Fine-tuning Technologies for Reducing the FER Bias Across Various Distributions

Zhisong Liu

Department of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, 610054, China
2021090914013@std.uestc.edu.cn

Abstract. Lacking sufficient data has become a serious problem in the field of Facial Expression Recognition (FER), since the cost of collecting a large amount of facial expression images is huge and training a new FER model from the beginning is time-consuming. In this paper, the author trained a FER model based on a gray-scale dataset (FER2013) and found several shortages in both the dataset and the model. In order to achieve better accuracy and reduce the bias in the previous training domain, the author searched for a new dataset and applied transfer learning to transfer the FER model to the new domain. More specifically, this study was based on the MobileV2 Convolution Neuron Network (CNN) model and the author adjusted the top layers to match the FER classification task, the special inverted residual blocks in the MobileV2 accelerate the training process while ensuring the high accuracy. Since the data were all labeled, this study applied model fine-tuning and froze the weights of the first few layers in the model which were trained to detect the special features in the images. Thus, by adjusting the weights of the fully connected layers, the model successfully transferred to a similar domain. Experimental results indicated that after applying the model fine-tuning, the FER model performed much better while recognizing colorful images of faces from different human races and the new model reduced the bias created by the previous training dataset.

Keywords: Computer Version, Facial Expression Recognition, Transfer Learning.

1 Introduction

Facial Expression Recognition (FER) is about detecting human emotion states due to the specific movement of their muscles on their face, it is significant for computers to understand human emotion and affective states [1]. The recognition process requires computer to classify expressions from photos of human faces and separate them into seven basic expressions including “angry”, “disgust”, “fear”, “happy”, “sad”, “surprise” and “neutral”, the first 6 expressions were defined by American psychologist Ekman and Friesen in the 1970s [2], and the “neutral” was added later by other researchers. According to psychology research, facial expression shows more communi-

Y. Wang (ed.), *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, Advances in Computer Science Research 115,
https://doi.org/10.2991/978-94-6463-540-9_92

cation information than any other non-verbal communication ways [3]. From the technology point of view, since facial expressions carry all kinds of this information, FER can be significant for fields like Computer Vision (CV) and Human Computer Interaction (HCI) and provide ways for machine-human communication.

In the early field of FER, many studies introduced simple neural networks to judge the expression, but dealing with pictures which may contain 10 thousand or even more pixels, the training process was long, and the accuracy was still not satisfying. In 2013, the champion of the competition of FER got 67.48% accuracy by using the Local Learning Bow, but this record was easily broken since researchers started to use the Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). CNN models like ResNet, VGGNet and GoogleNet all present their own advantage in classifying facial expressions. CNN models often have a large number of parameters and layers, and the common result of training with CNN is that the deeper the network is, the better the model will perform in classification and capture the most representative features [4-7]. But the deeper layers always bring problems to the computing process and network like VGG16 has about 138 million parameters which may take researchers a week or even more time to train and reach the final result. In order to avoid the time and resources researchers may waste on these huge CNN models, transfer learning comes out. Transfer learning focuses on helping researchers to transfer their training result from one domain to a different but similar domain. In the case of using a CNN model which may take about weeks to compile, researchers can use a pre-trained network and make a fine-tune to adjust it to a new dataset. Most of the FER datasets only contain a few hundreds or thousands of data, and the data insufficiency is considered to be one of the major challenges for FER, transfer learning can provide help to save the time and solve the problem of data insufficiency [1, 8].

After building and training a MobileNetV2 model with the dataset named Facial Expression Recognition 2013 (FER2013) collected from Kaggle [9], the author discovered that the accuracy of predicting expression was still not satisfactory, so this study aimed to improve the model with transfer learning. After evaluating the FER2013 dataset, the author found several problems with the dataset. Data from FER2013 are all in gray scale and most of the faces are from white people, thus when using the model to predict colorful pictures or faces from the other human races, the recognition performance will become worse. In order to reach a higher accuracy and adjust the model to have less bias on human races, the author did a transfer learning with a more balanced dataset. Based on the previous model, the author blocked the convolution layers which were used to capture features and made a fine-tune to the top layers. This time the test accuracy on the FER2013 increased and had a better performance while predicting colorful faces.

2 Methodology

2.1 Dataset

This study includes 2 datasets, and the first dataset is Facial Expression Recognition 2013 (FER2013) [9]. This dataset contains 7 expressions: angry (3,995 pictures), disgust (436 pictures), fear (4,097 pictures), happy (7,215 pictures), neutral (4,975 pictures), sad (4,830 pictures), surprise (3,171 pictures). The size of FER2013 pictures is all 48×48, and all the data are in gray scale.

The second dataset is Facial Expression Training Data (FETD) [10]. This dataset contains 8 categories, but since this study plans to do the transfer learning and improve the prediction accuracy on dataset FER2013 which only has 7 expressions, the author removed expression "contempt" from the original FETD dataset, the rest classes are: angry (3,218 pictures), disgust (2,477 pictures), fear (3,176 pictures), happy (5,044 pictures), neutral (5,144 pictures), sad (3,091 pictures), surprise (4,039 pictures). Data from FETD is composed of RGB channels and has the size of 96×96. Fig. 1 provides some examples from the collected datasets.

Fig. 1. Sample images from the collected datasets .

Since the amount of data is still not large enough and the data sizes are different in two datasets, the author did augmentation for both datasets. The augmentation preprocess included changing the size of data (to 224×224), implementing horizontal flips, zooming (at range 0.3) and height shifting (at range 0.2).

2.2 Transfer Learning Technical Details

Basic convolution neural network contains convolution layers, pooling layers and flatten layers. Convolution layers can easily detect spatial features such as edges and corners in images. Pooling layers can constantly decrease the spatial size of data, so the number of parameters and computation will also decrease, which can somehow prevent the model from overfitting. Flatten layers can convert multi-dimensional output from the layers above into a one-dimensional array, which can be used as the input of the fully connected layer. While building CNN models, researchers always make a complex combination of these three kinds of layers, and this combination always contains several convolution layers. The first CNN layers can learn the basic feature detection

filters like edges and corners. The middle layers can learn filters which detect parts of objects, in the case of FER, these middle layers might learn to response to eyes and noses. The last layers have higher representations, they may recognize the full face in different shapes and positions. In one word, each CNN layer learns features of increasing complexity.

This study's model is based on a CNN model named MobileNetV2 [11], which is lighter than most of the other typical models like VGG19 [12] or AlexNet [13]. The Inverted Residuals and the Linear Bottlenecks help to reduce the weight of MobileNetV2 and ensure high accuracy at the same time. Unlike the normal residual block shown in Fig. 2, which is made of a 1x1 dimensionally reduced convolution layer a 3x3 feature detecting convolution layer and a 1x1 dimensionally increased convolution layer, the inverted residual block first increases the dimension with a 1x1 convolution layer, and then applies 3x3 Depthwise Convolution layer to detect features, at last, it decreases the dimension by using a 1x1 convolution layer [11, 14]. The linear bottlenecks represent the linear activation function used on the projection convolution layer. While converting a high dimension to a lower one, using ReLu activation may cause the loss of information and damage the output. During the process of convolution, the lower the dimension of tensor is the less computation the convolution will cause, but if the whole convolution layers are using the low dimension to detect features, the information gain from the network will not be enough, so the MobileNetV2 model uses inverted residual blocks to insure that features are detected in a higher dimension and compress the data to make sure the out tensor's dimension remain small.

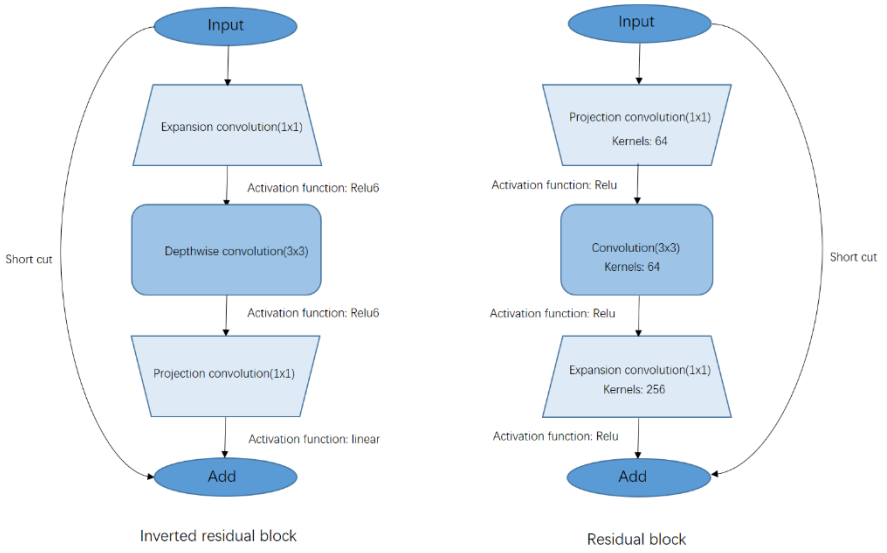

Fig. 2. Classical residual block structure and the inverted residual block structure .

Besides the basic feature detection layers, the author built fully connected layers after the original convolution layers of the MobileNetV2. The top layers include: A Dropout layer (dropout rate=0.5), an Average Pooling layer (pool size = (3,3)), a Flatten

layer, a Batch Normalization layer, a Dense layer with 1024 units uses Relu as activation function and uses L2 as kernel regularizer, a Dropout layer (dropout rate = 0.3), and an output Dense layer with 7 units and softmax activation function. The Average Pooling layer and the Flatten layer are used to convert the multi dimension output from the convolution layers above into a 1D array. Dropout layers and the Batch Normalization layers can help the model to avoid overfitting [15]. Since this study's task is to recognize and classify all the data into 7 classes, the final dense layer should contain 7 units and the softmax activation function.

Since the transfer learning process can transfer a pre-trained model from its original domain to a different but similar one, transfer learning was provided to be a vital solution when facing the problem of lacking sufficient data for the network to learn. Researchers have proved that the pre-learned features from the previous model have the robustness to handle all the problems that augmentation operates may bring, so transfer learning have shown promising results in the field of FER when dealing with the issues like head position problem, rotation problem and the other problem that augmentation may cause [1]. Generally, there are 4 ways to do transfer learning: Model Fine-tuning, Domain-adversarial training, zero shot learning, and Self-taught Learning [16]. In this study, both the source data (FER2013) and the target data (FETD) are labeled, model fine-tuning can be the fastest solution to do the transfer learning task. Model fine-tuning means taking a model which has already been trained with the old dataset and refreshing the classification layers' weights by training the model with a new dataset [17]. Since this task is to draw recognition on images, the author froze the convolution layers' weights which were used to detect features [18], and only did fine-tune on the fully connected layers to classify the expressions.

2.3 Implementation Details

This study's backend is based on Tensorflow and the Keras. In order to reach a better result in classification, this study chose Adam [19] as the optimizer and set the learning rate to 0.001. Besides, the loss function was Categorical Crossentropy. Trying to avoid the local minima and the problem of overfitting, this study applied early stopping function and reducing learning rate on plateau function as callback functions during the training process. The patience of early stopping function was set to 8 epochs and the patience of reduce learning rate function was 4 epochs.

The whole training process includes the first training process with dataset FER2013 which was set to train for 50 epochs and the second fine-tune training process which was about 20 epochs.

3 Results and Discussion

The original model, which was trained with the FER2013 dataset early stopped at about 35 epochs, reaching a training accuracy of 83.75% and a validation accuracy of 64.85%. By loading the weights of the original model and retrained the fully connected layers

on the mixed dataset of FETD&FER2013 for 20 epochs, the new model reached a training accuracy of 84.34% and a validation accuracy of 68.58% on FER2013 test dataset.

Fig. 3. Test result between the original model and the fine-tuned model.

After training the original model and the fine-tuned model, this study made a comparison between these two models by testing their accuracy on the FER2013 test dataset and the FETD test dataset shown in Fig. 3. The result claimed that by using transfer learning and making fine-tune on the original model, the new model predicted higher accuracies on both the FER2013 test dataset (1% higher) and the FETD test dataset (40% higher), especially when predicting the colorful pictures and facial expressions from different human races in the FETD dataset. Fig. 4 shows the confusion matric when predicting with the FETD test set.

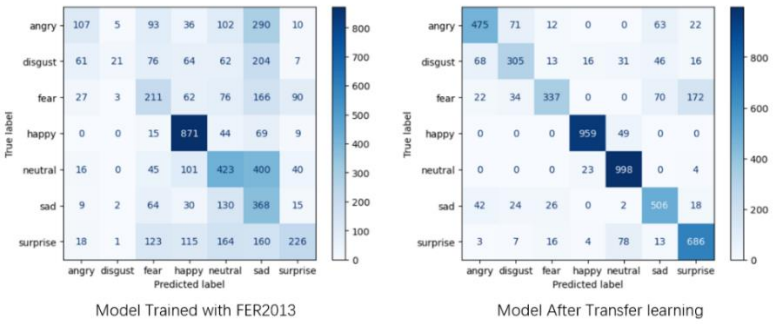

Fig. 4. Confusion matrices predict on the FETD test set .

From the confusion matrices in the Fig. 4, this study found that the original model which was trained with FER2013 had problem with recognizing the expression of "disgust", since the FER2013 dataset was imbalanced and the training data for "disgust" was less than the other expressions. Besides, in the FER2013 dataset, facial expression

data like “fear” “sad” and “surprise” not only contain the human faces, but they also have hands on these faces. These particular expressions are predominantly associated with the appearance of hands in these images and lead to the inaccurate learning to correlate and presence of hands with these expressions. But with the help of transfer learning, this study successfully overcame these shortages by transferring the old domain (FER2013) to a similar but less bias domain (FETD). On the one hand, FETD dataset enlarged the training data for the “disgust” expression, on the other hand FETD has less faces which are covered by hands and provides faces from different human races. Model Fine-tuning skills in transfer learning successfully helped the original model to learn new weights in the fully connected layers and made a huge improvement when predicting all these 7 expressions. Beyond all these achievements, there are still some deficiencies. Model fine-tuning requires both the source data and the target data to be labeled, otherwise, fine-tuning may not reach the expected result. Under those circumstances, different transfer learning skills should be applied depending on whether the data is labeled or not. In cases that the source data is labeled but the target data is unlabeled, domain adaptation training can handle these transfer process well by lying to the domain classifier layers in the model and forcing the feature extractor layers to detect the Generic features from these both the old and the new domains [20].

4 Conclusion

In this study, the author focused on improving the FER model trained with FER2013 to reduce the bias on human races and reach a better result when predicting colorful pictures. This study applied transfer learning skills, especially the model fine-tuning to make adjustment to the previous model weights and reach a higher accuracy in both the previous FER2013 dataset and the colorful FETD dataset. Experimental results showed that after applying transfer learning and model fine-tuning, the new model performed higher accuracy on predicting both the FER2013 dataset and the FETD dataset. Especially when predicting the colorful RGB images, the transferred model advanced about 40% in test accuracy comparing with the original model. In the future, the author plans to do extensive experiments like using other transfer learning skills to conduct a higher accuracy on gray-scale dataset and adjust this FER model to do real-time facial expression recognition tasks.

References

1. Ekundayo, O. S., Viriri, S.: Facial Expression Recognition: A Review of Trends and Techniques. IEEE Access 9, 136944-136973 (2021).
2. Ekman, P., Friesen, W. V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17(2), 124–129 (1971).
3. Hall, J. A.: Voice, tone, and persuasion. Journal of Personality and Social Psychology, pp. 924–934 (1980). Petterson, M. L.: Function of nonverbal behavior in social interaction. In: Giles, H., Robinson, W. P. (eds.) Handbook of Language and Social Psychology. John Wiley & Sons, New York (199).

4. Qiu, Y., Hui, Y., Zhao, P., Cai, C. H., Dai, B., Dou, J., Bhattacharya, S., Yu, J.: A novel image expression-driven modeling strategy for coke quality prediction in the smart cokemaking process. Energy 294, 130866 (2024).
5. Ye, X., Wu, P., Liu, A., Zhan, X., Wang, Z., Zhao, Y.: A deep learning-based method for automatic abnormal data detection: Case study for bridge structural health monitoring. International Journal of Structural Stability and Dynamics 23(11), 2350131 (2023).
6. Liu, Y., Bao, Y.: Intelligent monitoring of spatially-distributed cracks using distributed fiber optic sensors assisted by deep learning. Measurement 220, 113418 (2023).
7. Qiu, Y., Wang, J., Jin, Z., Chen, H., Zhang, M., Guo, L.: Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control 72, 103323 (2022).
8. Zhuang, F., et al.: A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE 109(1), 43-76 (2021).
9. Sambare, M.: FER-2013. Available at https://www.kaggle.com/datasets/msambare/fer2013 (Accessed on 20 May, 2024).
10. Segal, N.: Facial Expression Training Data. Available at https://www.kaggle.com/datasets/noamsegal/affectnet-training-data (Accessed on 20 May, 2024).
11. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510-4520 (2018).
12. Mascarenhas, S., Agarwal, M.: A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In: 2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON), Bengaluru, India, pp. 96-99 (2021).
13. Chen, H.-C., Widodo, A. M., Wisnujati, A., Rahaman, M., Lin, J. C.-W., Chen, L., Weng, C.-E.: AlexNet Convolutional Neural Network for Disease Detection and Classification of Tomato Leaf. Electronics (2022).
14. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251-1258 (2017).
15. Gonçalves, C. F., Dos Santos, J. P. P.: Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. ACM Computing Surveys, Volume 54, Issue 10s Article No.: 213pp 1–25 (2021).
16. Hosna, A., Merry, E., Gyalmo, J., et al.: Transfer learning: a friendly introduction. Journal of Big Data 9(1), 102 (2022).
17. Church, K. W., Chen, Z., Ma, Y.: Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering 27(6), 763-778 (2021).
18. Yosinski, J., Clune, J., Bengio, Y., et al.: How transferable are features in deep neural networks? Advances in Neural Information Processing Systems 27 (2014).
19. Melinte, D. O., Vladareanu, L.: Facial Expressions Recognition for Human–Robot Interaction Using Deep Convolutional Neural Networks with Rectified Adam Optimizer. Sensors 20, 2393 (2020).
20. Rangwani, H., Aithal, S. K., Mishra, M., et al.: A closer look at smoothness in domain adversarial training. In: International Conference on Machine Learning. PMLR, pp. 18378-18399 (2022).