Facial Expression Recognition through Machine Learning

Nikhil Jawade
Queen Mary University of London,
School of Electronic Engineering and Computer Science
n.jawade@se21.qmul.ac.uk

1. Introduction

This study is an extension of the literature review and data preparation and pre-processing that was performed for the Facial Expression Recognition (FER) domain. In the previous paper, we looked at the history as well as recent progressions in FER. As a part of the task, a subset of the Aff-Wild2 dataset was received to perform required data processing techniques to avail it for training the machine learning model. In this paper, we propose two machine learning models, Main and Baseline, for the purpose of multi-class classification of six basic human expressions. The results obtained from these models are discussed and critically analyzed, and alternate approaches are suggested. Later in the study, we pick the most effective of the two models and run an ablation study by adjusting three hyperparameters. To extensively test the capabilities of the machine learning techniques used, another subset of the Aff-Wild2 dataset was provided as Production Data. The outcomes of this dataset are presented, and approaches for further improvement are suggested. The Machine Learning model which showed the optimal results is made available here.

2. Task 1: Main Machine Learning Model

Deep Convolutional Neural Networks (CNNs) demonstrated great success in previous Facial Expression Recognition (FER) studies. One particular net that came into everyone's recognition, submitted for ImageNet Challenge 2014 by Visual Geometry Group of the University of Oxford [2], is VGG16. In the face recognition study [1], VGG16 was utilized over a large-scale dataset VGGFace of 2.6M images containing 2.6K subjects (celebrities), and later the model was stored with its weights. This makes the VGG16 model with weights initialized from the VGGFace dataset a perfect choice for the main model for recognizing six basic facial expressions. Table 1 shows the architecture and layer configuration of the main model.

A quick recap on the data at hand (received in the previous assignment), data preparation and pre-processing performed over it: Provided data is a subset of the Aff-Wild2 in-the-wild dataset containing 43,435 images of dimensions 128x128x3 labelled with six basic facial expressions. This

Table 1. Main Model VGGFace-VGG16: Architecture and Layer Configuration

LAYER (type)	OUTPUT SHAPE	PARAM #
13 - Convolutional Layers		
input (InputLayer)	[(None, 96, 96, 3)]	0
conv1_1 (Conv2D)	(None, 96, 96, 64)	1792
conv1_2 (Conv2D)	(None, 96, 96, 64)	36928
pool1 (MaxPooling2D)	(None, 48, 48, 64)	0
conv2_1 (Conv2D)	(None, 48, 48, 128)	73856
conv2_2 (Conv2D)	(None, 48, 48, 128)	147584
pool2 (MaxPooling2D)	(None, 24, 24, 128)	0
conv3_1 (Conv2D)	(None, 24, 24, 256)	295168
conv3_2 (Conv2D)	(None, 24, 24, 256)	590080
conv3_3 (Conv2D)	(None, 24, 24, 256)	590080
pool3 (MaxPooling2D)	(None, 12, 12, 256)	0
conv4_1 (Conv2D)	(None, 12, 12, 512)	1180160
conv4_2 (Conv2D)	(None, 12, 12, 512)	2359808
conv4_3 (Conv2D)	(None, 12, 12, 512)	2359808
pool4 (MaxPooling2D)	(None, 6, 6, 512)	0
conv5_1 (Conv2D)	(None, 6, 6, 512)	2359808
conv5_2 (Conv2D)	(None, 6, 6, 512)	2359808
conv5_3 (Conv2D)	(None, 6, 6, 512)	2359808
pool5 (MaxPooling2D)	(None, 3, 3, 512)	0
3 - FC Layers		
flatten (Flatten)	(None, 4608)	0
dense_1 (Dense)	(None, 4096)	18878464
dense_2 (Dense)	(None, 4096)	16781312
dense_3 (Dense)	(None, 6)	24582

Table 2. Baseline Model: Architecture and Layer Configuration (Only Classifier)

LAYER (type)	OUTPUT SHAPE	PARAM #
13 - Convolutional Layers		
1 - FC Softmax Layer		
flatten (Flatten)	(None, 4608)	0
dense (Dense)	(None, 6)	27654

data was divided into three sets: Train, Test and Validation at a split of 60-20-20 percentage containing 26,061, 8,704 and 8,670 respectively. As part of pre-processing, the image size was reduced to 96x96x3, and pixel values were normalized in the range [0,1]. Further, data augmentation techniques are included in the sequential execution of the model. The techniques performed over random samples are horizontal flip, rotation by 20°and zoom by 10%.

Cross-Entropy is best suited for multi-class classification problems, and as label encoding was performed in pre-processing, Sparse Categorical Cross-Entropy is selected as the loss function for training. For the best results, the Adam optimizer is selected with a learning rate of 0.0001 and the model is trained over 20 epochs. The batch size is set to 256. Selection of these hyperparameters is further discussed in **Section 4**.

The performance evaluation metrics considered for this study are the accuracy, loss and average F1 score across six categories (macro average of F1), as shown below:

$$P_{EXPR} = \frac{\sum_{expr} F_1^{expr}}{6} \quad (1)$$

From the learning curves depicted in Figure 1a it could be observed that the model starts learning in early epochs and starts converging around $5^{th} - 6^{th}$ epoch gaining training accuracy around 99.03%. The reason behind this early learning could be VGGFace weights and the right learning rate. The results obtained for this setting were optimal. Training and Validation accuracy of **99.64%** and **99.75%** respectively, and loss was the least at **0.011**.

Although the findings from the test data were ideal, it was possible to see that 10 out of 518 samples from the class DISGUST were assigned to the SADNESS category. The bias in favour of SADNESS, which had the greatest number of samples in the Training dataset, maybe the cause of this.

The main model findings are summarized in Table 3 along with results from the baseline model.

3. Task 2: Baseline Machine Learning Model

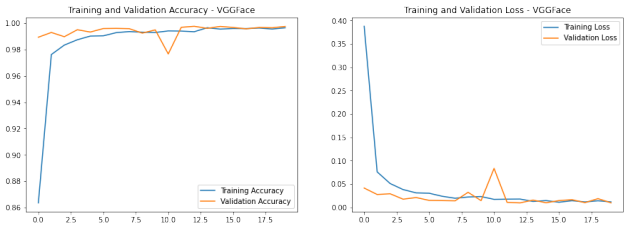
The purpose behind the creation of the baseline model was to create a CNN which generates near indistinguishable results to our main model on the subset of the Aff-Wild2 dataset, at the same time being computationally much lighter to train. The architecture selected for this model utilizes the 13 Convolutional layers from VGG-16 and later feeds their output to a softmax classifier. Table 2 describes the structure of the baseline model. While the underlying architecture of the baseline model is similar to the main, it differs from the main model significantly in two aspects. First, the classifier of this model consists of only a single fully connected softmax layer. Second, the model is initialized with weights obtained from the model trained over the classification of 1000 objects for the ImageNet Large Scale Recognition Challenge (ILSVRC). The baseline model has around 14.75M trainable parameters as opposed to the 50.4M of the main model.

Sparse Categorical Cross-Entropy was chosen as the loss function and Adam as the optimizer, keeping these choices consistent with the main model. The model is trained using a learning rate of 0.001 over 20 epochs.

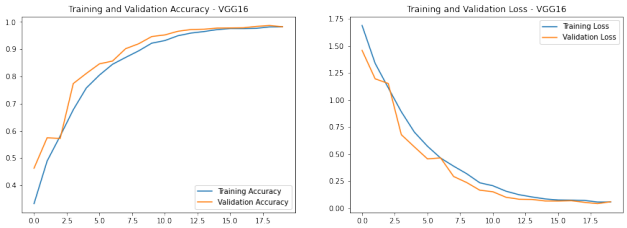
Figure 1b displays the learning and validation curves. It is clear that the model begins to learn somewhat later and begins to converge around the 15^{th} epoch, obtaining a training accuracy of about 97.19%. The accuracy during training and validation was **98.22%** and **98.25%**, respectively, and the loss was the least at **0.055**.

Model	Accuracy		Loss	F1-Score
	Training	Validation		Test
Main	99.64%	99.75%	0.011	99.83%
Baseline	98.22%	98.25%	0.055	97.66%

Table 3. Results of Machine Learning Models



(a) Main



(b) Baseline

Figure 1. Learning Curves of Main and Baseline Model

Hyperparameters			Training		Test
Optimizer	Learning Rate	Epochs	Accuracy	Loss	F1-Score
SGD	0.001	10	92.36%	0.242	95.5%
Adam	0.001	15	98.19%	0.057	97.6%
Adam	0.0001	20	99.64%	0.011	99.8%

Table 4. Ablation Study Results

The scenario between the classes DISGUST and SADNESS that was seen in the main model's confusion matrix could also be seen in this case. Additionally, around 32 samples of HAPPINESS out of 1856 were classified as SURPRISE.

4. Task 3: Ablation Study

The ablation study is performed on the main model. A series of experiments were performed with gradual changes in the hyperparameters to reach the settings which depicted optimal results. The three hyperparameters that were majorly affecting the performance of the main model were: Optimizer, Learning Rate and number of Epochs. Three experiments were performed with a different set of chosen hyperparameters maintaining the batch size at 256.

Initially, the Stochastic Gradient Descent (SGD) optimizer was selected to train with a learning rate of 0.001 over 10 epochs. This experiment yielded decent results but failed to reduce the loss. As SGD evolved a little slower throughout the training process, in the following experiment, the optimizer was changed to Adam, and the number of epochs was increased to 15 to allow the algorithm to learn longer. This setting returned satisfactory results, but the loss is still not completely mitigated. For the third experiment, the learning rate was reduced by a 10^{th} of previous, that is, to 0.0001 and as it is lowered, the number of epochs was increased to 20. This showed the best results and even performed to near perfection on the test data.

The results of the ablation study are presented in Table 4, and how accuracy and F1-score were increased and loss was lowered over these experiments is depicted over the graph in Figure 2.

5. Task 4: Production Data

We received another portion of the Aff-Wild2 dataset as the Production data for this project. Figure 3 displays the distribution of the data, which consists of 26,124 photos divided into six types of facial expressions. From the graph, it could be observed that the trend followed by the Production data is much different from the Training data. While SADNESS category still was the dominant class of all, HAPPINESS images received was nearly a 12^{th} of the Training portion for that category.

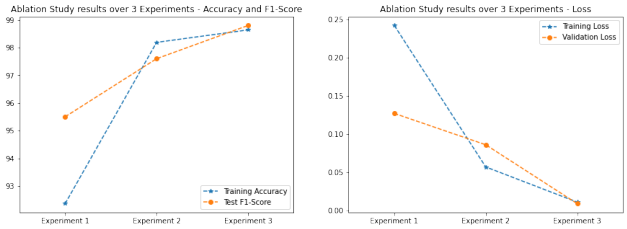


Figure 2. Evolution of Accuracy, F1-Score and Loss throughout the experiments

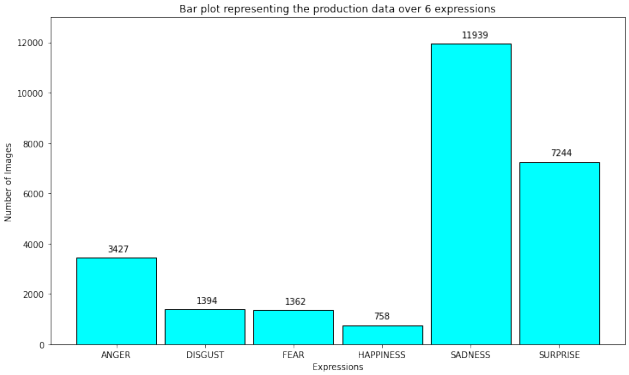


Figure 3. Production Data - Data Distribution

The performance of the main model was tested over this data, and the confusion matrix and classification report obtained are depicted in Figure 4 and Table 5 respectively.

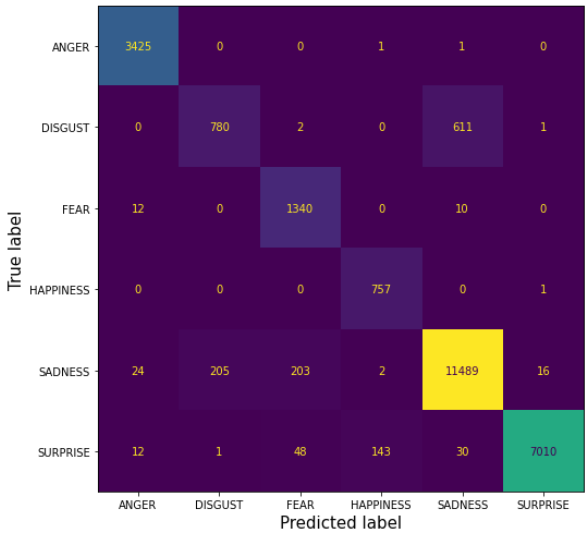


Figure 4. Production Data - Confusion Matrix

The overall accuracy of 95% and F1-Score (Macro Avg)

Expression	Precision	Recall	F1-Score	Samples
Anger	0.99	1.00	0.99	3427
Disgust	0.79	0.56	0.66	1394
Fear	0.84	0.98	0.91	1362
Happiness	0.84	1.00	0.91	758
Sadness	0.95	0.96	0.95	11939
Surprise	1.00	0.97	0.98	7244
Accuracy			0.95	26124
Macro Avg	0.90	0.91	0.90	26124
Weighted Avg	0.95	0.95	0.95	26124

Table 5. Classification Report of Production Data

of 90% was achieved over this data. While the model excelled in five of the six classes, it correctly predicted just 56% of the samples for the class DISGUST. Another observation is that around 400 samples of SADNESS class are being classified as DISGUST and FEAR (200 in each), which ends up lowering the precision of both classes.

There could be several reasons behind this. One of them is the data distribution shift which is expected when the model is in production as the images of different subjects and settings that are novel to the model will be fetched to it for results. Based on the results, this is a case of covariate/label shift which happens because of biases during the training process. The training data received was imbalanced. Three classes: ANGER, HAPPINESS and SADNESS are dominant in data containing 70% of the total, while two classes: DISGUST and FEAR are the smallest in data containing only 13% of complete data. The bias towards SADNESS could be the major reason behind classifying 611 DISGUST samples as SADNESS.

There are a couple of ways of handling this sort of distribution shift and mitigating the potential issues of bad performance. One is training the model over massive data containing most of the distribution. Another way is to retrain the model with the data from when it started to shift, in this case, re-train the model with data from classes DISGUST and FEAR.

References

[1] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 1

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1