Computing and Informatics, Vol. 43, 2024, 926–943, doi: 10.31577/cai_2024_4_926

IMAGE STRUCTURED ANNOTATION BASED ON DEEP NEURAL NETWORK NATURAL LANGUAGE PROCESSING

Jing Jia, Jing Hua

School of Software
Jiangxi Agricultural University
330000 Nanchang, Jiangxi, China
e-mail: {jiajing0120, kevinjia0330}@163.com

Abstract. The image structuring process was mainly divided into three stages: model training, model prediction, and report structuring. In the report structure stage, based on the feature annotation sequence, this paper associated the text sequence with the corresponding table structure and stored the text sequence in the corresponding database in the background. In dataset 1, the accuracy rate of removing visual information submodel was 30 %, and that of removing semantic information submodel was 50 %. The scheme proposed in this paper was to better perform automatic image annotation and meet the requirements of image annotation in the era of Big Data.

Keywords: Image structured annotation, natural language processing, deep neural network, image annotation technology, Big Data

1 INTRODUCTION

With the continuous updating and updating of the network technology, the way people receive information has also developed from a single text to a diversified and concrete image, video, and other aspects. A large amount of multimedia information, such as images and videos, would be generated on the network every moment. In front of a large amount of data, how to correctly process and obtain the key information in these data has become an urgent problem to be solved. With the development of image annotation technology, people begin to pay attention to the semantic mapping of images and expect to perform semantic retrieval based on automatic annotation of images. The accuracy of image semantic annotation depends on image visualization and semantic mapping. In recent years, people have had a lot of applications in speech recognition, machine translation, image recognition, and other aspects. Image structural annotation can recognize images, extract better features, and reflect the visual content of images to the maximum extent, which directly affects the semantic mapping effect and accuracy of the next step. It is difficult to describe a single pixel, contour, region and other features, and adopting better image characteristics is the difficulty and focus of current research. Semantic mapping, differentiation and generation mode, respectively, realize the mapping from visual characteristics to text.

The early semantic annotation of images was mainly based on the pattern of discrimination, while the automatic annotation of images was a supervised traditional classification problem. Chacko and Tulasi believed that accurate annotation was the key to efficient image search and retrieval. Semantic image annotation refers to adding meaningful metadata to an image to infer additional knowledge from the image. He proposed an image annotation technology combining deep learning and semantic annotation [1]. Xu et al. believed that due to the semantic gap between high-level semantic concepts and low-level visual representations, automatic image annotation was still a challenge in the practical application of computer vision [2]. Mojoo et al. believed that the task of image annotation was becoming increasingly important for efficient retrieval of images from the network and other large databases [3]. Xiao et al. believed that image annotation has always been a research hotspot in the field of computer vision. Most of the previous research work focused on labeling images with a fixed number of tags. He annotated all images with the same number of labels without considering the rationality of the image content [4]. Hou et al. believed that the recent shadow detection algorithm has shown initial success in small data sets of images from specific fields. Due to the lack of labeled training data, shadow detection in the broader image domain is still challenging. He proposed “lazy marking”. This was an effective labeling method. The announcer only needs to mark important shadow areas and some non-shadow areas [5]. Their research did not mention the accuracy of image detection.

Image feature extraction is an important basic research in image annotation. Srivastava and Srivastava proposed the overall framework of image annotation, including saliency target detection, feature extraction, feature selection, and multilabel classification [6]. Li et al. believed that in order to learn a good image annotation model, a large number of label samples were usually required. Although unmarked samples were easy to obtain and abundant in number, it was a difficult task for humans to manually label a large number of images [7]. Zhang et al. believed that the current research on two-dimensional image annotation methods lacked the annotation of historical and cultural information (such as dynasties, regions, etc.). He proposed an image annotation method based on visual attention mechanism and graph convolution network [8]. Ghostan Khatchatoorian and Jamzad believed that automatic image annotation was an image retrieval mechanism that extracted relevant semantic tags from visual content [9]. Nemade and Sonavane believed that image automatic labeling was a way to find appropriate labels for images, to obtain a method suitable for image data search and indexing. Automatic image annotation plays an important role in image retrieval and image management [10]. Jin and Jin believed that the multilabel automatic image annotation method based on machine learning has been widely used and developed. He proposed a new distance metric learning method based on cost-sensitive learning to reduce the impact of sample category imbalance [11]. Markatopoulou et al. proposed a deep convolution neural network architecture to solve the problem of video/image concept annotation by using two different levels of concept relations [12]. They need to make further research on semantic annotation of images.

When labeling an image, the trained "Inception-ResNet-V2" (ResNet's full name is Residual Network) mode can be used to obtain the visual feature vector of the labeled image. Then, based on the existing multilevel perception model, the paper extracted the semantic vector of the image by labeling the candidate labels and combined it with the visual features. Then, according to the robustness characteristics, it enters the probability of multiple labels in each label to complete the final label labeling. At the same time, on the basis of computer vision technology, the paper used CNN (Convolutional Neural Network) to extract visual information from images. Compared with the traditional recommendation algorithm, the pixel information contained in the image itself is very useful and can be mined in many places. Therefore, this paper attempted to maintain the visual information of the image and better realized the automatic labeling of the image, reducing the scarcity of image information. The scheme proposed in this paper had high accuracy when extracting the features of ImageNet images, and the overall average accuracy reached 82.4 %.

2 EXPLORATION METHOD OF IMAGE STRUCTURAL ANNOTATION

2.1 Image Processing of Natural Language

After recognizing and understanding the image, the computer must rely on natural language processing to accurately describe the image. NLP (Natural Language Processing) is an important branch of computer science, which can automatically analyze and express human natural language. The tasks of natural language processing mainly include three aspects: speech processing, machine translation, and intelligent conversation. This technology has been deeply rooted in daily life. In terms of finance, it can provide a variety of analysis data for securities investment, including hot spot mining, fraud identification, etc. In terms of law, it can assist in case retrieval, judgment, prediction, legal text translation, etc. In medical treatment, natural language processing technology has good application prospects, such as auxiliary input, query, analysis, etc.

Automatic image annotation is a complex multimode work, which combines computer images with natural language processing [13]. In recent years, with the emergence of a large number of advanced deep learning networks, automatic image annotation technology has gradually developed, which is a cross-mode work. Its development would promote the development of artificial intelligence, so that artificial intelligence can better integrate with human consciousness and learning abilities.

At present, most image description algorithms are based on coder-decoder, and the mainstream image automatic labeling algorithms are based on maximum likelihood estimation. During the training, the language generation model is used to fit the posterior probability distribution of the words in the data, and finally the cross entropy loss function is obtained through the optimal algorithm, to obtain the best parameter solution. Image description is a deep learning that combines image processing and natural language processing. It is an important subject in computer science at present. The existing processing methods generally adopt the coder-decoder structure. The main problems are the sparse text data, the exposure deviation between the encoder and decoder, and the overfitting in the model training process [14].

Automatic image labeling technology is to use the data of the training set to construct an image label that can automatically generate keywords, sentences, and other text descriptions. When performing a relevant search, users can get the corresponding search results by inputting the keyword of the image and usually use CNN to achieve labeling [15]. Through the automatic annotation technology of images, it realizes the description and retrieval of images, avoids the cumbersome manual annotation, and improves the efficiency of using text information, while taking into account the characteristics of text-based and the advantages of content-based image retrieval. Therefore, its application range is very broad, but there are still many problems in the current automatic image annotation technology:

1. Most of the image features extracted by computer are based on the basic characteristics of the image, which cannot correctly express the high-level semantic information in the image, resulting in semantic differences. In thousands of years, human beings have experienced countless times of history and knowledge, and their understanding of things is much more than the pictures they see with their eyes. At the same time, people can transform the basic visual information in the image into higher-level semantics through association and other ways, which also reflects the biggest difference between people and machines. This is also the focus of image processing technology research.
2. The impact of factors such as the size of the training set and the accuracy of the description on the training set: an accurate annotation would make the description of the model more accurate. If there is no correct training, the result would be very poor. Moreover, with the increase of the number and scale of training, the effect of automatic labeling needs to be better and better.

In image labeling, label imbalance is a common phenomenon, especially when there is a large number of labels in the labeled thesaurus, the label imbalance means that the number of labels would vary greatly. The common image label is based on the label of similar images, and the non-uniformity of the label often leads to poor labeling effect. In other words, if a label is in the collection only a few times, the possibility of labeling the label would be greatly reduced.

Due to its high flexibility, many image automatic labeling technologies based on CNN mode have achieved good results. First of all, image labeling based on CNN model would be affected by the size of the training set, and the labeling effect is not ideal when the number of training sets is small. Since the similarity between training samples and labeled samples mainly depends on the visual characteristics of the image, it is obvious that the more training samples, the better the performance of the model [16]. Secondly, when retrieving similar images, the labeling results are very sensitive, so people need to find a correct path to make the labeled image get the correct image.

2.2 Image Structured Annotation

2.2.1 Image High-Level Semantic Feature Extraction

In order to describe the image effectively, people must extract strong semantic characteristics from the image and combine them with visual characteristics to form a strong high-level feature. On this basis, it needs to use a deep neural network to learn high-level image markers, to obtain efficient semantic features. Specifically, a group of candidate tags can be constructed from the adjacent region of the image, and then the semantic characteristics of the image can be obtained by using the set of candidate tags and a simple multilevel perceptron model.

Image feature extraction is the most basic and critical part of deep learning. It can be simplified into one-dimensional data, and feature extraction is the process of converting image data into one-dimensional vectors. The feature of deep learning is to use convolution neural network to realize. When extracting full features, operations such as activation function, pooling, and fully connection are required. The activation function is used to compensate for the nonlinear factor that has poor expression ability of the linear model. The purpose of pooling is to reduce the characteristic graph, reduce the computational complexity of the network, and avoid overfitting to a certain extent. On the premise of preserving salient features, reducing feature dimensions, and increasing the perception range, the full connection layer can refit the features, thus reducing the loss of feature information.

Build candidate label set. The most fundamental principle is that the higher the similarity of two images, the higher the probability of sharing annotations between two images. Then, a group of candidate image labels can be constructed according to this idea. Using similar tags in the labeled image, a group of candidate tags to be labeled can be constructed, which would be related to semantic features. It mainly includes the following steps [17]:

The image closest to the labeled image can be obtained by using the similarity of the image. It can select the most tags from the closest images as candidate tags and select the remaining candidate tags that appear most frequently.

2.2.2 Multi-Layer Perceptron Model

Multi-layer perception (MLP), also known as multilayer neural network, can be divided into an input layer, hidden layer, and output layer, including the structure of multilayer perceptron with two hidden layers. On this basis, a nonlinear hidden layer is added to the hidden layer. In theory, as long as there are enough nodes in its hidden layer, it can fit any function. At the same time, with the increase of hidden layers, the fitting of complex functions becomes easier [18].

In the image annotation algorithm, MLP is expressed by a mathematical formula:

$$\omega(x) = H(P(P(MP + b) - c)) + b^3, \tag{1}$$

where $\omega(x)$ is the activation function. On this basis, the paper establishes two hidden layers to obtain the semantic characteristics of the image and then combines them with the visual characteristics of the image to form a powerful high-level feature. In the multi-level perception model, the most widely used ReLU (Rectified Linear Unit) function in recent years is applied. Its mathematical formula for image annotation is as follows:

$$\vartheta(x) = \max(0, x). \tag{2}$$

Clearly, if the input is less than 0, all outputs are 0; if the input is greater than 0, the output is x. There are two main reasons why ReLU function is so popular: first, it can effectively reduce the problem of gradient diffusion during reverse transmission and can also quickly update the initial parameters of the neural network. Second, the calculation speed is fast. In forward transmission, ReLU only needs to set a threshold to obtain the activation value.

2.2.3 Multi-Target Classification and Tag Number Prediction

Because each picture has one or more marks, this paper establishes a multiclassification mode for automatic marking of images. Specifically, the visual features in the CNN network and the semantic features in the MLP model are connected through a complete link layer, the probability operation is performed through the sigmoid function, and the probability value is combined with the predicted number of tags.

In the training of the multitarget classification module, the visualization and semantic characteristics of the image are combined first, and the output is obtained through the complete connection of the connection layer, and then the parameters of the image annotation model are trained using the cross entropy as the loss function:

$$Y_T = \mu + \Sigma t_b(\overline{\sigma}(x) - I)^3, \tag{3}$$

where t_b represents the actual annotation of the image; $\overline{\sigma}(x)$ represents sigmoid function.

On this basis, this paper proposes a tag number prediction model to achieve better results. On the one hand, it is expected to achieve better labeling effect by predicting the number of tags without increasing the computational complexity; on the other hand, it would want to improve the performance without increasing the computational complexity.

The image-based retrieval method is based on data and does not depend on models. Its basic idea is to map pictures and corresponding statements into a specific vector space, and then search according to the vector similarity [19]. This method is generally used to save the intermediate information of pictures and corresponding text. However, the method based on retrieval depends largely on the data in the database. In the case of insufficient or inaccurate data, the efficiency of retrieval-based methods would be greatly reduced. In the part of tag quantity prediction, the visual characteristics based on CNN network and the semantic characteristics of images obtained from MLP mode are used, and feature input needs to be carried out through a complete link layer.

$$Y_T = \Sigma(\overline{m} - m)^2, \tag{4}$$

where $\overline{m}$ is the predicted number of tags, and m is the actual number of tags of the image.

The model training and annotation process is shown in Figure 1. The training process of the whole model is as follows:

Step 1: Through the annotation of the training set image, the parameters of the Inception-ResNet-V2 model can be fine-tuned to obtain the visual characteristics of the image;

Step 2: The candidate annotation set to be labeled can be used to train the multilevel perceptron model to obtain the semantic characteristics of the image;

Step 3: The image visual characteristics obtained by Inception-ResNet-V2 mode and the semantic characteristics obtained by multilevel perceptron can be used to train the number of multitarget classification and labeling, and adjust its parameters.

When labeling the labeled image, the trained Inception-ResNet-V2 model can be used to obtain the visual feature vector of the labeled image. On this basis, people can use the existing multilayer perceptron model and then use the candidate annotation set to extract the semantic vector of the image and form a robust upper feature together with the visual characteristics. Then, according to the characteristics of robustness, the multitarget recognition model can be input into the probability of each label. Finally, people can use the dimension prediction model to predict each dimension, thus completing the final step of annotation.

Compared with traditional computer vision tasks such as image classification and target detection, automatic image labeling is more challenging and easier to

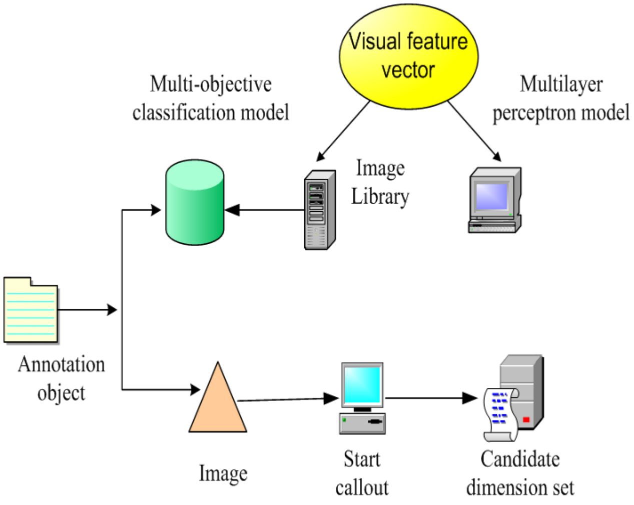


Figure 1. Model training and annotation process

understand work. The purpose of image automatic labeling is to induce semantic features with the same meaning and easy to understand from images. In the automatic labeling of images, people should not only solve the problem of object identification, but also analyze the relationship between objects in detail and express it with natural language. Therefore, for a long time, the automatic image labeling technology has been a big problem. This is a major challenge for machine learning, because it is like imitating a human's extraordinary ability to compress a large amount of visual information into a descriptive language [16].

2.2.4 Image Structured Mapping

In the structured mapping rule, the text sequence and feature mark sequence are used as input, and the output result is a group of entities with focus as the unit. The operation process is as follows:

1. Input the text sequence into the structure mapping algorithm, and divide the text into a group of clauses;
2. Based on the pathological entity generated in the unit matching in the set, it is applied to each unit in the set;
3. Based on the pathological feature descriptors generated in the set, match them with the attributes belonging to pathological entities;
4. Add a feature descriptor to the lesion entity as an attribute;
5. Add pathological entities to the pathological entity group and return structured results.

In the field of vision and speech recognition, the deep neural network technology has begun to take shape, which can combine NLP with neural networks. On this basis, it is necessary to initialize the features of the input layer using the pretraining vector, and then optimize the parameters. Deep learning is based on neural network, and its structure includes multiple inputs, output, and hidden layers, so it is called "depth". Each level outputs the input information to the next level as a special processing method, and then carries out multilevel processing, and finally obtains a specific work.

$>_m$ is the partial order relationship of the image, α represents the parameter vector in any image labeling model, and there are:

$$\prod P(>_m |\alpha) = \prod p(\mu_l >_m u_n|\alpha. \tag{5}$$

Image annotation is an interdisciplinary subject involving computer vision and natural language processing, and its research is of great significance. Most of the traditional annotation technologies adopt template-based annotation and retrieval-based annotation, which have certain defects and cannot generate flexible and smooth annotation. In the era of deep learning, the development of image annotation technology has always been based on a certain way of annotation. However, the current image annotation technology has not met people's expectations. Therefore, people need to focus on how to introduce the deep neural network mechanism into image annotation in the deep learning environment. The probability of correlation between image n and label u_n relative to label u_l is defined as:

$$P(u_l >_m u_n|\alpha) = \sigma(\alpha_i(\vartheta)). \tag{6}$$

Connect the weight Q_Z on the edge of the image label entity:

$$Q_Z = n_{(t,k)}/\Sigma n_{t,q}. \tag{7}$$

The structured process is mainly divided into three stages: model training, model prediction, and report structuring. The image structured business process is shown in Figure 2.

Step 1: In the model training stage, the image is preprocessed and manually annotated to convert it into a tagged corpus, and then the model training is carried out according to the tagged corpus.

Step 2: In the model prediction period, that is, the extraction stage of the report label, the unmarked image is preprocessed, and then input into the model and outputs a series of prediction labels.

Step 3: In the report structure stage, according to the obtained feature tag sequence, the text sequence is associated with the corresponding table structure, and the obtained feature tag sequence is associated with the structure, and then the text sequence is stored in the corresponding database in the background of the article.

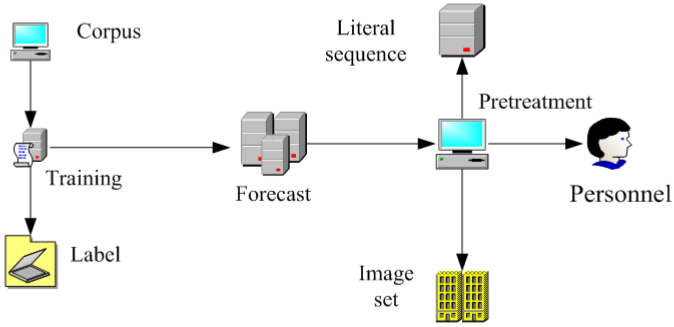


Figure 2. Image structured business process

In the process of image annotation, it is necessary to process the image, endow it with features, and associate it with keywords or text descriptions. Its essence is the mutual conversion between image and natural language. In the image processing of deep learning, image annotation occupies a large proportion. On this basis, this paper proposes an efficient and effective image annotation tool, which can effectively shorten the working time and reduce the collection of labeled data sets.

In multimedia and computer vision, automatic image annotation is a very promising research direction. In automatic image annotation, the most important technology is to reduce the "semantic gap". Computer vision, neural networks, artificial intelligence, and other technologies can effectively reduce the inconsistency between visual information and users' semantic information of images, thus reducing the gap between visual characteristics and advanced retrieval requirements. Automatic image labeling is a multidisciplinary research achievement, including data mining, semantic analysis, natural language processing, pattern recognition, machine learning, biology, and statistics.

3 RESULTS OF IMAGE STRUCTURAL ANNOTATION

With the continuous development of multimedia technology, people can express more information through pictures, and the information conveyed by pictures is clearer and more vivid than the simple text description. In today's society, there are thousands of video data every day. Therefore, how to effectively process and manage massive image data has become an urgent problem to be solved. Image annotation technology can effectively reduce human interference and reduce labor costs. Extracting features from images for semantic description is a very useful method, which can facilitate image retrieval and management. The current automatic image annotation technology has not reached the expected level, and there is a semantic gap.

On this basis, the paper studies the optimal matrix decomposition method based on CNN for the first time and combines it with the comparison model of KNN (K-

Nearest Neighbor) and LFM (Latent Factor Model) algorithms. The accuracy comparison results on image retrieval multilabel dataset 1 and image retrieval multilabel dataset 2 are shown in Figure 3 (the results of image retrieval multilabel dataset 1 and dataset 2 are shown in Figure 3 a) and 3 b), respectively). The accuracy rates of CNN on image retrieval multilabel dataset 1 and dataset 2 are 92.8 % and 90.3 %, respectively.

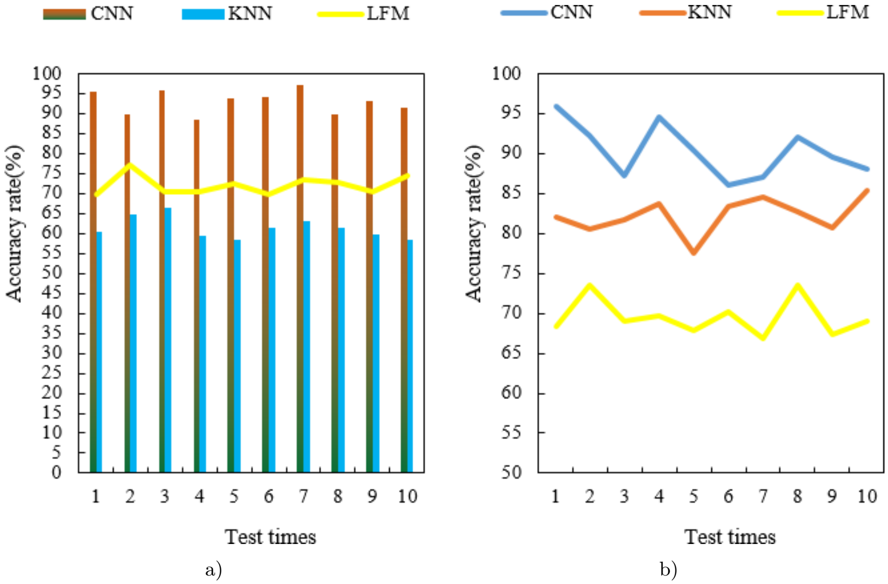


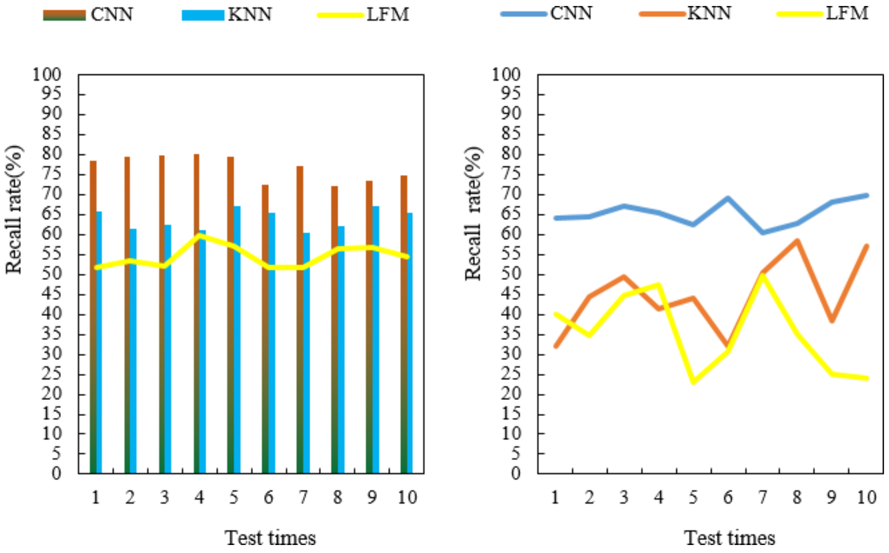
Figure 3. Comparison results of accuracy between dataset 1 and dataset 2

In the automatic labeling of images, the relationship between image-image, image-label, label-label, and image content, or restrictions are generally used to find the most suitable label for the image. The research based on graph theory can only explore the correlation of data at a deeper level, but can not fully mine out useful information. In the recommendation system, although there are many auxiliary information available, they can fully consider the potential interests of users.

Image, as an efficient and vivid information carrier, is receiving increasing attention, and the research and discussion of image retrieval has gradually become the current hot spot. At present, most of people's image retrieval focuses on the information on the image surface (low-level features and target layer), while the extraction of deep information (high-level features) is very few. The semantics of images can better reflect the viewers' subjective preferences for images, and also better reflect the users' retrieval requirements for images.

Therefore, the natural language problem can be used as a query entry to solve this problem. In terms of input, the natural language model inputs natural language problems that can better reflect the needs of users, rather than discrete keywords. In terms of output, users would get more accurate answers, such as words, phrases, and paragraphs, rather than irrelevant and complex pages. In terms of retrieval technology, the natural language model analyzes the problem from the semantic level, rather than just keyword matching. In addition, natural language can be used for retrieval, so that users can directly express their needs through subjective emotional descriptions. This can meet more needs without being limited to specific industries.

The comparison result of recall rate between image retrieval multilabel dataset 1 and image retrieval multilabel dataset 2 is shown in Figure 4 (the results of image retrieval multilabel dataset 1 and dataset 2 are shown in Figure 4 a) and 4 b), respectively). In the result of image retrieval multilabel dataset 1 and dataset 2, the CNN recall rates are 76.7 % and 65.3 %, respectively.



a) b)

Figure 4. Comparison results of recall rates on dataset 1 and dataset 2

The medical image annotation data set is built on the basis of deep learning of medical images. When establishing, people need to consider the quality of data and the efficiency of labeling. This data set is obtained through a series of processing of the original images. Deep learning technology is a semi-automatic tool, which can help researchers in deep learning obtain high-quality annotation data. It must have the following aspects: strict quality management and the quality of annotation data would directly affect the effect of subsequent deep learning models. Therefore, it is necessary to carry out strict quality control on the establishment of labeled data sets; it is easy to use and can effectively help users mark and save the time of marking. When creating the annotation data set, people can manage all aspects, including data annotation, personnel management, annotation task management, etc. However, most of the published labeling technologies still have problems such as inadequate management, inadequate quality control, and inefficient labeling of labeling personnel.

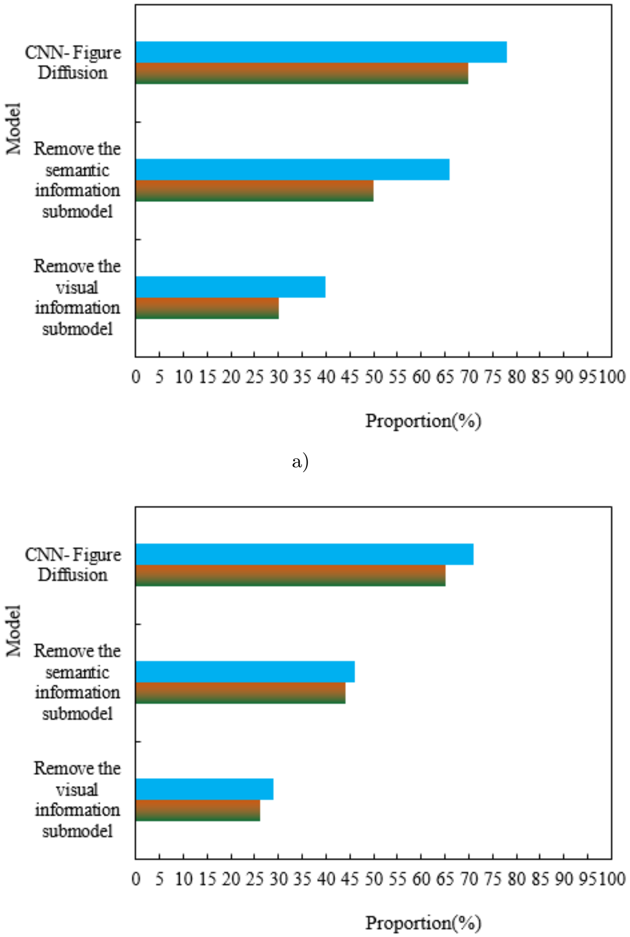
Deep learning has many advantages. It can use deep neural network and powerful computing ability to integrate data sets containing marked information into the training process. The training process is to properly train each parameter so that an unknown data can get complete information, which can make up for the shortage of human resources. Some existing machine learning systems can learn classification from machines according to specific conditions, which is beyond the scope of human capabilities.

In this part, the paper would analyze the role of each main module in the model and analyze the performance of the model after removing the image semantic information and the submodel of image visual information. The comparison of accuracy and recall rate on image retrieval multitag dataset 1 and 2 is shown in Figure 5 (the results of image retrieval multitag dataset 1 and dataset 2 are shown in Figure 5 a) and 5 b), respectively). The accuracy and recall rate of removing visual information submodel is lower than that of removing semantic information submodel. In dataset 1, the accuracy rate of removing visual information submodel is 30 %, and the accuracy rate of removing semantic information submodel is 50 %. Compared with the semantic information of images, the visual content information of images is more important. Semantic information can play an auxiliary role and enhance the role of the model. Therefore, a good semantic information extraction mode is also essential.

In recent years, with the rapid development of deep learning technology, significant breakthroughs have been made in computer vision. Different from the traditional machine learning algorithm, it needs to design and input features manually. It can only learn the model and match the initial data with the target through the deep learning method. Therefore, simulating the deep structure of human brain learning and cognitive self-learning through deep learning can better mine the correlation between image features and semantics, thus providing algorithm support for effectively narrowing the "semantic gap".

The deep structure of deep learning and the automatic learning method of the model can be used to obtain strong features, and on this basis, the correlation between image and semantic markers can be further mined. With the rapid development of deep learning technology, there have been many automatic image recognition technologies based on deep learning technology. Compared with traditional manual selection and traditional machine learning mode, under the "end-to–to-end" learning mode, the application of deep learning technology in automatic image annotation has made remarkable achievements.

The precision of feature extraction of ImageNet image using ResNet model is shown in Figure 6 (data sets are 200 and 400, as shown in Figure 6 a), and the data sets are 600 and 800, as shown in Figure 6 b)). The results show that the



a)

b)

Figure 5. Comparison of accuracy rate and recall rate on image retrieval multilabel data sets 1 and 2

model has a high accuracy when extracting the features of ImageNet images, with an overall average accuracy of 82.4 % The image content of the labeled data set is similar to that of the ImageNet data set, so the migration learning method is applied to the CNN image extraction model of the article. On the one hand, it can ensure the accuracy of the feature extraction of the model, on the other hand, it can also reduce the number of model parameters and avoid the overfitting problem caused by the small labeled data set.

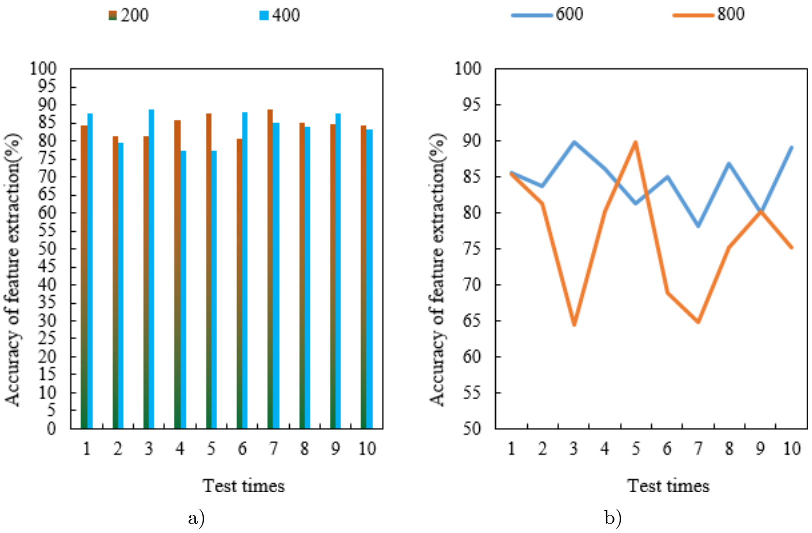


Figure 6. Accuracy of feature extraction of ImageNet image using ResNet model

4 CONCLUSIONS

In the process of automatic image labeling, the selected features must be classified. If the selected features are not representative, it is difficult to distinguish the selected objects and their relationships. Now, the best way to use traditional processing flow is to use multiple feature extractors and combine them to obtain better features. To solve this problem, this paper proposed a new learning method based on deep convolution neural network. In computer vision, some images can be generated by using antagonistic networks to make up for the lack of training data. However, in the automatic labeling of images, in addition to the image data, there should also be objective human marking, and manual marking is not only time-consuming and labor-intensive, but also cannot eliminate subjective factors. In this case, if the data encountered by the model in the article is different from the images in the training data set, it is difficult to use the coder-decoder mode to make a reasonable explanation. Therefore, the code-decoded mode needs to be optimized in the future to enhance its generalization.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61861021) and by Science and Technology Project of Education Department of Jiangxi Province (No. 190194).

REFERENCES

[1] Chacko, J. S.—Tulasi, B.: Semantic Image Annotation Using Convolutional Neural Network and WordNet Ontology. International Journal of Engineering & Technology, Vol. 7, 2018, No. 2.27, pp. 56–60.

[2] Xu, H.—Huang, C.—Huang, X.—Huang, M.: Multi-Modal Multi-Concept-Based Deep Neural Network for Automatic Image Annotation. Multimedia Tools and Applications, Vol. 78, 2019, No. 21, pp. 30651–30675, doi: 10.1007/s11042-018-6555-7.

[3] Mojoo, J.—Zhao, Y.—Kavitha, M. S.—Miyao, J.—Kurita, T.: Completion of Missing Labels for Multi-Label Annotation by a Unified Graph Laplacian Regularization. IEICE TRANSACTIONS on Information and Systems, Vol. E103-D, 2020, No. 10, pp. 2154–2161, doi: 10.1587/transinf.2019EDP7318.

[4] Xiao, F.—Chen, Y.—Zhang, Y.—Gong, X.—Gao, X.: Adaptive Image Annotation: Refining Labels According to Contents and Relations. Neural Computing and Applications, Vol. 34, 2022, No. 9, pp. 7271–7282, doi: 10.1007/s00521-021-06866-y.

[5] Hou, L.—Vicente, T. F. Y.—Hoai, M.—Samaras, D.: Large Scale Shadow Annotation and Detection Using Lazy Annotation and Stacked CNNs. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 43, 2019, No. 4, pp. 1337–1351, doi: 10.1109/TPAMI.2019.2948011.

[6] Srivastava, G.—Srivastava, R.: Design, Analysis, and Implementation of Efficient Framework for Image Annotation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol. 16, 2020, No. 3, Art. No. 89, doi: 10.1145/3386249.

[7] Li, Z.—Lin, L.—Zhang, C.—Ma, H.—Zhao, W.—Shi, Z.: A Semi-Supervised Learning Approach Based on Adaptive Weighted Fusion for Automatic Image Annotation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), Vol. 17, 2021, No. 1, Art. No. 37, doi: 10.1145/3426974.

[8] Zhang, S.—Chen, S.—Zhang, J.—Cai, Z.—Hu, L.: Image Annotation of Ancient Chinese Architecture Based on Visual Attention Mechanism and GCN. Multimedia Tools and Applications, Vol. 81, 2022, No. 28, pp. 39963–39980, doi: 10.1007/s11042-022-12618-4.

[9] Ghostan Khatchatoorian, A.—Jamzad, M.: Architecture to Improve the Accuracy of Automatic Image Annotation Systems. IET Computer Vision, Vol. 14, 2020, No. 5, pp. 214–223, doi: 10.1049/iet-cvi.2019.0500.

[10] Nemade, S.—Sonavane, S.: Refinement of CNN Based Multi-Label Image Annotation. Turkish Journal of Computer and Mathematics Education, Vol. 12, 2021, No. 13, pp. 1935–1941.

[11] Jin, C.—Jin, S. W.: Multi-Label Automatic Image Annotation Approach Based on Multiple Improvement Strategies. IET Image Processing, Vol. 13, 2019, No. 4, pp. 623–633, doi: 10.1049/iet-ipr.2018.5371.

[12] Markatopoulou, F.—Mezaris, V.—Patras, I.: Implicit and Explicit Concept Relations in Deep Neural Networks for Multi-Label Video/Image Annotation. IEEE

Transactions on Circuits and Systems for Video Technology, Vol. 29, 2018, No. 6, pp. 1631–1644, doi: 10.1109/TCSVT.2018.2848458.

[13] Tang, C.—Liu, X.—Wang, P.—Zhang, C.—Li, M.—Wang, L.: Adaptive Hypergraph Embedded Semi-Supervised Multi-Label Image Annotation. IEEE Transactions on Multimedia, Vol. 21, 2019, No. 11, pp. 2837–2849, doi: 10.1109/TMM.2019.2909860.

[14] Amhmed, B.: Distributed System Design Based on Image Processing Technology and Resource State Synchronization Method. Distributed Processing System, Vol. 2, 2021, No. 4, pp. 28–35, doi: 10.38007/DPS.2021.020404.

[15] Ke, X.—Zou, J.—Niu, Y.: End-to-End Automatic Image Annotation Based on Deep CNN and Multi-Label Data Augmentation. IEEE Transactions on Multimedia, Vol. 21, 2019, No. 8, pp. 2093–2106, doi: 10.1109/TMM.2019.2895511.

[16] Tan, Y.—Liu, M.—Chen, W.—Wang, X.—Peng, H.—Wang, Y.: Deep-Branch: Deep Neural Networks for Branch Point Detection in Biomedical Images. IEEE Transactions on Medical Imaging, Vol. 39, 2019, No. 4, pp. 1195–1205, doi: 10.1109/TMI.2019.2945980.

[17] Chen, J.—Ou, S.: Development and Application of the Semantic Annotation Framework for Digital Images. The Electronic Library, Vol. 39, 2021, No. 6, pp. 824–845, doi: 10.1108/EL-07-2021-0131.

[18] Diwakar, M.—Singh, P.—Shankar, A.: Multi-Modal Medical Image Fusion Framework Using Co-Occurrence Filter and Local Extrema in NSST Domain. Biomedical Signal Processing and Control, Vol. 68, 2021, Art. No. 102788, doi: 10.1016/j.bspc.2021.102788.

[19] Hou, Y.—Wang, Q.: Research and Improvement of Content-Based Image Retrieval Framework. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 32, 2018, No. 12, Art. No. 1850043, doi: 10.1142/S021800141850043X.

Jing Jia studied in the North China Institute of Technology from 1999 to 2003 and received his Bachelor's degree in 2003. From 2004 to 2006, he studied in the University of Abertay Dundee and received his Master's degree in 2006. Currently, he works in the Jiangxi Agriculture University. He has published five papers, three of which have been indexed by SCI. His research interests include deep learning and Big Data.

Jing Hua is Associate Professor at the Jiangxi Agricultural University. Her research interests include machine learning and embedded systems.