Detection of Key Points on Face Images and Recognizing Emotions Using Convolution Neural Network

Nishant Agarwal UFID: 61991874 Email-Id: nishantagarwal@ufl.edu

Abstract-- **Developing highly accurate Face Recognition techniques is one of the many active areas of research in the field of Computer Vision. The main objective of the project is to detect emotions from face images, which can be used in applications related to education, gaming, tracking faces in real time images and videos, medical diagnosis of facial abnormalities, online customer customization system. The project also demonstrates the use of the features learnt by the key point detection model to further train a Emotion Detection CNN model using the concept of transfer learning. The performance of these tasks is largely dependent on the accuracy of the facial point detectors. Detecting facial key points is a very challenging problem as facial features shows high variance, from one individual to another, due to varying size, position, brightness parameters. Recent study suggests that training a Convolutional Neural Networks, will be a reasonable approach for the problem in hand as neural networks seems to be the stable tool in image classification which shows reasonable immunity towards distortion and vertical and horizontal shift of subject in images. An Ensemble of models is implemented for the project which gives an accuracy of 62.1% on the test data and 62.4% on the validation set.**

INTRODUCTION

Human Communication is a expression of various verbal and non verbal cues, body gestures, and recent study suggests that about 80% of time communication is made using non verbal expressions, and thus understanding and interpreting the emotion without any human intervention can be used in various industries such as online stores to better serve their customers. This paper describes model based on Convolutional Neural Network (CNN) which categorizes individual faces into one of the seven categories of Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral.

The Emotion detection model proposed can be broken down into two phases:
Phase 1: Detecting Facial Key Points using 4 layer CNN.
Phase 2: Emotion detection using 4 layer CNN and features learned from phase 1.

Phase 1:
The problem is predominantly to predict exact coordinate locations of points on an image of a face. The model is expected to predict 30 numerical values from 15 facial points. Each predicted key point is represented by an (x, y) coordinates in pixel space. There are 15 key points, which represent the following elements of the face such as right_eyebrow_inner_end,left_eye_inner_corner,mouth_center _top_lip.

Accuracy of the model is evaluated on the root mean squared error which is very common and is a suitable general-purpose error metric for regression based problems:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

.

The CNN model generated for key point detection has four convolutional layers of varied number of filters for each layer, followed by two fully connected dense layers of 256 and 512 neurons each and finally an output layer which is predicting 30 key points, representing the x, y coordinates. Each convolutional layer is followed by max pooling of size 2 X 2 and dropouts and application of RELU (Rectified Linear Unit) activation functions.
The model was trained for 1000 epochs on AWS servers which resulted in the training error of 1.07X 10^{-3} and validation error of 1.56X 10^{-3}

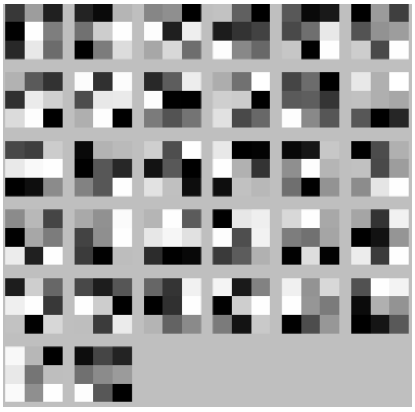


Fig1: Represents the 32 filters from the first layer of the convolutional layers

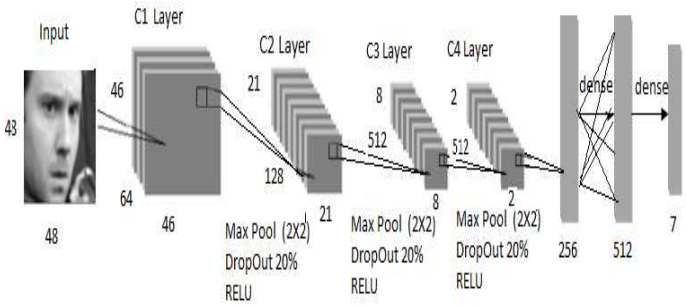


Fig2: CNN model used for emotion detection

Phase 2:

The task is to predict the relevant emotions, represented by numeric code ranging from 0 to 6 from a dataset consisting of 35887 grayscale images of 48X48 dimensions each, obtained from a Kaggle competition. The ratio between the training and test data split was chosen as 9:1. Two models were trained having the same architecture. The difference between the two models lies in the strategy used to initialize the weights for the layers. While one had its weights initialized using the GLOROT_STYLE initialization which is the default in Lasagne, the other was set to the weights and biases learned from the key point detection model trained in phase 1.

The CNN model used for the above task had four convolutional layers having 64, 128, 512, 512 number of filters of size 3 X 3 respectively. This was followed by two fully connected layer of 256 and 512 neurons each and an output layer classifying into one of the 7 classes with is activated using the softmax classifier. Each convolutional layers was followed by max pooling of size 2 X 2 and dropout of 20% with RELU activation applied in each layer. The model was trained for 100 epochs and **Nesterov_Momentum** was used as the optimization method.

Description

Phase1: Facial key point Detection

1) Data Preprocessing: The data set consists of 7049 training and 1783 test grey scale images of 96X96 dimension, where each row contains the (x, y) coordinates for 15 key points. As a first step we remove rows from the data set which does not contain all the labels.

left_eye_center_x	7039
left_eye_center_y	7039
right_eye_center_x	7036
right_eye_center_y	7036
left_eye_inner_corner_x	2271
left_eye_inner_corner_y	2271
left_eye_outer_corner_x	2267
left_eye_outer_corner_y	2267
right_eye_inner_corner_x	2268
right_eye_inner_corner_y	2268
right_eye_outer_corner_x	2268
right_eye_outer_corner_y	2268
left_eyebrow_inner_end_x	2270
left_eyebrow_inner_end_y	2270
left_eyebrow_outer_end_x	2225
left_eyebrow_outer_end_y	2225
right_eyebrow_inner_end_x	2270
right_eyebrow_inner_end_y	2270
right_eyebrow_outer_end_x	2236
right_eyebrow_outer_end_y	2236

Fig 3:The figure shows that the data set have 7039 labels for left_eye_center_x and only 2271 for left_eye_inner_corner_x.

Therefore to maintain consistency among the data we remove rows having missing labels.

Various Advantages in using CNN for Image Processing are:

- Local Connectivity and weight sharing
- Pooling
- Dropouts

These features reduces the number of learnable parameters and makes the training process much easier and faster.

The general architecture of the model is explained in fig2.
Data read from input file contains a flat vector consisting of pixel intensities, but input to lasagne convolutional layers has to converted into a three dimensional matrix of the form (C, X, Y) where C represents the color channel and X and Y the dimensions of the image. Since we use gray scale images the value of C is reduced to only 1 and the final shape of the image fed into the network is (1, 96, 96).

The model is trained for 1000 epochs with a learning rate of 0.01 and update momentum of 0.9. Since this is a regression based problem, no non-linearity is applied to the output layer.

Some of the results obtained from the above model with an RMSE of 1.7899:

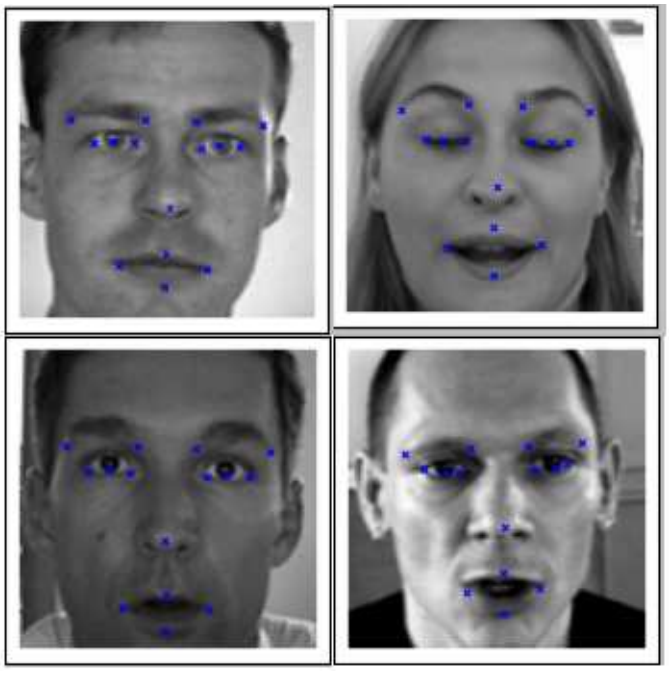


Fig 4 : The blue spots represents the various facial key points.

Phase 2: Emotion Detection

1) Data Preprocessing : The data set consists of 35887 grey scale images. To prevent the CNN model from over fitting and to generalize the model, artificial data set are created by Data Augmentation using the process of

transformation and noise introduction. Data Augmentation not only increases the number of training samples but also introduces variability in the data set , which the model can learn to predict with greater accuracy on test set. The data split ratio between the training and test data was set to 8:2.

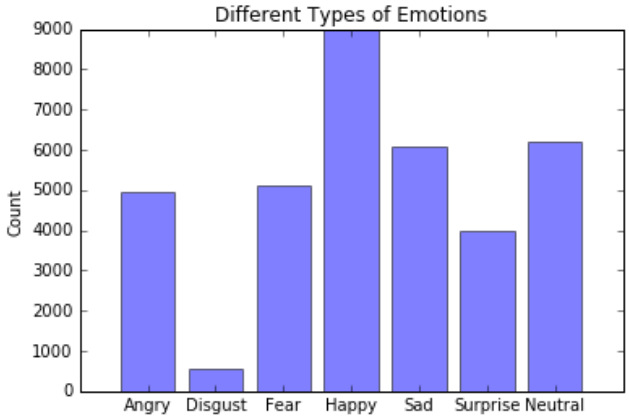
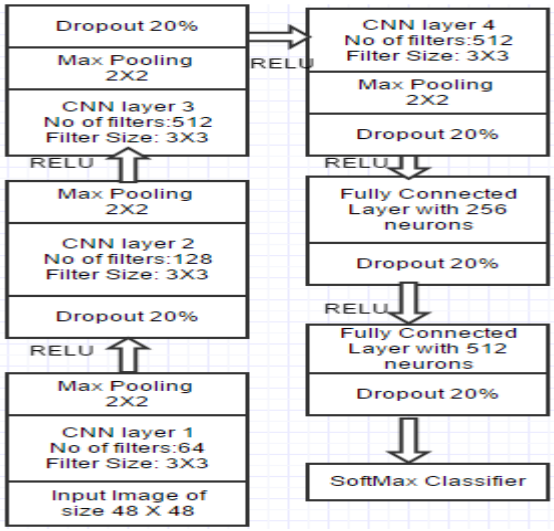


Fig 5: Bar chart showing the count of different emotions present in the dataset.

Sample image showing the data augmentation process where the image has been flipped horizontally:

Fig 6: Images has been transformed to create mirror images of each other.

Augmented input images are reshaped into the form of (1 X 48 X 48) and fed into the following CNN network



Learning rate was initialized with value 0.01 which was decreased with every epoch, since at the beginning of the training phase the model is quite far from the global minimum. and using a higher learning is justifiable., but as the model converges towards the optimum value a lower learning rate ensures better and faster convergence. The dropout was kept fixed at 20% since convolutional filter of 3X 3 itself helps in regularizing the model.

The model with weights initialized using GLOROT_STYLE gave an accuracy of 55% after 100 epochs.

Transfer Learning: The second model uses the concept of transfer learning to take advantage of the features learned from the key point detection model and further train on the input images to classify them into 7 emotion categories. This enables the application of pre trained models to be used as an initialization or feature extractor for the current problem in hand.

Different scenarios involved in Transfer Learning:

- We retrain the classifier on top of the CNN model obtained from phase 1 on the new data set and update the weights of the pre trained network using the process of backpropagation. Sometimes the weights from the initial layers are kept constant and only weights from recently added layers are backpropagated.
- Another use case of this concept involves the removal of the last layer of fully connected neurons of the pre trained model and using it just as a feature extractor for the new dataset.

The intuition behind using transfer learning is as follows:

- Eyebrow shape Analysis: Facial features gives different shapes of eyebrows for different emotions which can be considered as one of features in the classifier.

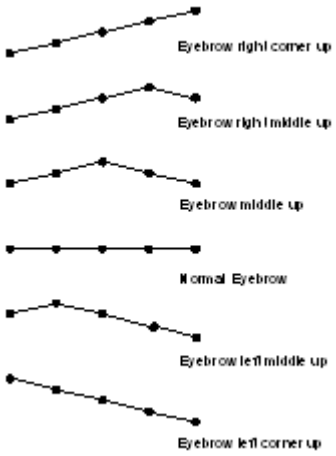


Fig 7 : Different shapes for eyebrows

- Mouth Structure Analysis: Detecting the coordinates of mouth_left_corner,mouth_right_corner,mouth_top_center, mouth_botton_center can unveil the relationship

between the emotions and mouth structure which can be treated as yet another feature.

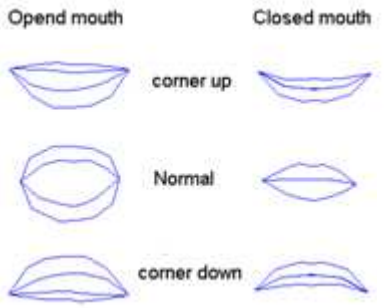


Fig 8 : Different Shapes of mouth

The model trained with weights initialized from facial key point detection model gives an accuracy of 62.2% on the validation set and converges in 50 epochs, which is 50% faster than the first model.

Evaluation

Phase 1 : Facial key point Detection

No of Epochs	Training Error	Validation Error	RMSE
1000	0.00107	0.00156	1.7899

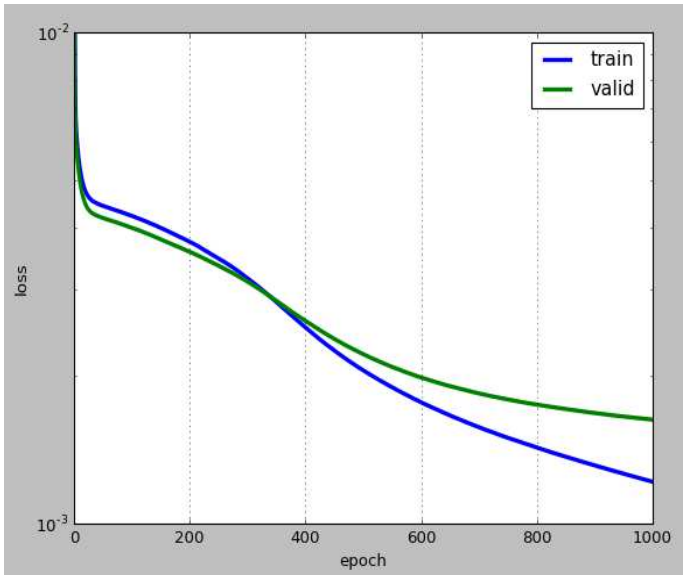


Fig 9: The learning curve for the model , showing the change of training and validation error with number of epochs

Phase 2 : Emotion Detection

No of Epochs	Data Augmentation	Accuracy	Transfer Learning
100	No	46%	No
100	Yes	55%	No
50	Yes	62%	Yes

A 5 fold cross validation SVM has also been implemented which showed an accuracy of only 32%

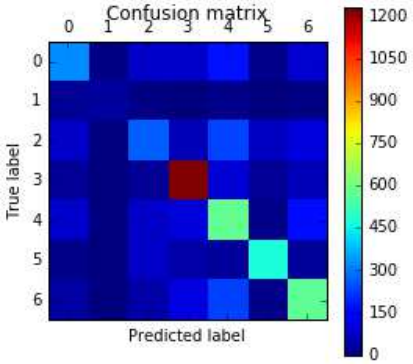
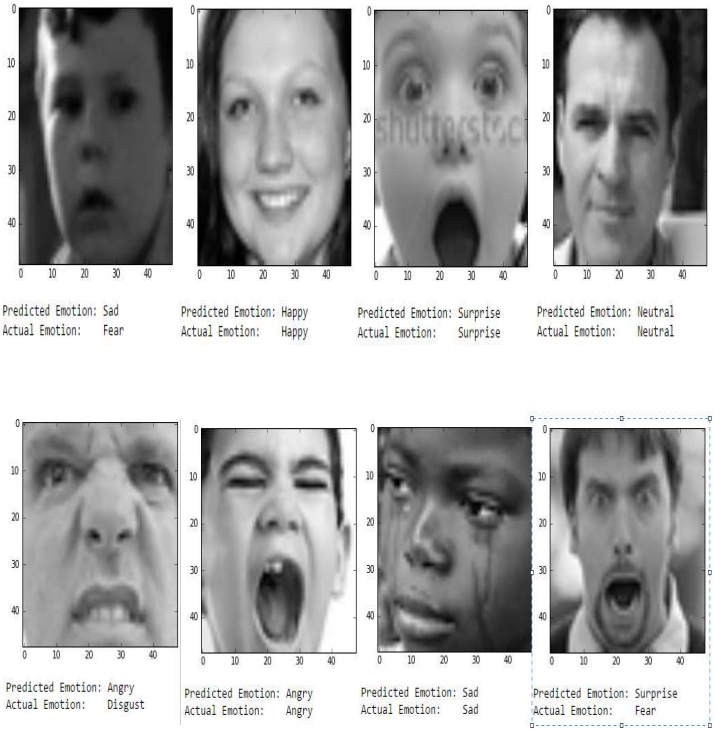


Fig 10: Confusion matrix showing the prediction values for different classes on test set. The elements on the diagonal are correctly classified

```
array([[ 326,    9,   81,   83,  175,   17,   89],
       [  27,   35,    7,    3,   11,    2,    5],
       [  81,    3,  272,   58,  237,   73,  104],
       [  26,    0,   27, 1232,   95,   25,   63],
       [  83,    2,   81,   97,  578,   14,  173],
       [  14,    1,   77,   46,   29,  476,   30],
       [  43,    3,   48,  110,  231,   10,  575]])
```

Fig11: Confusion matrix showing the prediction accuracy

Results from test set along with predicted and actual emotion label are shown below:



The above results points to certain important facts:

- Emotions such as Disgust and Angry are difficult to differentiate even for humans.
- Neutral emotions are often confused with Happy or

Sad, due to lack of variability in the facial key points for these emotion types.

- Happy , Sad and Surprise are emotions which are best detected by the CNN model.

Related Work

Facial and Emotion recognition has become an key source in multi cultural visual communication system, and lot of researchers have proposed new ideas to achieve higher accuracy. Alex P Pentland et al [6] described an approximated optical flow method, fused with geometric, physical characteristics describing the facial structure which probabilistically characterize physical actions to decipher the emotion category. This model is based on Facial Action Coding System which allowed to code expression from static pictures. Liyanage C. DE SILVA suggested a way of combining auditory and visual reception and using these hybrid multi modal information in detecting correct emotional state. Zhengyou Zhang [8] showed that multi orientation Gabor wavelet coefficients extracted from key points on the face are more robust than coordinate positions and with the use of two layer perceptron they achieved good recognition rate. Peter Burkert [9] has developed an architecture where in facial patches in the form of eyes or lips are marked and features are extracted taking account of the variance between two images. Further the dimensionality of these features is reduced using Principal Component Analysis (PCA) and subsequently fed into Support Vector Machine classifier, working on 10 fold cross validation scheme which resulted in F1 score of 0.87.

Summary and Conclusions

As a part of the project, learnt to build a complete CNN network in accordance to the problem and input data in hand and deciphered ways to affectively set the hyper parameters such as learning rate, size of pooling and convolutional layers.

Learnt about ways to improve the accuracy of existing model through generating more training samples via data augmentation. Regularization is an essential aspect in designing a CNN network and to make the model generalize better towards unseen data. A model with low training error cannot guarantee high accuracy, hence various regularization techniques such as L1 and L2 should be applied to prevent the model from over fitting. Dropout layers used in the CNN model architecture is used for similar purpose.

Studied upon the concept of Transfer Learning and implemented the same in the project to achieve better results. Incorporating the concept of transfer learning can be regarded as the most interesting part of the project since no other researchers have proposed any idea on these lines.

Familiarized with deep learning library of Lasagne and Theano for training the CNN models on AWS servers having CUDA's GPU support, which provides greater computational capabilities and is orders of magnitude faster than CPU. Finally made comparisons between the different models created for the above tasks and fine tuned them to prevent its early saturation.

References

[1] Sun, Yi, Xiaogang Wang, and Xiaoou Tang "Deep convolutional network cascade for facial point detection." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[2] Amberg, Brian, and Thomas Vetter. "Optimal landmark detection using shape models and branch and bound." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.

[3] Belhumeur, Peter N., et al. "Localizing parts of faces using a consensus of exemplars." Pattern Analysis and Machine Intelligence, IEEE Transactions.

[4] Thesis Research Manoj Gyanani (SJSU) http://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1470&context=etd_projects

[5] http://cs231n.stanford.edu/reports2016/010_Report.pdf

[6] Irfan A Essa, Alex P Pentland et al Coding, Analysis, Interpretation, and Recognition of Facial Expressions.

[7] Nouri, Daniel. 2015. Kaggle Facial Keypoints Detection.https://www.kaggle.com/c/facial-keypoints-detection#description

[8] Liyanage C. DE SILVA, I Tsutomu MIYASATO, Ryohei NAKATSU et at Facial Emotion Detection using multi modal information.

[9] CS231N: Facial Expression Recognition Hsiao Chen Chang, Emilien Dupont, William Zhang

[10] Zhengyou Zhang; Michael Lyons; Michael Schuster; Shigeru Akamatsu et al Comparison Between Geometry-Based and Gabor-Wavelets-Based Facial Expression Recognition Using Multi-Layer Perceptron.

[11] DeXpression: Deep Convolutional Neural Network for Expression Recognition et al Peter Burkert, Felix Trier, Muhammad Zeshan Afzal, Andreas Dengel and Marcus Liwicki.

[12] Face expression detection and synthesis using statistical models of appearance et al H. Kang, T. Cootes and C. Taylor.