Mask Recognition Method Based on Graph Convolutional Network

Qing Ye*1, Rui Li1
School of Computer Science and Technology, North China University of Technology,
No.5, Jinyuanzhuang Road, Shijingshan District, Beijing 100144, P.R.China
yeqing@ncut.edu.cn

Abstract. The rapid spread of COVID-19 worldwide has exacerbated the health crisis and affected our daily lives. Medical researchers have shown that wearing masks in public places is essential to reduce the spread of COVID-19 infection. The increase in the accuracy of the identification of masks in public places can effectively prevent the further spread of the epidemic. In this paper, we propose a mask recognition network based on graph convolutional network, D-GCN. The network adopts the method of combining convolutional neural network and graph convolutional neural network. First, DenseNet101 is used to extract features of the real-time image to be tested, and then GCN processes the training label information to form a directed graph through the word embedding vector, it forms a new classifier based on the label connection between the training pictures, and processes the extracted feature vectors. Finally the classification result is output and completing the recognition. The experiment is carried out on the mask dataset and MAFA dataset, and the recognition accuracy is significantly improved compared with previous methods.

1. Introduction

In 2020, COVID-19, as a pandemic disease, spread rapidly around the world. Medical researchers have shown that wearing masks is an effective prevention and control measure. Wearing masks in public places is particularly important. Research on mask recognition technology has also become particularly hot. Vinh et al. [1] presented a real-time face mask detector which utilized Haar cascade classifier to detect the face and YOLOv3 algorithm to detect the mask. Jiang et al. [2] proposed to use an improved YOLOv3 algorithm, also called YOLOv3-Slim, for mask recognition. In addition to the YOLOv3 algorithm, other convolutional neural networks(CNN) were also used for mask recognition, such as MobileNet-V2[3] and ResNet50V2[4]. This paper proposes a new mask recognition network, D-GCN. It uses the graph convolutional neural network for mask recognition, using DenseNet101 to perform feature extraction on the image to be tested, sending the feature vector output by the global max pooling layer into the classification and recognition part of the network. the classifier is trained by GCN for feature classification and recognition. Finally, the classification result of the image is output by the softmax network. This method has achieved good accuracy on both the mask dataset and the MAFA dataset.

2. Network structure

The network structure is shown in Figure 1. The network is divided into a feature extraction part and a classification recognition part. The feature extraction part uses a convolutional neural network, this
The paper uses DenseNet101. The classification recognition part is a classifier composed of a graph convolutional neural network (GCN). The label information is formed into a directed graph through word embedding vectors, and it is sent to GCN for learning, so that the association between the labels is reinforced, and the classification is device for training. The general process is as follows: The images to be tested are input into the DenseNet-101[5] feature extraction network for feature extraction, the extracted features are used as the input of the classification recognition part, and the graph convolutional network classifier is used for classification and recognition, finally the softmax network outputs the classification results.

![Image](convolution_polling_dense_block_global_max_polling.png)

**Figure 1. D-GCN network architecture**

### 2.1. Feature extraction network

The feature extraction part uses convolutional neural network for image feature extraction. This article uses DenseNet101 for feature extraction. DenseNet, as another convolutional neural network with a deeper number of layers, has fewer parameters than ResNet. The bypass enhances the reuse of features, it is easier to train, and has a certain regular effect, which can be relieved The problem of vanishing gradient and overfitting to a certain extent.

For a $L$-layer network, DenseNet contains a total of $L(L+1)/2$ connections. In DenseNet, all the previous layers are connected as input. If expressed by the formula, the output at layer $l$ is as follows:

$$X_l = H_l ([X_0, X_1, \ldots, X_{l-1}])$$  \hspace{1cm} (1)

Among them, $H_l(\cdot)$ represents non-linear transformation, it is a combined operation, which may include a series of BN (Batch Normalization), ReLU, Pooling and Conv operations.

The network structure of DenseNet is mainly composed of Dense Block and Transition. In Dense Block, the feature maps of each layer have the same size and can be connected in the channel dimension. The nonlinear combination function $H(\cdot)$ in Dense Block uses the structure of BN+ReLU+3x3 Conv. Since the input of the latter layer will be very large, the bottleneck layer can be used inside the Dense Block to reduce the amount of calculation. The main reason is that 1x1 Conv is added to the original structure, that is, BN+ReLU+1x1 Conv+BN+ReLU+3x3 Conv. The Transition layer can play a role in
the compression model. It mainly connects two adjacent Dense Blocks and reduces the size of the feature map. The Transition layer includes a 1x1 convolution and 2x2 AvgPooling, the structure is BN+ReLU+1x1 Conv+2x2 AvgPooling. The specific network structure of DenseNet101 is shown in Figure 2.

![Figure 2. DenseNet101](image)

2.2. Classification recognition network

Megvii Research Institute [6] proposes to establish a directed graph between tags through a data-driven approach, and map the category tags to corresponding category classifiers by Graph Convolutional Network to model category relationships for multi-label classification.

In this paper, the GCN part is used to extract the correlation information between face and face_mask, and the correlation coefficient matrix A is constructed through the mutual dependence relationship existing in the training samples, so as to realize GCN information transmission between nodes.

As is shown in Figure 3, the target label consists of the word embedding vector \( Z \in \mathbb{R}^{C \times d} \). Indicates that C represents the number of categories, and d represents the dimension of the word embedding vector. According to these labels, a directed graph is created, where each node represents a label. Using label maps to train stacked GCNs, and map these label representations to a set of interdependent target classifiers, namely \( W \in \mathbb{R}^{C \times D} \).

This paper has further analyzed the principle of GCN-classifier, the principle formula is as follows:

\[
H^{l+1} = h (\tilde{A} H^l W^l) 
\]

(2)

\( H^{l+1} \) is the output of the l-th GCN network, so \( H^0 \), the input of the GCN network, it is the Word Embedding [7] of the classification category, which is trained in the paper using the Glove method. \( h() \) represents the activation function. \( \tilde{A} \) is a correlation matrix that has been obtained by conditional probability matrix A. \( H^l \) is the input of the first layer GCN network, that is, the output of the l-1th layer GCN network. \( W^l \) (transition matrix) is a parameter that can be learned by the first layer GCN network, which can be initialized randomly.

Suppose the true label of an image is \( y \in \mathbb{R}^C \), \( y^i = \{0, 1\} \), indicating whether there are labels i is in the image, then the entire network can be trained using the loss function of traditional multi-label classification, as follows:
\[
L = \sum_{c=1}^{C} y^c \log\left(\sigma(y^c)\right) + (1 - y^c)\log \left(1 - \sigma(y^c)\right)
\]

where \(\sigma(\cdot)\) is the sigmoid function, \(\hat{y}\) is the predicted scores.

3. Experiment

3.1. Dataset

This experiment uses two datasets, namely the mask dataset and MAFA dataset.

1. The mask dataset is a self-built database. The images are crawled from the Internet. There are 4028 face images and 3154 face_mask images respectively, and 245 of them coexist.

2. MAFA dataset [8] is a commonly used mask detection network training dataset, which has 4141 face images and 27536 face_mask images, of which 866 coexist.

![Figure 4. Dataset images](image)

The above two data sets are divided according to the training set: validation set: test set at a ratio of 8:1:1, and they are processed into the format of the VOC2007 dataset[9] for training. Setting image size to 448, batch size to 8, and lr to 0.001 for training.

3.2. Experimental results

1. The mask dataset

After processing the data with the above method, the dataset is trained in D-GCN. In order to compare the experimental results, the data is also processed and trained in YOLOv3[10]. In other papers, YOLO algorithm is often used as a mask detection network. The experimental data obtained are shown in Table 1.

|                | Precision(%) | Recall(%) |
|----------------|--------------|-----------|
| YoloV3         | 91.13        | 94.12     |
| D-GCN          | 98.71        | 95.78     |

As shown in Table 1, on the mask dataset, the precision of YOLOv3 is 91.13% and the recall is 94.12%, the precision of D-GCN is 98.71% and the recall is 95.78%. The training results of the same data set in D-GCN are obviously better than the training results in YOLOv3. The precision increases about 8% and the recall increases about 2%. In order to make the experimental results more intuitive, we draw a comparison chart of the accuracy of the two methods during training in Figure 5.
Figure 5. Comparison of the accuracy on the mask dataset

(2) MAFA dataset

The experimental data of MAFA database are shown in Table 2. Truong et al. [4] proposed detecting proper mask usage with soft attention, and the comparative experimental data of D-GCN comes from their experiment.

Table 2. The experimental data of MAFA dataset

| Model            | Precision(%) | Recall(%) |
|------------------|--------------|-----------|
| DenseNet121      | 78.86        | 73.94     |
| InceptionResNetV2| 86.16        | 59.01     |
| InceptionV3      | 86.88        | 25.29     |
| MobileNetV2      | 85.55        | 36.56     |
| ResNet50V2       | 84.76        | 74.38     |
| VGG16            | 84.91        | 2.95      |
| Xception         | 83.52        | 72.02     |
| D-GCN            | 88.26        | 79.53     |

By comparison, we can see that D-GCN has the best experimental results on the same dataset, achieving the highest values in both Precision and Recall. In the comparison paper data, we can see that the highest precision is 86.88%, which is the result of the improved network training of InceptionV3, and the highest Recall is 74.38%, which is the result of the improved network training of ResNet50V2. However, the precision of D-GCN is 88.26%, and the recall is 79.53%, which has improved significantly in both evaluation indicators. Especially compared with DenseNet121, we can see that Precision has increased by nearly 10%, and Recall has also increased by about 6%.

4. Conclusions.

In this paper, a method for mask recognition using a combination of convolutional neural network and graph convolutional neural network is proposed. Specifically, DenseNet101 is used for feature extraction of the image, and the feature vector is output from the global maximum pooling layer. In the classification and recognition part, the label correlation training is carried out through the graph convolutional neural network, and the output result and the feature vector are processed by dot multiplication. Finally, the softmax network outputs the classification result to complete the image
Experiments show that this method has obvious gains in improving the recognition accuracy, but whether other feature extraction networks can get better results still needs further research.

**Acknowledgment.**
This paper is supported by Beijing first-class professional construction project, Technology Project of Beijing Municipal Education Commission (No.SQKM201810009002), National Natural Science Foundation of China (No.61371143), National Natural Science Foundation of China (No.61806008), Beijing Natural Science Foundation(No.4192022), Ministry of Education Science and Technology Development Center Project (No. 2018A03029) . The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

**References**

[1] T. Q. Vinh and N. T. N. Anh, "Real-Time Face Mask Detector Using YOLOv3 Algorithm and Haar Cascade Classifier," 2020 International Conference on Advanced Computing and Applications, QuyNhon, Vietnam, 2020, pp. 146-149, doi: 10.1109/ACOMP50827.2020.00029.

[2] X.M. Jiang, F.G. Xiang, M.H. Lv, W.Wang, Z.H. Zhang, Y.Yu. “YOLOv3-Slim for Face Mask Recognition”, Journal of Physics: Conference Series, v 1771, n 1, February 25, 2021, International Conference on Artificial Intelligence and Cloud Computing, ICAICC 2020

[3] K. Suresh, M. Palangappa and S. Bhuvan, "Face Mask Detection by using Optimistic Convolutional Neural Network," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1084-1089, doi: 10.1109/ICICT50816.2021.9358653.

[4] T. Truong, D. Lalseta, R. Ittyipe and S. Yanushkevich, "Detecting Proper Mask Usage with Soft Attention". 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 2020, pp. 1745-1750, doi: 10.1109/SSCI47803.2020.9308430.

[5] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 2261-2269, doi: 10.1109/CVPR.2017.243.

[6] Z. Chen et al. "Multi-Label Image Recognition With Graph Convolutional Networks," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5172-5181, doi: 10.1109/CVPR.2019.00532.

[7] S. Dlamini et al."isiZulu Word Embeddings," 2021 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 2021, pp. 121-126, doi: 10.1109/ICTAS50802.2021.935011.

[8] S. Ge, J. Li, Q. Ye and Z. Luo, "Detecting Masked Faces in the Wild with LLE-CNNs," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 426-434, doi: 10.1109/CVPR.2017.53.

[9] H. Ibrahim, A. D. A. Salem and H.-S. Kang, "Real-Time Weakly Supervised Object Detection Using Center-of-Features Localization," in IEEE Access, vol. 9, pp. 38742-38756, 2021, doi: 10.1109/ACCESS.2021.3064372.

[10] Z. Zheng, J. Zhao and Y. Li, "Research on Detecting Bearing-Cover Defects Based on Improved YOLOv3," in IEEE Access, vol. 9, pp. 10304-10315, 2021, doi: 10.1109/ACCESS.2021.3050484.