Facial Expression Recognition with Transfer Learning from Deep Convolutional Networks

Jiazong Gong

Research School of Computer Science, Australian National University
u6660171@anu.edu.au

Abstract. In this paper, we explored facial expression recognition methods based on transfer learning from deep convolutional neural network with ResNet backbone where the image-based data is collected from Static Facial Expressions in the Wild database. Apart from the original database, we also produced a new database where the face images are recognised and generated with FaceNet and made a comparison between the accuracy of models on both of the databases. After a thorough evaluation on the newly-generated database, we found that this new database could not lead to performance increase but actually decrease the accuracy of the original database by more than 10%. Regarding this unexpected result, we proposed some hypothesis and suggestions for future facial expression recognition task.

Keywords: Convolutional Neural Network, FaceNet, Transfer Learning, ResNet, Static Facial Expressions in the Wild

1 Introduction

Facial emotion recognition is a popular topic in computer vision, capable of recognizing emotion types from video-based or image-based data where the challenge lies in face detection and recognition [1]. There are various techniques to solve the challenges, such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), K-nearest neighbor (KNN) and Support Vector Machine (SVM) [2–4].

1.1 Problem Description

The purpose of the investigation is to apply transfer learning with pretrained convolutional neural network (CNN) on the images and recognize the face emotion from given seven classes since there are insufficient images in the database. After the transfer learning approach, an additional FaceNet recogniser will be implemented to extract the faces from the images and then these faces will be further used in the emotion recognition task. Results from the original transfer learning approach and the one with cropped face approach will be later compared.

1.2 Dataset

The dataset is Static Facial Expressions in the Wild (SFEW) which is extracted from a temporal facial expressions database Acted Facial Expressions in the Wild (AFEW) [5], a static facial database containing 700 images in 7 classes including Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. For this task, there are 75 images in class Disgust instead of 100.
The reason why we choose this database for facial expression recognition is that the database is balanced and there are small size of facial expressions from abundant movie scenes to make the database representative.

1.3 Transfer Learning

The purpose of transfer learning is to utilise existing models acquired from previous task on another relevant task [6]. For the task of image recognition, convolutional neural network (CNN) is a common transferable technique which is mostly built based on ImageNet, a database that contains more than 14 million images for image recognition [7]. In the pretrained CNN, there are convolutional layers, pooling layers and fully connected layers where the fully connected layer will be retrained in transfer learning and other layers will be used as feature extractor in transfer learning.

1.4 FaceNet

FaceNet is a face recognition system that was developed by research group from Google, which builds deep convolutional neural networks to directly optimize its image embedding rather than use bottleneck layers as previous deep learning approaches [8]. It has achieved state of the art results on a variety of face recognition databases and thus proves its effectiveness in face recognition. [9-11]

2 Method

Fig. 1 shows the whole process for the investigation: SFEW image-based data is first loaded as training and testing data and then preprocessed before the training data is fed into the neural network which is ResNet pretrained on Imagenet by transfer learning. During the process of transfer learning, the parameters from the fully connected layer will be adjusted to task-specific. After finishing the transfer learning process, the trained model will be evaluated on the testing data for further performance comparison, during which the parameters within the model should be tuned for better performance. Additionally, there will be face preprocessing by FaceNet and faces from each image will be extracted and saved for expression recognition later.

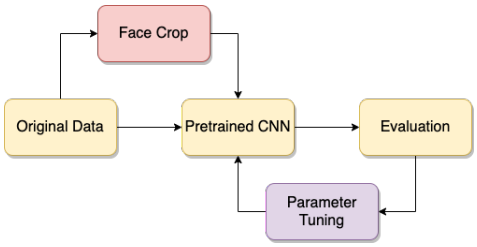

Fig. 1: Experiment Process

2.1 Data Loading

Since the backbone for the transfer learning is ResNet-50, the size of input images should then be scaled to $224 \times 224 \times 3$. Apart from resizing the original images, normalization is also required for the preprocessing, such as min-max normalization (1) or standardization methods (2).

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

$$Z = \frac{X - \mu}{\sigma}, \quad where \;\; \mu = \sum_{i=1}^{N} X_i, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} \tag{2}$$

For this task, we choose standardization over min-max normalization since min-max normalization is more sensitive to outliers and thus not robust and standardization is more commonly used in image-based dataset. Since it is always necessary to split the data into training data and testing data, the raw data is therefore split into training data by 5-fold cross validation, with 80% proportion and 20% for testing randomly to guarantee the reliability of the evaluation results. Because the dataset is not large enough to prevent potential variance during training or testing, we use stratification in the split process to guarantee the proportion of class distribution within training and testing data is the same as that within the original dataset, which can make each folder representative for the whole dataset to prevent potential large bias and variance from splitting small dataset.

2.2 Network Structure

As mentioned above, the data will be firstly preprocessed and then loaded into the pretrained model. The backbone is ResNet-50. To make it more illustrative, the topology of a basic residual block is shown below.

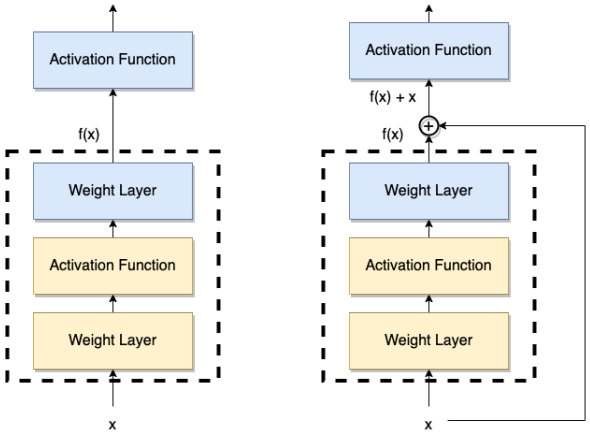

Fig. 2: A Normal Block (left) and a Residual Block (right)

The intuition behind the residual block is that sometimes identity mapping is optimal and it would be difficult for a set of non-linear layers to fit an identity mapping so that the residual blocks make it easier to optimize the residual function $f(x) + x$ than $f(x)$. During the training process, only the fully connected layer will be updated in each epoch and the hidden layers will be frozen.
Firstly, we will load the data into data loaders with preprocessing methods mentioned above for further classification task. At each epoch, the loss for the epoch will be calculated and the optimization function will be applied to reduce the loss. After enough epochs, the loss will be reduced to a low level and the model will be evaluated next.

2.3 Face Crop

Although the scenes in the images are abundant for better generalization, such surrounding abundancy could also be large noise in the dataset and leads to large influence on the expression recognition. Considering this, we introduce FaceNet to crop the faces from original dataset first and saved the faces for further recognition.

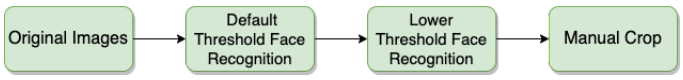

Fig. 3: Face Crop Process

However, the accuracy of FaceNet could not be 100% and thus some of the faces not recognised by FaceNet will require cropping manually. The principle of cropping faces manually is similar to that with FaceNet for the uniformity within the dataset. After saving all the cropped faces, they will be processed and loaded into the network the same as that of the original images.

2.4 Parameter Tuning and Implementation Specification

After the network and technique are decided and implemented, it is important then to tune the parameters in the network to increase the performance of the model. However, grid search is extremely time-consuming so that we only evaluate the performance on the original dataset with different batch size and it is found that 256 is better and faster than 128 and 64 so we decide to keep 256 as the batch size.
Therefore, learning rate is 0.01, batch size is 256, and training epoch is 200. As for the parameters in the network, activation functions are all ReLU and the backbone is ResNet-50 pretrained on ImageNet. For the optimizer, we choose Adam with L2 regularization where we set weight decay as 0.001 to prevent overfitting on such small dataset. Then we train and test our model on the original dataset and the cropped face dataset separately and compare their performance.

3 Results and Discussion

3.1 Transfer Learning on Original Images and Face Images

At first, a pretrained ResNet-50 is loaded and the fully connected layer is retrained on the original dataset and following evaluation is made to test the performance of the model. To make the evaluation results more reliable and convincing, we perform stratified 5-fold cross validation on the original dataset and each split will be trained and tested separately. In order to make the results solid to compare, we set the same random state when performing cross validation.

Table 1: Evaluation Results on Original Dataset

	Folder 1 (%)	Folder 2 (%)	Folder 3 (%)	Folder 4 (%)	Folder 5 (%)	Average (%)
Accuracy	38.5	40.0	41.5	41.5	43.0	40.9

As shown in the table above, the average accuracy is 40.9%. The results also show that the accuracy between each cross validation is roughly the same, which means the stratification in cross validation split really helps to reduce the variance and bias in the original images.
After the evaluation of transfer learning results on the original images, the results of transfer learning on the face images has also been collected as below with the same parameters mentioned previously.

Table 2: Evaluation Results on Face Dataset

	Folder 1 (%)	Folder 2 (%)	Folder 3 (%)	Folder 4 (%)	Folder 5 (%)	Average (%)
Accuracy	43.0	33.3	31.1	37.8	34.1	35.9

As shown in the table above, the average accuracy of the 5 tests is 35.9% and it is 12.2% lower than the accuracy on the original dataset. Besides, it also shows that the accuracy between each cross validation can be very different and such large difference might be due to the potential overfitting on certain set of cross validation or the face crop process is not effective on this dataset. Either way, we need to do some more experiments since the accuracy on the processed face dataset is supposed to be higher than that on the original dataset since it reduces the influence from the surrounding or the scenes in the original images. In order to make sure whether the model could actually recognise expressions on face dataset well or not, we tested different ResNet with different epoch number:

Table 3: Further Evaluation Results on Face Dataset

	Folder 1 (%)	Folder 2 (%)	Folder 3 (%)	Folder 4 (%)	Folder 5 (%)	Average (%)
(ResNet-50, 100)	44.4	36.3	35.6	40.7	37.0	**38.8**
(ResNet-50, 50)	39.3	35.6	35.6	42.2	37.8	38.1
(ResNet-34, 100)	37.8	38.5	36.3	31.1	40.7	36.9
(ResNet-34, 50)	36.3	43.0	32.6	35.6	40.7	37.6
(ResNet-18, 100)	39.3	36.3	30.4	32.6	38.5	35.4
(ResNet-18, 50)	42.2	34.8	35.6	34.1	34.1	36.1

The table above clearly shows that the model should have been overfitting with larger epoch number previously and it also shows that the ResNet-50 with 100 epochs could achieve higher accuracy than other models on face dataset. Accordingly, we tested ResNet-50 with 100 epochs on the original dataset and found that the overall average accuracy is the same as before, only that the accuracy between folders are slightly different so we do not record them here. Besides, it is found that the accuracy between folders can be really different compared with that on original dataset, which shows that the face cropping might not be very stable due to some reasons such as no uniform metrics and a lack of quality control. And the low accuracy on both datasets indicate that the pretrained model might need fine tuning for the last few transferred layers.

3.2 Comparison between Two Approaches

From the previous evaluations and discussions, it is found that recognising directly on cropped face images might not contribute to the performance increase and the performance on the original dataset is also not satisfying.

In order to obtain a better understanding the performance on each dataset, we record the confusion matrix for each of them. In order to make the matrix more representative, we choose the split that could achieve similar accuracy to the overall average accuracy.

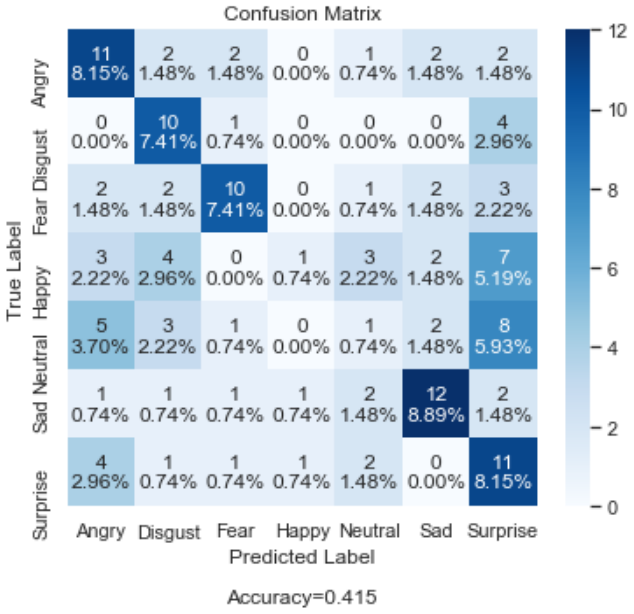

Fig. 4: Confusion Matrix on Original Dataset

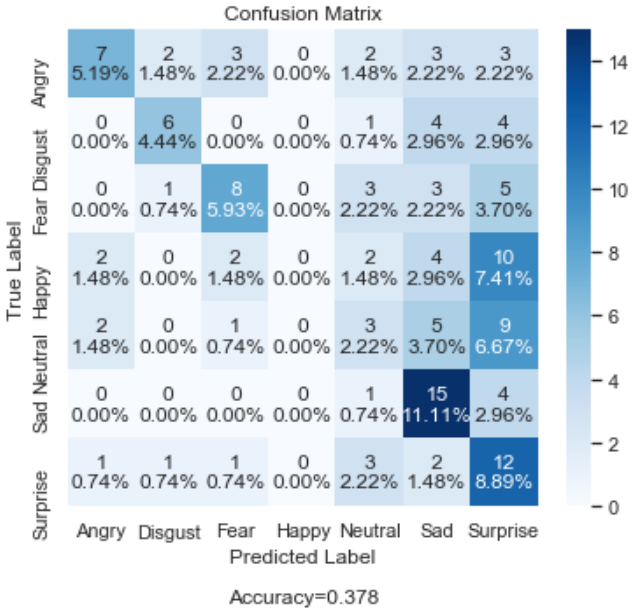

Fig. 5: Confusion Matrix on Face Dataset

As shown in the two confusion matrixes above, the models tend to recognise less "Happy" and "Neutral" and the model does not classify any expression as "Happy" in the face dataset. Instead, both models tend to classify more instances in the testing folder as "Surprise". And this could also be due to the fact that the last few layers

in the pretrained model are task-specific and thus could not extract high-level features to help the current task [13]. However, the model trained on face dataset is still supposed to perform better than that on the original dataset because the influence from the surrounding scenes has been relieved. Therefore, it is necessary to check the images in the database and here are some examples I found.

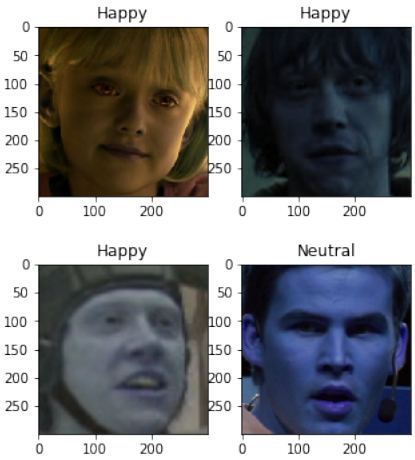

Fig. 6: Some Examples

From the examples, we realise that there could be some expressions not well-labelled like the three expression labelled as "Happy" which looks pretty "Neutral" to me or these expressions just cannot be correctly labelled to one specific emotion and thus could "mislead" the model to a wrong direction in the training process. Sometimes, it is really difficult to say whether a person is angry being silent or just neutral and it is also like Mona Lisa's smile, which has confused people whether she is smiling or frowning, or maybe "Neutral" for this task.

3.3 Overall Comparison with Statistics from Paper

The evaluation results from different methods have been collected and analyzed in previous sections. It is significant to make comparison with these approaches and the approach from the paper.

Table 4: Comparison between Our Approach and Paper

Method	Average Accuracy (%)
Transfer Learning on Original Dataset	**40.9**
Transfer Learning on Face Dataset	38.8
SPI Baseline [5]	19.0
LPG [5]	43.7
PHOG [5]	**46.3**

As shown in the table above, it is found that transfer learning is not a good solution for this task despite there could be some other reasons within the dataset since the performance of transfer learning is 11.7% lower than classify on pyramid of histogram of oriented gradients (PHOG) features which is a histogram representing image shape where each bin in the histogram represents the number of edges that have orientations within a certain angular range [12]. Such performance weakness also shows that the model might need fine tuning or it is not very suitable for facial expression recognition in the first place.

4 Conclusion and Future Work

4.1 Conclusion

Facial expression recognition is still a difficult and challenging topic and approaches like transfer learning could be a potential good solution to the task. This paper explored transfer learning on SFEW database for expression recognition task and we have also applied FaceNet to extract faces from the database and apply another transfer learning on these face images. Although both methods have not achieved a better performance than that mentioned in the paper, it is still worth considering whether facial expression recognition is really a well-defined task since there could be some expressions that might not be well labelled or it would be difficult to strictly label certain expression as one specific emotion.

4.2 Limitation and Future Work

As mentioned previously, the model might not achieve a better performance on the newly-generated database with FaceNet due to the fact that the database is of small size and thus could be largely influenced by certain misleadingly labelled data. Even if the model was pretrained on ImageNet which means it could extract good high-level features from the dataset, such small size of data might still make it difficult for the model to reach a good generalization because the fully connected layer could possibly be influenced by the variance caused by the data split and certain mislabelled data. Regarding this, it is necessary to fine tune the last few layers from the pretrained model which we do not perform due to limited computation resources and time. Apart from that, it is also necessary to control the quality of the faces collected by FaceNet, which we did not include, so that the new data would be more uniform and achieve a more stable performance with different data split. Also, it is difficult to really recognise the expression of a person and the label might not represent the actual expression of a person so the model could be largely affected by certain strong emotions like "Surprise".
About the potential future work, self-supervised learning is considered to be a better solution for such small database where it could relieve the influence brought by the small size of database since the model would less likely to be affected much by small portion of mislabelled data and could focus on more the intrinsic features within the images. Besides, the quality control of face preprocessing should be evaluated with more robust metrics rather than some recognition threshold and such metrics could guarantee the stability of model performance on the recognised faces. Apart from introducing valid metrics to achieve good quality control in the generated face images, it is also worth trying other face processing methods such as face alignment to position the faces and guarantee their uniformity. In addition, the collected data related to facial expression recognition area in the future could include both labelled data and unlabelled data so that the model would not be largely affected by certain possibly mislabelled data.

References

1. D. A. AL CHANTI and A. Caplier, "Deep Learning for Spatio-Temporal Modeling of Dynamic Spontaneous Emotions," in IEEE Transactions on Affective Computing.
2. Li, Jiequan, and M. Oussalah. "Automatic face emotion recognition system." 2010 IEEE 9th International Conference on Cyberntic Intelligent Systems. IEEE, 2010.
3. Kahou, Samira Ebrahimi, et al. "Combining modality specific deep neural networks for emotion recognition in video." Proceedings of the 15th ACM on International conference on multimodal interaction. 2013.
4. Fan, Yin, et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks." Proceedings of the 18th ACM International Conference on Multimodal Interaction. 2016.
5. 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.
6. Singh Virk, Jitender, and Deepti R. Bathula. "Domain Specific, Semi-Supervised Transfer Learning for Medical Imaging." arXiv (2020): arXiv-2005.
7. J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.
8. Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
9. Wan, Weiguo, and Hyo Jong Lee. "FaceNet Based Face Sketch Recognition." 2017 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2017.
10. Korshunov, Pavel, and Sébastien Marcel. "Deepfakes: a new threat to face recognition? assessment and detection." arXiv preprint arXiv:1812.08685 (2018).
11. Ming, Zuheng, et al. "Simple triplet loss based on intra/inter-class metric learning for face verification." 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, 2017.
12. Harfiya, Latifa Nabila Wahyu Widodo, Agus & Wihandika, Randy. (2017). Offline signature verification based on pyramid histogram of oriented gradient features. 23-28. 10.1109/ICICOS.2017.8276332.
13. Pan, Sinno Jialin, and Qiang Yang. "A survey on transfer learning." IEEE Transactions on knowledge and data engineering 22.10 (2009): 1345-1359.