CNN with Bimodal Distribution Removal on faces-emotion dataset

Jialei Zhang
Research School of Computer Science
Australian National University ,ACT, Australia
u6965305@anu.edu.au

Abstract. The faces-emotion dataset used in this assignment is some pictures based on face emotions. They contain seven different emotions and usually difficult to classify accurately. With the help of CNN, we choose a popular network named Resnet in order to solve the problem of neural network classification performance degradation when the number of layers is large through residual learning. In addition, the bi-modal distribution elimination technology (BDR) is also used in this article. With this technology, we can delete the abnormal data in the database, thereby further improving the training speed and accuracy of the entire database. In summary, I applied the BDR algorithm and ResNet and compared its accuracy with the accuracy of the Resnet without BDR. The results show that the improved data accuracy can reach 41%,. Finally, the conclusion drawn in this article is that with the help of CNN neural network and BDR, we can improve the accuracy of training based on SFEW database.

Keywords: Bimodal Distribution Removal, CNN, Resnet, face-emotion

1. Introduction

1.1 background

In the previous work, I tried three different ways to classify face emotions: Simple neural network without the BDR, simple neural network with the BDR on only training set and simple neural network with the BDR on the whole dataset. The data set was SFEW[1] which has extracted the features by LPQ[2] and PHOG[3]. The results showed that accuracy in the test set was 21%, 24% and 48% respectively.

In this paper, I will try to make establish a more complicate neural network in order to get a better performance and find out what can be improved. After the comparison of DenseNet[10], Alexnet[14], VGG[15],Resnet[16],BN[4] and dropout[5] ,I choose a classical and popular classification CNN model named Resnet , it is the most widely used CNN feature extraction network.

1.2 Dataset

The dataset face-emotion is a dataset contain 7 different folders which are “Angry”, “Disgust”, “Fear”, “Happy”, “Neutral”, “Sad” and “Surprise, each one include some pictures drawn from different movies. I split this dataset into two different sets, the training set and the validation set. I put 295 pictures in the training set and the remain 379 pictures in the validation set. In the code, I split the validation set into validation and test set half by half.

2. Methodologies

2.1 Convolutional neural networks (CNN)

Convolutional neural networks are very similar to ordinary neural networks. They are composed of neurons with learnable weights and bias constants. Each neuron receives some input and does some dot product calculations. The output is the score of each category. Some calculation techniques in ordinary neural networks are still applicable here.

So what makes it different? The default input of a convolutional neural network is an image, which allows us to encode specific properties into the network structure, making our feedforward function more efficient and reducing a large number of parameters.

Convolutional neural networks usually contain the following layers:

1. Convolutional layer. Each convolutional layer in a convolutional neural network is composed of several convolutional units. The parameters of each convolutional unit are optimized by a back propagation algorithm. The purpose of the convolution operation is to extract different features of the input. The first convolutional layer may only extract some low-level features such as edges, lines, and corners. More layers of the network can iteratively extract more complex features from the low-level features. feature.

2. Linear rectification layer. The activation function of this layer of nerve uses linear rectification.

3. The pooling layer usually obtains features with large dimensions after the convolution layer, cuts the features into several regions, and takes the maximum or average value to obtain new features with smaller dimensions.

4. Fully connected layer, which combines all local features into global features, is used to calculate the score of each last category.

However, we found that after CNN network reaches a certain depth, blindly increasing the number of layers does not bring further improvement in classification performance, but will cause the network convergence to become slower and the classification accuracy of the test dataset to become worse. After excluding the problems of over-fitting the model caused by the small data set, we found that too deep networks still reduce the classification accuracy (compared to shallower networks).

For example, Figure2.1.1 below shows how it happened in the paper[7] I found. Obviously ,it shows the decrease in classification accuracy due to the increase in the number of layers in the later stage of the conventional CNN network.

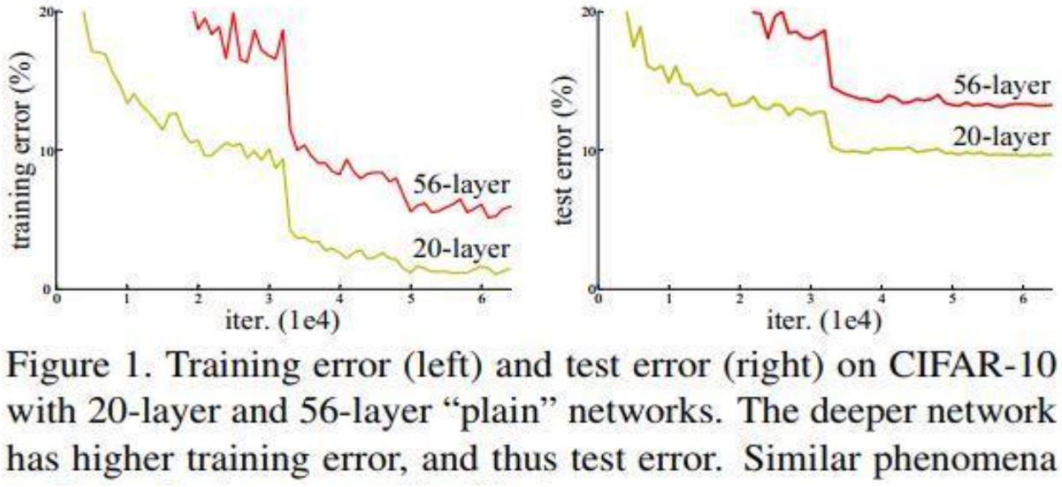

Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena

Fig2.1.1

2.2 Resnet

To solve the problem of CNN, the author of the Resnet network thought of the concept of residual representation commonly used in the field of conventional computer vision, and further applied it to the construction of the CNN model, so there was a basic residual learning block. It learns the residual representation between input and output by using multiple parameter layers instead of using parameter layers to directly try to learn the mapping between input and output, as in the general CNN network (such as Alexnet / VGG, etc.). Experiments show that it is much easier to learn residuals directly with parameterized layers in the general sense than to directly learn the mapping between input and output (faster convergence speed), and it is also much more effective (higher can be achieved by using more layers Classification accuracy).

The residual module helps the network to achieve identity mapping.

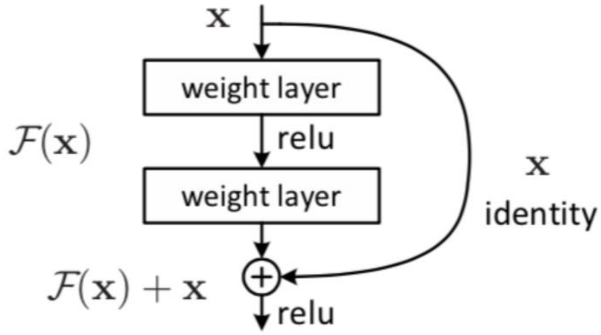
x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig 2.2.1

According to the figure 2.2.1 above, it copy the output of a shallow network to the output of the deep layer, so that when the network feature reaches optimal, the task of deeper identity mapping is released from the original stacked layer into the newly created identity mapping relationship. The tasks in the original layer are changed from identity mapping to all zeros.

F (x) = H (x) −x, x is the output of the shallow layer, H (x) is the output of the deep layer, F (x) is the transformation represented by the two layers sandwiched between the two, when the shallow layer The feature represented by x is mature enough. If any change to feature x will increase the loss, F (x) will automatically tend to learn to become 0, and x will continue to pass from the path of identity mapping. In this way, the initial goal is achieved without increasing the computational cost: in the forward process, when the output of the shallow layer is sufficiently mature (optimal), the layers behind the deep network can achieve the role of identity mapping.

2.3 Biomodal Distribution Removal

The general BDR progress can be shown into the following steps:

1.Implement the training process with all training sets

2.When the loss is reduced to a certain extent, calculate the corresponding loss of each pattern. 3.Calculate the mean loss as M1

4.Put those potential outliers whose loss is greater than M1 into a subset.

5.Calculate the mean of subset as M2, the standard deviation as std2.

6.Remove those patterns whose loss are greater than M2 + k*std2 (k =1)

7.Repeat these steps until the new subset's standard deviation < 0.01

BDR uses error distribution to perform anomaly detection when training neural networks. In some cases, he is a suitable method to detect outliers. But if the initial error distribution in the first few hundred periods does not form a double peak as in the case, when the model continues to train, as the BDR runs, the chance of overfitting is greatly increased. It forms a large peak in the error distribution. This is because the error distribution of the initial neural network training is normally distributed. Therefore, the central distribution may be very close to 50% in the early stage, and the variance is less than 0.1 (the trigger condition of BDR), which will trigger BDR to detect outliers. Since the erroneous normalized values may be very close to each other, it causes the BDR to incorrectly recognize the outlier pattern.

3. Result and Discussion

3.1 Resnet without BDR

At first, I trained it without BDR and I draw the accuracy and loss of both training set and validation set.

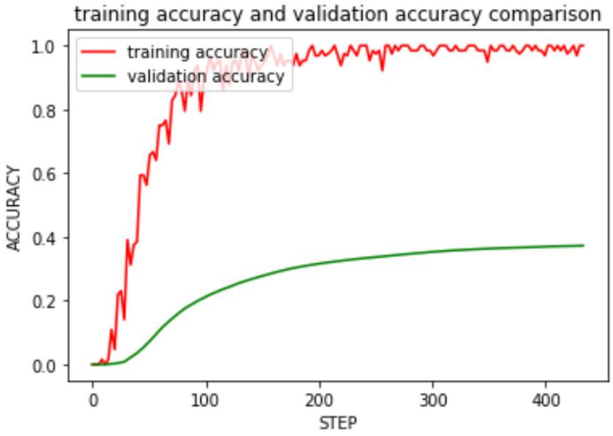

Fig3.1.1

As is shown in Fig3.1.1 above, the accuracy of both the training set and validation set increased sharply in the first 100 step. It keep shacking before 200 step, after that, it is close to its peak and more stable than before. The training accuracy nearly close to 100, I think it is an obvious overfitting because of the size of training set. The dataset I have only contain hundreds of pictures and it's easy to cause the overfitting problem. It should perform better if I could have a dataset larger than the dataset I have now..

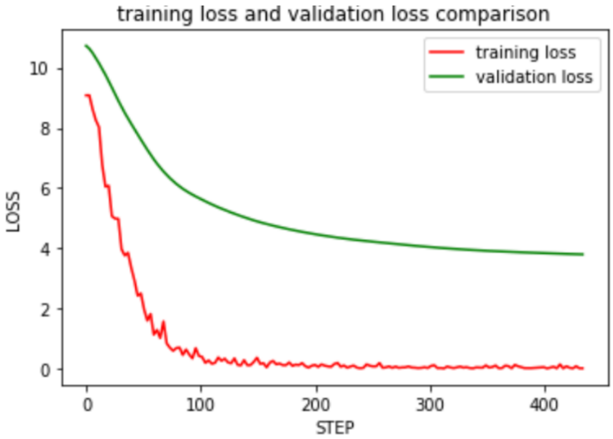

Fig 3.1.2

The figure 3.1.2 above shows that the loss of both training set and validation set decrease sharply in the first 100 step. After 100 step ,the loss of training set is close to 0 and it also prove the problem of overfitting.

3.2 Resnet without BDR

After that ,I trained the dataset again with BDR. The accuracy and loss is shown as Fig3.2.1 and Fig3.2.2.

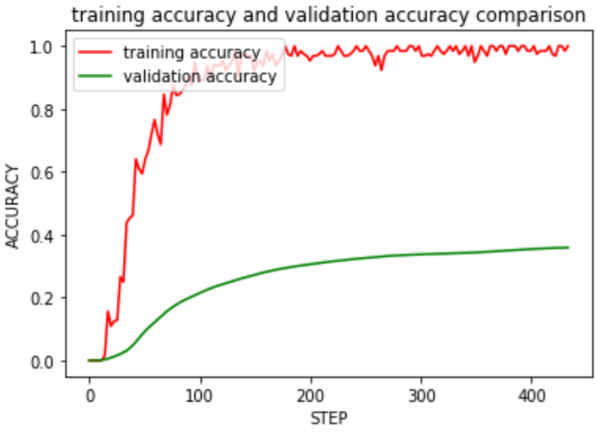

Fig 3.2.1

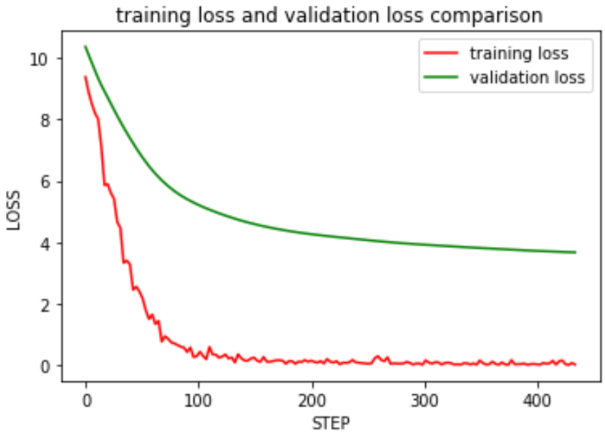

Fig3.2.2

It is no doubt that both figures above are quite similar with the Fig 3.1.1 and Fig3.1.2.BDR is unable to have a big change in the accuracy and loss. But we can find that during the first 200 step ,the curve become more smooth than before. It's contributed by the BDR which helps to remove the outliers.

To find out why the performance of BDR is not good enough, we make some extra work. We know that BDR is built based on statistics. It regards the occurrence of bimodal distribution as a signal that outliers are trying to mislead ANN. Hence, we start to prune them to improve model quality. To examine if BDR is functional, we design two modes. The difference is whether deploy BDR, while other settings remain the same. After contrasting results, we conclude if BDR is efficacious and explain in details. As instructed in [18], we plot error distribution when epoch is 50. From Figure 3.2.3, it is clear that error distribution is bimodal. The cluster deviating from the main one is suspect. As training proceeds, BDR remove noises and we can find scattered points disappear. It is ideal to be a unimodal distribution after several executions. However, this bimodal pattern maintains invariant till termination.

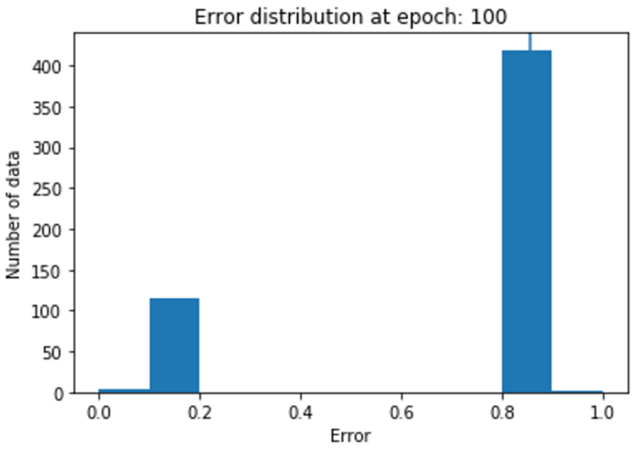

Fig3.2.3

Since training set is relatively clean, small in size and entirely labelled, BDR is inapplicable in this kind of dataset. However, the operation of removing outliers is only done through the training data. But in fact, both training data and test data should have abnormal values. If we first perform BDR operations on the entire data set, then the neural network model should have a high accuracy score.

3.3 Comparison to Back-Propagation Neural Network with BDR

I did some work about SFEW dataset before and it contains the LPQ and PHOG features of the figure we have in face emotion dataset. Apparently, there are some common place between the SFEW dataset and the face emotion dataset. Both datasets have a small size and all labelled, and BDR has little contributions on both of them as well. Since both datasets are small ,clean and entirely labelled ,it is hard to find enough outliers during the training, and that is the reason of poor behavior of BDR during the whole training.

Table 1

MODEL	ACCURACY
BP NN with SFEW	19-21%
BP NN with BDR and SFEW	18-24%

As for my previous implementation, which is shown in Table 1,I used the Back-Propagation Neural Network without BDR which achieved 21% accuracy in test sets and Back-Propagation Neural Network with BDR which achieved 24% accuracy .

Tabel 2

MODEL	ACCURACY
CNN	41%
CNN with BDR	43%

Meanwhile,in this paper, as table 2 shown below, the CNN can achieve 41% accuracy in validation sets . The CNN with BDR can achieve 43% accuracy in validation sets as well .

4. Conclusion and Future work

In summary, this paper demonstrates a convolutional neural networks functioning as facial emotion estimator. This paper firstly designs a special neural network named resnet and then analyses the malfunction of BDR statistically, and we find out the reason of error. Lastly, this paper provides a comparison between the performance of CNN and BPNN.

The results imply that CNN has a higher accuracy .

If we are to conduct a further research in depth, we should consider the effects of hyperparameters more carefully. Even though the exploration of the optimal configurations is tedious, tuning parameters is an indispensable step in designing a neural network. It would be preferable if more experiments can be conducted to optimise parameters. Furthermore, we need to derive a comprehensive evaluation method to effectively avoids overfitting problem.

References:

[1] 4. F. E. Grubbs, “Procedures for detection outlying observations in samples,” Technometrics 11(1), 1–21 (1969).

[2]Azme Khamis, Zuhaimy Ismail, Khalid Haron, and Ahmad Tarmizi Mohammed. “The Effects of Outliers Data on Neural Network Performance”. Journal of Applied Sciences, 5: 1394-1398 (2005)

[3] D. Coursineau, “Outliers detection and treatment: a review,” Int. J. Psychol. Res. 3(1), 58–67 (2010).

[4]Slade, P. and Gedeon, T.D., 1993, June. “Bimodal distribution removal.” In International Workshop on Artificial Neural Networks (pp. 249-254). Springer, Berlin, Heidelberg.

[5]Chen, L., Gedeon, T., Hossain, M. Z., & Caldwell, S. (2017, November). : “Are you really angry?: detecting emotion veracity as a proposed tool for interaction.” In Proceedings of the 29th Australian Conference on Computer-Human Interaction (pp. 412-416). ACM.

[6] L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621, 2017.

[7] Deep Residual Learning for Image Recognition, He-Kaiming, 2015

[8]. Slade, P., and Gedeon T.D.: Bimodal Distribution Removal, vol. 686 (1993)]

Quan, K.: Bimodal distribution removal and genetic algorithm in neural network for breast cancer diagnosis. arXiv preprint arXiv:2002.08729 (2020)

[9] Rodriguez, J.D., Perez, A., Lozano, J.A.: Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE transactions on pattern analysis and machine intelligence 32(3), 569–575 (2009)

[10]Engineering; Studies from Xi'an University of Posts and Telecommunications Update Current Data on Engineering (Df-ssd: an Improved Ssd Object Detection Algorithm Based On Densenet and Feature Fusion)[J]. Journal of Mathematics,2020.

[11] Sola, J., Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on nuclear science 44(3), 1464–1468 (1997)

[12]Srivastava, N.: Improving neural networks with dropout. University of Toronto 182(566), 7 (2013) [13] Suresh, S., Sundararajan, N., Saratchandran, P.: Risk-sensitive loss functions for sparse multi-category classification problems. Information Sciences 178(12), 2621–2638 (2008)

[14] Yani Ioannou,Duncan Robertson,Jamie Shotton,Roberto Cipolla,Antonio Criminisi.Training CNNs with Low-Rank Filters for Efficient Image Classification. (2015)

[15] Hachim El Khiyari, Harry Wechsler.Face Recognition across Time Lapse Using Convolutional Neural Networks DOI: 10.4236/jis.2016.73010, PP. 141-151

[16] Jiahe Yan, Emily Tucci, Nathaniel Jaffe.Detection of t(9;22) Chromosome Translocation Using Deep Residual Neural NetworkDOI: 10.4236/jcc.2019.712010, PP. 102-111