Image Guidance Encoder-Decoder Model in Image Captioning and Its Application

Zhen Yang, Ziwei Zhou, Chaoyang Wang, Liang Xu

***Abstract*—This paper introduces a new network model - the Image Guidance Encoder-Decoder Model (IG-ED), designed to enhance the efficiency of image captioning and improve predictive accuracy. IG-ED, a fusion of the convolutional network VGGNet-16 and the long short-term memory network (LSTM), is designed based on the encoder-decoder structure. The image captioning performance sees significant enhancements when leveraging the IG-ED network model. The network training process unfolds in a series of steps. Initially, the input image undergoes convolution via the VGGNet-16 network, producing a 512-dimensional vector. Concurrently, each word in the image's caption is encoded to generate a corresponding 512-dimensional vector consistent with the image feature dimension. These two vectors form the input for the decoding process. Subsequently, the vectors are fed into the redesigned fusion LSTM (F-LSTM) network at different time steps to gradually train the parameters of the IG-ED framework. The training process is completed by utilizing a loss function for determining convergence. Evaluation of the IG-ED model's performance is conducted using CIDEr and seven other evaluation metrics on the MSCOCO 2014 dataset.**

The results exhibit substantial improvements over the "Adaptive Attention Mode" network and "Neural Talk" network. Additionally, the parameter count of the IG-ED architecture is significantly reduced compared to the "Adaptive Attention Mode" network, leading to decreased computational resource requirements and enabling edge computing on the neural network.

***Index Terms*—Image Captioning, VGGNet-16, LSTM**

I. Introduction

Image captioning, one of the core issues in computer vision, is closely related to image semantic analysis and image annotation technology [1], [2], [3], [4]. The research in this area plays a crucial role in various applications such as image search, information dimension reduction, video captioning [5]. video tracking [6], and human-computer interactions. The main goal of image captioning is to automatically generate coherent and accurate sentences that describe the content of an image. While humans can easily describe an image in words, achieving similar results with a computer poses several challenges. Several key factors must be addressed in image captioning, including leveraging image features effectively, translating comprehension into descriptive text, and converting these processes into logical code.

This task is highly complex and cannot be achieved easily through traditional computer algorithms. Early image captioning methods [7], [8] used to extract object information from images by merging image processing and SVM classification. The process involved inferring object information and attributes obtained in the previous stage, followed by utilizing CRF or other custom rules to create an image description. These methods heavily depended on explicit rules for sentence generation. In response, Fei-Fei Li [9] proposed a shift towards using a standard encoder-decoder framework. The method involved leveraging a high-performing convolutional neural network as the encoder to capture image feature information, followed by employing a recurrent neural network as the decoder for image description generation.

Inspired by the attention mechanism mentioned by Volodymyr [10], attention mechanism has attracted increasing attention due to its ability to leverage image characteristics effectively. Traditionally, machine translation relies on extracting feature information from the source language uniformly, which limits its capacity to analyze specific contexts. Bahdanau [11] tackled this limitation by integrating attention mechanism into machine translation, allowing the translation model to assign varied attention levels to different words during the translation process. As a result, the generated target language output becomes more logically coherent. Furthermore, Cheng J P [12] incorporated attention mechanism into Long Short-Term Memory (LSTM) networks for machine reading tasks, while Lin Z [13] utilized attention mechanism for image feature extraction and sentence sentiment analysis. Additionally, Shen T [14] leveraged attention mechanism for enhancing language understanding. Finally, [15-17] applied attention mechanism to text classification tasks, further demonstrating the versatility and effectiveness of attention mechanism across a range of applications in natural language processing.

Researchers have extensively explored attention mechanisms and their applications in image captioning. Xu K [18] introduced both soft and hard attention mechanisms, utilizing different types of attention in various regions. Lu J [19] proposed an adaptive attention mechanism, employing middle layers as visual guidance information after convolution to guide the inference process based on current

Manuscript received December 14, 2023; revised May 30, 2024.

Zhen Yang is an Associate Professor of School of Applied Technology, University of Science and Technology Liaoning, Anshan, China. (e-mail: xunlian@126.com).

Ziwei Zhou is an Associate Professor of the School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, China. (corresponding author to provide phone: 086-139-4125-5680; e-mail: 381431970@qq.com).

Chaoyang Wang is a postgraduate student of the School of Computer and Software Engineering, University of Science and Technology Liaoning. Anshan, China. (e-mail: 1669106430@qq.com).

Liang Xu is a postgraduate student of the School of Computer and Software Engineering. University of Science and Technology Liaoning. Anshan, China. (e-mail: 1606648814@qq.com).

circumstances. Previous studies [20-24] have successfully employed attention mechanisms to enhance image understanding. Additionally, [25-26] developed a "top-down" image captioning method that utilizes partial labels to generate detailed captions, thereby improving the effectiveness of image captioning. Furthermore, Ashish et al. [27] introduced a novel recurrent neural network structure known as Transformer, while Devlin et al. [28] proposed the BERT architecture. These innovative network structures leverage attention mechanisms extensively and aim to replace LSTM networks for improved captioning performance; however, they have not yet been widely adopted in image captioning applications.

While these methods have enhanced the quality of image captioning to some degree, the neural network approach still faces challenges such as high parameter complexity, algorithm intricacy, and significant CPU/GPU resource consumption during implementation. These issues pose economic and energy consumption obstacles, hindering the practical application of research outcomes.

A novel neural network architecture IG-ED is proposed to address the issues outlined above in this paper. Unlike traditional attention mechanisms, the IG-ED model leverages global image information in lieu of local image details to optimize network parameter inference for improved image captioning performance at reduced computational expense. Functionally, the encoder processes input images and corresponding annotation sentences into a standardized vector format, while the decoder extrapolates core features from these input vectors. Specifically, the encoder utilizes the VGGNet-16 convolutional neural network and a data dictionary to unify the representation of the input images and labeled sentences. In contrast, the decoder employs a "fusion long short-term memory network" (F-LSTM) to generate textual descriptions of images based on the unified vector representation. Notably, the decoder outputs are influenced by both the image vector and the labeled sentence vector, ensuring a more precise depiction of the image content. Experimental findings demonstrate that, in comparison to traditional "attention mechanisms" [19], the IG-ED model boasts a streamlined network architecture, necessitates fewer parameters, and yields superior descriptive performance.

The paper is structured as follows: section 2 presents related research on image captioning, section 3 details the IG-ED model and F-LSTM network algorithm, section 4 presents experimental results and analysis, and section 5 provides the final conclusions.

II. Encoder-Decoder Structure of Image Captioning

This framework bifurcates the image captioning process into two distinct steps: encoding and decoding. In the encoding step, a unified multi-dimensional vector is constructed to encompass all pertinent information, such as image features and labeled sentences. This vector facilitates processing by consolidating various types of data. On the other hand, the decoding step involves generating a linguistic description that effectively conveys complex information. This is accomplished by leveraging a multi-dimensional vector that encapsulates diverse information elements, such as image data, spatial information, and language aspects. In practice, the encoding step is predominantly executed by convolutional neural networks, while the decoding step primarily relies on recurrent neural networks.

A. Encoder Structure

To encode the input image, the encoding algorithm employs VGGNet-16. This process involves the removal of the classification layer of VGGNet-16 since the decoder does not rely on the image's classification results. Consequently, upon inputting the image, a 4096-dimensional global feature information is extracted. Subsequently, this global feature information undergoes conversion using the method described in [9].

$$\mathrm{V}_g = VGG(I_b) * W_{\mathrm{m}} + b_g \tag{1}$$

where, $VGG(I_b)$ is used to convert images to multi dimension vector , if the number of dimension is L, the dimension of W_m is $4096 \times L$, the dimension of b_g is L , V_g is the output vector of the convolutional neural network.

The input object labeled as a sentence needs to be vectorized. This involves transforming each word in the labeled sentence into a canonical vector using a probability dictionary to prepare for the decoding process.

B. Decoder Structure

LSTM excel in context processing and are commonly employed as decoders. With three gates, LSTM capture long-term and short-term semantic information effectively, thereby preventing gradient issues. [29-32]

The expression of each gate at time t is as follows:

$$i_t = \sigma(x_t * W_{\mathrm{x}} + h_{t-1} * W_{\mathrm{h}} + b_i) \tag{2}$$

$$f_t = \sigma(x_t * W_x + h_{t-1} * W_h + b_f) \tag{3}$$

$$o_t = \sigma(x_t * W_{\mathrm{x}} + h_{t-1} * W_{\mathrm{h}} + b_o) \tag{4}$$

Where σ is the activation function $1/(1+e^{-x})$;

$*$ means multiply of matrix;

i_t is input gate, f_t is forget gate, and o_t is the output gate. x_t means word vector input of time t, $h_{t\text{-}1}$ is the hidden state of time t-1, W_{x} and W_{h} is the weight of the network.

The decoding structure progressively extracts vector information through a recurrent neural network, revealing the input vector content over time. The cost function decreases with each iteration until it reaches a specific threshold, prompting the iteration to cease.

III. Encoder-Decoder Model Based on Vision Guidance

To train this network effectively, we utilize both image data and their corresponding language labels. The initial step in this process involves converting the image data and labels into vectors, a crucial encoding process. Subsequently, the decoder uses the encoded information to generate the output sentence corresponding to the input image. Throughout this decoding process, the network's parameters are iteratively refined. Once the network has learned from the training data and its parameters are fixed, it can accurately generate the language description given the image features. These training and testing processes are essential components of the network's operation.

The network model IG-ED is designed to achieve encoding and decoding work to enhance image captioning effectiveness. It is based on an encoder-decoder structure, with the encoder part utilizing the convolutional neural network VGGNet-16. Through this network, the image is processed into a 512-dimensional vector, while the sentence is vectorized using a probability dictionary. Each word in the labeled sentence corresponding to the image is also vectorized into a 512-dimensional vector, allowing for a unified format of vectors for subsequent decoding.

The decoder part of the model employs F-LSTM, a recurrent network that generates the language description of the image based on the unified vectors derived from both the image and the text, which are obtained from the encoder. The IG-ED network framework consists of distinct training and testing phase network models, each comprising encoding and decoding processes. Subsequent sections outline the architecture of the training stage network and the testing stage model.

Training Process of IG-ED Structure: The training stage aims to update network parameters continuously using the loss function as an indicator, based on input training images and corresponding annotation sentences. This iterative process leads to the gradual convergence of the loss function, ultimately resulting in the acquisition of all model parameters at the conclusion of training. Conversely, the encoding stage involves the conversion of each input image and each word in the associated tagged sentence into a unified vector. The convolutional neural network completes the vectorization of image information, with the VGGNet-16 network playing a key role in transforming images into 4096-dimensional vectors. Following linearization, a 512-dimensional vector is produced and forwarded to the subsequent decoding phase.

Similarly, the vectorization of labeled sentences is achieved through a vector dictionary, with every word within each labeled sentence yielding a 512-dimensional vector post-passage through the dictionary.

The second interface of image information that enters the F-LSTM network is the V_g interface of the F-LSTM, and the dedicated interface for receiving image vector input. At any time, step of the network training, the image vector V_g is used as the input information of F-LSTM to provide inferring information, then the effect of image generation sentences can be significantly improved.

The F-LSTM model is essentially a language inferring model, which combines unordered and discrete words into sentences. The function of the image vector V_g is to gradually deduce the way of network parameter combination through the image information. Therefore, the guiding role of the image vector V_g is very important.

A. The Internal Structure of F-LSTM

Inspired by the work of Lu [19], this paper introduces a new design for the network. In contrast to Lu's model, the proposed model utilizes global feature information obtained after convolution as a guiding visual cue for the decoding process. The global image information is reduced to a 512-dimensional vector post-vectorization, significantly lower than the dimensionality of the convolutional middle layer. Consequently, the parameter count in this model is notably decreased. By transforming the convolutional global information into a 512-dimensional vector, the data aligns in dimensions with the labeled language, simplifying subsequent decoder design. This reduction in overall model complexity and computational burden enhances the efficiency of the model, shown in Fig. 1.

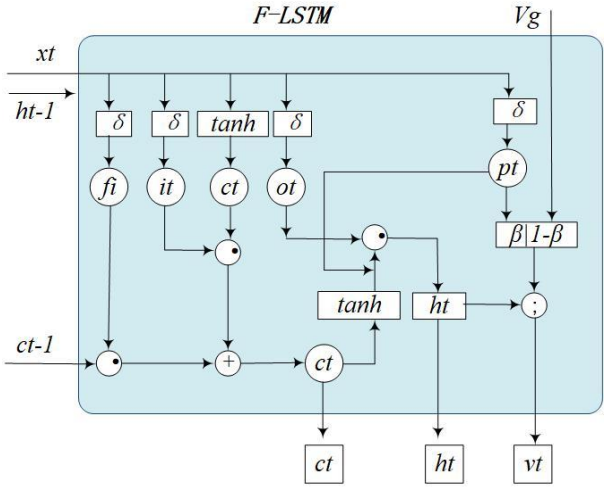

Fig. 1. Internal parameters of F-LSTM

In F-LSTM model, a monitoring gate p_t designed controls the use degree of image vectors according to the current situation. If the value of p_t is 1, it means that the global image feature information needs to be judged according to the current situation, the expression is as follows:

$$p_t = \sigma(x_t * W_x + h_{t-1} * W_h + b_p) \tag{5}$$

This model uses the ideas in the [19], the hidden state information is used at the current time step, and combined with the global image feature information and the semantic vector information controlled by the monitoring gate, then the respective proportions $\hat{\alpha}_t$ of the two are obtained. The

softmax function is used to limit the ratio to $[0,1]$, the expression of $\hat{\alpha}_t$ as follows:

$$a_t = (p_t \bullet \tanh(c_t)) * W_{ac} + h_t * W_{ah} \quad (6)$$

$$e_t = V_g * W_v + h_t * W_h \quad (7)$$

$$\hat{e}_t = \tanh(\mathrm{e}_t; a_t) * \mathrm{W}_{\hat{e}} \quad (8)$$

$$\hat{\alpha}_t = \mathrm{soft\,max}(\hat{e}_t) \quad (9)$$

Where $\tanh$ activation function is $(1-e^{-2x})/(1+e^{-2x})$, symbol ; means joining two matrices together along dimension.

Assume that $\beta = 1-\hat{\alpha}_t[1]$ is the proportion of the global image features, the contex vector u_t is:

$$u_t = (1-\beta)\cdot(p_t \bullet \tanh(c_t)) + \beta \cdot V_g \quad (10)$$

The vector h_t' is:

$$h_t' = [h_t; u_t] * W_{h'} \quad (11)$$

The main features of the newly designed F-LSTM recurrent network can be summarized as follows: Based on LSTM architecture, the neural network incorporates input gates, output gates, forget gates, and memory units to store and manage information. Additionally, it introduces new components. The memory units contain fused and distributed information, enabling efficient processing of input data.

The parameter a_t in (9) reflects the proportion of language information in the input vector, the parameter e_t in (10) reflects the proportion of the visual information. The global image information V_g is transmitted to the recurrent network at each time step. Parameter β reflects the proportion of the current input image vector V_g in the final output. This proportion has a direct impact on the generation of the final description result and plays an important role in improving the description structure.

The F-LSTM described above differs from the approach in reference [9], where the image input is sent to the LSTM only at time step 0. With prolonged calculations, the reduction of image information may no longer effectively guide the construction of the inferred sentence.

B. Design of Loss Function

In the training process, the loss function is used to evaluate the results of the network output, and to adjust the network parameters. The F-LSTM model uses the cross-entropy loss function to calculate the loss. The cross-entropy loss function is defined as follows:

$$l(\Theta) = \frac{1}{n}\sum_{i=1}^{n} H(y^{(i)}, \hat{y}^{(i)}) \quad (12)$$

Where Θ is the parameters of the model, $H(y^{(i)}, \hat{y}^{(i)})$ is cross entropy, which is defined as:

$$H(y^{(i)}, \hat{y}^{(i)}) = -\sum_{j=1}^{q} y_j^{(i)} \log \hat{y}_j^{(i)} \quad (13)$$

Where $y^{(i)}$ is the i-th probability distribution of real labels, $\hat{y}^{(i)}$ is the i th probability distribution of prediction labels, q is size of dictionary, $y^{(i)} \in \mathrm{R}^q$, $\hat{y}^{(i)} \in \mathrm{R}^q$.

C. Testing Process of IG-ED Structure

After the training process, both the encoder and decoder parameters of the network are established. The main objective of the IG-ED during training is to derive sentence information based on the input image. In this process, the encoder generates a 512-dimensional image vector, which is then fed into the F-LSTM network.

The decoder subsequently generates a language description corresponding to the image. During the testing phase, the IG-ED model receives input solely in the form of image information without any additional textual content. Following the encoding of the image, the output vector is directly input into the F-LSTM network.

The decoder then produces the textual description of the image. In the inference process, the F-LSTM network predicts one word at each time step. This prediction serves as the output from the network interface and becomes the input for the subsequent time step. These iterative steps are repeated to generate the complete sentence. The network structure in the testing stage is illustrated in Fig. 2.

IV. Experiment and Analysis

The MSCOCO 2014 dataset, established by Microsoft, is a comprehensive image dataset widely used for various computer vision tasks such as object detection, semantic segmentation, and image captioning. This dataset contains a substantial number of examples specifically tailored for Image Captioning, making it a popular choice for researchers. Each image in the dataset is associated with 5 English labels, providing rich training data for neural networks. In this study, researchers have utilized several samples from the MSCOCO 2014 dataset, The dataset comprises 82,783 training samples, 40,504 testing samples, and 40,775 validation samples. Additionally, there are separate sets of 270,000 and 886,000 images utilized for segmenting people and objects, respectively. Notably, the MSCOCO 2014 dataset serves as a crucial benchmark for evaluating the performance of models such as the IG-ED model.

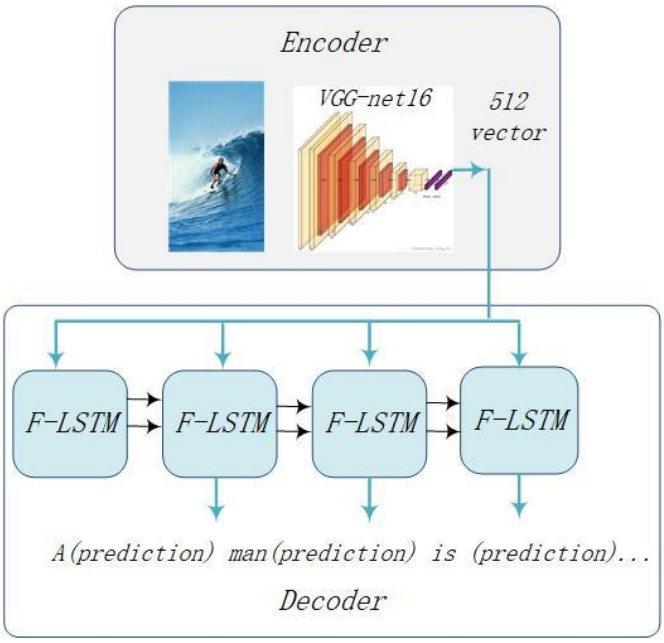

Fig. 2. Architecture of neural network in testing

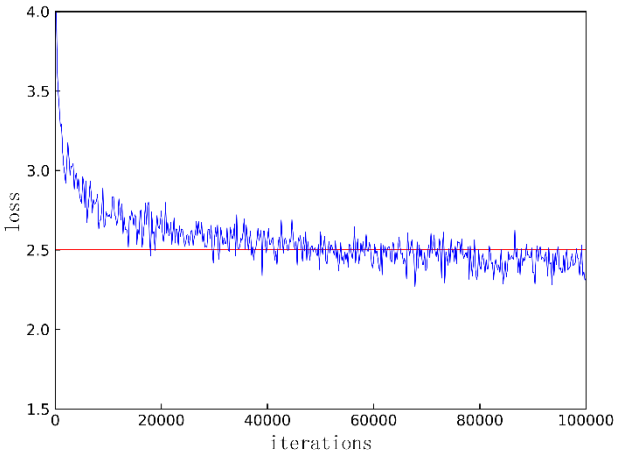

Fig. 3. Loss curve of model training

A. Training Process and Evaluation Method

Different configuration servers were used in the experiment to verify the functions of different stages. The higher configuration NVIDIA Titan X GPU server was specifically utilized during the training and evaluation stage of the dataset to obtain accurate training and evaluation results. Subsequently, in the verification stage of image captioning, the lower configuration server was employed to test the prediction results of various models. This step aimed to identify the minimum computational requirements for the network structure. Throughout the experiment, a range of network improvement methods were explored. While many of these approaches did not yield ideal results, they provided valuable insights to steer towards a more promising direction and deepen the understanding of image captioning.

The IG-ED network model undergoes training with 100,000 iterations, utilizing a batch size of 64 in each iteration. The total training duration spans approximately 87 hours. During training, the Adam optimization algorithm [35] is employed with a learning rate of 4e-4. To prevent overfitting, a dropout layer is appropriately included in the model. Notably, in all experiments, the network's parameters remain unaltered to avoid any convolutional neural network interference. The model is trained using a combination of supervised learning and end-to-end approaches, enabling faster and superior performance outcomes.

Upon completion of model training, the training loss is logged every 200 iterations, with validation performed using 3200 images to assess model performance, and average results calculated every 10,000 iterations. The training concludes after 100,000 iterations, as illustrated in Fig. 3.

Showcasing the loss curve. Following training, the model's weights at 90,000 iterations outperform those at 100,000 iterations based on observation and comparison. Thus, the weights at 90,000 iterations are chosen for evaluating experimental outcomes.

In the experimental phase, a greedy search strategy, which selects only the word with the highest score each time, fails to produce the optimal sentence description. Hence, employing the beam search method guarantees the selection of words with the highest probabilities iteratively until the final sequence is constructed.

The final bundle search size chosen is 3, achieved by continuously adjusting the parameters to improve sentence description. During model training, verification and evaluation are conducted using evaluation code support provided by coco-caption [36].

Eight indexes including BLEU1-4 [37], ROUGE [38], METEOR [39], SPICE [40], and CIDEr [41] are calculated for evaluation purposes. BLEU focuses on accuracy by analyzing text similarity through comparing the appearance of -tuples in prediction sequences with real labels. On the other hand, Rouge emphasizes recall rate by comparing the absence of -tuples in real labels within prediction sequences.

METEOR considers both recall and precision together in its evaluation. SPICE utilizes probabilistic Context-Free Grammar (PCFG) to encode predicted sequences and real labels into semantic dependency trees, applying specific rules for evaluation. CIDEr employs TF-IDF and cosine similarity to predict the similarity between description and reference sentences, making it a more suitable index for evaluating sentence quality. Fig. 4 and 5. display the progression of evaluation indices throughout model training.

B. Experimental Results

The three typical network models selected for comparison in the test and evaluation of the model results are respectively the Google NIC network [42], Neural Talk network [9], and the attention mechanism model named Attention Model [19].

These choices were made because the aforementioned papers offer classical and feasible theoretical methods. To effectively verify the IG-ED network model in the experiment, the encoder used in all the compared network models is the VGGNet-16 convolutional neural network model for image understanding, with no fine-tuning applied to the VGGNet-16 network. This approach ensures a more accurate evaluation under these specified conditions. During the test process, 1000 images from the test set are utilized as input. Each iteration involves reading one of these images,

and the evaluation results are displayed in Table 1.

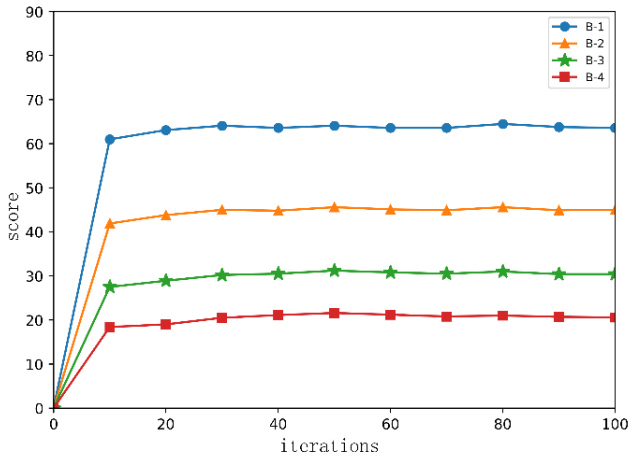

Fig. 4. Curve of the first 4 indexes of IG-ED

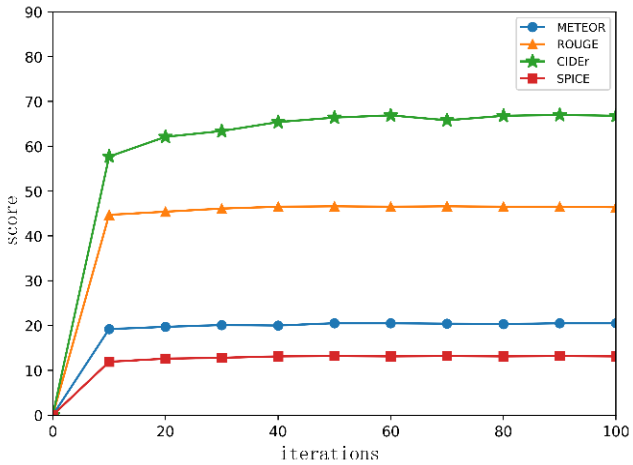

Fig. 5. Curve of the last 4 indexes of IG-ED

TABLE 1
RESULTS OF EXPERIMENT

Methods	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	CIDEr
Google NIC	66.6	46.1	32.9	24.6	-	-
Neural Talk	62.5	45.0	32.1	23.0	19.5	66.0
Attention Model	56.7	37.1	24.5	17.0	16.7	46.5
IG-ED	67.5	47.0	36.5	23.8	21.9	77.9

After comparing various evaluation indicators, it is evident that our model exhibits significant improvements over the Neural Talk [9] model. Specifically, the BLEU_1, BLEU_2, BLEU_3, METEOR, and CIDEr scores have all witnessed improvements of 3.2%, 2.2%, 1.24%, 7.18%, and 16.52% respectively when compared to the baseline model. These enhanced scores collectively demonstrate the superior predictive capabilities of the model we have developed. Analyzing the model parameters further reveals key differences between our IG-ED model and the Attention Model. The IG-ED model comprises 13,747,553 parameters, whereas the Attention Model encompasses 17,684,320 parameters, signifying a reduction of 3,936,767 parameters or 22.1% when compared to the latter. This reduction in the number of parameters is indicative of the streamlined architecture of our model in comparison to the Attention Model.

The experimental outcomes corroborate the superiority of our IG-ED model over traditional network structures in terms of prediction accuracy. Notably, our model outperforms traditional models, showcasing its advanced predictive capabilities. Additionally, to validate the effectiveness of our model across diverse datasets, predictions were made on four images from the MSCOCO 2014 dataset and two real-life images, highlighting the model's generalizability and robustness in various settings. The prediction results are shown in Fig. 6 to 7.

A Train traveling next to a forest

Fig 6. Prediction result I

A Group of boy playing foot ball

Fig 7. Prediction result II

.

V. CONCLUSION

The IG-ED network model has been developed to address the challenge of image captioning by employing an encoder-decoder architecture. The encoder leverages the VGGNet-16 network to perform convolution and generate a vector representation corresponding to the image, while the decoder utilizes the newly designed F-LSTM network structure.

By integrating image spatial information with sentence temporal information, this framework aligns with the inherent structural characteristics where images are intertwined with language. According to the findings, when utilizing the same VGGNet-16 encoder, the image captioning evaluation metric, CIDEr, demonstrates a 65.37% improvement compared to the attention mechanism. Our

IG-ED model comprises 13,747,553 network parameters, which is a 22.1% reduction from the 17,684,320 parameters in the attention mechanism. Furthermore, our model achieves real-time performance with predictions completed in under 0.5 seconds, showcasing its strong efficiency and network design advantages. The emergence of the novel language model, BERT, has introduced fresh perspectives in natural language processing. Subsequent advancements in image captioning techniques are anticipated to leverage BERT to enhance performance and deliver superior outcomes in the field.

REFERENCES

[1] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." ACM Computing Surveys (CSUR) 51.6 (2019): 1-36.

[2] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An Overview of Image Caption Generation Methods." Computational Intelligence and Neuroscience 2020 (2020).

[4] Park, Cesc Chunseong, Byeongchang Kim, and Gunhee Kim. "Towards personalized image captioning via multimodal memory networks." IEEE transactions on pattern analysis and machine intelligence 41.4 (2018): 999-1012.

[5] Lee, Sujin, and Incheol Kim. "Multimodal feature learning for video captioning." Mathematical Problems in Engineering 2018 (2018).

[6] Núñez-Marcos, Adrián, Gorka Azkune, and Ignacio Arganda-Carreras. "Vision-based fall detection with convolutional neural networks." Wireless communications and mobile computing 2017 (2017).

[7] Kulkarni, Girish, et al. "Babytalk: Understanding and generating simple image descriptions." IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

[8] Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

[9] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[10] Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." Advances in neural information processing systems. 2014.

[11] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[12] Cheng, Jianpeng, Li Dong, and Mirella Lapata. "Long short-term memory-networks for machine reading." arXiv preprint arXiv:1601.06733 (2016).

[13] Lin, Zhouhan, et al. "A structured self-attentive sentence embedding." arXiv preprint arXiv:1703.03130 (2017).

[14] Shen, Tao, et al. "Disan: Directional self-attention network for rnn/cnn-free language understanding." arXiv preprint arXiv:1709.04696 (2017).

[15] Wang, Linlin, et al. "Relation classification via multi-level attention cnns." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.

[16] Zhou, Peng, et al. "Attention-based bidirectional long short-term memory networks for relation classification." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.

[17] Yang, Zichao, et al. "Hierarchical attention networks for document classification." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016.

[18] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

[19] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[20] Pedersoli, Marco, et al. "Areas of attention for image captioning." Proceedings of the IEEE international conference on computer vision. 2017.

[21] Xu, Huijuan, and Kate Saenko. "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering." European Conference on Computer Vision. Springer, Cham, 2016.

[22] Yang, Zichao, et al. "Stacked attention networks for image question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[23] Yao, Ting, et al. "Boosting image captioning with attributes." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[24] Yun, Jing, ZhiWei Xu, and GuangLai Gao. "Gated Object-Attribute Matching Network for Detailed Image Caption." Mathematical Problems in Engineering 2020 (2020).

[25] Rennie, Steven J., et al. "Self-critical sequence training for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[26] Yang, Zhilin, et al. "Review networks for caption generation." Advances in neural information processing systems. 2016.

[27] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[28] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[29] Wang, Xuanhan, et al. "Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length." IEEE Transactions on Multimedia 20.3 (2017): 634-644.

[30] Song, Jingkuan, et al. "Self-supervised video hashing with hierarchical binary auto-encoder." IEEE Transactions on Image Processing 27.7 (2018): 3210-3221.

[31] Wang, Xuanhan, et al. "Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition." IEEE Signal Processing Letters 24.4 (2016): 510-514.

[32] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[33] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[34] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[35] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[36] Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." arXiv preprint arXiv:1504.00325 (2015).

[37] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[38] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

[39] Denkowski, Michael, and Alon Lavie. "Meteor universal: Language specific translation evaluation for any target language." Proceedings of the ninth workshop on statistical machine translation. 2014.

[40] Anderson, Peter, et al. "Spice: Semantic propositional image caption evaluation." European Conference on Computer Vision. Springer, Cham, 2016.

[41] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[42] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Zhen.Yang received the M.Sc. degree in University of Science and Technology Liaoning in 2006. He holds the position of Professor and serves as School of Applied Technology, University of Science and Technology Liaoning. His research interests include artificial intelligent, robotics, intelligent manufacturing, and 3D vision.

Ziwei Zhou obtained B.Sc. and M.Sc. degrees in computer science from University of Science and Technology Liaoning in 1997 and 2005, respectively, and Ph.D. degree in control science from the Harbin Institute of Technology in 2013. He holds the position of Professor and serves as a master's Supervisor at the School of Electronic and Information

Engineering, University of Science and Technology Liaoning. His primary areas of research focus encompass deep learning, image processing, and robot control systems.

Chaoyang Wang is a postgraduate student of the School of Computer and Software Engineering, University of Science and Technology Liaoning, His primary areas of research focus in deep learning, image processing.

Liang Xu is a postgraduate student of the School of Computer and Software Engineering, University of Science and Technology Liaoning. His primary areas of research focus in deep learning, NLP processing.