SFEW DATABASE ANALYSIS: FACE EMOTION CLASSIFICATION WITH CNN

Chen Yang

School of Computer Engineering
The Australian National University
u7201888@anu.edu.au

Abstract. Facial detection and recognition have been used in wide range of applications in recent years, so human face related research is vital important. Face image is a kind of high dimension data, simple Neural Network and other algorithms cannot handle such high dimension data. Therefore, principle component analysis is a common face detection and recognition method and performs well in lab environments. However, this method is not robust to misalignment, background variation and pose, scale, rotation, lighting changes. So PCA is not robust to varied realistic environments. This paper did a facial expression classification task on Static Facial Expression in the Wild (SFEW) database, which contains various unconstrained facial expression extracted from movies. Previous research has shown that PCA algorithm performed poorly on this database, so this paper used Convolution Neural Network (CNN) for the classification task, and compared the result with previous research.

Keywords: Facial expression classification· CNN

1 Introduction

Human facial expression databases have been captured in lab environments instead of realistic environments. This paper used a static facial expression data set, which covers unconstrained facial expressions, varied head poses, large age range, different face resolutions, occlusions, varied focus and close to real word illumination.The data set named Static Facial Expressions in the Wild (SFEW), which contains 675 different facial expression images extracted from movies. All images have been labelled for seven basic expressions: angry, disgust, fear, happy, neutral, sad and surprise[1].

Fig. 1. Sample images from the SFEW database.

In previous research, the data set was preprocessed with Kernel PCA method to reduce the dimension of the data. Previous research used top 5 principle components of Local Phase Quantization (LPQ) descriptor and Pyramid of Histogram of Gradients (PHOG) descriptor[2], so all images in the data set were projected into 5-D space. The classification accuracy of the support vector machine(SVM)[3] and the simple neural network (neural networks

that have less than 3 hidden layers) are 22.70% and 20.78%, respectively. The prediction has 7 different classes, if we predict by chance, then the accuracy should be about 14.29%, so the performance of SVM and NN are not outstanding. The low accuracy is mainly due to the complexity of the data set.

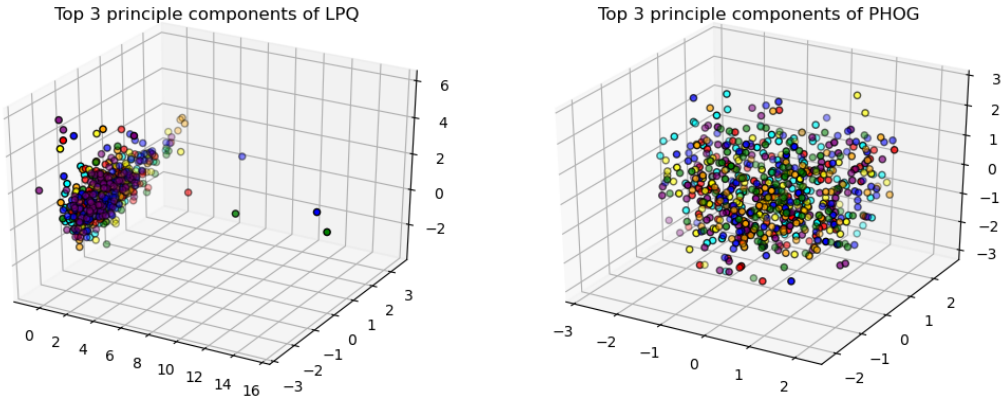

Fig. 2. The top 3 principle components of LPQ features and PHOG features. Red, orange, yellow, green, cyan, blue and purple points represent angry, disgust, fear, happy, natural, sad and surprise respectively.

The figure above shows that the data points cannot be separated by plane or simple surface after the projection. That is one reason why SVM and NN performed poorly on this data set. So in this experiment we first used Deep Neural Network wich can learn and fit more complex non-linear model to classify the data set. Another reason why PCA performed poorly is the data points lost some information after projection. If we can do the classification with the original image, the performance could be better. So in this experiment we used Convolution Neural Network (CNN) for the classification task, and compare the result with previous research.

2 Classification with Deep Neural Network

The training accuracy of NN with one hidden layer in previous research is less than 30%, the result shows that simple NN can not fit such a complex model. So we first used a DNN with 3 hidden layers to classify the LPQ features.

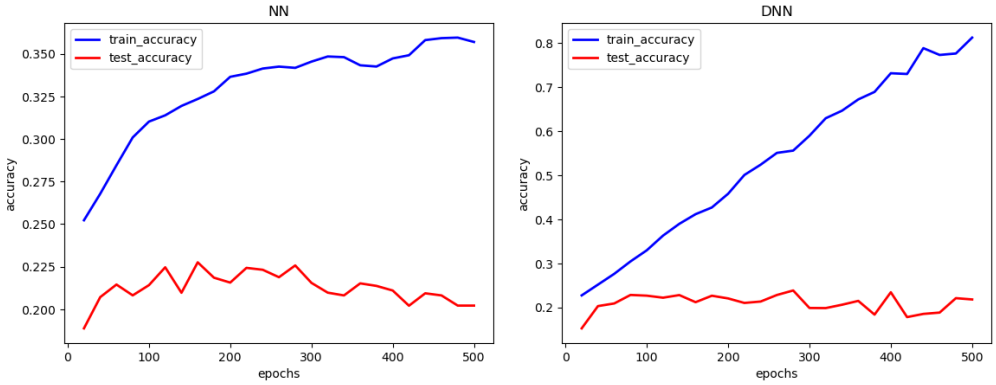

Fig. 3. Changes of training accuracy and testing accuracy when epochs increase. The simple network topology was 5-5-7, being 5 inputs, 5 hidden neurons, and 7 output neurons. And the deep network topology was 5-24-48-24-7, being 5 inputs, 24, 48, 24 hidden neurons in the first, second and third hidden layer respectively and 7 output neurons.

The result shows that the training accuracy of DNN can reach more than 80% after training, which means that the DNN has the ability to fit the training set quite well. But the highest testing accuracy is just 22.86%, that is just a little bit higher than the accuracy of a simple NN model (22.76%). The DNN preformed well on the training set but performed poorly on the testing set. The low accuracy is mainly due to the data points lose some information after projection. In PCA algorithm, we try to maximize the variance in the low-dimensional representation of the data to retain as much information as possible, but those information may not all helpful for face emotion classification. The low-dimensional representation may contain the information about lighting, pose and background, which are useless for face emotion classification.

3 Classification with Convolution Neural Network

The mainly different between CNN and the method above is that CNN used convolution layer to capture the features, while another method used PCA to capture the features. Note that CNN also need a DNN calssifier, but in this paper DNN represents using PCA algorithm to the original image then using DNN to classify.
First we used a simple CNN, and this network had 3 convolution layers, 3 pooling layers with max-pooling, 4 linear layers and used ReLU as the activation function.

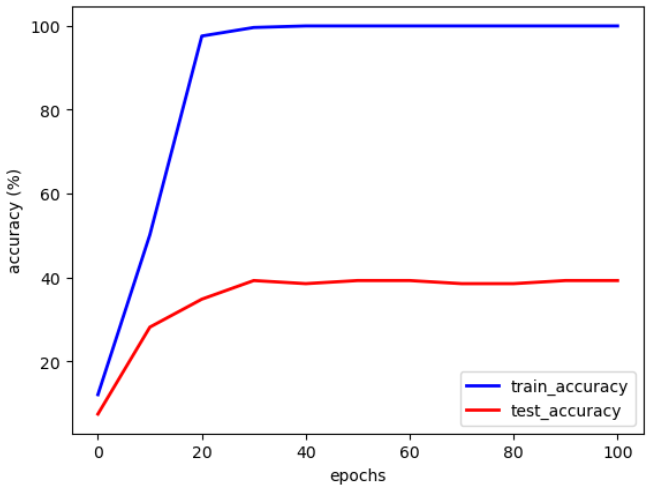

Fig. 4. Changes of training accuracy and testing accuracy when epochs increase. The final training accuracy and testing accuracy were 100% and 39.26% respectively.Here we used the accuracy in epoch 40 to represent the performance of the model, but we trained the model with more epochs to test the performance of this model when overtraining.

This figure shows that CNN performed much more better than DNN. The high training accuracy and testing accuracy illuminate that CNN can capture the features which are important for face emotion classification.
But this model is quite easy to overfit the training set. Firstly, if we train the model more epochs, the testing accuracy and training accuracy will not change a lot, which means the model is easily to struggle in local minimal. Secondly, this model has very high training accuracy, but the testing accuracy is relatively lower than the training accuracy, that denote that the model learned the training set too well so it lost generalization.

3.1 Dropout

Dropout is a recently introduced algorithm for training neural networks by randomly dropping units during training to prevent their co-adaptation[4][5]. This method adds noise to hidden units, so it can improve the generalization ability of the model.

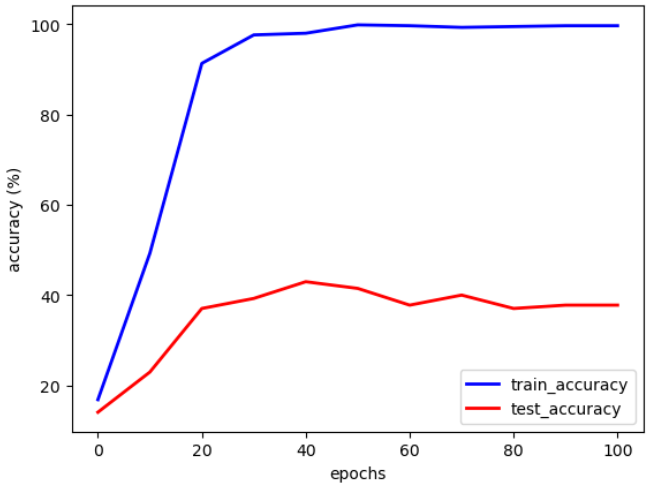

Fig. 5. Applying dropout. The training accuracy and testing accuracy were 97.96% and 42.96% respectively.

We applied dropout to the CNN model. Compared with simple CNN model, the training accuracy dropped but the testing accuracy increased. The result shows that compared with the model without dropout, this model has more generalization. When overtraining, the testing accuracy still changed, that denote this model is less likely to struggle in local minimal.

3.2 Batch Normalization

In Batch Normalization[6], the data distribution has the attribute that the mean of the data distribution is 0 and the variance is 1. It could reduce the internal covariate shift, and it also implicitly regularizes the model due to the noise in the batch estimates for mean and variance.

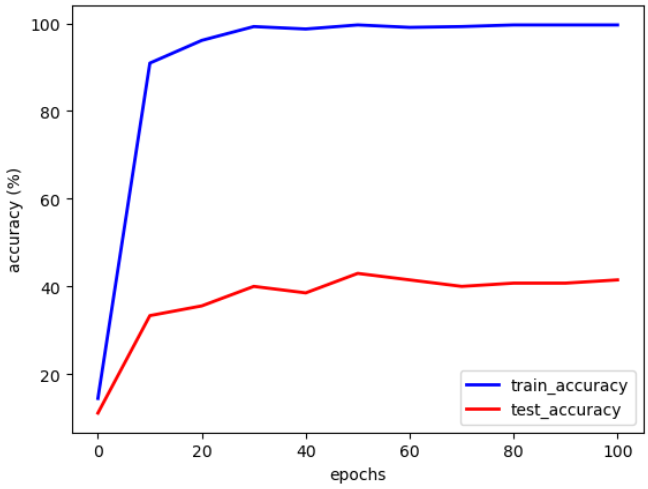

Fig. 6. After applying batch normalization (no dropout), the testing accuracy increased to over 40%, which is higher than the model without batch normalization.

3.3 Sparse batch normalization CNN model

Sparse Batch Normalization CNN (SBN-CNN) facial expression recognition model is an improved model of convolutional neural network based on VGGNet structure combined with facial expression image features[7]. SBN-CNN model uses continuous convolution at the begining of the network, and batch normalization layers are sparsely added in the network.This model performed outstanding in Japanese Female Facial Expression (FAFFE) and the Extended Cohn-Kanade (CK+) data set[7]. But the images in those data set were captured in lab environments, so we tested this model with the SFEW database.

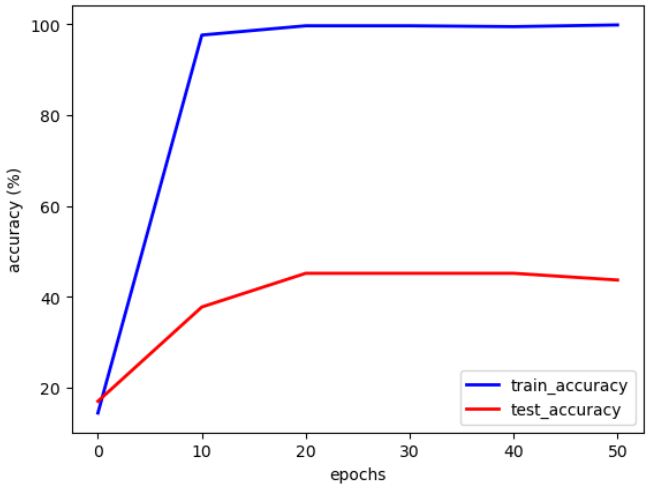

Fig. 7. Accuracy curve of SBN-CNN model. The training accuracy and testing accuracy are 99.63% and 45.19% respectively.

The result shows that SBN-CNN can get higher testing accuracy than simple CNN, CNN with dropout and CNN with batch normalization. In previous research[7] SBN-CNN could get more than 95% accuracy under JAFFE and CK+ data set, but the testing accuracy was only more than 45% under SFEW data set. The main reason that caused this result is the complexity of SFEW data set.

4 Conclusion

This thesis aimed to classify facial expression of images captured from realistic environments and we used two different classification methods. The first one is to project the data to lower dimension and then classify with a DNN, the second one is to classify the original image with a CNN. The result shows that CNN can preform much more better than DNN under the SFEW database.
We also used different methods to optimize the CNN. Firstly we added dropout layer to the CNN, secondly we added batch normalization layers to the CNN and finally we used a more complex CNN model. SBN-CNN performed best in those models, while both of CNN with drop and CNN with batch normalization performed better than simple CNN model.

5 Future work

Although CNN performs well in lab environments, and can get over 90% accuracy in face emotion classification task. Due to the complexity of the SFEW data set, CNN model did not show very good performance. If we want to get higher testing accuracy, we may need a more complex model. We could use a human face detector to detect the face from the image, and then use the detected human face as the input of CNN model. Therefore, the CNN model would not learn features related to background, lighting or something useless for face emotion classification.

Another possible future work is to find a method that can compute the contribution of inputs of a CNN. In previous research we introduced Q scores which could be extended to networks with large numbers of hidden layers and features, but this method has two drawbacks, firstly it can only compute the contribution of inputs for linear layers, secondly it is hard to compute for a network with too much neurons. The DNN classifier in CNN usually has thousands of input features, so Q scores is not suitable for CNN. In CNN the convolution layer capture the features in different region of the image. What we need is a method that can find which region of the image is the most significant for the output.

References

1. Dhall, A., Goecke, R., Lucey, S., Gedeon, T. (2011, November). Static facial expressions in tough conditions: Data, evaluation protocol and benchmark. In 1st IEEE International Workshop on Benchmarking Facial Image Analysis Technologies BeFIT, ICCV2011.
2. Bosch, A. , Zisserman, A. , Munoz, X. , Zisserman, P. . (2007). Representing shape with a spatial pyramid kernel. ACM international conference on Image and video retrieval. University of Girona, Girona, Spain;University of Oxford, Oxford, UK;.
3. C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/cjlin/libsvm
4. Baldi, P., Sadowski. (2014). The dropout learning algorithm. ARTIFICIAL INTELLIGENCE -AMSTERDAM-ELSEVIER-.
5. Hinton, G. E. , Srivastava, N. , Krizhevsky, A. , Sutskever, I. , Salakhutdinov, R. R. . (2012). Improving neural networks by preventing co-adaptation of feature detectors. Computer Science, 3(4), págs. 212-223.
6. Ioffe, S. , Szegedy, C. . (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. JMLR.org.
7. Jun, Cai, Quan, Chang, Xian-lun, Tang, et al. (2018). Facial Expression Recognition Method Based on Sparse Batch Normalization CNN. The 37th Chinese Control Conference.