Classifying Face Images: Comparing Neural Network, Deep CNN and Conventional Machine Learning Models

Hang Zhang[1]
u6921112@anu.edu.au

Research School of Computer Science, Australian National University, Australia

Abstract. Human facial expression is a very important part in human social life and communication, to recognize face expression, many features might be used, however, in most cases, reasonable data compression will not cause a large loss of data features and greatly improve computing efficiency. Therefore, LPQ and PHOG features are used as features to classify facial expression images with conventional machine learning methods such as logistic regression, decision tree and SVM(support vector machine) [Lindaand et al., 1995] and feed-forward neural networks. Besides, some experiments are conducted to check whether adjusting threshold of feed-forward neural network can improve the overall performance. I will also make good use of the modern deep convolutional neural network(CNN) to evaluate whether the deep CNN could extract facial expression features better than PHOG and LPQ, which may lead to better classification results.

Keywords: Image Classification · Deep Learning · Neural Network.

1 Introduction

Facial expressions are the most direct emotional expression in people's daily life and communication. Facial expression recognition is also an important part of machine learning and computer vision. In current people's lives, human-computer interaction is becoming more and more frequent, and correct recognition of human facial expressions becomes more important. To train a well-performed model, I used two datasets from Acted Facial Expressions in the Wild (AFEW)[Abhinav et al., 2011]. The first dataset is compressed data which is compressed to overall ten features of LPQ and PHOG descriptor to largely preserve original image information and largely improve computational efficiency. The second dataset is raw facial expression data set of the aforementioned one, therefore, it has a one-to-one correspondence with the compressed one. This paper includes experiments using both compressed and raw facial expression dataset on some conventional machine learning models, feed-forward neural networks and deep CNN to evaluate and analyze their performance.

2 The Data

2.1 Data Content

The compressed dataset contains 675 samples of human facial expressions in SFEW database.[Abhinav et al., 2011] The features of the data set are first 5 principle components of Local Phase Quantization (LPQ) descriptor features and the first 5 principal components of Pyramid of Histogram of Gradients (PHOG) descriptor features. LPQ features provide a lot of facial details which is common in facial analysis[Zhang et al., 2016], and PHOG features describe the local

gradient information which is widely used in computer vision.

The raw dataset contains 675 images, and each image has a one-to-one correspondence with a piece of data in the above compressed dataset. The both datasets are separated into 7 categories of facial expression which are angry, disgust, fear, happy, neutral, sad and surprise.

2.2 Data Distribution

The scale and distribution of features is important in training machine learning models, too large difference of features may cause the model's slow convergence and bad performance. According to the density curve of features the scales of features do not differ largely from each other. The histogram(Fig. 1) of label shows the data set is balanced enough. There is no need for further processing on the data set.

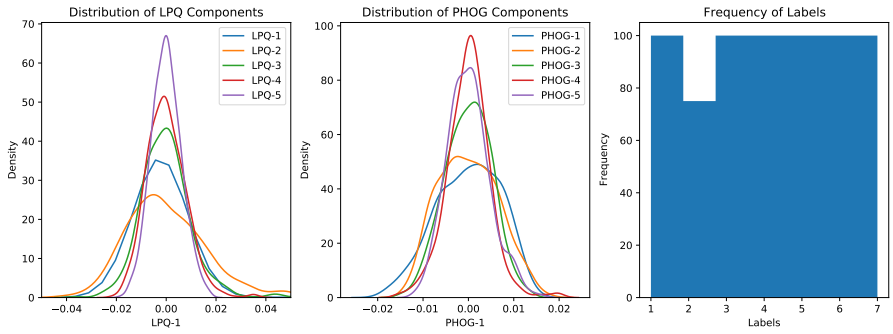


Fig. 1: Distribution of Features and Label

3 Conventional Machine Learning Classification

When we started a classification task we always need baseline for the task, and baseline is the classification results from some simple and common machine learning models such as decision tree, these learning algorithms are highly versatile and robust to data. Five-fold cross-validation is used to reduce the impact of data set division on the model fitting effect.

3.1 Decision Tree Classification

Decision tree classifier will derive a set of rules based on training data for classification and apply these rules to testing data for model validation. The result is shown in Table. 1, the average validation accuracy of the model is **20.9%**.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.234234	0.106061	0.275510	0.27957	0.153846	0.204301	0.181818
Recall	0.260000	0.093333	0.270000	0.26000	0.160000	0.190000	0.200000
F1-score	0.246445	0.099291	0.272727	0.26943	0.156863	0.196891	0.190476

Table 1: Decision Tree Classification Report

3.2 Logistic Regression Classification

Logistic regression algorithm is the most basic linear classification method which is robust to different kinds of data, therefore, it is always used as the baseline of classification tasks. The result of logistic regression is shown in Table. 2, the average validation accuracy of the model is **13.2%**. It shows that logistic regression can not fit this model well. Note that there may NaN appear for precision and F1-Score values, this happens when the model predicts zero instance to be that class.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.125000	NaN	0.240	0.125926	0.135135	0.121212	0.130435
Recall	0.250000	0.0	0.060	0.170000	0.150000	0.080000	0.180000
F1-score	0.166667	NaN	0.096	0.144681	0.142180	0.096386	0.151261

Table 2: Logistic Regression Classification Report

3.3 Support Vector Machine Classification

Support vector machine is a relative complex model compared to the previous two models, however, it is efficient and guaranteed to find the optimal solution. Besides, in some cases, SVM is equivalent to shallow neural network, which provides heuristics for neural network classification. The result of SVM with rbf non-linear kernel[Gao and Zhang, 2007] classification is shown in Table. 3, the average validation accuracy of the model is **22.1%**.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.213793	0.250000	0.273504	0.281690	0.181034	0.267606	0.263158
Recall	0.310000	0.026667	0.320000	0.400000	0.210000	0.190000	0.200000
F1-score	0.253061	0.048193	0.294931	0.330579	0.194444	0.222222	0.227273

Table 3: SVM Classification Report

4 Neural Network Classification

Neural network is a more complex model for classification task, different topologies can give different results. Neural networks with more hidden layers and more hidden nodes may lead to overfitting on training set, however, NN with too simple topology may lead to insufficient expressiveness of the model and unable to fit complex data. Therefore, the NN model is built from simple structure to complex ones. The following configurations(Table. 4) remain unchanged in all attempts. **The classification reports displayed afterwards are all generated by the best performing model during the training process.**

Optimizer	Adam
Learning Rate	0.01
No. Epochs	4000
Loss Function	Cross Entropy Loss
Hidden Layer Activation	ReLU
Output Layer Activation	Sigmoid

Table 4: Neural Network General Configurations

4.1 NN with One Hidden Layer

For the simplest case, only one hidden layer with ten hidden nodes are used in this neural network. Even though it is a shallow and simple neural network, it is expected to perform better than conventional machine learning algorithms. The result of one hidden layer neural network is shown in Table. 5, the average validation accuracy of the model is **15.1%**. The result shows that it performs worse than most of previous models which may due to simple architecture and lack of expressiveness. More complex neural network models may help improve performance.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.200000	0.140741	0.196078	0.147420	0.189655	NaN	0.500000
Recall	0.040000	0.253333	0.100000	0.600000	0.110000	0.0	0.020000
F1-score	0.066667	0.180952	0.132450	0.236686	0.139241	NaN	0.038462

Table 5: One-Hidden-Layer NN Classification Report

4.2 NN with More Hidden Layers

First try neural networks with two hidden layers and each layer has ten hidden nodes, this model is expected to perform better than the one-hidden-layer neural network due to better expressiveness. Final result report of two-hidden-layer neural network is shown in Table.6, the average validation accuracy of the model is **21%**, which indicates that more complex

model may lead to better performance.

Therefore, more complex models are tested, accuracies results comparisons are shown in

Emotion	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
Precision	0.164384	0.200000	0.255474	0.260116	0.265306	0.236559	0.25
Recall	0.120000	0.133333	0.350000	0.450000	0.130000	0.220000	0.25
F1-score	0.138728	0.160000	0.295359	0.329670	0.174497	0.227979	0.25

Table 6: Two-Hidden-Layer NN Classification Report

Table. 7, which indicates that increase complexity of the model is not necessarily leading to improvement of performance. As the result shows, two-hidden-layer neural network model is the most appropriate model.

No. Layers	**1**	**2**	**3**	**4**	**5**
Accuracy	0.151	0.211	0.212	0.225	0.207

Table 7: Multi-Hidden-Layer NN Classification Accuracy

4.3 NN with Customized Threshold

In all experiments, five-fold cross-validations are used. Therefore, for each round of training, all the validation sets are combined to form the complete original data set. Therefore, the correct label distribution is 100, 75, 100, 100, 100, 100, 100. However, from the results of many experiments, the model is always more biased towards predicting a few of them. Fig. 2 shows the sum of the distributions of the predicted results and ground truth of twenty trials. It clearly presents that the model prefer to make predictions on **Happy** and **Fear**, while making few predictions on **Disgust**, therefore, customized threshold will be used to make the model be less likely to make predictions on **Happy** and **Fear**, and be more likely to make predictions on Disgust.

The original prediction step is as follow: The model outputs a vector, then the one with the greatest value will be selected as the predicted class. To add threshold in this multi-label circumstance, simply add some fixed numbers to the model's output vectors then take the maximum value as the predicted class. According to prediction frequency from Fig. 2, adding a large number for **Disgust** class is needed. Then perform some experiments using different additional vectors to find the most suitable value for this model.

Based on the previous frequency distribution, the first additional vector [0.15, 0.25, 0.05, 0, 0.1, 0.1, 0.1] for experiment is used to adjust the model's output vector, especially increase the prediction preference for the **Disgust** category and decrease the prediction preference for the **Happy** category. Classification report is as Table. 8, the average accuracy is **22.3%**. The overall accuracy does not improve significantly, however, the evaluation on other metrics are more balanced than the original situation.

The Frequency distribution of prediction is shown as Fig. 3. It is clear that the prediction

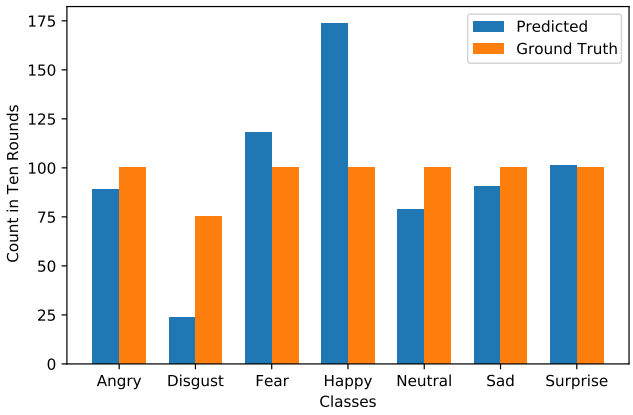


Fig. 2: Initial Frequency of Prediction and Ground Truth

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.205357	0.166667	0.240741	0.245763	0.227848	0.225490	0.225490
Recall	0.230000	0.120000	0.260000	0.290000	0.180000	0.230000	0.230000
F1-score	0.216981	0.139535	0.250000	0.266055	0.201117	0.227723	0.227723

Table 8: Classification Report with the First Threshold Set

distribution becomes more balanced than original situation, however, due to I added a too large value to **Angry** category, and a value not great enough to **Disgust** making too many predictions on **Angry** and too few predictions on **Disgust**.

Therefore, I try to use greater addition value for Disgust and smaller addition value for **Angry**, then another additional vector [0.05, 0.35, 0, 0, 0.1, 0.1, 0.1] is used in the a new experiment. Classification report is as Table. 9, the average accuracy is **23.8%**.

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Precision	0.187303	0.250000	0.255814	0.265487	0.201493	0.264706	0.281553
Recall	0.180000	0.160000	0.220000	0.300000	0.270000	0.270000	0.290000
F1-score	0.183578	0.195122	0.236559	0.281690	0.230769	0.267327	0.285714

Table 9: Classification Report with the Second Threshold Set

The report table and average accuracy have shown that with change of threshold, the overall performance proved a little, with more balanced predictions, which is supported by Fig. 4, prediction distribution is closer to ground truth distribution.

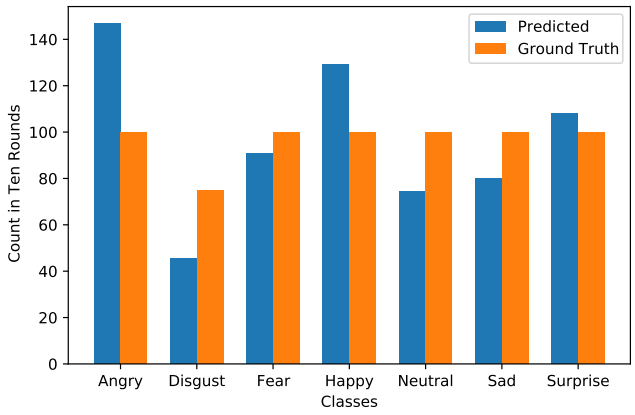


Fig. 3: Frequency of Prediction and Ground Truth with the First Threshold

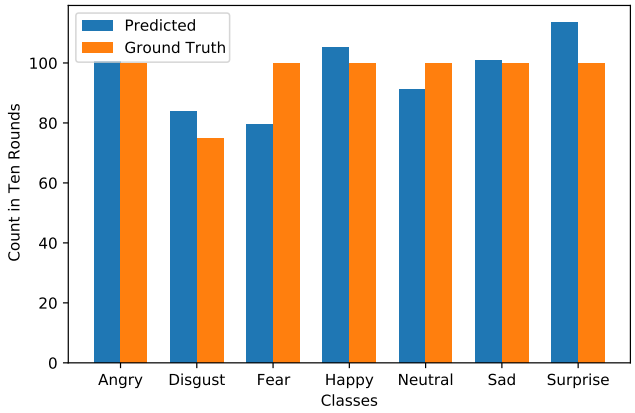


Fig. 4: Frequency of Prediction and Ground Truth with the Second Threshold

5 Deep Convolutional Neural Network Classification

In recent years, convolutional neural network has been proved to be superior to conventional computer vision methods in many image related problems, such as image classification, semantic segmentation, target detection, etc. [Zheng et al., 2017], especially the emergence of deep convolutional neural network, which makes use of the characteristics of residual to further improve the image information processing ability of convolutional neural network, such as, ResNet and MobileNet [He et al., 2015; Sandler et al., 2019].

In this section, we will conduct some experiments to compare the performance of a deep convolutional neural network trained with raw datasets and the best classification model in Subsection. 4.3 trained with the compressed dataset. In essence, this is also a comparison of image features extraction ability of deep CNN and conventional computer vision methods which is LPQ and PHOG in the compressed dataset. Therefore, we chose ResNet18 and

ResNet50 as our experiment models. The model training configuration details are shown in Table a. In this experiment, due to the limitation of computing resources, I did not use five-fold cross-validation in the experiment.
Therefore, we chose ResNet18 and ResNet50 as our experiment candidate models. The model

Optimizer	Adam
Learning Rate	2e-5
No. Epochs	20
Loss Function	Cross Entropy Loss
Hidden Layer Activation	ReLU
Output Layer Activation	Sigmoid

Table 10: CNN General Configurations

training configuration details are shown in Table. 10. In this experiment, we randomly choosing one fifth of the data as the validation set. After training, we evaluated our model on the validation set. Finally, we got **51.1%** validation accuracy with ResNet18 and **50.4%** validation accuracy with ResNet50. Detailed classification reports are shown in Table. 11 and 12.

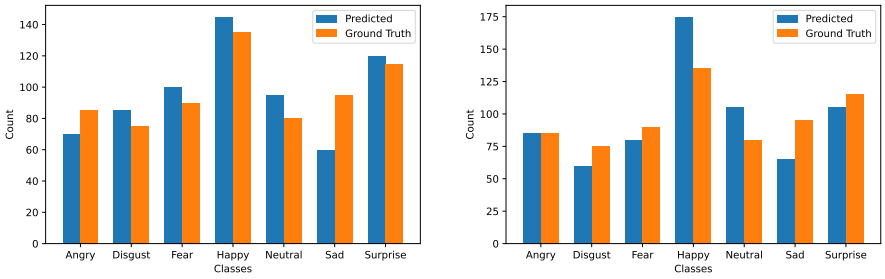
Emotion	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
precision	0.500000	0.352941	0.650000	0.620690	0.263158	0.916667	0.333333
recall	0.411765	0.400000	0.722222	0.666667	0.312500	0.578947	0.347826
f1-score	0.451613	0.375000	0.684211	0.642857	0.285714	0.709677	0.340426

Table 11: Classification Report with ResNet18

Emotion	**Angry**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**	**Surprise**
precision	0.411765	0.583333	0.625000	0.571429	0.333333	0.923077	0.285714
recall	0.411765	0.466667	0.555556	0.740741	0.437500	0.631579	0.260870
f1-score	0.411765	0.518519	0.588235	0.645161	0.378378	0.750000	0.272727

Table 12: Classification Report with ResNet50

The classification reports have shown that the performance of deep CNN is much better than the previous methods that based on conventional computer vision feature extraction. Besides, in Subsection. 4.3, we have discussed how the customized threshold could possibly have influence on the model's performance and prediction distribution. While for deep CNN we trained we noticed that the prediction distribution is similar enough to the ground truth distribution(Fig. 5), therefore, there is no evidence that we need to adjust the customized threshold to the original model.



(a) ResNet18 Distribution (b) ResNet50 Distribution

Fig. 5: Frequency of Prediction and Ground Truth of Deep CNN Models

6 Future Work and Conclusion

By comparing with different classification algorithm, it has shown that performance does not differ greatly for some conventional machine learning algorithm and neural network. Logistic regression has the worst performance on the validation set. Then some experiments are conducted on neural network have shown that setting appropriate threshold can improve the overall performance of the model on this Face-emotion dataset.

We also train two deep CNN with the raw face emotions data to compare the ability of deep CNN and traditional computer vision methods, which are LPQ and PHOG, in extracting image features. We found that with the excellent image analysis ability of deep CNN, we can greatly improve the performance of the downstream classification task.

So far, we have known that a deep convolution neural network can extract image features better than conventional computer vision methods. However, in this paper, we do not discuss in detail why does this happens and how these features are different. In addition, we have not compared the feature extraction ability of shallow CNN and deep CNN and to what extent they are different. Further, we can find a suitable method to make them visual and more intuitive.

References

Abhinav, D., Roland, G., Simon, L., and Tom, G. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. pages 2106–2112.

Gao, D. and Zhang, T. (2007). Support vector machine classifiers using rbf kernels with clustering-based centers and widths. pages 2971–2976.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Lindaand, M., Tom, G., and Andrew, S. (1995). Classifying dry sclerophyll forest from augmented satellite data : Comparing neural network, decision tree and maximum likelihood.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2019). Mobilenetv2: Inverted residuals and linear bottlenecks.

Zhang, Z., Li, F., Liu, M., and Yadav, P. K. (2016). Image matching based on local phase quantization applied for measuring the tensile properties of high elongation material. *Mathematical Problems in Engineering*, 2016:1–10.

Zheng, L., Yang, Y., and Tian, Q. (2017). Sift meets cnn: A decade survey of instance retrieval.