Facial Emotion Recognition with ResNet Transfer Learning

Daniel Farahani[1]

Research School of Computer Science, Australian National University
u5800723@anu.edu.au

Abstract. Facial recognition has advanced with the development in learning models which allow for a more accurate and mobile applicability. We compare a basic newly created Convolutions Neural Network to a fine tuned pre-trained model. We apply these models to the Static Facial Emotion in the Wild data set which emulates real world conditions - acting as a baseline for the field. We find pre-trained models provide a much better result in comparison to newly created model but also traditional models used on the data set.

1 Introduction

Facial Emotion Recognition (FER) models consist of video and image analysis. Although video has been proven [1] to be better for facial expression analysis, it is not always feasible given the available data. Image based analysis itself also has many applications regarding medical analysis for psychology to smile detection in consumer products [2]. In the past, image based FER methods mapped pixel spaces, performed mathematical transforms and optimised based on feature extraction. However, recent models are able to capture more facial emotion nuances by using the raw image and large amounts of data to train on. Modern day training based methods can identify patterns from large sets of data and achieve a higher performance and often these trained models perform better when faced with realistic and more challenging data sets.

The Static Facial Expression in the Wild (SFEW) data set is a robust collection of images consisting of emotions portrayed by actors. The movie scenes where these actors are placed in, mimics the real world conditions much better than laboratory constructed image sets. As a result the performance of computer vision methods decreases due to small changes in shading, rotations and turning, subject variance and occlusion [1]. The SFEW data set contains 675 images categorised in anger, disgust, fear, happy, neutral, sad, or surprised emotions. SFEW data set has a baseline accuracy score set by LPQ [8] and PHOG [9] feature modelling which we will compare our results to. Give the data consists of raw screen shot of movie scenes, we will pre-processing through cropping, filtering, resizing and normalising before training/ fine tuning our networks.

Although, the SFEW data set is very robust it is quite small for training an entire network on. Models such as ResNet18 [5] are a pre-trained CNN with

16-30 hidden layers that have been trained on databases like ImageNet which consists of 1.2 Million data points for over 1000 categories. Although ImageNet consists of nature related object [3], it has many elements that that are required for image classification. Although our data is different, we have quite a small sample and can still leverage the ResNet architecture by fine-tuning the network to achieve a better result
. We will train a simple CNN on the SFEW data as a base model and compare to the fine tuned ResNet18.

2 Method

The data set is categorised in to anger, disgust, fear, happy, neutral, sad, or surprised emotions, where the raw screen shot of the movie is given. To input these into these into our models we crop the faces from the images, resize them, normalise the values and convert them into Tensors. We then split the data 80-20 and kept consistent through out when training and testing our models.

For face extraction we use the Viola-Jones [4] algorithm giving us coordinates of a box surrounding the face. Upon inspection of the first iteration it appeared that some of the cropped images were too small, leaving out the chin and the forehead. Additionally, some were false positives where the cropped areas were not a faces but the algorithm thought otherwise. To adjust for the first issue 20 pixels were added to the boxes boundary which captured most of the head for all images. The second issue appeared to be consistent with additional face detection attempts, proving an algorithm downfall. The false positives were manually removed from each emotion category (both train and test), Table 1 shows these values per category where there are a few rare cases the face was not detected at all. Figure 1 gives a sample of the falsely detected faces.

Emotion	Raw	Train	FP	Tot	Test	FP	Tot	Total
Angry	100	70	9	79	19	1	20	99
Digust	75	57	2	59	15	1	16	75
Fear	100	72	7	79	19	2	21	100
Happy	100	65	15	80	18	1	19	99
Neutral	100	78	1	79	20	1	21	100
Sad	100	72	7	79	20	0	20	99
Surprise	100	75	4	79	20	1	21	100

Table 1. Comparison of NN model to SFEW baseline score

We will present results in the next section for the base model, with and without the false positives to get a sense of how much of an effect it would have. Helping us gauge the effect of any other noise such as blurriness, small crops, questionable emotion categorisation etc. in the data set.

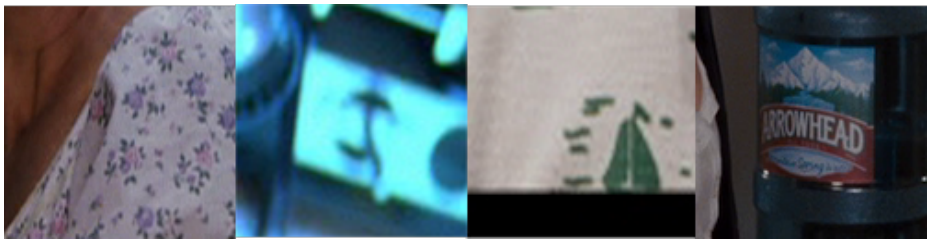

Fig. 1. Example of false positive face detection with the Viola-Jones algorithm

The images are then resized to 120 x 120 as it is around the median of all the images, which avoids shrinking or expanding the other images by too much. They were also kept in their original RGB colours as required for the ResNet. However each channel was normalised using the Mean and and STD of the channels.

The base model is organised as two convolutions layers, a dropout layer and two fully connected layers leading to the final classification demonstrated in Figure 2. ReLU activation function was used in the hidden layers, with a Negative Log Likelihood loss function as it is a classification model, 5x5 kernel at the convolution layers, learning rate of 0.1 and momentum of 0.5. Max-pooling is applied at the first two layers as its help with noise reduction, given blur was quite common in the images during inspecting. Then two dropout operation on the second convolution layer and first fully connected layer to prevent over fitting.

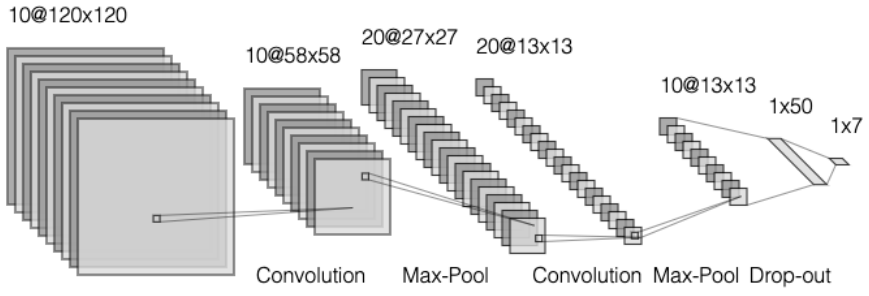

Fig. 2. Base model CNN architecture

The ResNet is loaded as pre-trained in the experiment and fine tuned with our data. Layers in deep image analysis CNNs extract different features out of the image for example edges, shapes, object, etc. which we can leverage for our data set. To further improve the performance we fine tune the model. This allows

us to adjust the weights by performing back propagation to adapt it to out data set. This is particularly important in this instance as our data set consists of human faces but, the data set ResNet was trained on consisted of 1000 objects relating to nature objects [6]. ResNet uses normalisation after each convolution layer, learning rate of 0.1 and momentum of 0.9, however it does not have any dropout layers[5].

We perform multiple trials and report the median testing score when the model is considered to have stabilised in the last few epochs. We compare the results to the SFEW benchmark set by LPQ and PHOG features.

3 Results and Discussion

The base model was not able to perform very well even with noise manually removed. However, there is a slight difference showing there may be an effect given the sample size. Conversely, when fine tuning the ResNet we are able to achieve results that are comparable to the base line set for the SFEW data. Comparing the results to the 7 emotion classification by LPQ and PHOQ features in Table 2 we see the ResNet with small amount of effort performs very well.

Model	Train Acc %	Test Acc %
Base CNN w/ FP	19.2	15.8
Base CNN w/o FP	15.4	14.1
ResNet	98.3	51.5
PHOG	-	43.7
LPQ	-	46.3

Table 2. SFEW classification accuracy of 7 emotions for different models

We see the ResNet was able to outperform the traditional LPQ and PHOG feature modelling comfortably, as the model was not modified too much. Although, we see a very large discrepancy between the training and testing accuracy. This demonstrates and over fitting is occurring in the training phase. This is possibly due to fine tuning all the layers on the ResNet, alternative approach is to freeze some of the earlier layers and only back propagate to a certain level.

Given a better performance of the model over the SFEW baseline, it is still quite low. This is possibly due to face extraction, noise and data labeling. The falsely detected faces are forgone faces in the image, where there was a face present but the algorithm did not detect it. Additionally, there were cases where the faces of the person in the back ground was detected instead of the subject person. This made for noisy images where the images were very blurry. Finally, during manual inspection it was evident that the same images was classified with two different emotions or the classified emotion was arguable from a humans perspective. These factors if addressed may improve the accuracy of the classification.

4 Conclusion and Future Work

Development in CNNs have greatly advanced Facial Emotion Recognition as we found from experimentation. A simple CNN model was able to classify 7 emotions categories with 15% accuracy, but with fine tuning a pre-trained model the emotions were classified at an 51% accuracy. That was also higher than the ≈45% accuracy achieved by traditional methods such as LPQ and PHOG used on the SFEW. This demonstrates the amount performance gain possible by leveraging the mobility of CNN models to other data sets.

Although, the performance beats the SFEW base line is it not accurate enough to use for real world applications. To do so more effort allocated to customising the ResNet model and better processing of the data set for future work would provide substantial improvements.

References

1. Z. Ambadar, J. Schooler, and J. Cohn. Deciphering the enigmatic face: The importance of facial dynamics to in- terpreting subtle facial expressions. Psychological Science, 16(5):403–410, 2005.
2. Derntl, B., Seidel, E.M., Kryspin-Exner, I., Hasmann, A. and Dobmeier, M., 2009. Facial emotion recognition in patients with bipolar I and bipolar II disorder. British Journal of Clinical Psychology, 48(4), pp.363-375.
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
4. Jones, M. and Viola, P., 2003. Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96, 3(14), p.2.
5. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
6. Ng, H.W., Nguyen, V.D., Vonikakis, V. and Winkler, S., 2015, November. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction (pp. 443-449).
7. Reyes, A.K., Caicedo, J.C. and Camargo, J.E., 2015. Fine-tuning Deep Convolutional Networks for Plant Recognition. CLEF (Working Notes), 1391, pp.467-475.
8. Rahtu E., Heikkilä J., Ojansivu V., Ahonen A.: Local phase quantization for blur-insensitive image analysis, Image and Vision Computing, Volume 30, Issue 8, 2012, Pages 501-512, ISSN 0262-8856.
9. Bai, Y., Guo, L., Jin, L. and Huang, Q., 2009, November. A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 3305-3308). IEEE.