Facial Expression Recognition using Network Ensembles

Anirudh Bhupalarao

Birla Institute of Technology and Science, Hyderabad

1. **Abstract**

Facial Expression Recognition (FER) is a growing field of research and has been recognized as one of the most important topics for human-computer interaction. Hence, there has been active research on expression analysis utilizing CNNs and other deep learning techniques. All these differ significantly in terms of CNN architecture and methods of feature extraction. This paper focuses on employing preprocessing techniques like facial landmarking and alignment, along with building an ensemble of various CNN models to better the results of previous novel models. With the help of modern deep CNNs, I obtained a test accuracy of around 82% on the FER 2013 dataset, outperforming previous works by almost 10-12% without requiring any auxiliary training data.

1. **Introduction**

In the vast field of Artificial Intelligence, recognizing emotions on humans' faces is one of the most sought-after goals. Doing this accurately strengthens the ties between AI and humans and only brings the two closer. Knowing what expression/emotion a person is displaying can tell us so much during human interaction and opens a whole new world that can be reached by AI solutions. On this account, there has been an emphatic rise in the study of Facial Expressions and how to detect them. Computer Vision is the forerunner in this field and has enabled many novel architectures to come into the picture.



When we talk about basic expressions, we are generally talking about seven primary emotions: Anger, Neutral, Sadness, Surprise, Fear, Happiness and Disgust. These are the fundamental emotions of a human being and studies show that they can be detected by learning about Action Units (AUs) and how they trigger these basic expressions in the face. These action units further give rise to 68 points on the face called Facial Landmarks. This computer vision technique deals with detecting various features of the face (eyes, mouth, and nose). While there are many other flavors of facial landmarks detection, some of which use a 194-point model trained on the HELEN dataset, we use a 68-point model trained on the iBUG 300-W dataset. Despite all this, it is not easy to detect expressions under naturalistic circumstances. That is attributed to head posture, lighting differences, occlusions, and the fact that unposed expressions are always subtle.

Deep Convolutional Neural Networks (DCNNs) make up for this inaccuracy. CNNs have made substantial advances in efficiency and associated tasks and some recent studies on FER have successfully used DCNNs for feature extraction and analysis. Factors that impact performance - model architecture, preprocessing, training and test procedures, vary considerably in all these previous works.

1. **Literary Survey**

For CNN-based FER, we analyze various state-of-the-art approaches, highlight methodological discrepancies, and examine the performances recorded. Several of these approaches were tested on many datasets, the most common dataset being FER-2013 [4]. FER-2013 is a publicly available dataset containing 35,887 different face cropped images. It is split into 28,709 training images and 7,178 test images. They are segmented into different sections based on the expression each image shows. All these images are taken “in the wild” so, preprocessing is necessary most of the time. They are all grayscale images of size 48 by 48.

Mollahosseini et al. [1] trained a CNN based on the Inception architecture [2] on compiled data from various naturalistic datasets to obtain a generalized model. The EmotiW2015 winners employ a large group of CNNs [3]. To achieve more diverse models, specific properties like input preprocessing of the individual networks differ. Kim et al. [5] used both registered and unregistered versions of each face image during the training and test phases and showed that it was beneficial. There was a small increase in performance (0.5%) after utilizing pose information taken from the deep networks used for registration. Shan Li and Weihong Deng [6] also provide an extensive survey on various FER papers and highlight the different datasets used, preprocessing techniques, and deep learning models and architectures. They review existing novel neural network models and weigh the advantages and limitations for both static and dynamic FER datasets. There are many such previous and ongoing works on FER and many competitions to go along with them to figure out the best possible models for this. We consider all these works and try to improve the performance of our model.

1. **Methodology**

We break down the method used into three parts: (i) Data preprocessing, (ii) CNN architecture and (iii) CNN training.

Data preprocessing: Preprocessing data when it comes to facial expression recognition is primarily face detection and registration, facial alignment and facial landmarking. In unconstrained scenarios, patterns that are unrelated to facial expressions, such as surroundings, illuminations, and head poses, are reasonably common. Therefore, preprocessing is necessary to align and normalize semantic information transmitted by the face before training the deep neural network to learn meaningful features.

One of the main preprocessing techniques used for expression recognition is Facial Landmarking. They are used to localize and represent salient features of the face (nose, mouth, eyebrows, eyes, and outline of face). Landmarking helps in facial alignment which we have used later as another preprocessing technique. Detecting facial landmarks consists of localizing the face and detecting the key facial features in the ROI (Region of Interest). A training set of labeled facial landmarks of images is used. These images are manually labeled, specifying (x, y) coordinates to each facial landmark on the face. Along with this, we also calculate the probability of distances between pairs of input pixels. Given this training data, an ensemble of regression trees is trained to estimate the landmark positions from the pixel intensities. There are 68 such coordinates that map to features on the face.

A picture containing light, decorated, close, day

Description automatically generated

Facial alignment is also a commonly used preprocessing step in FER related work. It can be used as a form of data normalization. This is done with the help of facial landmarks where the angle of the line connecting the landmarks of the eyes is set to zero i.e., horizontal.

A picture containing text

Description automatically generatedA picture containing text, head covering

Description automatically generated

CNN architecture: We mainly work with two novel models – Inception v1 and VGG. Firstly, the Inception model is a convolutional neural network which is 27 layers deep. This model contains a layer called the inception layer, which is a combination of 1x1, 3x3, 5x5 convolution layers with their outputs concatenated into a single input for the next hidden layer. It helps the internal layers choose which filter size is important for the information needed to be learned. So, even though the ROI (face) is different, the layers function to recognize the face accurately. As we are dealing with grayscale images of size 48x48, we removed the 5a and 5b layers to prevent feature loss.

Table

Description automatically generated

The second model we are going to use is based on the VGG model. The VGG-Face model has 22 layers and 37 deep units. We decrease a few layers from this for the same reason as before i.e., to avoid feature loss, however, the basic structure of the Deep CNN remains the same.

Table

Description automatically generated

CNN training: We use a network ensemble of both the models presented in the previous section. Research has shown that assemblies of multiple networks can outperform individual networks. For feature level ensembles, concatenated input features learned from different networks is the most common strategy. In our final model, we use a network ensemble to weight and average the outputs of both the CNN models to get a better result. The image shown in the next page is an accurate representation of how modern network ensembles work and are mainly used along with Deep CNNs to outperform novel individual models.

Diagram, schematic

Description automatically generated

The models are trained for around 25-30 epochs each based and hyperparameter tuning is done to ensure that there is no overfitting. The initial learning rate, batch size and decay are fixed at 0.0005, 256 and 0.96, respectively. An optimal dropout rate and softmax activation are used for both the models while training.

1. **Discussions and Conclusions**

Over several runs of the final model, the average training accuracy was around 80-82% with the preprocessing techniques and around 67-68% without preprocessing. This outperforms many state-of-the-art models by almost 10-12% under similar conditions. The overall loss was dropping to around 0.499 – 0.6. A batch size of 256 for a dataset this large seemed to be the sweet spot. Below are the graphs of the results obtained. The training time on a CPU averaged about 15-20 mins simply because of the number of parameters being trained.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Without using additional training data, face registration or extensive augmented data, this ensemble network has obtained state-of-the-art results. This FER approach is conceptually easier than previous approaches by not requiring face registration. However, in further works, utilizing facial registration can significantly better the results. Nevertheless, findings obtained on this and other FER datasets due to dataset bias are only representative of FER in the real world. This restriction applies not only to this study but to FER research in general. With an emphasis on FER data augmentation, I intend to explore ways to solve various bottlenecks and further use this FER project’s results to achieve something bigger. With such a vast area for development, facial expression recognition can lead to incredible areas of research and broaden the reach of Artificial Intelligence.

1. **References**

[1] A. Mollahosseini, D. Chan, M. H. Mahoor, “Going Deeper in Facial Expression Recognition using Deep Neural Networks”, CoRR, 2015.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going Deeper with Convolutions”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] B.K. Kim, J. Roh, S.Y. Dong, S.Y. Lee, “Hierarchical committee of deep convolutional neural networks for robust facial expression recognition”, Journal on Multimodal User Interfaces, 2016

[4] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, Y. Bengio, “Challenges in representation learning: A report on three machine learning contests”, Neural Networks, 2015.

[5] B.K. Kim, S.Y. Dong, J. Roh, G. Kim, S.Y. Lee, “Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach”, IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, 2016.

[6] S. Li, W. Deng, “Deep Facial Expression Recognition: A Survey”, IEEE Transactions of Affective Computing, 2020.

[7] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, ICLR, 2015.