# Emotion Classification


Alan Hui
Big Data Technology
Hong Kong University of Science and Technology
Clearwater Bay, N.T., Hong Kong
slhui@ust.hk

Ray Li
Big Data Technology
Hong Kong University of Science and Technology
Clearwater Bay, N.T., Hong Kong
kyliaq@ust.hk

Owen Tin
Information Technology
Hong Kong University of Science and Technology
Clearwater Bay, N.T., Hong Kong
pwtin@ust.hk

Alex Chow
Information Technology
Hong Kong University of Science and Technology
Clearwater Bay, N.T., Hong Kong
tkchowad@ust.hk



## Abstract

Facial emotion recognition (FER) is a popular topic in the computer vision and artificial intelligence areas because of their significant academic and commercial potential. Although facial emotion recognition can be conducted by many methods including multiple sensors, devices and different machine learning algorithms, our project is focusing on applying convolutional neural network (CNN) to process facial emotion recognition from real time video or images and employ a 5-fold cross-validation to evaluate different CNN models with their precision and recall. In our work, we trained and compared two different models, Mini-Xception Model and Alexnet Model, using images from Kaggle facial expression challenge in 2013 [6], and ultimately achieving an accuracy of 75.7% in a seven emotion categories classification test.



## Categories and Subject Descriptors

facial emotion recognition, conventional FER, deep learning-based FER, convolutional neural networks

## General Terms

Algorithms, Documentation, Design, Theory

## Keywords

Deep Learning, Neural networks, Convolution Neural Network (CNN)


## 1. Introduction

Facial emotions can help us understand the intentions of others in human communication. In general, people deduce the emotional states of other people by facial expressions and vocal tone. However, according to the survey conducted by Mehrabian A. [5], verbal components only can explain one-third of human communication and the remaining two-third need nonverbal components to express. In nonverbal components, facial expressions are the main information sources which carry important emotion meaning in human communication. Therefore, in our project, we focused on the detection of human facial emotion and we applied convolutional neural network (CNN) to carry out the recognition. In general, emotion recognition on human faces consists of three steps: face detection, face modelling and classification. Our final goal is to detect the emotion on bounded human face and classify them into seven most basic human expressions: Anger, Disgust, Fear, Happy, Neutral, Sad and Surprise.

## 2. Theoretical Background

There are thousands of artificial neural networks proposed by researchers. Some are whole new approaches while some are modifications to existing approaches. In general, there are three classes of artificial neural networks:

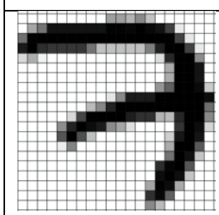
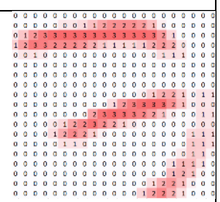
- Multilayer Perceptrons (MLP)
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)

We will only focus on CNN.

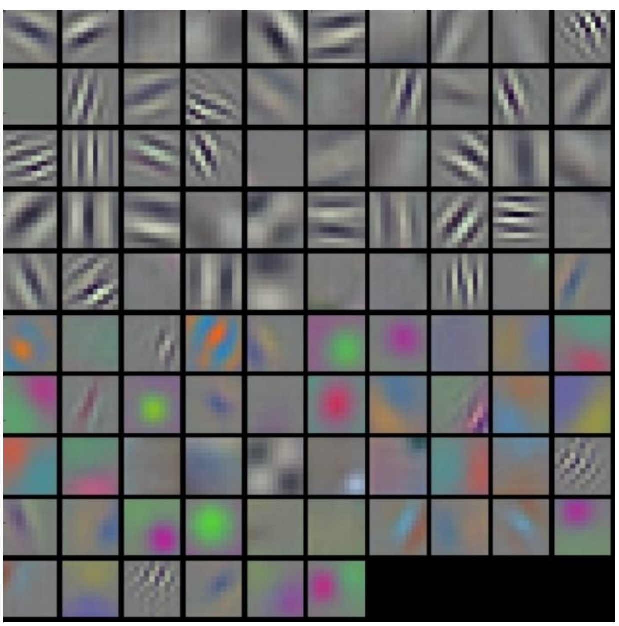
### 2.1 Convolutional Neural Networks (CNN)

Traditional feedforward neural network requires a 1d input weights. It has difficulties to deal with the problem that the input has spatial relationship. Flattering the image from pixel matrix to long vector of pixel values will lose the spatial structure in the image [1].

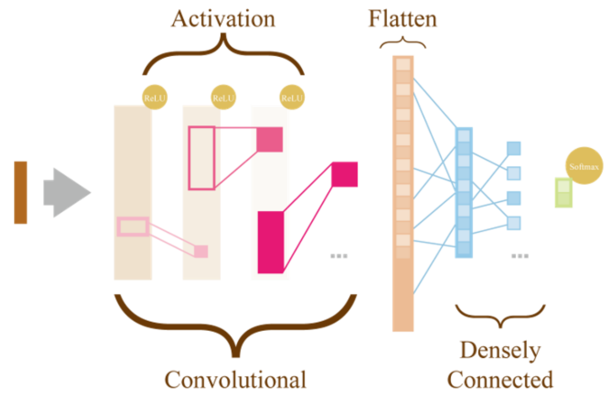
CNN overcomes this issue by learning internal features using small squares of input data. For example, a filter for detecting horizontal lines is applied to the image. The area with horizontal line will have large activation value while other activation values are small.

| Input | Filter | Output |
|---|---|---|
|  | $\begin{array}{\|c\|c\|c\|} \hline 1 & 1 & 1 \\ \hline 0 & 0 & 0 \\ \hline -1 & -1 & -1 \\ \hline \end{array}$ |  |

In the convolution layer (conv2D), a bunch of filters were applied. Each feature is learnt from one filter.



After some pooling layers and fully-connected layers, the image is mapped to output variable.




## 3. Data

### 3.1 Data Description

In our work, we used facial expression recognition (FER) dataset from Kaggle challenge in 2013. The data consists of 48×48 pixel grayscale images of faces and it contains 35,888 records in csv format. The csv file contains two columns, "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image. The emotion and numeric code mapping table is as below.

| emotion | numeric code |
|---|---|
| Angry | 0 |
| Disgust | 1 |
| Fear | 2 |
| Happy | 3 |
| Sad | 4 |
| Surprise | 5 |
| Neutral | 6 |

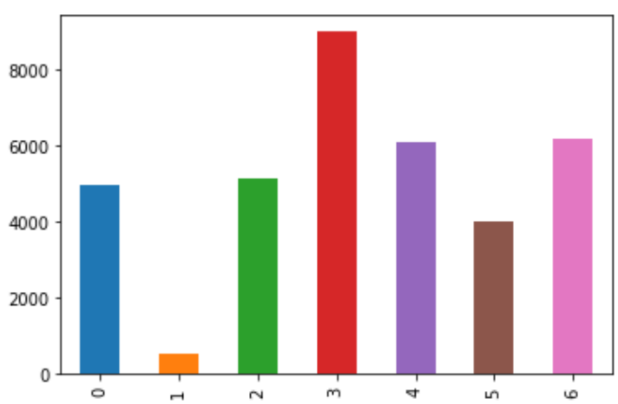
The "pixels" column contains a string surrounded in quotes for each image. The contents of this string are space-separated pixel values in row major order. Sample data are shown below:

| emotion | pixels |
|---|---|
| 3 | 70 80 82 72 58 58 60 63 54 58 60 48 89 115 121 119 115 110 98 91 84 84 90 99 110 126 143 153 158 171 169 172 169 165 129 110 113 107 95 79 66 62 56 57 61 52 43 41 65 61 58 57 56 69 75 70 65 56 54 105 146 154 151 151 155 155 150 147 147 148 152 158 164 172 177 182 186 189 188 190 188 180 167 116 95 103 97 77 72 62 55 58 54 56 52 44 50 43 54 64 63 71 68 64 52 66 119 156 161 164 163 164 167 168 170 174 175 176 178 179 183 187 190 195 197 198 197 198 195 191 190 145 86 100 90 65 57 60 54 51 41 49 56 47 38 44 63 55 46 52 54 55 83 138 157 158 165 168 172 171 173 176 179 179 180 182 185 187 189 189 192 197 200 199 196 198 200 198 197 177 91 87 96 58 58 59 51 42 37 41 47 45 37 35 36 30 41 47 59 94 141 159 161 161 164 170 171 172 176 178 179 182 183 183 187 189 192 192 194 195 200 200 199 199 200 201 197 193 111 71 108 69 55 61 51 42 43 56 54 44 24 29 31 45 61 72 100 136 150 159 163 162 163 170 172 171 174 177 177 180 187 186 187 189 192 192 194 195 196 197 199 200 201 200 197 201 137 58 98 92 57 62 53 47 41 40 51 43 24 35 52 63 75 104 129 143 149 158 162 164 166 171 173 172 174 178 178 179 187 188 188 191 193 194 195 198 199 199 197 198 197 197 197 201 164 52 78 87 69 58 56 50 54 39 44 42 26 31 49 65 91 119 134 145 147 152 159 163 167 171 170 169 174 178 178 179 187 187 185 187 190 188 187 191 197 201 199 199 200 197 196 197 182 58 62 77 61 60 55 49 59 52 54 44 22 30 47 68 102 123 136 144 148 150 153 157 167 172 173 170 171 177 179 178 186 190 186 189 196 193 191 194 190 190 192 197 201 203 199 194 189 69 48 74 56 60 57 50 59 59 51 41 20 34 47 79 111 132 139 143 145 147 150 151 160 169 172 171 167 171 177 177 174 180 182 181 192 196 189 192 198 195 194 196 198 201 202 195 189 70 39 69 61 61 61 53 59 59 45 40 26 40 61 93 124 135 138 142 144 146 151 152 158 165 168 168 165 161 164 173 172 167 172 167 180 198 198 193 199 195 194 198 200 198 197 195 190 65 35 68 59 59 62 57 60 59 50 44 32 |

## 3.2 Data Pre-processing

### 3.2.1 Imbalance of dataset

Number of disgust image in FER2013 dataset is very small while comparing to other emotions. With this imbalance dataset, it is impossible for our model to recognize disgust expression.



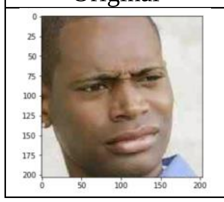
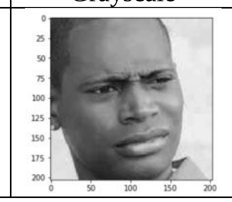
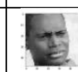

### 3.2.2 Data Augmentation

#### 3.2.2.1 New Dataset

AffectNet is a dataset of facial expressions created by Mohammad [7], a CE professor of University of Denver. It contains more than 1 million facial images either collected from the internet or manually annotated. Due to storage and network speed limitation, we only able to download 10% of the database. It contains around 400 disgust images.

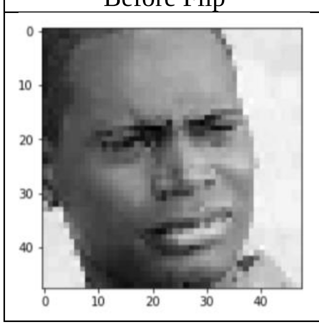
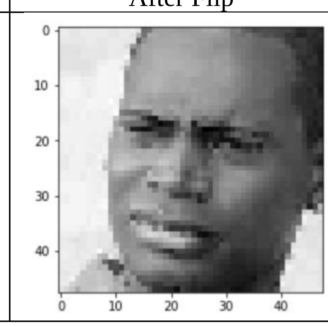
Procedures of data pre-processing:

1. AffectNet images are color images, we need to convert them into grayscale
2. AffectNet images are in different dimensions, we need to resize the image into 48 x 48
3. Labelling of expression are different, we need to map the expression according to FER2013.

| Original | Grayscale | Resize |
|---|---|---|
|  |  |  |

#### 3.2.2.2 Other Augmentation Technique, Flip

Another augmentation technique is to flip the image. We strongly recommend to do the horizontal flip only. Vertical flip of human face always confuse the CNN models.

| Before Flip | After Flip |
|---|---|
|  |  |

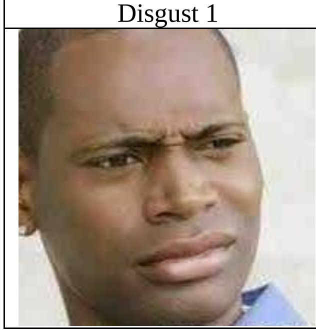
There are other techniques as well like rotation, scale, crop, gaussian noise, etc. We did not apply those due to the result of only using flipping is encouraging enough.

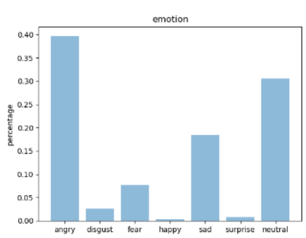
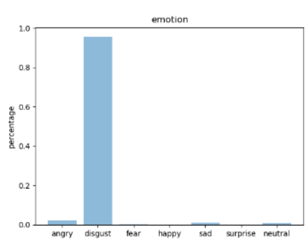
### 3.2.2.3 Data Augmentation Result

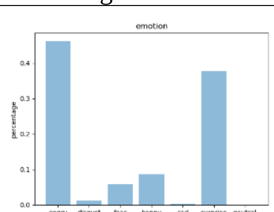
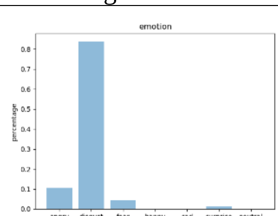
After data augmentation, we obtained almost triple sample data size of disgust image.

We trained the same mini-Xception model twice, one with FER2013 data only while another one is trained with augmented data.

We picked two images from AffectNet which manually annotated as disgust. We applied these 2 images to both of our models.

| Disgust 1 | Disgust 2 |
|---|---|
|  |  |

| Disgust 1 | |
|---|---|
| Model without data augmentation | Model with data augmentation |
|  |  |

| Disgust 2 | |
|---|---|
| Model without data augmentation | Model with data augmentation |
|  |  |

First model without data augmentation always return two or more emotions with similar probability, i.e. the model is not able to clearly classify the emotion. Second model with data augmentation returns a very definite emotion that is disgust.

## 4. Models

There are quite a few well-known models for image processing:

- Alexnet – winner of ImageNet ILSVRC in 2012
- VGGNet – runner-up of ImageNet ILSVRC in 2014
- Inception – winner of ImageNet ILSVRC in 2014
- ResNet – winner ImageNet ILSVRC in 2015

The general trend in these models is the increasing number of layers. However, simply stacking the layers does not guarantee lower testing error as the gradients are difficult to propagate back to lower layers.

Another trend is reducing number of parameters which can lower computation power.

### 4.1 Model 1 - mini-Xception

### 4.1.1 Overview

The two main inspirations of this proposed model, mini-Xceptions [2] are:

- Removal fully connected layers
- Inclusion of combined depth-wise separable convolutions and residual modules

### 4.1.2 Removal fully connected layers

Usually most of the parameters in CNN are concentrated in the fully connected layers. E.g. 90% of parameters of VGG16 are in the FC layers. By completely removing the fully connected layers, we reduced the number of parameters to 600,000 while comparing to 60M parameters in AlexNet.

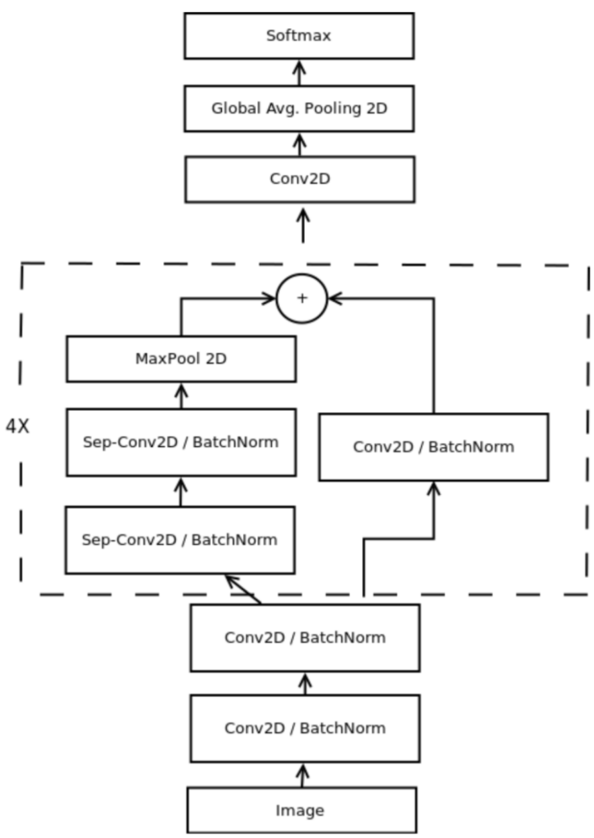
### 4.1.3 Inclusion of combined depth-wise convolutions and residual modules

By introducing depth-wise convolutions, the spatial cross-correlations are separated from the channel cross-correlations. Therefore the number of parameters is further reduced in convolutional layers

By referencing ResNet, residual modules enable the gradients better back-propagate to lower layers.

### 4.1.4 Architecture of mini-Xception

This model contains 4 residual depth-wise separable convolution layers. Each convolution layer is followed by a batch normalization operation and ReLU activation function. The last layer is the global average pooling and soft-max activation function.



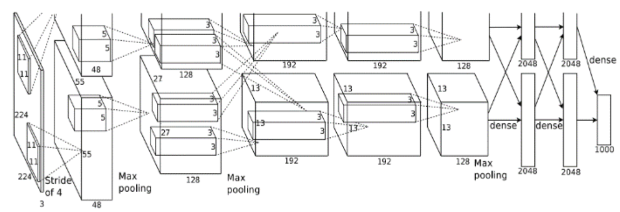

Proposed Mini_Xception architecture for emotion classification

## 4.2 Model 2 - AlexNet

### 4.2.1 Overview

AlexNet is one of the network that implemented convolutional neural network (CNN). The capacity of CNN can be controlled by varying their depth and breadth, and strong and mostly correct assumptions about the nature of images are made by them. Therefore, compared to the standard feedforward neural networks, CNN has much fewer connections and parameters so that it is much easier to train and reduces the training time.

### 4.2.2 Network Architecture




[3]AlexNet Network Architecture

AlexNet has 60 million parameters in eight layers. Five are convolutional layers and three are fully connected layers. It attached ReLU activation function after every convolutional and fully connected layers with a final softmax activation. In our modified version, only 5 million parameters are used.

### 4.2.3 Overfitting Problems

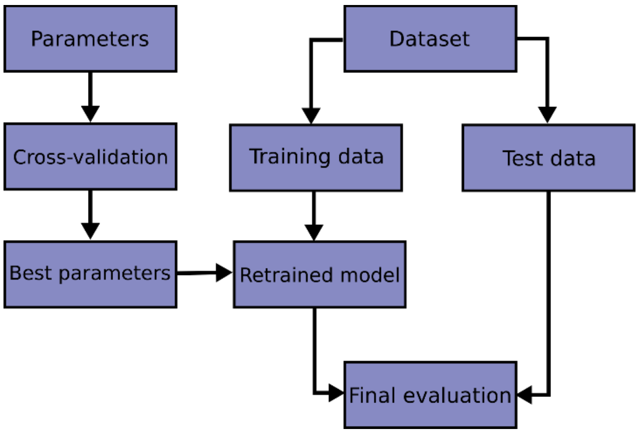
Since AlexNet has many parameters and only 7 classes are used as the output, it is insufficient to learn so many parameters without considerable overfitting. The solution will be proposed in Experiments and Results section

## 5. Experiments and Results

## 5.1 Cross-validation

To evaluate the model results, both models will be undergone two stage of training and evaluation. First, we will implement the cross-validation technique to make sure it does not overfit the training date.

The flow chart of cross-validation:




The training data are divided into 5 different parts. The 5 different parts will train into 5 different models. Then, they will be evaluated accordingly to retrieve their performance in terms of accuracy, precision, recall and F-score. The averaged and the whole set of evaluation attributes will be used to detect whether there is overfitting in the trained model.

In the second part of evaluation, both of the models will predict against a new set of testing data retrieved from the web. The result will also be evaluated with attributes accuracy, precision, recall and F-force. This evaluation is to validate the model performance handling photos in different type of formats and structures.

Expectation for the experiments is to capture whether there are overfitting and the model has captured the generic pattern and linkage for human facial movement to emotion.

## 5.2 Reduce Overfitting

Dropout is one the the effective method to reduce overfitting. It is applied in the fully connected layer, and it randomly drop some units in the neural network in AlexNet.

Using a large dataset is another way to reduce overfitting in a model. In our experiment, a dataset from Kaggle is utilized to train the model. However, seems that the training dataset is not large enough and overfitting still exist, hence regularization by dropout is not be so useful for AlexNet.

## 5.3 Result - mini-Xception

In 5-fold cross validation, with epoches equals to 8, the average accuracy for the model is 0.588154.

The average precision, recall, F-score and support for different emotions are shown as below:

| | angry | disgust | fear | happy | sad | sur prise | neutral |
|---|---|---|---|---|---|---|---|
| precision | 0.46 | 0.57 | 0.49 | 0.81 | 0.54 | 0.72 | 0.49 |
| recall | 0.60 | 0.31 | 0.27 | 0.83 | 0.37 | 0.70 | 0.66 |
| F-score | 0.51 | 0.36 | 0.33 | 0.82 | 0.43 | 0.70 | 0.56 |
| support | 990.6 | 109.4 | 1024.2 | 1797.8 | 1215.4 | 800.4 | 1239.6 |

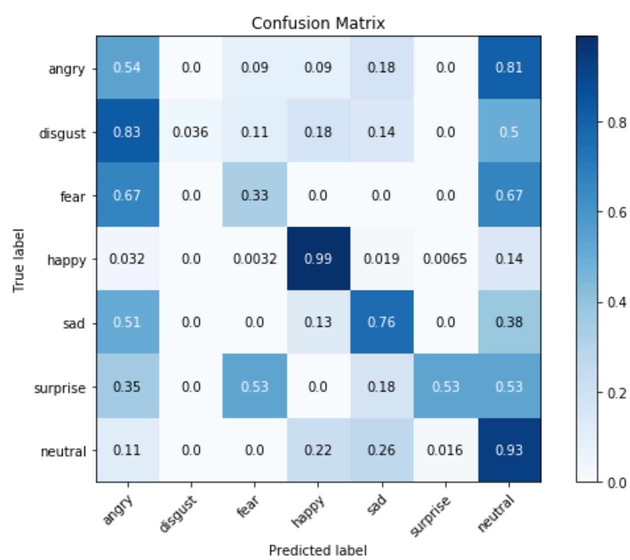
For the evaluation result against a new set of testing data, 565 photos are used in the process. The model has been run for 100 epochs with full set of training data.

The accuracy of the model is 0.67612.

The average precision, recall, F-score and support for different emotions are shown as below:

| | angry | disgust | fear | happy | sad | surprise | neutral |
|---|---|---|---|---|---|---|---|
| precision | 0.11 | 1.00 | 0.11 | 0.94 | 0.17 | 0.50 | 0.45 |
| recall | 0.32 | 0.02 | 0.20 | 0.83 | 0.43 | 0.25 | 0.61 |
| F-score | 0.16 | 0.04 | 0.14 | 0.88 | 0.24 | 0.33 | 0.52 |
| support | 19.0 | 50.0 | 5.0 | 366.0 | 14.0 | 12.0 | 99.0 |

The normalized confusion matrix result is as followed:



### 5.4 Result - AlexNet

In 5-fold cross validation, with epoches equals to 8, the average accuracy for the model is 0.62241.

The average precision, recall, F-score and support for different emotions are shown as below:

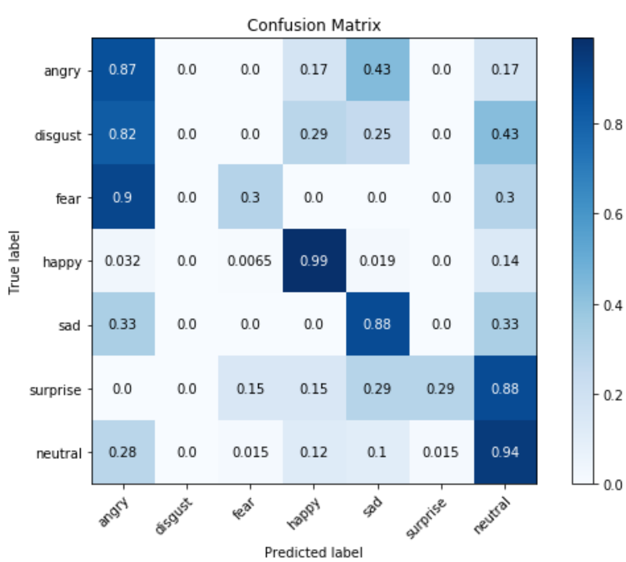
| | angry | disgust | fear | happy | sad | surprise | neutral |
|---|---|---|---|---|---|---|---|
| precision | 0.49 | 0.24 | 0.59 | 0.95 | 0.42 | 0.68 | 0.52 |
| recall | 0.62 | 0.14 | 0.30 | 0.84 | 0.53 | 0.31 | 0.71 |
| F-score | 0.55 | 0.18 | 0.40 | 0.89 | 0.47 | 0.43 | 0.60 |
| support | 990.6 | 109.4 | 1024.2 | 1797.8 | 1215.4 | 800.4 | 1239.6 |

In the experiment above, we used a dataset, which contains 565 images and are distributed in seven classes, to evaluate the performance of the model.

The percentage accuracy of the model is 0.6902655 in 100 epochs training.

| | angry | disgust | fear | happy | sad | surprise | neutral |
|---|---|---|---|---|---|---|---|
| precision | 0.15 | 0.00 | 0.20 | 0.94 | 0.23 | 0.67 | 0.49 |
| recall | 0.53 | 0.00 | 0.20 | 0.84 | 0.57 | 0.17 | 0.64 |
| F-score | 0.23 | 0.00 | 0.20 | 0.89 | 0.33 | 0.27 | 0.55 |
| support | 19.0 | 50.0 | 5.0 | 366.0 | 14.0 | 12.0 | 99.0 |

The normalized confusion matrix result is as followed:




### 5.5 Result Comparison Between Two Models

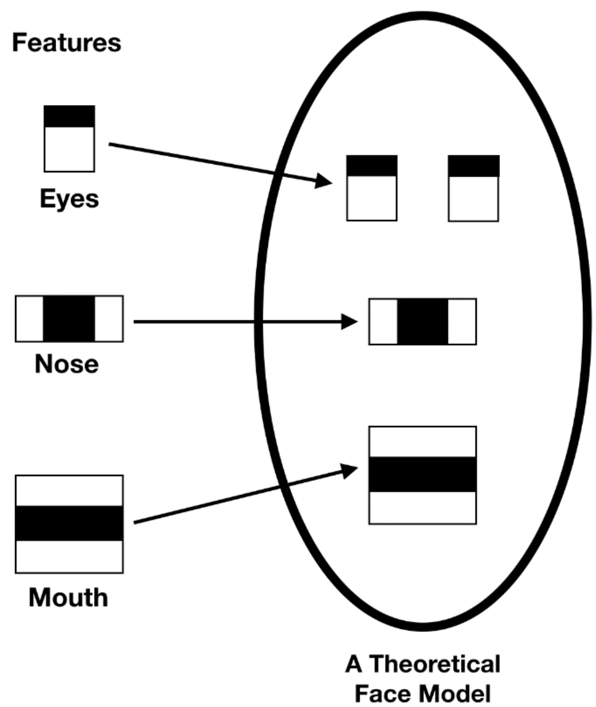
From the above results, AlexNet and Mini-Exception models have similar performance on classifying Happy, Neutral, Angry, Fear and Surprise facial expression. The performance of Disgust in Mini-Exception is better than AlexNet, whereas, the performance of Sad in AlexNet is better than Mini-Exception.

## 6. Emotion Classification in Webcam

Workflow of the real-time emotion classification in webcam:

- Face detection in webcam [4]
- Capture the detected face into image
- Predict the emotion from the captured image using our emotion classification model

We adopted Haar Cascade classifier in OpenCV for Face detection. It uses different filters to extract features like eyes, nose and mouth.



## 7. Conclusion and Future Work

In this paper, we presented 2 different CNN models, Mini-Xception and Alexnet for facial emotion detection. In our experiment, we evaluated the models by 5-fold cross-validation and using new set of testing data retrieved from the web. The AlexNet model reached 69% test accuracy in our new testing data which is better than Mini-Xception which is 67.6%. On the other hand, for cross validation accuracy, AlexNet having 62.2% also better than Mini-Xception which is 58.8%. In general, emotion with strong facial expression such as Happy and Surprise are getting better performance.

Future work might focus on trying out other types of CNN models, like VGGNet, Inception and ResNet or tuning more parameters and layers to construct an optimal model.

## 8. Reference


[1] https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/

[2] Octavio Arragia et al, Real-time Convolutional Neural Networks for Emotion and Gender Classification 2017.

[3] Alex Krizhevsky, ImageNet Classification with Deep Convolutional Neural Networks

[4] https://appliedmachinelearning.blog/2018/11/28/demonstration-of-facial-emotion-recognition-on-real-time-video-using-cnn-python-keras/

[5] Mehrabian A. Communication without words. Psychol. Today. 1968;2:53–56.

[6] https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data

[7] http://mohammadmahoor.com/affectnet/