Research Article

Natural Language Description Generation Method of Intelligent Image Internet of Things Based on Attention Mechanism

Jianquan Ouyang⁴,² and Haoyang Yu⁴,²

¹Key Laboratory of Intelligent Computing & Information Processing, Ministry of Education, Xiangtan University, Xiangtan 411105, Hunan, China
²School of Computer Science & School of Cyberspace Science, Xiangtan University, Xiangtan 411105, Hunan, China

Correspondence should be addressed to Haoyang Yu; 201905720420@smail.xtu.edu.cn

Received 18 February 2022; Revised 11 March 2022; Accepted 28 March 2022; Published 11 April 2022

Academic Editor: Muhammad Arif

Copyright © 2022 Jianquan Ouyang and Haoyang Yu. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of Internet of Things technology, the image data on the Internet are growing at an amazing speed. How to describe the semantic content of massive image data is facing great challenges. Attentional mechanisms originate from the study of human vision. In cognitive science, due to bottlenecks in information processing, humans selectively attend to a portion of all information while ignoring the rest of the visible information. This study mainly discusses the natural language description generation method of Internet of Things intelligent image based on attention mechanism. In this study, a CMOS sensor based on Internet of Things technology is used for image data acquisition and display. FPGA samples cis16bit parallel port data, writes FIFO, stores image data, and then transmits it to host computer for display through network interface. In order to minimize the value of cross-entropy loss function, maximum-likelihood estimation is used to maximize the joint probability of word sequences in the language model when sentence descriptions are generated using the encoder-decoder framework. At each moment, in addition to image features, additional text features are input. Image feature vector and text feature vector are weighted and summed by attention mechanism at each time. In decoding, the attention mechanism gives each image region feature weight, and the long-term and short-term memory network decodes in turn, but the long-term and short-term memory network has limited decoding ability. We use bidirectional long-term and short-term memory network instead of long-term memory network, and dynamically focus on context information through forward LSTM and reverse LSTM. The specificity of the proposed network is 5% higher than that of the 3D convolution residual link network. The results show that the performance of image description model is improved by inputting image context and text context into long-term memory network decoder.

1. Introduction

For a long period of time, with the rapid development of mobile imaging technology and emerging social networking sites, visual data such as photos and videos have shown an explosive growth trend. These visual data contain a huge amount of information, and more and more people care about how to manage and use these data, and efficient visual data retrieval algorithms have already aroused research enthusiasm in academia and industry. With social development, the demand for information resources continues to increase, and the rapid development of communication technologies such as the mobile Internet and the popularization of Chinese text-based social software have made a large amount of Chinese text information created on the Internet almost at all times. If this massive Chinese text information is allowed to stay on the Internet disorderly, it will waste and squeeze the operating resources of the Internet, and will reduce the efficiency and utilization of the information resources hidden in the massive Chinese text information.

Huge image data resources, such as natural scene images, medical diagnosis and treatment images, aerospace images, and satellite remote sensing images, can reflect the real world more objectively and accurately. The different visual states and different changes they produce are all hidden, and rich
semantic information provides sufficient conditions for perceiving the real world. However, with the rapid growth in the quantity and value of image data information, it also makes the semantic description and understanding of massive image data content face great challenges. At the same time, it also makes the intelligent learning and understanding of image content have great theoretical research value and extensive application value and huge potential market value. To make rational use of the limited visual information processing resources, humans need to select a specific part of the visual area and then focus on it.

The attention mechanism plays a very important role in image processing. Liu et al. believe that during matching, hand-made descriptors and existing learning-based feature descriptors will limit the rendered image. In order to learn robust and invariant 128-D local feature descriptors for ground cameras and rendered images, they proposed a novel network structure SiamAM-Net, which embeds an autoencoder with attention mechanism into Siam Network. His research lacks appropriate experiments [1]. Ning et al. are inspired by the ability of organisms to quickly select the information of interest from a noisy environment and process it with limited attention resources, and introduce biological attention mechanism to design a novel selective perception framework, called attention mechanism inspired selective perception (AMiSS). In addition, they use video-based target tracking to implement proof-of-concept simulations to verify the feasibility and effectiveness of the verification. His research process is not novel enough [2]. Centenaro et al. introduced a new method of providing connectivity in IoT scenarios and discussed its advantages over established paradigms in terms of efficiency, effectiveness, and architecture design (especially for typical smart city applications). The method they proposed is not novel enough [3]. Lin et al. believe that research on IoT localization systems for smart buildings has attracted more and more attention. They proposed a novel positioning method that uses the strength of the received signal to build a fingerprint database and uses a Markov chain prediction model to assist positioning. Markov chains are stochastic processes in the probability theory and mathematical statistics that have Markovian properties and exist in discrete exponential sets and state spaces. In the LNM scheme they proposed, the historical data of pedestrian positions are analyzed to further reduce the unpredictable signal fluctuations in the intelligent building environment and at the same time to realize the calibration-free positioning of various devices, but not very practical [4]. Their research lacks experimental data. Zhou et al. introduced the architecture and unique security and privacy requirements of the next-generation mobile technology on the cloud-based Internet of Things, identified the improprieties of most existing work, and proposed solutions to resolve secure data packets forwarding and effective privacy protection, which are the challenging problems of identity verification, a new type of efficient privacy protection data aggregation that does not require public key homomorphic encryption. In the end, they put forward some interesting unsolved problems and promising ideas to trigger more research in this emerging field. His research has no practical significance [5]. The above studies provide a detailed analysis of the attention mechanism and the application of IoT technologies. It is undeniable that these studies have greatly contributed to the development of the corresponding fields. We can learn a lot of lessons from the methods and data analysis. However, there are relatively few studies on intelligent images in the field of IoT, and there is a need to fully apply these algorithms to the research in this field.

This research uses a CMOS sensor based on the Internet of Things technology to collect and display image data. When using the encoder-decoder framework to generate sentence descriptions, in order to minimize the value of the cross-entropy loss function, the maximum-likelihood estimation is used in the language model to maximize the joint probability of the word sequence. In addition to the input image features at each moment, additional text features are also input. The image feature vector and text feature vector are weighted and summed by the attention mechanism at each moment. When decoding, the attention mechanism gives each image area feature weight, and the long-short-term memory network decodes in turn, but the long-short-term memory network has limited decoding and expression ability. On the other hand, this article also intends to explore the mechanism by which attention plays a role, and to design a new explicitly guided spatial attention mechanism to improve the interpretability of the model.

2. Image Natural Language Description Generation Method

2.1. Internet of Things Technology. The Internet of Things is an application expansion based on the Internet. The Internet of Things sends out the information periodically obtained by various sensors installed in the environment through the Internet, because the number of sensors is extremely large. The core and foundation of the Internet of Things is the Internet, which is a network extended and expanded on the basis of the Internet. Its user side extends and expands to allow information exchange and communication between items and objects, which means that things are connected to each other. The Internet of Things uses various sensors to collect data on various physical changes in the environment. The growth of big data will undoubtedly challenge the ubiquitous sensing in the Internet of Things (IoT) paradigm because of limited sensing resources. Processing a large amount of sensing data requires a huge and unnecessary resource pool. Both of these reasons strongly support the idea of using selective sensing solutions to handle the mapping between physical space and cyberspace and reduce the burden of data processing in IoT applications, and the applications of IoT are broadly focused on smart home, smart transportation, smart agriculture, smart industry, smart logistics, smart electricity, smart medical, smart security, and other fields [6, 7].
Connectivity may be the most basic building block of the IoT paradigm. So far, the two main methods of providing data access to things are based on multihop mesh networks using short-range communication technologies in unlicensed spectrum, or based on long-range traditional cellular technologies mainly used for 2G/GSM/GPRS [8]. Recently, these reference models have been challenged by a new type of wireless connection, which is characterized by a low-speed, long-distance transmission technology in the license-free sub-gigahertz frequency band, which is used to achieve a star topology (called low power consumption) access network wide area network (LPWAN). The Internet of Things is increasingly becoming a ubiquitous computing service that requires a lot of data storage and processing. Unfortunately, due to the unique characteristics of resource constraints, self-organization, and short-distance communication in the Internet of Things, it has always relied on the cloud for outsourcing storage and computing, which has brought a series of new challenging security and privacy threats. When sensing the same scene, different domain image sensors or imaging mechanisms will provide cross-domain images [9, 10]. There is a domain offset between cross-domain images, so the image gap between different domains is the main challenge to measure the similarity of feature descriptors extracted from different domain images. Specifically, the ground camera image is matched with the image rendered by the UAV 3-D model. These are two extremely challenging cross-domain images that are indirectly established between 2-D and 3-D. Space is a method of spatial relations. This provides a solution for the virtual reality registration of augmented reality (AR) in outdoor environments [11].

2.2. Attention Mechanism. The idea of the attention mechanism comes from the observation and analysis of the visual processing mechanism of animals and humans [12, 13]. Attention is generally divided into two types: top-down conscious attention, called focused attention, and bottom-up unconscious attention, called salience-based attention. Focused attention is attention that has a predetermined, task-dependent, active, and conscious focus on an object. When humans observe a thing, it can be found that the processing of visual information by humans is to selectively pay attention to part of the area and ignore the rest of the information. In the field of image processing, although there are many different forms of attention modules combined with deep learning networks, their overall functions and structures are not much different. Modeling in neural networks is the most advanced model for solving multiple tasks, and outside of improving performance on the main task, they are widely used to improve the interpretability of neural networks. According to the role of its attention mechanism in neural networks, it can be divided into two types: the first is the spatial attention mechanism, and the second is the channel attention mechanism. Spatial attention is to train to find the areas that need to be paid attention to in the picture information. Channel attention makes the network focus on different filters, thereby improving the accuracy of network classification [14, 15]. The similarity between the spatial layout and the spatial layout of the most matching N regions in the image to be checked S(Q,T) can be defined as

$$S_{whole}(Q,T) = \sum_{i=1}^{n-1} \sum_{j=1}^{n} \left[ f(x_i^q - x_j^q)\text{sign}(x_i^q - x_j^q) \right] + f(y_i^j - y_j^j)\text{sign}(y_i^j - y_j^j)$$

Among them, $x_i^q$ and $y_i^j$ are the horizontal and vertical coordinates of the center of the $i$-th and $j$-th salient areas of a certain image T in the image library [16, 17], respectively.

2.3. Image Description Generation Based on Image Features and Text Features. In the image dataset, each image contains several sentence descriptions about the image. These descriptions are manually annotated. If some words in the text appear more times than other words, then it has a closer relationship with the image content [18, 19]. In the image description model based on the attention mechanism, the attention mechanism calculates the weight of each image area at the current moment according to the hidden state of the long and short-term memory network, and these areas can be associated with the words given in the title [20, 21]. With the increasing number of network information resources, a large amount of text content information has been accumulated, and how to efficiently use this potential text information, and mine users’ emotions and satisfaction with products, services, and news based on this text information, and then carry out accurate assessment of relevant users’ recommendation has become the direction that more and more researchers pay attention to [22, 23]. For example, Tmall Mall can evaluate the products that have been purchased, and according to the user evaluation, it can formulate the next sales plan according to the user’s preference; in the catering service of Meituan Dianping, the catering service can be ordered by the user. By conducting evaluations, we can dig out some service and catering quality based on these evaluations, so as to make relevant recommendations to other users; hot events that appear in the news can be analyzed by sentiment analysis of users’ hot comments, so that the management department can accurately grasp the guidance of public opinion. Avoid the intensification and occurrence of contradictions [24, 25]. Therefore, text sentiment analysis uses natural language processing technology to fully mine the sentiment contained in the text, so as to be able to conduct opinion analysis and sentiment judgment. Natural language processing technology is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that enable effective communication between humans and computers in natural language [26]. The multilayer perceptron calculates the average of the image context vector and the text context vector to initialize the cell unit $c_0$ and hidden state of the long- and short-term memory network $h_0$.
\[ c_0 = f_{\text{init},c} \left( \frac{1}{L} \sum_{i=1}^{L} a_i + \frac{1}{M} \sum_{j=1}^{M} b_j \right), \]
\[ h_0 = f_{\text{init},h} \left( \frac{1}{L} \sum_{i=1}^{L} a_i + \frac{1}{M} \sum_{j=1}^{M} b_j \right), \]

where \( i \) represents the input gate and \( f \) represents the forget gate [27].

2.4. Bilinear Interpolation. Generally, the principle of backward mapping is used to scale the image, and the bilinear interpolation method is used to calculate the gray value of each pixel. The gray value calculation method used in the process of determining the gray value is bilinear interpolation. When bilinear interpolation is applied to four adjacent pixel points, the resulting surfaces are coincident in the neighborhood, but the slopes do not match. And the smoothing effect of bilinear grayscale interpolation may make the details of the image degraded, and this phenomenon is especially obvious when image enlargement is performed. In the process of using the backward mapping algorithm, the pixel points of the output image are often mapped to noninteger points when they are mapped back to the input image. This requires the use of the gray value of the input pixel to perform gray-level interpolation to achieve the output. The gray value of the pixel is difficult to grasp, so the choice of interpolation method is also very important. According to the bilinear interpolation method, the pixel value of the coordinate point \( U \) in \((x_{ij}, y_{ij})\) is obtained [28, 29]:

\[ V_{c,j} = \sum_{i=1}^{W} \sum_{j=1}^{H} U_{c,j} \cdot \left( i - x_{ij} \right) \left( j - y_{ij} \right), \]

where \( V_{c,j} \) is the pixel value of the corresponding point. The forward mapping method is also called pixel handover mapping. Its gray value is mainly distributed by the four pixels of the target pixel through interpolation. The calculation of gray value also needs to rely on interpolation.

2.5. Long Short-Term Memory Network. Long short-term memory network is a kind of temporal recurrent neural network, which is specially designed to solve the long-term dependence problem existing in general recurrent neural networks, and all RNNs have a chained form of repeating neural network modules. The first step of long-short-term memory network (LSTM) is to delete certain information in neural network modules. \( L \) the first step of long-short-term networks, and all RNNs have a chained form of repeating dependence problem existing in general recurrent neural network, which is specially designed to solve the long-term memory network is a kind of temporal recurrent neural network. 

\[ f = \sigma \left( W_f \cdot [h_{t-1}, x_t] + b_f \right), \]

Then, use tanh to process the state of nerve cells, and get the output of information through the output gate [31, 32]. To minimize the training error, a gradient descent method can be used such as applying a temporal inverse transfer algorithm, which can be used to modify the weights each time based on errors:

\[ O_f = \sigma \left( W_0 \cdot \left[ h_{t-1}, x_t \right] + b_0 \right) + f_t \cdot C_{t-1} + i_t \cdot C_t, \]

where \( O_f \) is the output information.

3. Image Natural Language Description Generation Experiment

3.1. Experimental Environment Configuration. The operating system used in the experiment is Ubuntu 14.04, the GPU model is GeForce GTX TITANX, and the video memory is 12G. In addition, the implementation of the image description algorithm is based on the open source deep learning framework PyTorch, and its version is PyTorch-03; the python version is 3.6. The training and testing are completed by modifying the framework. And use VOC07 to complete the performance evaluation.

3.2. Establishment of Image Detection Model. The basic framework of image natural language description generation is shown in Figure 1. In order to make full use of the visual features of images and establish an image description model that can effectively process text sequences, the algorithm of this research starts from two aspects of image features and language models, designs the network structure of the image description algorithm, and proposes a method based on multiple attention. The graphic description algorithm combines force and multigranularity features [33, 34]. The language model based on the multi-attention mechanism designed in this research uses the target detection model to extract different levels of features of the image and connects the above features to the language model through the attention structure, so that the model can simultaneously focus on the overall semantic representation of the image and local details. Through the above methods [35], the language model in the image description generation algorithm can independently determine the weight of each feature, and through an adaptive method, it can generate a natural language text description consistent with the image content. The algorithm of this research consists of four parts: dataset selection, image feature extraction, language model establishment, and image description generation.

3.3. Construction of Identification Network. The recognition network is a fully connected neural network, and it accepts the feature matrix (fixed length) of the candidate region from the dense positioning layer. Its function is to stretch the feature of each candidate region into a one-dimensional column vector and make it pass through two fully connected layers, each time using the ReLu activation function and the dropout optimization principle. Finally, for each candidate region, a one-dimensional vector with length \( D = 4096 \) will be generated. This contains the visual information in this area. The information of all positive samples will be stored to form a matrix of BxD size, which is then transferred to the
natural language model. At the same time, the recognition network will refine the confidence score and location information of the candidate area twice to generate the final confidence score and location information of each candidate area. It is just another boundary regression algorithm for the vector of this length.

3.4. Image Processing. ARM11 and its follow-up products, Cortex-A8, are cost-effective. They have strong system application capabilities and image processing. Therefore, ARM11 is chosen as the node processor solution. The processor module sends an image acquisition command to the image acquisition module through a 20-pin CIF (Camera Interface) interface and reads the collected data into the memory, then performs image JPEG compression, and finally sends it to the control center through a wireless communication module based on the Internet of Things.

3.5. Collection and Display of Image Data. In the Internet of Things devices, applications need to use facilities outside the device for data collection and device control. For example, data collection needs to be connected to sensors. In this research, CMOS sensors are used for image data acquisition and display. Among them, FPGA samples CIS16bit parallel port data, writes FIFO, and stores image data, while DSP can only read FIFO and read data. DM6467 processes the received image data and then transmits it to the host computer for display through the network interface.

CIS completes the photoelectric conversion process, and after internal analog-to-digital conversion, the obtained digital signal is output to the FPGA through the parallel port, a total of 16 bits, including 12bit image signals, PIX_CLK, LINECLK, PIC_CLK, and SIGN signals. The image pixel of this CMOS image sensor is 1080×720 and the transmission rate is 64M.

DM6467 realizes the control and communication of CIS and FPGA through IIC bus. The control signal of FPGA and DM6467 also includes some reset pins. In the design, in order to make the test between different versions of the chip universal, the modular design concept is adopted in the design, and the image output part and the image data acquisition part of the CMOS image sensor are designed separately, which does not affect each other. In this way, when the chip is revised, only part of the design of the image output part and the FPGA code of the image acquisition part can be basically changed, which improves the efficiency of the test.

3.6. Feature Vector Representation. After sampling, the candidate regions we get are rectangular boxes with different sizes and aspect ratios. In order to establish a connection with the fully connected layer and the natural language model, we must extract the candidate region into a fixed-size feature representation vector. The bilinear interpolation method is used to obtain a grid map with a fixed small grid, and then according to the principle of maximum pooling, the maximum value of the pixel in the small grid is used as the feature pixel of the grid, and finally, a fixed feature vector can be obtained.

3.7. Extraction of Image Feature Vectors. Use VGGNet19 pretrained on the image net dataset to extract features of the image. Image net is a large-scale dataset containing more than 1.2 million images in 1,000 categories. It is used in research fields such as image classification, target location, and target detection. After pretraining on image net, the image features extracted by VGGNet19 have a good generalization ability, which can solve visual problems such as image description generation. The convolutional layer in VGGNet19 performs convolution operation on the input image to get the image feature vector.

3.8. Design of Image Description Generative Model. The previous image description generation model only focused on the image once, and input the feature vector obtained by the fully connected layer of the LSTH neural network to the decoder, which lost a lot of useful information of the image. Therefore, it is necessary to introduce the attention mechanism into the image description model to pay attention to some specific areas at each moment. The model mainly includes three parts: encoder (convolutional neural network, VGGNet19), attention mechanism, and decoder (long- and short-term memory neural network, LSTM). In the coding stage, VGGNet19 uses the features extracted from the lower layer (con5_3 layer) as image features and establishes the relationship between the features and the position of the picture. The attention mechanism calculates the image context vector at the current moment based on the hidden state and output at the previous moment, instead of encoding the entire picture as a vector input from the beginning. In the decoding stage, input the image context vector calculated by attention, the feature weight of the salient area of the image at the current moment is larger, and the probability of the output word is calculated according to
the hidden state and the image context vector at each moment. The most important part of image description generation is to generate reasonable sentences. The relationship between objects in the image is detected by computer vision and language models. When the image description method uses the encoder-decoder framework to generate sentence descriptions, in order to minimize the value of the cross-entropy loss function, the maximum-likelihood estimation is used in the language model to maximize the joint probability of the word sequence. However, when generating sentence descriptions, the output of the decoder depends on the words previously generated by the model. If one of the word predictions is biased, then the words generated after it will also be affected, and the final image description may deviate from the true content of the image. Therefore, this chapter introduces text features to solve this problem. In addition to the input image features at each moment, additional text features are also input. The image feature vector and text feature vector are weighted and summed by the attention mechanism at each moment. Always have a high degree of attention to certain image areas or certain words. After the introduction of text features, when predicting words at each moment, not only the current image area with larger weight but also the words in the text is considered. Therefore, the model can generate sentence descriptions that are more accurate, comprehensive, and closer to the real content of the image.

3.9. Decoder Based on Attention Mechanism and Two-Way Long- and Short-Term Memory Network. In the image description model based on the attention mechanism, when decoding, the attention mechanism gives each image area feature weight, and the long and short-term memory network decodes in turn, but the long and short-term memory network has limited decoding and expression ability, and only pays attention to the first few moments. The following information is not paid attention to. To solve this problem, we use bidirectional long-short-term memory network instead of long-short-term memory network, and dynamically focus on context information through forward LSTM and reverse LSTM.

3.10. Evaluation Index. In the experiments in this chapter, the performance of the image description algorithm is mainly evaluated from two aspects: artificial subjective sampling evaluation, by comparing the natural language description results generated by different algorithms on the same image, and evaluating the quality of the image description model; objective quantitative scoring methods, by using CIDEr, BLEU, ROUGEL, METEOR, and other evaluation criteria derived from natural language processing such as machine translation, score the generated image descriptions.

3.11. Optimization of Neural Network. Neural network may produce overfitting phenomenon during training. In order to prevent overfitting, a dropout method is added to the neural network. The discard method can be used to prevent the results of the neural network from relying too much on certain weights. When using the discarding method to train the neural network model, the neural network randomly discards some of the parameters, which is equivalent to the network being divided into multiple parts and training in turn. Because different parts of the network are trained, both the methods and values of overfitting are different. Therefore, the discarding method can average the influence of the overfitting of these different parts of the network on the overall network and, to a certain extent, can reduce the overfitting. In the process of using the discarding method to train the neural network model, the weight and bias of each update are obtained by deleting a certain proportion of hidden layer neurons. Therefore, after the training is completed, the output from the hidden layer is often neurons shrink by a corresponding proportion. During model training, this study sets the dropout to 0.6. In the model, the number of hidden nodes in the long- and short-term memory network and the hidden nodes of the attention structure in the language model is set to 1000. In order to select a more reasonable text description, in the training of the image description generation model, the training method of BeamSearch is added, and the size of Beam is set to 3.

3.12. Model Training. Long- and short-term memory networks use ReLu activation function to perform nonlinear transformations for neural networks to improve the sparse expression ability of the network. The degree of network convergence and the number of network iterations together determine whether the training can be terminated. If the accuracy of both is very low and the loss value is very large, then the network model can be judged as under-fitting. At this time, you can change the overfitting state by selecting the appropriate optimization algorithm, reducing the complexity of the network model, adjusting the number of convolution kernels, or appropriately increasing the training data samples. If the correct rate of the validation set is much less than the correct rate of the training set, then the model can be judged as an overfitting state, and regularization techniques are usually needed to increase the complexity of the network or reduce the dimensionality of the input features. In addition, if the loss value gradually decreases and finally stabilizes, the network stops training at this time. In the experiment, the Microsoft Context Common Objects (MSCOCO) dataset is used to train the model. And each image also contains 5 text descriptions of the image. The MSCOCO dataset is divided into a training set containing 6000 images, a verification set with 4000 images, and a test set with 4000 images.

In the model training in this chapter, the batch size (BatchSize) of each read data is set to 10. In order to make the model get better results, this study trains the model in stages. In the first stage, the learning rate is set to 1 × 104 and 25 rounds (Epoch) are trained. The second stage uses the self-criticism sequence training method (SCST). This
method can effectively improve the model effect by feeding back the CIDEr indicator of the algorithm model during the test as a reward to the network model. This chapter uses this method and sets the learning rate to Train 50 rounds after 1X106. Finally, change the learning rate to 1X10-6 to optimize the model parameters for 10 rounds. The experimental environment is the PyTorch learning framework based on Linux, which supports GPU computing. The specific experimental parameter settings are shown in Table 1.

4. Generation Method of Image Natural Language Description

4.1. Model Performance Analysis. Four indicators (LSTM, SEN, SPE, AUC) are selected to judge and evaluate the quality of the model. The comparison results of each index are shown in Table 2. From the comparison results, it can be found that the 3D convolutional densely connected attention network proposed in this chapter has better classification performance than the widely used ResNet and DenseNet networks. For the classification of AD and NC, the accuracy, sensitivity, and specificity of the 3D convolutional backbone network are all lower than 90%. The long- and short-term memory network has obtained better results than the 3D convolutional backbone network, because the residual module used in the long-term and short-term memory network can directly combine the information of the previous layer with the features of the latter layer after convolution, making full use of information in the network. At the same time, the information of the latter layer can be directly backpropagated to the previous layer through the reverse function, which solves the problem of gradient disappearance to a certain extent. Because of these advantages, the accuracy of the 3D convolutional residual connection network is increased to 93.43%, which is 8.36% higher than that of the 3D convolutional backbone network.

The long-short-term memory network uses direct connections from any layer to all subsequent layers, thereby increasing the accuracy rate to 94%. The accuracy of the long- and short-term memory network is 91%, which is 2.19% higher than the accuracy of the 3D convolution densely connected network without the attention mechanism. For the classification of sMCI and cMCI, the classification accuracy of the 3D convolutional backbone network is less than 80%. The accuracy rates of the 3D convolution densely connected network and the 3D densely connected attention mechanism network are 87.53% and 82%, respectively. For the classification of sMCI and cMCI, the performance of the four networks is between 70% and 80%, and the worst is still the 3D convolutional backbone network. The accuracy is 73.43%, the sensitivity is 72.50%, and the specificity is 71.86%. The accuracy rates of the 3D convolution residual connection network and the 3D convolution dense connection network are 73.24% and 76.02%, respectively. The sensitivity of the four methods is similar, but the specificity of the long- and short-term memory network is 5% higher than that of the 3D convolution residual connection network. The 3D convolution dense connection attention network has a classification accuracy of 78.79% for sMCI and cMCI. The highest results have been achieved in all networks. When there is no attention mechanism, the effective receptive field of the final output feature of the network is concentrated in the center of the image, which may not be the most critical area to distinguish the target, and the spatial attention map obtained through pretraining can well point out the object of attention, where the final output feature of the network is guided to the location of the object through high weight, and more discriminative features are obtained.

4.2. Analysis of MSCOCO Score Results. Describe the effectiveness of quantitative and qualitative results of image description generation models based on image features and text features on the MSCOCO dataset. In the model settings of this research experiment, the performance of the model is represented by BLEU and METEOR scores. The score is calculated from the matching degree between the candidate sentence description generated by the model and the multiple artificially labeled sentences provided in the MSCOCO data set. The MSCOCO score results are shown in Figure 2. In Figure 2, B-1, B-2, B-3, B-4, and M represent BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR scores. The score is represented by BLEU and METEOR scores. From the research results, it can be seen that compared to the benchmark model, BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR scores increased by 4.10%, 5.49%, 8.14%, 9.47%, and 6.28%, respectively, indicating that the introduction of text features can effectively improve model performance. In order to compare with the experimental results of mainstream models, first select the model that performs best on the test set and then generate the results of the image description on the verification set. Among them, image + text is the model proposed in this chapter. Based on the soft-attention of the model, the text feature vector is used as the additional input of LSTM, and the cross-entropy loss function and image description optimization algorithm are used to make further optimization. Compared with popular image description algorithms, the performance of this model has been improved. The decoding only relies on image features. When the predicted word is wrong, the error will be transmitted and accumulated, resulting in lower quality of the generated sentence. In this study, we input the image context vector and text context vector into the long- and short-term memory network for decoding. The attention mechanism takes into account the degree of attention to the image and text at each moment, and the parameters are updated according to the image and text during decoding. The experimental results in this chapter show that inputting the image context and text context into the long-short-term

| Parameter name | Value | Parameter name | Value |
|----------------|-------|----------------|-------|
| Dictionary size | 10368 | Training set size | 6000 |
| Validation set size | 5000 | Test set size | 4000 |
| Optimization | ADAM | INN hidden node number | 1000 |
| Alpha | 0.9 | Beam size | 3 |
| Beta | 0.999 | Dropout | 0.5 |
memory network decoder improves the performance of the image description model. Using the text context can guide the image so that the model pays attention to a specific area of the image and the words associated with that area. The model we proposed can be applied to practical scenarios with rich text, such as news and websites.

4.3. Model Calculation Time Analysis. We calculate the average calculation time for each stage. When inputting $480 \times 640$ images, our model only needs about 1 second to complete the entire subtitle processing. When inputting $1600 \times 2400$ images, the running time is less than 5 seconds. Obviously, our model is computationally efficient. Our target detection model can calculate the convolutional features of the entire input image and then reduce the computational redundancy by sharing the extracted vector of the feature map. If the remote sensing image is large, such as $10000 \times 10000$ pixels, we recommend processing the image after slicing. Here, we use time1 to indicate the operation time of the target detection model, time2 indicates the operation time of the natural language model, and time3 indicates the total operation and GPU and CPU, respectively, indicate whether the graphics card acceleration operation is invoked. The calculation results of different network models are shown in Figure 3. It can be seen from Figure 3 that the recall rate and F value of the LSTM model proposed in this study are better than Char-MEMM, Rich-L, and Trigger-Mapping based on traditional methods on the task of element role classification. In the task of element recognition, F-values are all superior to this method. Word&CharC-BiLSTM + Erratatable and LSTM + CRF, our ATT-DBiLSTM model is better than this neural network-based method on the task of element role classification. In the task of element recognition, F-values are all superior to this method. Word&CharC-BiLSTM + Erratatable combines BiLSTM and convolutional neural network, and uses the classification method to extract Chinese event element roles. The Word&CharC-BiLSTM + Erratatable method is to connect the BiLSTM and the output vector of the convolutional neural network, and uses the classification method to extract Chinese event element roles. The LSTM + CRF model is similar to Word&CharC-BiLSTM + Erratatable. The difference is that the role of event elements is extracted through CRF after the features are extracted through the convolutional neural network and the long-short-term memory network. Our ATT-DBiLSTM model uses double-layer BiLSTM to extract feature information of sentence sequences at different stages, and also introduces an attention

| Net   | LSTM% | SEN%  | SPE%  | AUC% | LSTM% | SEN%  | SPE%  | AUC% |
|-------|-------|-------|-------|------|-------|-------|-------|------|
| AlexNet | 83.07 | 85.39 | 87.36 | 93.52 | 79.35 | 83.24 | 73.43 | 83.22 |
| ResNet | 93.43 | 94    | 95.39 | 93.00 | 85.99 | 82.39 | 84.39 | 86.36 |
| DenseNet | 95    | 93    | 93.43 | 96.54 | 87.33 | 76.29 | 88.92 | 90.82 |
| CWAN   | 92    | 97.20 | 97.93 | 99.70 | 87    | 87.36 | 88.84 | 95.83 |

Figure 2: MSCOCO score results.

Figure 3: Calculation results of different network models.
mechanism to better obtain the feature information of words and the feature information between words, and comprehensively consider each word for prediction importance of information output greatly improves the performance of the Chinese event element role extraction task.

4.4. Analysis of Accuracy Rate Effect. The AT-LSTM model based on forward and reverse sequences constructed in this article is trained on the Twitter text sentiment classification task of SemEval-2017Task4. During the training and testing process of 35 epochs, the accuracy effect obtained during the training process is shown in Figure 4. From Figure 4, it can be concluded that the forward and reverse sequence AT-LSTM model achieved the best test result value at 34 times during 35 epochs. The AT-LSTM model based on forward and reverse sequences constructed in this study will be compared with the LSTM model and the AT-LSTM model on the Twitter text sentiment classification task of SemEval-2017Task4 for emotional polarity classification. The word vectors of the LSTM model and the AT-LSTM model both use the same pretrained Glove word vector as the forward and reverse sequence. In addition to the network structure, \( w \) is the most important parameter that affects the experimental results, which is used to control the preference of the change detection results on the accuracy and recall rate. If you are more concerned about how many real changed pixels are detected, that is, the recall rate, then let \( w \leq 0.5 \), at this time \( 0 \leq w < 1 \). On the contrary, if we pay more attention to the proportion of real changed pixels in the image change detection result, that is, the accuracy of the detection result, let \( c > 0.5 \) and \( w > 1 \). According to this analysis, we set \( w \) to 0.5, 1.0, and 2.0, respectively. When \( w \) is equal to 0.5, the recall rate of detection results on all datasets is the maximum. But with the increase of \( w \), the recall rate is gradually decreasing, and the result is opposite to the accuracy. It can also be seen from Figure 4 that as the accuracy increases, the image details are better preserved. \( w = 1 \) is a trade-off between recall rate and accuracy.

4.5. Feasibility Analysis of Bilinear Interpolation. The feasibility analysis result of the bilinear interpolation method is shown in Figure 5. It can be seen from the figure that after the features are represented by bilinear interpolation, the classification accuracy of the LSTM model is about 1% higher than that of CNN3. The classification accuracy rate of 3CNN is about 3% lower than that of LSTM, and the average single-round training time is higher than that of LSTM. Compared with the above models, the bilinear interpolation model proposed in this study is significantly improved. It proves the effectiveness of the semantic understanding attention algorithm model (SUAT) and the multi-feature fusion model (bilinear interpolation) proposed in this research. From the comparative experimental data in Figure 5, it can be seen that compared to the monomer basic attention mechanism model (BATT), the classification accuracy of the feature difference enhanced attention algorithm model (FDEAT) is about 10.97 percentage points higher than that of BATT, which proves the proposed feature difference enhancement attention mechanism FDEAT, by appropriately amplifying the differences between text semantic features, makes the important Chinese text features that have greater significance and greater influence on text feature recognition, and play a greater role in the process of text recognition. The role of a good Chinese text classification effect is achieved, which proves the effectiveness of the proposed feature difference enhanced attention algorithm model (FDEAT). This research has achieved obvious advantages on the MSCOCO dataset, and on the NUS-WIDE dataset, it has a slight lead compared with the optimal results in the comparison method. Similarly, the bilinear interpolation method has a slight advantage over the nonbilinear interpolation method. This is because the bilinear interpolation method directly feeds the loss of similarity maintenance through the network to the base network responsible for extracting features, making the base network a large number of parameters can be adjusted according to the current task to extract features that are more discriminative to the current task; nonbilinear interpolation can only use the fixed and invariant depth features extracted by the pretraining base network, and the parameter capacity that the model can learn quite limited, thus limiting the performance of nonbilinear interpolation. The work done in this research is to introduce a spatial attention mechanism to make the network pay more attention to the
objects related to the retrieval task in the image when extracting features, taking into account the global and local information of the image to extract more discriminative features. Therefore, it can further enhance the advantages of bilinear interpolation to extract features and help it achieve leading results.

5. Conclusion

In the task of Chinese event detection and classification, an event detection and classification method LSTM, which combines attention mechanism and long-short-term memory neural network is proposed. This method uses LSTM to capture the context information of the sentence and the semantic information of the sentence sequence and at the same time uses the attention mechanism to obtain the information between words in the sentence and calculate the weight of the word, combined with the obtained global and local information to complete event detection and classification tasks. Moreover, it does not need to rely on manual extraction of features and a series of natural language processing tools, can automatically learn the deep features of the text, and does not require professional domain knowledge, which improves the generalization ability and migration ability of the model. In order to establish a connection with the fully connected layer and the natural language model, we must extract the candidate region into a fixed-size feature representation vector. The bilinear interpolation method is used to obtain a grid map with a fixed small grid, and then according to the principle of maximum pooling, the maximum value of the pixel in the small grid is used as the feature pixel of the grid, and finally, a fixed feature vector can be obtained.

The neural network may produce overfitting during training. In order to prevent overfitting, a discarding method is added to the neural network. The discard method can be used to prevent the results of the neural network from relying too much on certain weights. When using the discarding method to train the neural network model, the neural network randomly discards some of the parameters, which is equivalent to the network being divided into multiple parts and training in turn. Because different parts of the network are trained, both the methods and values of overfitting are different. Therefore, the discarding method can average the influence of the overfitting of these different parts of the network on the overall network and to a certain extent can reduce the overfitting. In the process of using the discard method to train the neural network model, the weight and bias of each update are obtained by deleting a certain proportion of hidden layer neurons. Therefore, after the training is completed, the output from the hidden layer is often neurons shrink by a corresponding proportion.

In the image description model based on the attention mechanism, when decoding, the attention mechanism gives each image area feature weight, and the long- and short-term memory network decodes in turn, but the long- and short-term memory network has limited decoding and expression ability, and only pays attention to the first few moments. The following information is not paid attention to. To solve this problem, we use bidirectional long-short-term memory network instead of long-short-term memory network, and dynamically focus on context information through forward LSTM and reverse LSTM. The work done in this research is to introduce a spatial attention mechanism to make the network pay more attention to the objects related to the retrieval task in the image when extracting features, taking into account the global and local information of the image to extract more discriminative features. Therefore, it can further enhance the advantages of bilinear interpolation to extract features and help it achieve leading results.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This research has been supported by Key Projects of the Ministry of Science and Technology of the People’s Republic of China (2020YFC0832401). The key project of National College Student Innovation and Entrepreneurship Training Program was approved by the Ministry of Education of the People Republic of China (202110530001).

References

[1] W. Liu, C. Wang, X. Bian et al., “Learning to match ground camera image and UAV 3-D model-rendered image based on siamese network with attention mechanism,” IEEE Geoscience and Remote Sensing Letters, vol. 17, no. 9, pp. 1608–1612, 2020.
[2] H. Ning, X. Ye, A. Ben Sada, L. Mao, and M. Daneshmand, “An attention mechanism inspired selective sensing framework for physical-cyber mapping in internet of things,” IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9531–9544, 2019.
[3] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, “Long-range communications in unlicensed bands: the rising stars in the IoT and smart city scenarios,” IEEE Wireless Communications, vol. 23, no. 5, pp. 60–67, 2016.
[4] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced fingerprinting and trajectory prediction for IoT localization in smart buildings,” IEEE Transactions on Automation Science and Engineering, vol. 13, no. 3, pp. 1294–1307, 2016.
[5] J. Zhou, Z. Cao, X. Dong, and A. V. Vasilakos, “Security and privacy for cloud-based IoT: challenges,” IEEE Communications Magazine, vol. 55, no. 1, pp. 26–33, 2017.
[6] E. Ronen, A. Shamir, A.-O. Weingarten, and C. O’Flynn, “IoT goes nuclear: creating a zigbee chain reaction,” IEEE Security & Privacy, vol. 16, no. 1, pp. 54–62, 2018.
[7] J. Duan, D. Gao, D. Yang, C. H. Foh, and H. H. Chen, “An energy-aware trust derivation scheme with game theoretic approach in wireless sensor networks for IoT applications,” IEEE Internet of Things Journal, vol. 1, no. 1, pp. 58–69, 2017.
[8] E. Hodo, X. Bellekens, A. Hamilton, P. L. Dubouilh, E. Lorkyase, and C. Tachtatzis, “Threat analysis of IoT networks using artificial neural network intrusion detection system,” *Tetrahedron Letters*, vol. 42, no. 39, pp. 6865–6867, 2017.

[9] H. Cai, B. Xu, L. Jiang, and A. V. Vasilakos, “IoT-based big data storage systems in cloud computing: perspectives and challenges,” *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75–87, 2017.

[10] L. Kong, M. K. Khan, F. Wu, G. Chen, and P. Zeng, “Millimeter-wave wireless communications for IoT-cloud supported autonomous vehicles: overview, design, and challenges,” *IEEE Communications Magazine*, vol. 55, no. 1, pp. 62–68, 2017.

[11] P. Schulz, M. Matthe, H. Klessig et al., “Latency critical IoT applications in 5G: perspective on the design of radio interface and network architecture,” *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[12] A. Brogi and S. Forti, “QoS-aware deployment of IoT applications through the fog,” *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1185–1192, 2017.

[13] R. S. Sinha, Y. Wei, and S.-H. Hwang, “A survey on LPWA technology: LoRa and NB-IoT,” *Ict Express*, vol. 3, no. 1, pp. 14–21, 2017.

[14] S. Alletto, R. Cucchiara, G. Del Fiore et al., “An indoor location-aware system for an IoT-based smart museum,” *IEEE Internet of Things Journal*, vol. 3, no. 2, pp. 244–253, 2016.

[15] O. Ibrahim Khalaf, C. Andrés Tavera Romero, A. Azghu Jaisudhan Pazhani, and G. Vinuja, “VLSI implementation of a high-performance nonlinear image scaling algorithm,” *Journal of Healthcare Engineering*, vol. 2021, Article ID 6297856, 10 pages, 2021.

[16] O. I. Khalaf and G. M. Abdulshahib, “Design and performance analysis of wireless IPv6 for data exchange,” *Journal of Information Science and Engineering*, vol. 37, pp. 1335–1340, 2021.

[17] O. I. Khalaf and G. M. Abdulshahib, “Optimized dynamic storage of data (ODSD) in IoT based on blockchain for wireless sensor networks,” *Peer-to-Peer Netw. Appl.*, vol. 14, no. 2, pp. 1–16, 2021.

[18] O. I. Khalaf and G. M. Abdulshahib, “Energy efficient routing and reliable data transmission protocol in WSN,” *International Journal of Advances in Soft Computing and Its Applications*, vol. 12, no. 3, pp. 45–50, 2020.

[19] O. Skarlat, M. Nardelli, and S. Schulte, “Resource provisioning for IoT services in the fog,” *Service Oriented Computing & Applications*, vol. 11, no. 4, pp. 427–443, 2016.

[20] X. Xiang, Y. Zhang, and A. El Saddik, “Pavement crack detection network based on pyramid structure and attention mechanism,” *IJET Image Processing*, vol. 14, no. 8, pp. 1580–1586, 2020.

[21] H. F. Sang, Z. Z. Chen, and D. K. He, “Human Motion prediction based on attention mechanism,” *Multimedia Tools and Applications*, vol. 79, no. 2, pp. 1–16, 2019.

[22] M. Adil, M. A. Jan, S. Mastorakis et al., “Hash-MAC-DSDV: mutual authentication for intelligent IoT-based cyber-physical systems,” *IEEE Internet of Things Journal*, 2021.

[23] M. Adil, H. Song, J. Ali et al., “EnhancedAODV: a robust three phase priority-based traffic load balancing scheme for internet of things,” *IEEE Internet of Things Journal*, 2021.

[24] D. P. Skorich, T. B. Gash, K. L. Stalker, L. Zheng, and S. A. Haslam, “Exploring the cognitive foundations of the shared attention mechanism: evidence for a relationship between self-categorization and shared attention across the autism spectrum,” *Journal of Autism and Developmental Disorders*, vol. 47, no. 5, pp. 1341–1353, 2017.

[25] X. Xiaozhuo Wang, S. Tian, L. Yu, and Q. Yang, “Identification of local adverse drug reactions in xinjiang based on attention mechanism and BiLSTM-CNN hybrid network,” *Automatic Control and Computer Sciences*, vol. 54, no. 2, pp. 117–127, 2020.

[26] F. Wang, J. Ni, and T. Dong, “No-reference image quality assessment method based on visual attention mechanism and sharpness metric approach,” *Journal of Applied Optics*, vol. 39, no. 1, pp. 51–56, 2018.

[27] Z. Lv and H. Ko, “Introduction to the special issue on recent trends in medical data security for e-health applications,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 17, 2021.

[28] P. Shan, “Image segmentation method based on K-mean algorithm,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, p. 81, 2018.

[29] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, “DDoS in the IoT: mirai and other botnets,” *Computer*, vol. 50, no. 7, pp. 80–84, 2017.

[30] M. Kumar, A. Venkataramaiah, and N. Ambuga, “Twitter sentiment analysis using aspect-based bidirectional gated recurrent unit with self-attention mechanism,” *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 5, pp. 97–110, 2020.

[31] L. Chen, D. Zhang, P. Li, and P. Lv, “Change detection of remote sensing images based on attention mechanism,” *Computational Intelligence and Neuroscience*, vol. 2020, no. 1, pp. 1–11, Article ID 6430627, 2020.

[32] W. Li and B. Xu, “Aspect-based fashion recommendation with attention mechanism,” *IEEE Access*, vol. 8, no. 8, pp. 141814–141823, 2020.

[33] C. Tian, X. Zhu, Z. Hu, and J. Ma, “Deep spatial-temporal networks for crowd flows prediction by dilated convolutions and region-shifting attention mechanism,” *Applied Intelligence*, vol. 50, no. 10, pp. 3057–3070, 2020.

[34] N. A. Andriyanov, V. E. Dementiev, and Y. D. Kargashin, “Analysis of the impact of visual attacks on the characteristics of neural networks in image recognition,” *Procedia Computer Science*, vol. 186, no. 12, pp. 495–502, 2021.

[35] H. Li, K. Ota, and M. Dong, “Learning IoT in edge: deep learning for the internet of things with edge computing,” *IEEE Network*, vol. 32, no. 1, pp. 96–101, 2018.