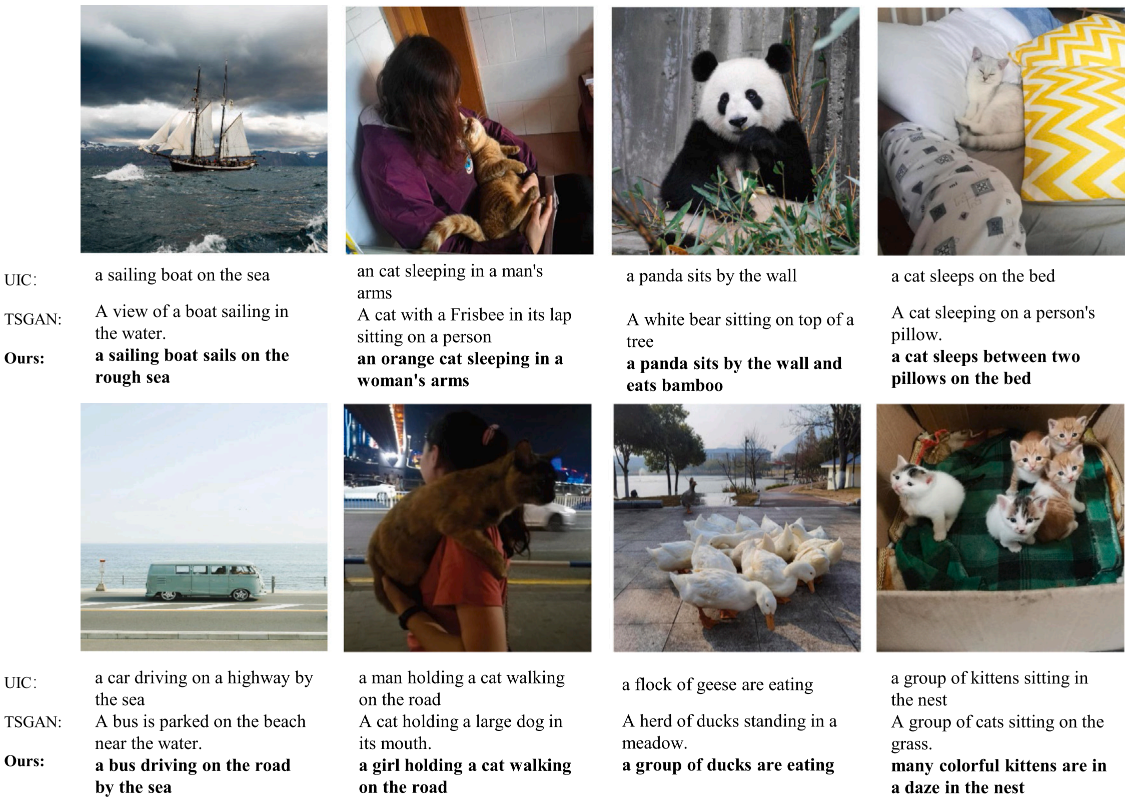
Table 2 shows the evaluation scores of each model on the MSCOCO split test set, including several indicators such as CIDEr, METEOR, and BLEU1-BLEU4, as well as R and S which refer to ROUGE and SPICE. We also included the model without pre-trained for comparison purposes. “UIC w/o init” and “Ours w/o init” represent the scores of the UIC model and our model without pre-trained, respectively. UIC refers to the UIC model presented in the paper by Feng et al. [2]. The scores under each indicator for our model and UIC are listed under “Ours”, “UIC” and “CLIPRe” [41], respectively, using the same data set. As shown in the table, the overall index score improves significantly after adding the pre-trained model, highlighting the importance of pre-trained. Comparing UIC and Ours, our model achieves higher scores in most indicators, demonstrating that our model generates more accurate captions than UIC. We also compared the CLIPRe was published in 2022, there are some indicators that our model's results are lower, however it can be seen that our BLUE score is also significantly higher.

4.3.2. Unpaired test set

The above comparison is based on the same test set, but due to the difference in language characteristics between the COCO caption and the captured image caption, the final score may not be entirely satisfactory. Therefore, we conducted horizontal comparisons with other models on different data sets to highlight the advantages of our model. Specifically, we used the MSCOCO training set without matching captions proposed by Gu et al. [42] and replaced them with sentences from our corpus.

In Table 3, we compare our model with the UIC model, the “Pivoting” model proposed by Gu et al. [42], and the TSGAN model proposed by Zhou et al. [43], which all use the same test set. Our model performs slightly better than the UIC model in several indicators, except for BLEU4 and CIDEr. Compared to the Pivoting and TSGAN models, our model outperforms them by a considerable margin. Additionally, we compared our unsupervised model to the supervised Google NIC model [14]. Although our unsupervised method is slightly inferior to the supervised approach, it is still impressive given the unsupervised nature of our model. This research on unsupervised image captioning shows the development potential of this area.

Fig. 9 presents a comparison between the UIC model, the TSGAN model, and the captions generated by our model. For instance, in the top left picture, while the original model generates a caption of “a sailboat on the sea”, our model describes “the rough sea”. Our model has a better ability to pay attention to the positional relationship of targets in images, such as “between two pillows”. Moreover, we can



UIC: a sailing boat on the sea
TSGAN: A view of a boat sailing in the water.
Ours: **a sailing boat sails on the rough sea**

UIC: an cat sleeping in a man's arms
TSGAN: A cat with a Frisbee in its lap sitting on a person
Ours: **an orange cat sleeping in a woman's arms**

UIC: a panda sits by the wall
TSGAN: A white bear sitting on top of a tree
Ours: **a panda sits by the wall and eats bamboo**

UIC: a cat sleeps on the bed
TSGAN: A cat sleeping on a person's pillow.
Ours: **a cat sleeps between two pillows on the bed**

UIC: a car driving on a highway by the sea
TSGAN: A bus is parked on the beach near the water.
Ours: **a bus driving on the road by the sea**

UIC: a man holding a cat walking on the road
TSGAN: A cat holding a large dog in its mouth.
Ours: **a girl holding a cat walking on the road**

UIC: a flock of geese are eating
TSGAN: A herd of ducks standing in a meadow.
Ours: **a group of ducks are eating**

UIC: a group of kittens sitting in the nest
TSGAN: A group of cats sitting on the grass.
Ours: **many colorful kittens are in a daze in the nest**

Fig. 9. Comparison of captions generated by different models.

Table 3
The performance distinctions on unpaired test set.

Method	B1	B2	B3	B4	M	R	C	S
Pivoting [42]	46.2	24.0	11.2	5.4	13.2	–	17.7	–
UIC [2]	58.0	40.3	27.0	18.6	17.0	43.1	54.9	11.1
TSGAN [43]	60.3	41.1	27.8	18.9	18.2	43.3	55.2	11.3
Ours	**60.9**	**43.2**	**29.3**	**18.9**	**18.7**	**45.6**	**55.8**	**11.5**
Google NIC	66.6	46.1	32.9	24.6	–	–	–	–

Table 4
Ablation contrast.

Improvement	B1	B2	B3	B4	M	R	C	S
UIC [2]	58.0	40.3	27.0	18.6	17.0	43.1	54.9	11.1
R2-Inception-V4	59.2	41.9	29.1	18.8	17.6	45.2	55.4	11.9
Bi-FGRU	59.3	42.5	28.6	18.5	18.4	44.5	55.3	11.5
Pre-trainer	58.8	41.4	27.4	18.2	16.5	42.9	51.7	10.4
Fast RF-UIC	60.9	43.2	29.3	18.9	18.7	45.6	55.8	11.5

focus on the actions of the targets, such as "eats bamboo" and "are in a daze". Due to the expanded corpus, the model can more accurately recognize "bus" instead of "car" and "duck" instead of "geese". These results demonstrate that our model can generate a descriptive sentence corresponding to an image with high performance in most cases.

4.3.3. Ablation contrast

To verify the effectiveness of our proposed method, we conducted ablation experiments on each innovation point in unpaired test set. We compared the results of our method with the original UIC model, and found improvements in all eight evaluation metrics when using the R2-Inception-V4 encoder, which demonstrates that adding the Res2net structure can effectively extract more image features and generate more descriptive image captions. Furthermore, the performance of the designed Bi-FGRU decoder was found to be one percentage point higher than that of the original model. Both the Pre-trainer and corpus expansion also improved some of the evaluation metrics. In particular, our proposed unsupervised image captioning model, Fast RF-UIC, achieved an average 2% increase in performance across all eight evaluation metrics compared to the original UIC model. These results demonstrate the effectiveness of our proposed improvements to the unsupervised image captioning model as shown in Table 4.

5. Conclusions

In this paper, we propose several improvements to the training cycle and performance of current unsupervised image captioning models. Firstly, we designed the Inception-Res2net-v4 for image feature extraction. Secondly, to bridge the gap between feature extraction and caption generation, we introduced a new decoder, the Bi-FGRU, which utilizes bidirectional gating mechanisms to capture more semantic information. Thirdly, we proposed a simple and effective application for the pre-trained model of unsupervised image captioning and we expanded the corpus with descriptive sentences which not only enhances the quality of our model but also increases its diversity and adaptability. Comparative experiments have demonstrated that our model greatly reduces the training cycle while also improving performance, and shown the development potential of unsupervised image captioning models in vision-aided systems. In the future, we will continue to conduct further research in this direction.

CRediT authorship contribution statement

Rui Yang: Conceptualization, Methodology, Writing – original draft. **Xiayu Cui:** Software, Visualization, Writing – original draft. **Qinzhi Qin:** Software, Investigation. **Zhenrong Deng:** Methodology, Resources, Validation. **Rushi Lan:** Writing – review & editing. **Xiaonan Luo:** Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 6202780103), by the Guangxi Science and Technology Project, China (Nos. AB20238013, AB22035052), Guangxi Key Laboratory of Image and Graphic Intelligent Processing Project (Nos. GIIP2211, GIIP2003), by the Innovation Project of GUET Graduate Education (No. 2023YCXB09).

References

[1] A. Oliva, A. Torralba, The role of context in object recognition, Trends in Cognitive Sciences 11 (12) (2007) 520–527.
[2] Y. Feng, L. Ma, W. Liu, J. Luo, Unsupervised image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4125–4134.
[3] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: European Conference on Computer Vision, Springer, 2010, pp. 15–29.
[4] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Babytalk: Understanding and generating simple image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (12) (2013) 2891–2903.
[5] P. Kuznetsova, V. Ordonez, T.L. Berg, Y. Choi, Treetalk: Composition and compression of trees for image descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 351–362.
[6] Y. Yang, C. Teo, H. Daumé III, Y. Aloimonos, Corpus-guided sentence generation of natural images, in: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454.
[7] A. Torralba, R. Fergus, W.T. Freeman, 80 Million tiny images: A large data set for nonparametric object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 30 (11) (2008) 1958–1970.
[8] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, Y. Choi, Collective generation of natural image descriptions, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2012, pp. 359–368.
[9] Y. Verma, A. Gupta, P. Mannem, C. Jawahar, Generating image descriptions using semantic similarities in the output space, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2013, pp. 288–293.
[10] R. Lan, X. Hu, C. Pang, Z. Liu, X. Luo, Multi-scale single image rain removal using a squeeze-and-excitation residual network, Appl. Soft Comput. 92 (2020) 106296.
[11] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, R. Lan, Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM, ACM Trans. Multimedia Comput. Commun. Appl. (TOMM) 17 (1s) (2021) 1–18.
[12] V. Ordonez, G. Kulkarni, T. Berg, Im2text: Describing images using 1 million captioned photographs, Adv. Neural Inf. Process. Syst. 24 (2011) 1143–1151.
[13] R. Lan, H. Zou, C. Pang, Y. Zhong, Z. Liu, X. Luo, Image denoising via deep residual convolutional neural networks, Signal Image Video Process. 15 (2021) 1–8.
[14] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.
[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, PMLR, 2015, pp. 2048–2057.
[16] J. Lu, C. Xiong, D. Parikh, R. Socher, Knowing when to look: Adaptive attention via a visual sentinel for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 375–383.
[17] T. Xian, Z. Li, C. Zhang, H. Ma, Dual global enhanced transformer for image captioning, Neural Netw.: Off. J. Int. Neural Netw. Soc. (148) (2022) 148–156.
[18] C. Long, H. Zhang, J. Xiao, L. Nie, T.S. 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, pp. 5659–5667.
[19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
[20] J. Lu, J. Yang, D. Batra, D. Parikh, Neural baby talk, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7219–7228.
[21] J. Wei, Z. Li, J. Zhu, H. Ma, Enhance understanding and reasoning ability for image captioning, Appl. Intell. 53 (3) (2023) 2706–2722.
[22] Y. Hua, R. Shi, P. Wang, S. Ge, Learning patch-channel correspondence for interpretable face forgery detection, IEEE Trans. Image Process. (2023).
[23] T. Xian, Z. Li, Z. Tang, H. Ma, Adaptive path selection for dynamic image captioning, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 5762–5775.
[24] S. Zhao, L. Li, H. Peng, Aligned visual semantic scene graph for image captioning, Displays 74 (2022) 102210.
[25] J. Feng, R. Liu, LRB-net: Improving VQA via division of labor strategy and multimodal classifiers, Displays 75 (2022) 102329.
[26] W. Jiang, Q. Li, K. Zhan, Y. Fang, F. Shen, Hybrid attention network for image captioning, Displays 73 (2022) 102238.
[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, Comput. Res. Repos. (2015) 1–9.
[28] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.
[29] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[30] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[31] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[32] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (6) (2016) 1137–1149.
[33] J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, X. Xue, Arbitrary-oriented scene text detection via rotation proposals, IEEE Trans. Multimed. 20 (11) (2018) 3111–3122.
[34] S. Li, X. Xia, S. Ge, T. Liu, Selective-supervised contrastive learning with noisy labels, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 316–325.
[35] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
[36] M. Denkowski, A. Lavie, Meteor universal: Language specific translation evaluation for any target language, in: Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 376–380.
[37] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81.
[38] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 4566–4575.
[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.
[40] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, Comput. Sci. (2014).
[41] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, N. Collier, Language models can see: Plugging visual controls in text generation, 2022.
[42] J. Gu, S. Joty, J. Cai, G. Wang, Unpaired image captioning by language pivoting, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 503–519.
[43] Y. Zhou, W. Tao, W. Zhang, Triple sequence generative adversarial nets for unsupervised image captioning, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 7598–7602.