

Deep Facial Expression Recognition Using ResNet34

Manling Zhao^{1, a, *}, Fang Yu², and Yuwei Dai³

¹College of Computer, Mathematical, and Natural Sciences, University of Maryland, College Park 20740, USA.

²Hangzhou Foreign Language School, Hangzhou 310023, China.

³Oxford International College of Chengdu, Chengdu 610051, China.

^aCorresponding author Email: mzhao99@terpmail.umd.edu

Abstract

How are computers affecting our social lives in day-to-day activities is a heating topic recently. Specifically, individuals attach lots of attention on human-computer interactions which progressed by the aid of computer AI and robots. One of the most important aspects in this field is computers' ability for recognizing humans' facial expressions, and in this project, we are presenting the facial expression recognition of six basic human emotions: angry, fear, happy, neutral, sad, and surprise. We used a slightly modified version of the Kaggle Facial Expression and Recognition Challenge dataset, and ResNet34 as our model. By the aid of constant training and multiple transfer learning techniques, we successfully increased the accuracy from 64% to 70%.

Keywords

Facial Expression Recognition, Transfer Learning, Resnet, Confusion Matrix, Cleaning Dataset, Fine Tuning, Pre-diction.

1. INTRODUCTION

Facial emotions are one of the most important aspects of our daily communication and interaction [1]. It is a form of nonverbal communication expressed the same way by all people around the world with different culture. Throughout the years, experts and researchers in this area have been working hard to let computers have the ability understanding human emotions. This would allow one to have access to more opportunities and applications, such as improving the human-computer interaction experiences and building up much well-designed targeted advertisements. This is one of the reasons why we choose to focus on this area in this paper.

In 1971, the first paper on classifying human emotions using a facial action coding system (FACS) was published [2]. The system was built by Ekman and Friesen, which identified the following six emotions by interacting with the participants during the experiment: anger, fear, disgust, happiness, sadness and surprise [3]. These emotions have been seen as a person's basic facial expressions and are still being used by current researchers in this area. This paper is also focused on research based on these expressions, except for the disgust attribute. Instead, our dataset includes another attribute called neutral.

To achieve our goal of recognizing facial expressions by computers, this project will be using ResNet34, a convolutional deep learning network as our model. This is one of the state-of-the-art architectures, while some adjustments such as transfer learning and deep learning techniques will also be made in the work. The dataset we will leverage in our research is from Kaggle's Facial Expression Recognition Challenge. This dataset is fairly representative in terms of its size and the structure of the faces. However, since we want the dataset to be relatively

uniformly distributed, and disgust is being the only underrepresented one within the dataset, the disgust attribute is removed the rest of the data is used.

To evaluate the performance of the model, the result will be focusing on the accuracies on the training and validation sets as well as the top losses. Moreover, further testing on the model by letting it make predictions on several images outside of the dataset will be done to see its performance. Our goal for the best model to achieve is at least 70% accuracy as the current average performance on this dataset is around 70% to 75% [4].

2. RELATED WORK

2.1. Area of Focus

Unlike years ago when recognizing human expressions in the context of video footage was popular in this area, nowa- days a large portion of existing studies on facial expression recognition are based on static images, since they are easier to be processed and to get training and testing done [4]. Convolutional neural networks that have been pre-trained on the ImageNet dataset like VGG-16 and ResNet50 are useful among the current deep learning classifiers [5], which are also parts of the main models being used to classify facial expressions.

2.2. FER Competitions

There are two major facial expression recognition competitions: FER2013, which is the dataset used for this project, and Emotion Recognition in the Wild Challenge (EmotiW), which took place in 2016 [4]. The winner of EmotiW used CNN while the winner of FER13 used a deep neural network (DNN) to detect emotions and apply classifications [6].

3. METHOD

Resnet is a residual learning framework presented in 2015, of which the main purpose is to make the training of networks that are substantially deeper than what we previously used easier [7]. The framework was evaluated on the ImageNet dataset with a depth of up to 152 layers, which was 8 times deeper than VGG nets while still having lower complexity, and at the same time, it won the 1st place on the ILSVRC 2015 classification task with an error rate of 3.57% on the ImageNet test set [7]. Similar to VGG nets, it is one of the state-of-the-art architectures in image recognition field, while it has additional identity mapping capability compared to VGG. It also takes advantages in the training time compared to other CNNs and allows for deeper networks [6]. ResNet34 means that it is a 34-layer residual network, and there are other variants like ResNet50 and ResNet101 also. ResNet34 will be used in the implementation of this project.

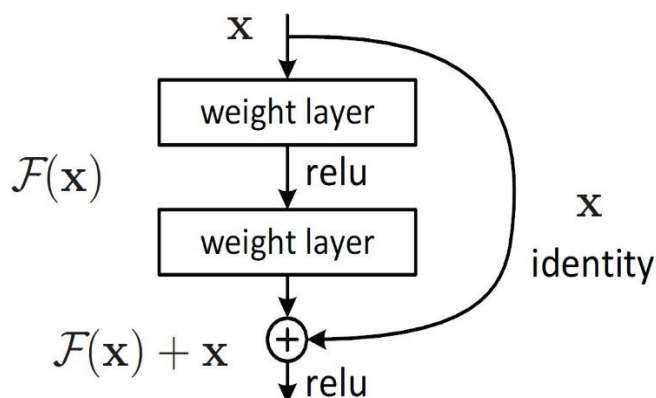


Fig 1. Residual learning: a building block [6]

4. DATASET AND FEATURES

4.1. Dataset Preparation

For our dataset, we aimed to choose the dataset that not only provide representative number of figures but also contain data with relatively even distribution across the emotions of human expressions included happy, angry, surprise, sad, neutral, and fear. Fortunately, the Kaggle dataset from the competition called Challenges in Representation Learning: Facial Expression Recognition Challenge (FER13) fits all the following attributes:

- (1) 35,352 figures
- (2) Figure Format: 7 x 6 units
- (3) Various individuals including different characteristics: race, age, gender and the pictures are taken at various angles.
- (4) Contain six fundamental expressions

4.2. Data Processing

A package called fast.ai is used in the work, which is built on top of PyTorch, and inside the package exists a function named ImageDataBunch, which reads the data from folders. Table 1 shows the number of images in each class for both training and validation sets, as well as their corresponding proportion among the whole dataset.

4.3. Changing Class Weight

As shown in Table 1, the training set of the dataset is imbalanced. The class happy clearly contains much more images than the other classes, so here a technique called class balancing is used, where we alter the weight of each training example carries to get the highest possible percentage accuracy. We put the weights to classes respectively to each class as follows: 10%, 30%, 5%, 15%, 20%, and 20%. This distribution is manually decided by us, which gives the best result after many experiments were conducted.



Fig 2. An example of part of the dataset, with their corresponding labels

Table 1. Number of images in the dataset by each classes with their corresponding proportion

Expression	Training Set	Validation Set	Corresponding %
Angry	3996	959	10%
Fear	4098	1025	30%
Happy	7216	1775	5%
Neutral	4966	1234	15%
Sad	4831	1248	20%
Surprise	3172	832	20%
Total	28279	7073	100%

4.4. Cleaning Dataset

One of the biggest challenges was cleaning the dataset. After plotting images that produced the top losses as shown in Figure 3, it is obvious that there were lots of labeling mistakes in the dataset (the expression label on the left is the predicted expression by the model, and the label on the right is the labeled emotion). Consequently, in order to decrease the error rate, we cleaned the dataset by:

- (1) Transferring the wrong labelled images to its corre- sponding folder.
- (2) Deleting some deviation figures.

Since there are a lot of images in the dataset, we were only able to clean up the data that causes top losses manually, while there still leaves some incorrectly labelled data. If the dataset could be further cleaned, there would be room for more improvement on the accuracy.

4.5. Fine Tuning

Finally, fine tuning is the key to the whole training process. Fine tuning goes through the neural network model that has been trained for a specified task and make the original weights to be re-tuned to suit the new dataset [8]. We unfrozen the trained model and fit it with four more epochs with learning rate from 1e-5 to 1e-4, getting a better performance result as shown in Table II.



Fig 3. Images that produces top losses. Ones enclosed with red squares are incorrectly labelled.

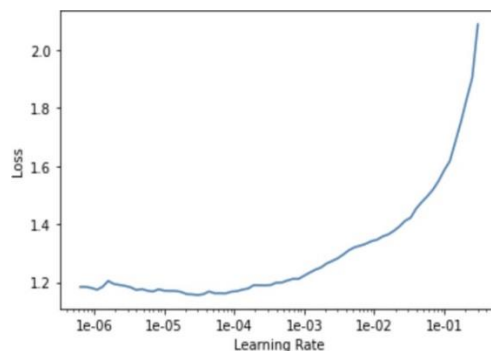


Fig 4. Learning rate of the model

5. RESULTS

To access the performance of the model, a matrix showing the accuracy the model gets before and after cleaning the dataset is constructed, as well as its performance before fine tuning and after, as shown in Table II. When the original dataset was used without cleaning out any noises, the best result the model could get even after fine tuning was an accuracy of around 63.3%. After manually cleaning the data, the error rate decreased by around 6%, with a current accuracy of 70%, meaning that the model correctly identified 70% of the facial images. There are still some incorrect and dirty data that were not able to be cleared out, and therefore there is still room for improvements.

The results also include interpretation of 20 images in the validation set that has top losses as shown in Figure 4. It shows the images that the model was confident about it getting wrong. The title of the image shows the predicted emotion, the actual emotion, loss, and probability of the expression to be the actual emotion [9].

Table 2. Dataset performance (error rate) of resnet34

	Test Normal	Fine Tuned
Original Data	40.62%	36.70%
After Cleaning	38.47%	30.03%



Fig 5. Predict/Actual/Loss/Probability interpretation

Figure 5 demonstrates the confusion matrix for the best expression on the dataset. In the confusion matrix, the line of diagonal for the confusion matrix indicates the number of images

that are correctly predicted and the other shown the wrongly predicted images. It reveals that sad, neutral, and fear tend to be miscategorized with each other. Qualitatively, the facial expressions of sadness and fear can be confusing by looking at the raw data as they share some commonalities, especially for the eyebrow and mouth part. On the other side, surprise and sadness don't share much commonalities, thus the recognition inaccuracy between the two is small.

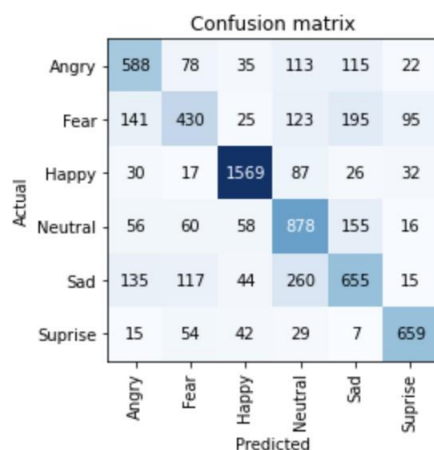


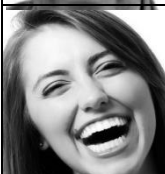




Fig 6. Confusion matrix with predicted emotion rows and actual emotion columns

Table 3. Making predictions on real world images [10]. the facial expression attribute indicates the expression of the face. The prediction class shows the probability of each class to be the predicted emotion. the model takes the class that has the highest probability to be the predicted result.

Testing Images	Facial Expression	Prediction Class					
		Angry	Fear	Happy	Neutral	Sad	Surprise
	Angry	0.1685	0.2264	0.0081	0.1147	0.0367	0.4456
	Fear	1.5451e-01	5.7619e-01	7.7906e-02	5.5693e-04	2.2645e-03	1.8857e-01
	Happy	8.0170e-03	3.8605e-03	9.7317e-01	4.3138e-04	3.7268e-04	1.4147e-02
	Sad	0.0357	0.4803	0.0194	0.2225	0.2288	0.0133
	Angry	0.4011	0.0538	0.0250	0.3534	0.1372	0.0295

Finally, we use the trained model to make prediction on images outside of the dataset. Table 3 shows the probabilities that the model predicted of each class for the corresponding images [10]. It predicts the expression by choosing the class that has the highest probability. It is clear that the model sometimes predicts incorrectly, which corresponds to the 70% accuracy of the trained model.

6. CONCLUSION

Our research explored the model ResNet34 for recognizing facial expressions. Multiple deep learning techniques such as fine tuning and changing class weights were introduced to improve the performance of the model. The results showed that we've achieved the goal of reducing the error rate down to 30% or less. For context, the winner of the Kaggle competition achieved an accuracy of 71.2% [6], and from a survey presented by IEEE this year, the performance summary on this dataset reveals that the average accuracy is around 70% to 75% [4]. So compared with others, our result is acceptable. Future work could be done on thoroughly clean the noises in the dataset to reduce the error rate.

ACKNOWLEDGMENTS

We would like to express our gratitude to the CIS program at Torhea Education Group for providing us the opportunity to do research on this topic. We would also like to thank Prof.

Pradeep Ravikumar and peers for reviewing our work and providing useful feedbacks.

REFERENCES

- [1] Frank, M. (2001) Facial expressions. In: N. J. Smelser and P. B. Baltes. (Eds.), International Encyclopedia of the Social Behavioral Sciences. Pergamon, Oxford. pp. 5230 – 5234.
- [2] Tettegah, S. Y., Gartmeier, M. (2015) Emotions, Technology, Design, and Learning. Academic Press. Cambridge.
- [3] Friesen, W.V., Ekman, P. (1983). EMFACS-7: Emotional Facial Action Coding System.
- [4] Li, S., Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, pp. 1-1.
- [5] Rosebrock, A. (2017). ImageNet: VGGNet, ResNet, Inception, and Xception with Keras. <https://www.pyimagesearch.com/2017/03/20/imagenet-vggnet-resnet-inception-xception-keras/>.
- [6] Savoiu A., Wong J. (2017). Recognizing Facial Expressions Using Deep Learning. <http://cs231n.stanford.edu/reports/2017/pdfs/224.pdf>
- [7] He K., Zhang X., Ren S., Sun J. (2016). Deep Residual Learning for Image Recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas. pp. 770–778.
- [8] Kandel I. and Castelli M. (2020). How Deeply to Fine-tune a Convolutional Neural Network: A Case Study Using a Histopathology Dataset. Applied Sciences, 10: 3359.
- [9] fast.ai. (2020). <https://docs.fast.ai/vision.learner.html>
- [10] Sharma G. (2018). Real Time Facial Expression Recognition. <https://medium.com/datadriveninvestor/real-time-facial-expression-recognition-f860dacfeb6a>