Optimized Faster-RCNN in Real-time Facial Expression Classification

Jiadi Bao¹, Shusong Wei², Jingfan Lv³ and Wenli Zhang⁴

¹,²,⁴Department of Information Technology, Beijing University of Technology, Beijing, China
³Department of Applied Science, Beijing University of Technology, Beijing, China

*Corresponding author

Abstract. In order to make convolutional neural network adapt to the mobile terminals which lack of hardware resources in facial expressions recognition. We modified the recent algorithm of real-time CNNs on facial expression. Firstly, we used Gaussian distribution to reduce the irrelevant data. Secondly, we used random forest to reduce the time complexity. We implemented the model and algorithm on raspberry pi. As a result, we reduce the amount of data by about 40% and the time complexity to logarithmic level. Thus, our system can run smoothly on mobile terminals with lack of hardware resources. We validate the accuracy of our system on raspberry pi which has the ability to detect faces and classify the emotion. The accuracy in facial expressions recognition remain stable as the original algorithm. In all, compared with the traditional algorithm, our optimize algorithm improve the number of frames remarkably without reducing the accuracy.

Keywords: Convolutional neural network; Gaussian distribution; facial expression recognition; decision tree; random forest.

1. Introduction

With the rapid development of machine learning, computer technology, artificial intelligence and neural network, the rate and accuracy of facial expression recognition have reached a very high level. The application of facial expression recognition technology is very wide, including optimizing human-computer interaction, pre-prediction of criminal behavior and patient emotion detection.

In recent years, the deep learning-based Convolutional Neural Network (CNN) has become the mainstream algorithm of facial expression recognition, given its advantages in the field of image recognition. The structure of the deep learning-based convolutional neural network is divided into input layer, convolution layer, fully-connected layer and output layer. This method greatly improves the rate of feature extraction and further simplifies the format of the input data. However, deep learning based convolutional neural network has simplified the process of facial expression recognition, but often needs to calculate a large amount of data, this requires the corresponding equipment to have strong data processing ability, so it is difficult to run on the mobile terminal (such as Raspberry Pi) with lack of data processing ability, which limits the application of facial expression recognition to a certain extent.

In order to further simplify the computing data, while ensuring the accuracy of facial expression recognition, allowing it to run on mobile terminals with relatively weak computing power, and then further expand the actual application of facial expression recognition system by taking advantage of the high portability and low cost characteristics of mobile terminals, the following method is proposed in this paper. That is, by incorporating the faster-rcnn algorithm and the random forest algorithm, using the global uniform pooling algorithm based on random forest to suppress the full zipper in CNN, so as
to achieve the goal of reducing the computation amount while ensuring the accuracy of identification, thereby reducing the time complexity successfully from $o(N)$ to $o(\log_2 N)$.

2. Related Work

The classification model of the deep learning based convolutional neural network (CNN) differs from the traditional model in that it can directly input a 2D image into the model and then present the classification results at the output. The network core is the network structure design and the network solution \[4\]. In general, the recognition process of the deep learning-based convolutional neural network requires the following steps: The convolution layer extracts the features initially; the pooling layer extracts the main features; the fully-connected layer summarizes the features of each part, and produces the classifier for prediction and recognition. The process can be roughly represented by Figure 1. At the fully-connected layer, it generally contains all the characteristic parameters, such as the VGG6 network structure, which contains about 90% of the characteristic parameters. In order to reduce feature parameters and thus simplify the calculation process, the idea of global average pooling is put forward, such as Inception V3, which simplifies each feature map into a scalar value by averaging from all feature parameters. This paper uses TensorFlow as the framework for deep learning-based convolutional neural networks (CNN) \[7\]. TensorFlow is a dataflow programming-based symbolic mathematical system that is widely used in the programming implementation of various types of machine learning algorithms, formerly known as Disbelief, Google's neural network algorithm library.

Faster R-CNN consists of two main steps. The first step, called the Region Proposal Network (RPN) which uses the recently popular neural network 'attention' \[8\], is proposed to improve Fast R-CNN. Fast R-CNN takes a significant portion of the time to extract the candidate area, and it is optimized by Faster R-CNN. The Region Proposal Network (RPN) can be thought of as a full convolution network that can be used for end-to-end training, with the ultimate role being to recommend candidate areas, which will significantly reduce the time spent. The second step is the general Fast R-CNN network \[10\], which performs pooling and full-linking based on the candidate areas selected by the above-mentioned Region Proposal Network (RPN), and ultimately achieves image recognition.

For face detection, we have adopted the Face Detection algorithm of Open CV. First of all, analyze the obtained video stream frame by frame, and convert a specific color image of one frame into a gray image, that is, convert it into a single channel to reduce the amount of calculation, and then draw a rectangle on the grayscale image, and finally use the training classifier to find the face and frame the position of the face. This algorithm can quickly frame the specific position of the face, simplifying the calculation for the next facial expression recognition \[11\].

For the data set part of the deep learning-based convolutional neural network (CNN), the Fer2013 data set from a facial expression contest on Kaggle is used \[12\]. Fer2013 facial expression data set is composed of 35,886 facial expression images, of which 28,708 are Training images, 3,589 are PublicTest images and PrivateTest images each. Each image is made up of 48x48 fixed gray scale images with 7 expressions, corresponding to the digital tag 0-6, the labels corresponding to the specific expressions are in Chinese and English. These training sets will be used as input to the above-described deep learning-based convolutional neural network (CNN) framework.

\[Figure 1.\] Step of Convolutional Neural Network.
3. Method

3.1. Introduction
Two mathematical models were built to optimize the Convolutional Neural Network (faster-RCNN), each model corresponding to a parameter in order to debug the accuracy of the network. The purpose of these two models is to improve the speed of the algorithm as much as possible without affecting the Faster-RCNN Algorithm, and to make the deep learning algorithm run smoothly on the Mobile Terminal(Raspberry pi), which is short of hardware resources.

By cutting some data and some parameters in the model, two goals are successfully achieved. First, the building of mathematical models, centered on the Gaussian distribution\textsuperscript{[13]}, helps reduce the amount of data by about 40%, allowing the convolutional neural network algorithm can run smoothly on mobile terminals such as small robots, small computers, and Raspberry Pi, and provide a better user experience in human-computer interaction. Second, the final classification using Decision Tree\textsuperscript{[14]} allows us to observe data much more visually which parameters have heavier weights. Our first model is built against the background of fully-connected layer, after two times of convolution, eight times of Separable Convolution and four times of pooling, the feature map group with a data volume of approximately 500,000 is obtained, showed in Figure 2.

![Optimized Faster-Rcnn Diagram](image)

**Figure 2.** Optimized Faster-Rcnn.

From the nature of convolution, the process of convolution is the process of image feature extraction, So after the previous processing, we completely removed the fully-connected steps, used our model instead to simplify the computation, then performed the logistic regression (softmax), and use the decision tree to classify the data. For gender, The algorithm's accuracy for gender recognition reached 90%. And for facial expression, The algorithm's accuracy for expression recognition reached 68%. We call this method Gaussian Classification (Normal Classification).
3.2. Gaussian distribution

In the process of CNN, each feature map is a target recommended by RPN algorithm after screening. In the RPN Algorithm, the target has been covered by the target detection box as far as possible, and the target detection box is the optimal solution. The target recommendation process is shown in *Towards real-time object detection with region proposal networks* [9]. The algorithm places the target as far as possible in the middle of the candidate box.

Our model is derived from the Gaussian distribution in statistics, and the probability density function of the Gaussian distribution is described in Equation 2. For the recognition frame recommended by the RPN algorithm, the target is already in the center, and all the pixels around the image are irrelevant. These pixels are not helpful for the training result, but it takes a lot of CPU time to fully zip and sort these extra pixels, and the processing of these pixels in a relatively resource-poor mobile device has a negative effect on processing efficiency. The traditional classification algorithm can choose bp neural network (BP neural network), a three bp neural network [15]. A bp neural network with a three-layer hidden layer of input N usually requires more than $N^* \log_2 N$ parameters, and it takes a lot of CPU time to calculate such a large number of neural networks.

$$f(x) = \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)$$ (1)

We establish a mathematical model, assuming that after many times of convolution, the characteristics of the pool after each map is consistent with the following mathematical relations. The feature map is composed of rows and columns, in which the contribution of each pixel to the final classification fitting can be approximately considered as a random variable $x$, and the whole feature map is a joint wide stationary random sequence, so it can be assumed that it follows the standard Gaussian distribution with $Dk/2$ as its mathematical expectation, as shown in Equation 2.

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(x-Dk/2)^2}{2} \right)$$ (2)

$$X \sim N(\frac{Dk}{2}, \sigma)$$ (3)

The Gaussian distribution probability density function is roughly as shown in Figure 3.

![Gaussian Distribution](image)

*Figure 3. Gaussian Distribution.*

It can be seen from the figure that the closer to the desired position, the higher the probability, that is, the greater the contribution in classification. The feature map obtained after the convolution and pooling process is shown in Figure 4.
Due to the properties of multiple convolution and pooling, since the convolution kernel is a matrix of 3 to 5 orders with the same length and width, as shown in Figure 4 Each three-dimensional square matrix, which is Dk in length and M in height, stores all the characteristic information of candidate boxes recommended by RPN Algorithm. In a three-dimensional square matrix, the matrix of each layer of length and width Dk is the result of a convolution operation of a constant step; such a matrix retains all the features of the original image, so we perform the following operation:

As shown in Figure 4, the feature map is Dk in width and Dk in length; N is the number of targets detected; M is the number of feature maps processed for a detection box; M is related to the step size of convolution.

In this paper, 3σ principle[16] is introduced, and the feature points with high contribution are assumed to be concentrated in the following confidence interval:

\[ \mu - \sigma < X < \mu + \sigma \]
\[ \mu - 2\sigma < X < \mu + 2\sigma \]
\[ \mu - 3\sigma < X < \mu + 3\sigma \]

\[ P(\mu - \sigma < X < \mu + \sigma) = 68.2\% \]
\[ P(\mu - 2\sigma < X < \mu + 2\sigma) = 95.45\% \]
\[ P(\mu - 3\sigma < X < \mu + 3\sigma) \] (4)

Among them, \( \mu \) is the mathematical expectation, \( \sigma \) is the standard deviation, and P is the probability of a random variable. According to the above rules, the data quantity in the confidence interval can be induced, and the feature map can be trimmed according to the model. By modifying the pre-modification constant, the part of the reject domain in the normal distribution is adjusted, and the amount of data to be trimmed is adjusted; thus, the data trained in the subsequent classification can be reduced.

3.3. Decision Tree & Random Forest

After preprocessing the feature map, the feature map as shown in Figure 5 is obtained. P of which is the proportional coefficient defined in Equation 4. Then, a filter as shown in Figure 5 is defined, and the convolution operation is carried out for each pre-processed feature map, each filter performs a convolution operation on a feature map, the feature vector obtained is shown in Figure 5. M refers to the number of filters [17]. We use these M filters to carry out convolution operations for feature map.

![Figure 5. Filter.](image)
For every array of length M, we call it the eigenvector; after all the preprocessing, we get n eigenvectors, and we use the random forest to classify these data. The time complexity of random forest is $O(n)$, while that of traditional bp neural network is $O(M)$, which reduces the time complexity to logarithmic level and reduces the CPU time complexity.

4. Result

Experimental steps:
Firstly, 427 sets of data are collected, of which 80 sets of expressions are neutral data, 63 for sadness data, 72 for happiness data, 60 for fear data, 79 for surprise data, and 63 for anger data. The image data are as follows:

Figure 7. Facial Recognition Result.

Based on two models of convolutional neural network (faster-RCNN) and random forest optimization algorithm, images are scanned into both models in real time and input image data. In the input process, the variables necessary are controlled to make the data comparable, such as control the same picture and control the same computer machine. The output results are divided into six categories: “neutral”, “sadness”, “happiness”, “fear”, “surprise” and “anger”. Finally, two sets of data based on two models are obtained.

Result analysis:
1. The statistical results of the obtained data are as follows:

|       | neutral | sadness | happiness | fear | surprise | anger |
|-------|---------|---------|-----------|------|----------|-------|
| neutral | 0.72 | 0.12 | 0.06 | 0.04 | 0.02 | 0.04 |
| sadness | 0.2 | 0.58 | 0.04 | 0.08 | 0.02 | 0.08 |
| happiness | 0.03 | 0.04 | 0.88 | 0 | 0.03 | 0.02 |
| fear | 0.1 | 0.18 | 0.03 | 0.5 | 0.09 | 0.1 |
| surprise | 0.02 | 0.02 | 0.04 | 0.05 | 0.85 | 0.02 |
| anger | 0.1 | 0.15 | 0.03 | 0.08 | 0.03 | 0.61 |

Figure 8. Result.
The above two figures are the experimental results based on the traditional algorithm and the improved algorithm. The vertical axis of the table indicates the categories of experimental graphs, namely, neutral, sadness, happiness, fear, surprise, and anger, while the horizontal axis represents the result type of the experimental model, and the data represents the proportion of a certain type of graph obtained by the experiment belonging to different types of results, that is, the sum of the results of the class divided by the total number of the graphs. So the focus is on the diagonal data, indicating the accuracy of the model. In order to see the difference between the two sets of data more clearly, a comparison bar chart is shown as follows.

![Figure 9. Result.](image)

The horizontal coordinates represent the proportion of different expressions, the sum is 100%, different colors represent different expressions, while the vertical coordinate can be divided into six groups, each group represents the difference between the two models to judge the same expression. From the contrast chart, it is clear that there is little difference between the two models. It is worth noting that the accuracy of the two models in determining sadness, fear, and anger did not reach 60%, and that the correct ratio was not high, and of course, it was closely related to the images. In addition, we found that when the image shows a clear signal such as “Frown”, both models can correctly reflect the expression as anger, and when the marker information is not obvious, both models are easier to classify the image as neutral, which is also in line with the main design ideas of the two models.

By calculation, the correct expectation of image expression recognition of convolutional neural network (faster-RCNN) is 0.6864, and that of improved model is 0.6722, which is about 98% of the original model. From the collected frame rate, the frame rate of the convolutional neural network (faster-RCNN) averages 6 frames per second, while the improved model frame rate averages 14 frames per second, which is increased by 233%. Therefore, the improved model greatly improves the recognition efficiency while maintaining the recognition accuracy.

The fully-connected layer in the convolutional neural network is completely removed, the process of classifying using the bp neural network in the traditional algorithm is also removed, and the random forest algorithm is used instead for classification. Compared with bp neural network, the structure of binary tree in random forest can make full use of space and reduce time complexity. The time complexity of random forest is \( O(\log_2 n) \), while that of traditional bp neural network is. By reducing the classification time without reducing the classification accuracy, the real-time recognition frame rate is improved by these two methods: the frame rate is increased from the original 6 frames to 14 frames, which is doubled, and the user experience is optimized.

5. Conclusion
We proposed and tested a real-time CNNs on raspberry pi. Our model was proposed in order to reduce the amount of data and parameters. We put the faster-RCNN as the previous steps, our own model began by completely remove the fully connected layers and by reducing the amount of parameters in the average pooling layer via the mathematical model we built. Along with that, we also use different
methods of classification to reduce the time complexity from $o(n)$ to $o(\log_2 n)$. We have shown that our proposed models on raspberry pi. The result is sufficient we makes the frame frequency doubled, from 7Hz to 14Hz, at the same time the accuracy remain stable. We believe that it will play its role in criminal investigation and medical field.

References
[1] V. D. Van, T. Thai and M. Q. Nghiem, “Combining convolution and recursive neural networks for sentiment analysis,” in Proc. 8th Int. Symp. Information and Communication Technology, Nha Trang City, Dec. 7-8, 2017, p.151-158.
[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
[3] Lei Zhang, Falk Huettmann, Xudong Zhang, Shirong Liu, Pengsen Sun, Zhen Yu, Chunrong Mi,”The use of classification and regression algorithms using the random forests method with presence-only data to model species’ distribution,” MethodsX, 2019, 6.
[4] S. Alizadeh and A. Fazel, “Convolutional neural networks for facial expression recognition,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, July 21-26, 2017.
[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2818–2826, 2016.
[6] Franc ois Chollet. Xception: Deep learning with depthwise separable convolutions. CoRR, abs/1610.02357, 2016.
[7] Liang Yu,” Research and Implementation of CNN Based on TensorFlow,” Proceedings of the 2nd International Seminar on Computational Intelligence, Engineering and Technology (SCIET2018), 2018:7.
[8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in Neural Information Processing Systems (NIPS), 2015.
[9] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015. 1
[10] R. Girshick. Fast R-CNN. In ICCV, 2015.
[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results,” 2007.
[12] I. J. Goodfellow et al., “2013. Challenges in representation learning: A report on three machine learning contests,” in International Conference on Neural Information Processing. Springer, pp.117-124.
[13] Wei Xu,Yingjie Liang,Wen Chen,Fajie Wang. “Recent advances of stretched Gaussian distribution underlying Hausdorff fractal distance and its applications in fitting stretched Gaussian noise,” in Physica A: Statistical Mechanics and its Applications, 2020, 539.
[14] Stephen S. Chen, Nick A. Zaborek, Amanda R. Doubleday, Sarah C. Schaefer, Kristin L. Long, Susan C. Pitt, Rebecca S. Sippel, David F. Schneider, Optimizing “Levothyroxine Dose Adjustment After Thyroidectomy With a Decision Tree,” Journal of Surgical Research, 2019, 244.
[15] Wang Liang, Guangwei Wang, Xiaojun Ning, Jianliang Zhang, Yanjiang Li, Chunhe Jiang, Nan Zhang,” Application of BP neural network to the prediction of coal ash melting characteristic temperature,” Fuel, 2020, 260.
[16] McCord J Fraser, Grey Nick J A, Winstanley Raymond B, Johnson Anthony,” A clinical overview of removable prostheses: 3. Principles of design for removable partial dentures,” Dental Update, 2003, 29(10).
[17] Xiaowu Zou, Zidong Wang, Qi Li, Weiguo Sheng, “Integration of residual network and convolutional neural network along with various activation functions and global pooling for time series classification,” Neurocomputing, 2019, 367.