FACIAL EXPRESSION RECOGNITION

Akash Mudhol
18HS20006

July 19, 2020

1 Introduction

Facial expressions are one of the most important features to reflect the human emotional state and they convey **55% of a communicated message** which is more than the part conveyed by the combination of voice and language. FER technique can be used for the development of **human–computer interaction systems**, **such as social robots, visual-interactive games, and data-driven animation.** Automatic facial expression analysis can be applied in many areas such as **emotion and paralinguistic communication, clinical psychology, psychiatry, neurology, pain assessment, lie detection, intelligent environments, and multimodal human computer interface (HCI).**

The goal of the project is to detect seven human emotions using facial expressions using Convolutional Neural Network. The datasets that we have used for this purpose are CK+ and JAFFE. The seven expressions are anger, disgust, fear, happy, sadness, surprise, neutral (in case of JAFFE) and contempt (in cased of CK+). We were able to produce an accuracy of 93.46% and 90.54% on JAFFE and CK+ dataset respectively for seven classes and an accuracy of 92.34% and 92.24% on the above mentioned datasets for six classes.

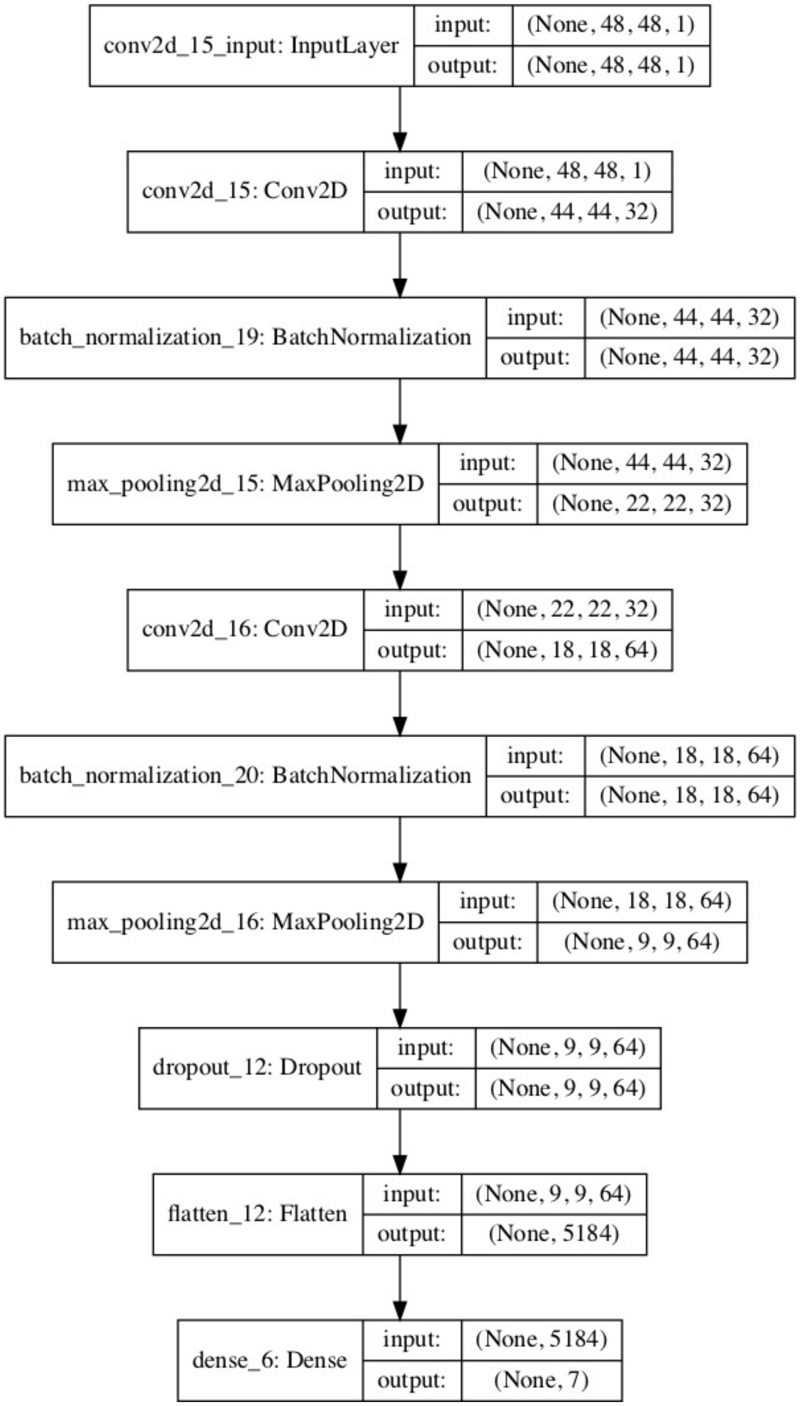
Further we have used the combined datasets to train a model for detecting single and multiple faces and their six different expressions in a realtime environment.

2 Research Paper

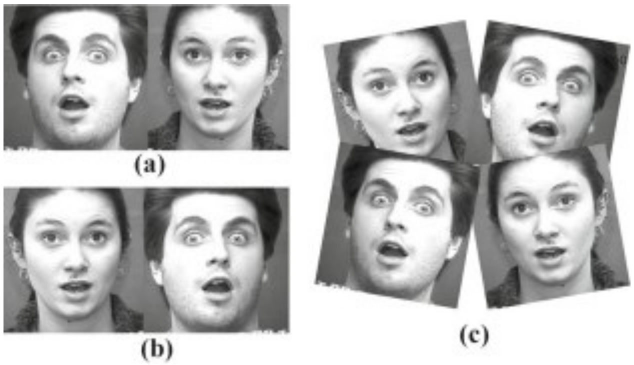
2.1 Overview

We used a **new face cropping and image rotation strategy** to improve the accuracy and simplify the CNN structure. By facial cropping we removed the emotionally inactive part of the region and **random rotations to cope up with data scarcity.** Also **Histogram equalization, Z-score normalization, and down-sampling** were applied to standardize the image data.The expanded training data thus obtained is used to train the CNN, and we got our best CNN model by **ten-fold cross validation.** During the validation or testing phase, the normalized testing images (without expansion) were sent to the CNN model from the training phase for prediction.

2.2 Model Framework


conv2d_15_input: InputLayer
input: (None, 48, 48, 1)
output: (None, 48, 48, 1)
conv2d_15: Conv2D
input: (None, 48, 48, 1)
output: (None, 44, 44, 32)
batch_normalization_19: BatchNormalization
input: (None, 44, 44, 32)
output: (None, 44, 44, 32)
max_pooling2d_15: MaxPooling2D
input: (None, 44, 44, 32)
output: (None, 22, 22, 32)
conv2d_16: Conv2D
input: (None, 22, 22, 32)
output: (None, 18, 18, 64)
batch_normalization_20: BatchNormalization
input: (None, 18, 18, 64)
output: (None, 18, 18, 64)
max_pooling2d_16: MaxPooling2D
input: (None, 18, 18, 64)
output: (None, 9, 9, 64)
dropout_12: Dropout
input: (None, 9, 9, 64)
output: (None, 9, 9, 64)
flatten_12: Flatten
input: (None, 9, 9, 64)
output: (None, 5184)
dense_6: Dense
input: (None, 5184)
output: (None, 7)

2.3 Preprocessing



2.3.1 Face Alignment

For this purpose mlxtend library is used which returns the facial landmark points by detecting the face in the image. It returns the required 68 points i.e 68 (x,y) coordinates which map to facial attributes like eyebrows, eyes, nose, mouth and jaw. The landmark coordinates for both eyes were used for rotating image to align the face horizontally for accurate cropping. The images were rotated by calculating the appropriate angle using the mean of left and right eyes whose coordinates were obtained after landmark detection. The formula that we used is mentioned below :

$$angle = \tan^{-1} \frac{\sum_{n=42}^{n=47} y_n - \sum_{n=36}^{n=41} y_n}{\sum_{n=42}^{n=47} x_n - \sum_{n=36}^{n=41} x_n} \qquad (1)$$

2.3.2 Cropping Image

WHY: The images obtained after face alignment contains of background which is of not required for detecting human expressions. Images were cropped to crop out the irrelevant parts of the face or the background and only have the closeup of face in image which is the area of interest.

2.3.3 Histogram Equalization:

Histogram Equalization was performed using the OpenCV library on the cropped and rotated images to equalize the intensities. This technique improves the contrast in images such that the resultant image obtained contains uniform distribution of intensities.

2.3.4 Z- score Normalization:

All the images were normalized with z-score method to get the data distribution centered at zero. The formula for the same is:

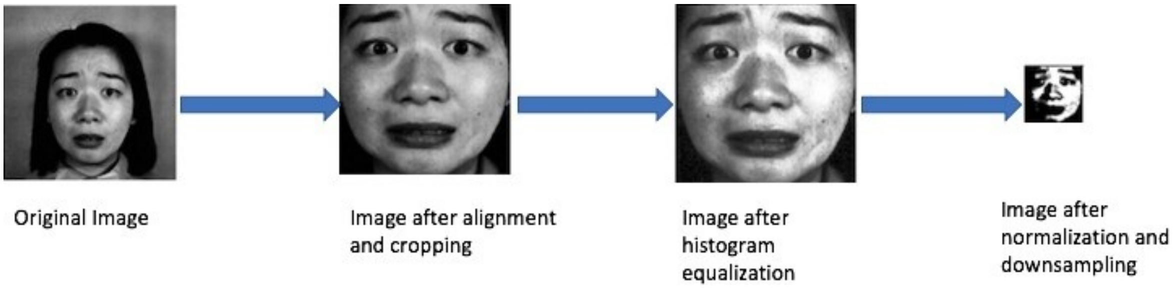
$$normalized\ value = \frac{value - mean}{standard\ deviation} \qquad (2)$$

2.3.5 Downsampling:

The images were downsampled to 48x48 pixels.

2.3.6 Data Augmentation:

The images in the training set are randomly flipped and rotated to increase the data set inorder to avoid overfitting on data. The rotation angle for the augmented images is 3°.


Original Image
Image after alignment and cropping
Image after histogram equalization
Image after normalization and downsampling

2.4 Architecture of the Model

Convolutional Neural Network is applied to extract the relevant features and detect human expressions. The architecture consists of two convolutional layers with ReLU activation and two max pooling layers. The first and third convolutional layers have 32 and 64 kernels having the size of 5 X 5. The second and fourth layers are the max pooling layers having kernel size of 2 X 2 and stide of 2. The output is then flattened into a 5184-dimensional vector and connected to an output layer which uses the softmax activation function.

2.4.1 Numerical Hyperparameters

Batch size is 16, learning rate (eta) is 0.001 and no. of epochs are 120.

2.4.2 Categorical Hyperparameters

We have used momentum optimiser, categorical cross entropy function and accuracy as metrics. We have used ten cross validation method for evaluation of our model.

2.5 Results

We achieved an accuracy of 93.46% on JAFFE (7 classes), 92.34% on JAFFE (6 classes), 90.54% on CK+ (7 classes), 92.24% on CK+ (6 classes) and 90.45% on JAFFE and CK+ combined (6 classes). The weights trained by the above model were then applied on live feed and it was able to detect 5 out of 6 emotions correctly. It even detected 2 different emotions for 2 different people present in the same frame and works for multiple faces in the image as well.

3 Datasets used

3.1 JAFFE

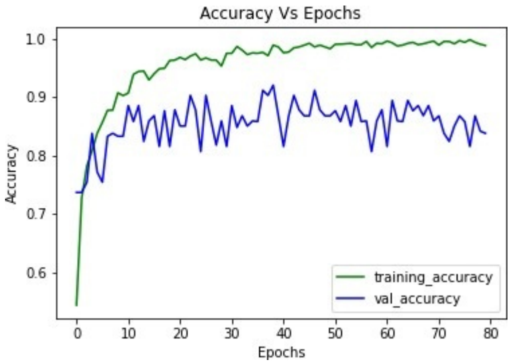
The database contains 213 images of 7 facial expressions (6 basic facial expressions + 1 neutral) posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects. The database was planned and assembled by Michael Lyons, Miyuki Kamachi, and Jiro Gyoba.

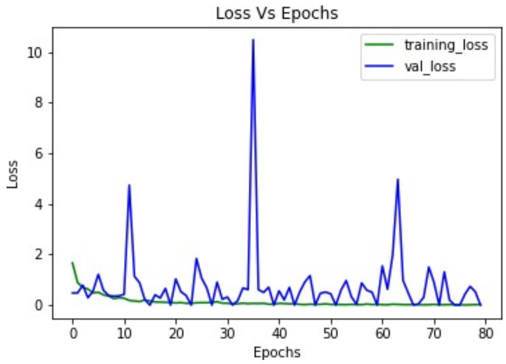
3.2 Extended Cohn-Kanade Dataset (CK+)

The dataset consists of a total of 327 images. The numbers of anger, contempt, disgust, fear, happiness, sadness, and surprise expressions are 45, 18, 59, 25, 69, 28, and 83, respectively. For 6 class experiment we have not considered the contempt expression and trained the model on remaining 309 images.

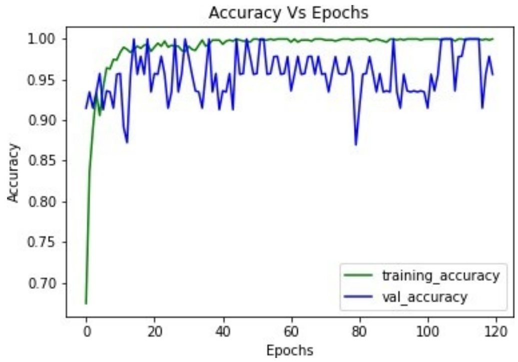
4 Relevant Graphs - 10 fold cross validation

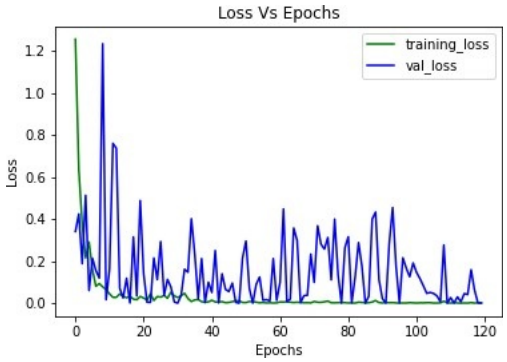
4.1 Combined JAFFE and CK+ (6 classes)



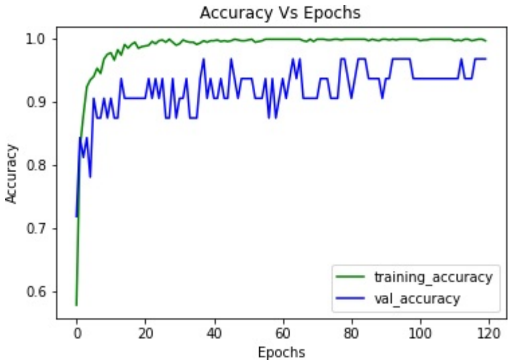


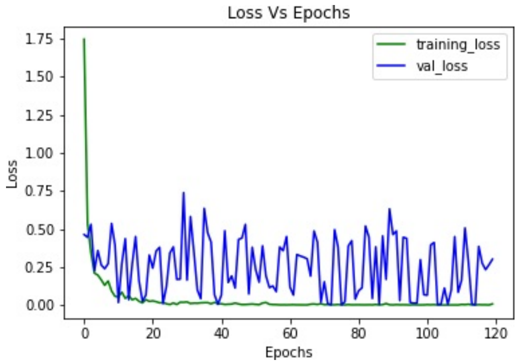
4.2 CK+ (6 classes)



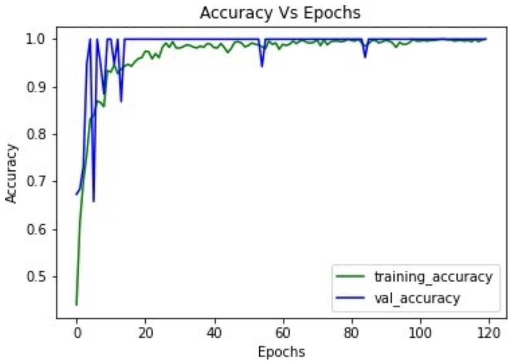


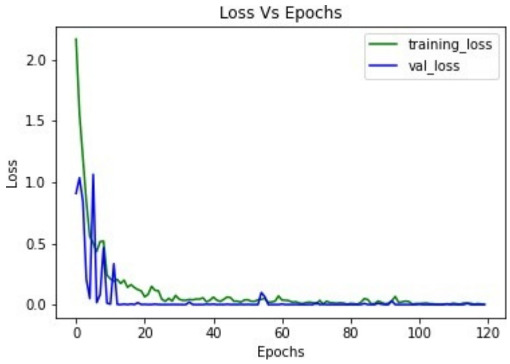
4.3 CK+ (7 classes)



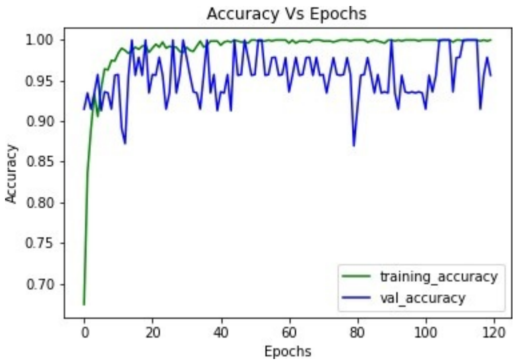


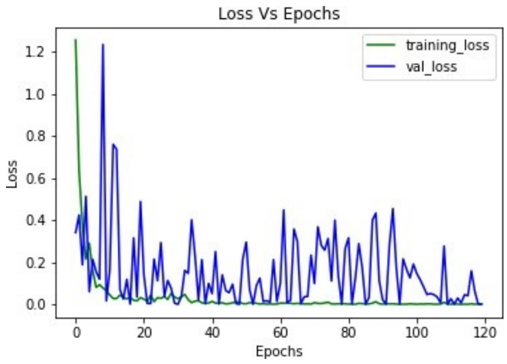
4.4 JAFFE (6 classes)



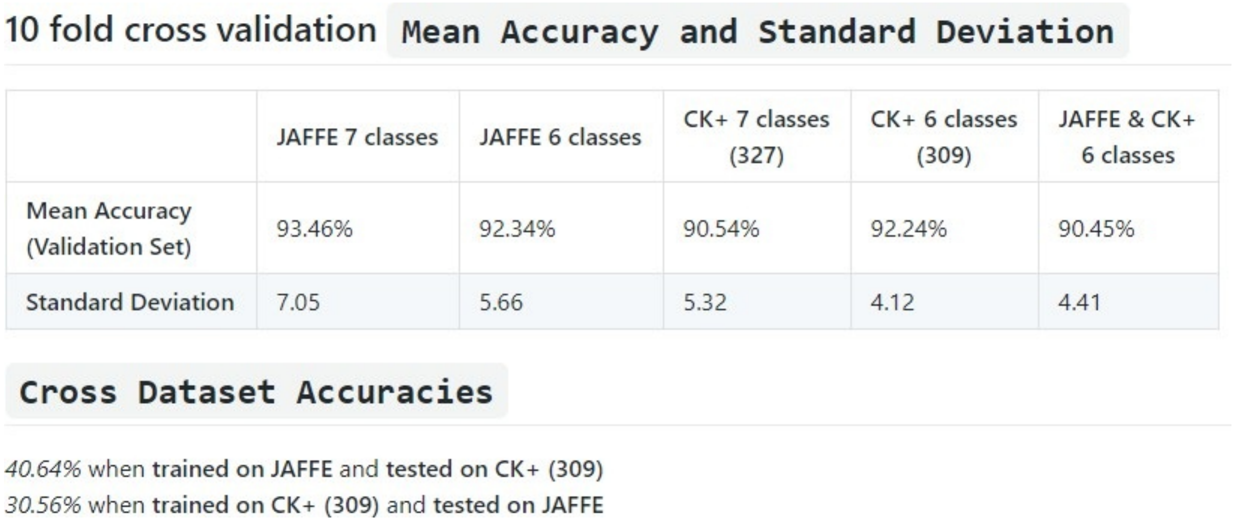


4.5 JAFFE (7 classes)





5 Accuracy and Standard Deviation Table



10 fold cross validation `Mean Accuracy and Standard Deviation`

	JAFFE 7 classes	JAFFE 6 classes	CK+ 7 classes (327)	CK+ 6 classes (309)	JAFFE & CK+ 6 classes
Mean Accuracy (Validation Set)	93.46%	92.34%	90.54%	92.24%	90.45%
Standard Deviation	7.05	5.66	5.32	4.12	4.41

`Cross Dataset Accuracies`

40.64% when **trained on JAFFE** and **tested on CK+ (309)**
30.56% when **trained on CK+ (309)** and **tested on JAFFE**

References

[1] Li, Kuan & Jin, Yi & Akram, Muhammad & Han, Ruize & Chen, Jiongwei. (2019). Facial expression recognition with convolutional neural networks via a new face cropping and rotation strategy. The Visual Computer. 10.1007/s00371-019-01627-4.

[2] JAFFE dataset : https://zenodo.org/record/3451524#.Xr0QGxMzbVo
CK+ dataset : http://www.consortium.ri.cmu.edu/ckagree/