Facial Emotion Recognition using Key-Points

Deepak Balaji Selvam
University of Central Florida
Orlando, Florida
deepak@knights.ucf

Kavin Arasu Balasubramanian
University of Central Florida
Orlando, Florida
kavinstanes@knights.ucf.edu

Rahul Anand Cheeti
University of Central Florida
Orlando, Florida
rahulanand@knights.ucf.edu

Abstract

In this paper we detail about the experimental results of facial emotion recognition with 2D key-points. The key-points or landmarks are generated for eyes, nose and lips of the facial regions for all the images in the FER-plus dataset and using this processed dataset with key-points, we proceed to train a convolutional neural network model which will classify the facial expression portrayed by the facial regions of a given image. Additionally, we use parameter reduction techniques like global average pooling and depth-wise separable convolution to reduce the total number of trainable parameters and thereby speeding-up the process of training and prediction.

1. Introduction

Facial Emotion recognition involves detecting the facial regions of a given image and we feed the facial regions to a CNN based classifier which finally gives the class of the emotion portrayed by this face. Thus, it involves an object detection and classification task. We use Histogram of Gradients methods to detect the facial regions of a given image and a CNN classifier will be used to compute the emotion classification task.

We started with the assumptions that the key-points generated will help the classifier distinguish the variation of emotions as the shape of primary key-point locations like lips and eyes will greatly vary in accordance to the emotion portrayed. We have primarily worked with two dataset the FER-2013 and FER-Plus dataset and both of them apparently use the same images but the labeling information is different.

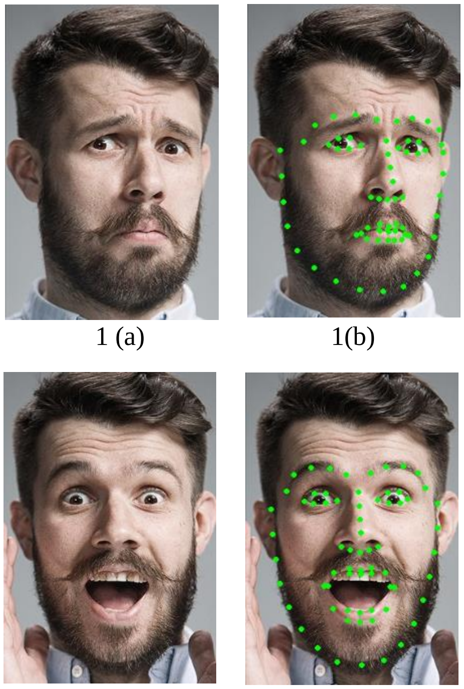
2. Related Work

The closely related work we could identify was the CNN model for emotion and gender classification task by Octavio Arriaga et al [1]. It involves using the FER 2013 dataset and achieved an emotion classification accuracy of about 65% and it employed the 'Xception' model which uses the is a residual connections with depth-wise separable convolution, they have also replaced the fully connected layer with Global Average Pooling leading to a 90% reduction in the parameter count compared to the former. They used Haar-Cascades to detect the facial regions of a given image. For gender classification they used the IMDB dataset.

3. Proposed Approach

3.1 Key-point Generation

The original dataset has the images and their corresponding label information, we have added facial landmarks or key-points to all the images in the dataset while the label information is preserved.



1 (a) 1(b)

2 (a) 2 (b)

Figure 1: Key-point information for Facial regions. (Sample image obtained from Toolshero.nl website)

From figure 1, the images on the left 1 (a) and 2 (a) will be fed to landmark generation algorithm which is based on ensemble of regression trees. Notice that the key-points or landmarks centered around the lips and eyebrows of 1(b) and 2(b) vary considerably in accordance with the emotion portrayed by the image. We assumed this to be one of the learned features by the CNN based classifier for emotion recognition task. Thus, we have overlapped the key-information to the raw images of the FER plus datasets and the labeling information was preserved.

3.2 Parameter Reduction

The motivation behind employing parameter reduction techniques is that it enables the model to predict with less computational resources and thus it takes reduced amount of time to train the network.

3.2.1 Depth-wise separable convolution

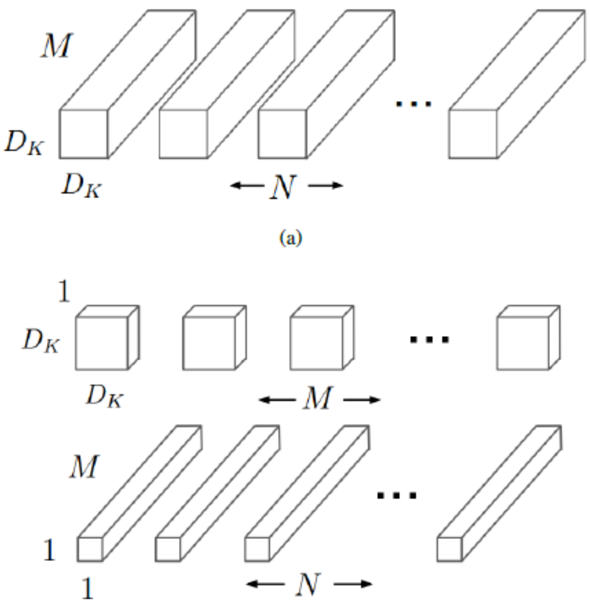

M
D_K
D_K
$\leftarrow N \rightarrow$
(a)
1
D_K
D_K
$\leftarrow M \rightarrow$
M
1
1
$\leftarrow N \rightarrow$

Figure 2[8]: Depth-wise Separable Convolution

Figure 2 compares the depth-wise separable convolution process with Standard Convolution. In standard convolution the filter dimension always extends to the full depth of the input image and the depth of the output dimension will depend on the number times the convolution filters are applied to the input volume.

Separable convolutions take more numbers of steps to essentially arrive at the same output as standard convolution network. There are two primary steps: First, we apply a two-dimensional filter to get an intermediate volume. The number of times the 2D filter is applied equals to the number of channels in the input volume. Secondly, one-dimensional filter is convolved with the intermediate volume to get the desired output, which will eventually be the same as the output of the standard convolution. The number of times this one-dimensional filter is applied will be equal to the number of channels of the output volume. By doing this, we actually reduce the number of multiplication operations that occurs and thereby it leads to a reduction in the number of parameters used. Thus depth-wise separable convolution increases the speed of prediction and training. The parameters are reduced by a factor of $(1/D^2 + 1/N)$, where D is the dimension of the 2-dimensional filter and N is the number of channels in the output volume.

3.2.2 Global Average Pooling

In traditional CNN models, the convolutional layers extract the feature and feed them into fully connected networks followed by SoftMax function for the classification task. The fully connected layers contribute to roughly 90% of the total number of parameters of the entire CNN model and higher parameter count also makes it more prone to overfitting and thereby not allowing the model to generalize properly. Thus, the fully connected layers can be replaced by Global Average Pooling.

Figure 3 shows the global average pooling's action, it takes feature maps as input and for every single feature map it produces a single scalar value as output. This scalar value for each feature map is obtained by averaging the entire feature map. Thus, the output of global average pooling is an array of scalar values. The depth of the output array of the global average pooling will be equal to the number of features maps given to the GAP as input.

Thus, GAP speeds the speed of training by several folds. The final model will also use lesser resource to compute the predictions.

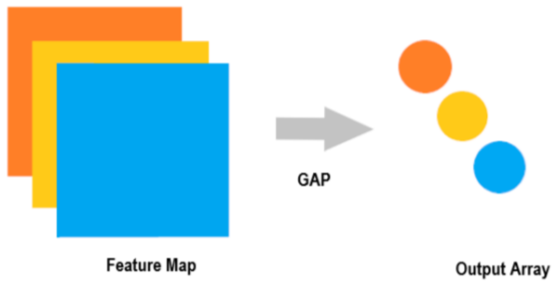
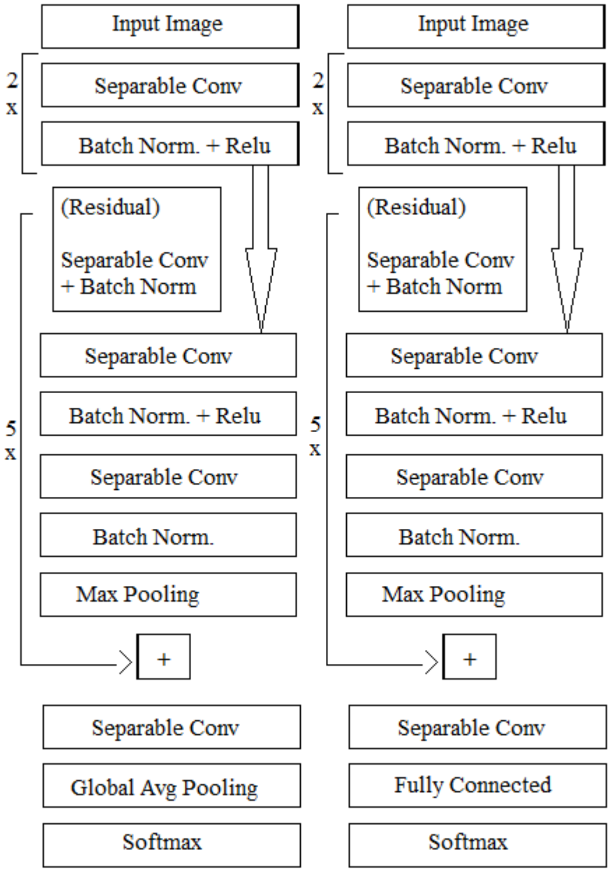


Figure 3: Global Average Pooling.

3.2.2 Classifier Architecture

The ‘Xception’ module uses residual methods in combination with the Depth-wise separable convolution. We train an architecture which entirely replaces the fully connected layers by using Global Average Pooling and we also compare this accuracy metrics to the same architecture with fully connected layers preceding the soft-max layer.



4(a) 4(b)

Figure 4: Model Comparison

Figure 4 summarizes the two primary models we worked with. Model depicted in 4(a) uses the global average pooling and has only about 0.2 Million parameters and the model depicted in figure 4(b) uses fully connected layers and has about 8 Million Parameters and weighs about 99 Megabytes. Thus its evident that GAP has produced a better model with a small compromise in the accuracy with a difference of about 1% compared to 4(b). Both the trained models 4(a) and 4(b) employ residual methods. It enables us to train much ‘deeper’ architecture and helped us arrive at good results. The relu activation function is used. The residual block is repeated 5 times by before the final classification is done. We have used softmax function as we are doing a classification task, both the models have an accuracy over 80% with FER plus.

4. Metrics and Results

The results here summarize the training and testing trend of the Fully connected model and the GAP model. In addition to this we have adopted a base method to which we will be comparing our result. The base method we have chosen is the emotion classification tasks accomplished by the same classification architecture expect that no key-points are generated. We plan to represent our results with the help of a confusion matrix. It gives a clear idea of the model’s performance for each class of label with the respect to the ground truth information. The confusion matrix is always synthesized from the test data results.

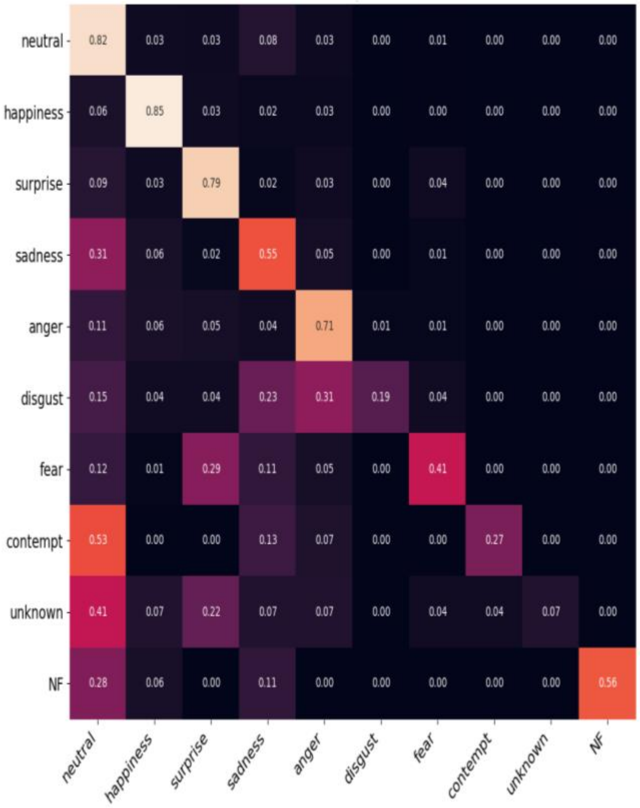


Figure 5: Confusion Matrix for GAP model.

The confusion matrix from figure 5 displays the likelihood of the images in the test set will be matched to the ground truth values for all the class of labels or emotions. We can infer that the model is best at predicting happiness, neutral surprise and anger emotions. Contempt, fear and disgust are likely to be mis-predicted.

Model	Accuracy	Parameter Count
Key-point based GAP model	75%	0.2Million
Key-point based Fully Connected model	78%	8 Million
No Key-point GAP model	80%	0.2Million
No Key-point Fully Connected Model	81%	8 Million

Table 1: Metric Information

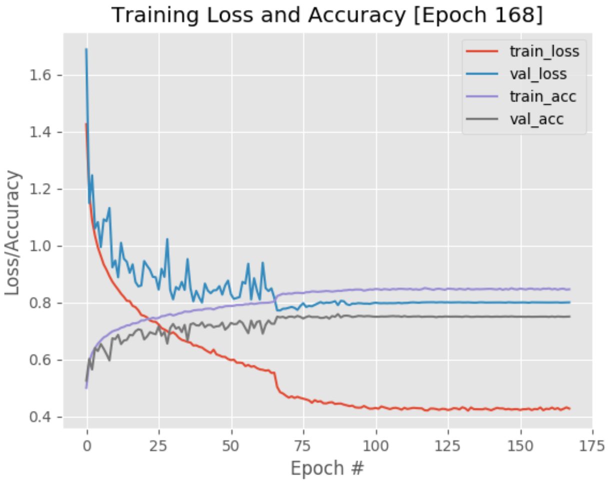
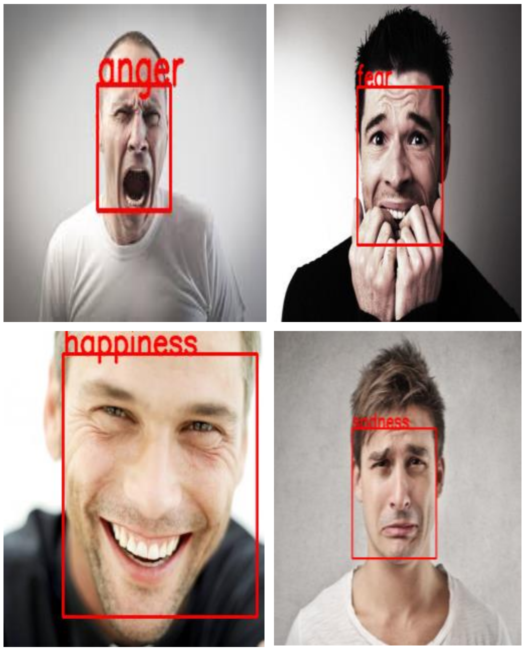


Figure 6: Training trend for GAP with Key-points.



(Source: Google Images)
Figure 7: Predictions of trained models

6. Conclusion

From the results and metrics obtained we can infer that the GAP model produces a major drop in the parameter count compared to the fully connected layers. The key-point based CNN models accuracy dropped slightly compared to the models which didn't use key-points at all. At-most a 5% accuracy difference was obtained in the results between the key-point based models and the base models.

7. Contribution

Deepak – I did a literature survey about Xception module & Global average pooling for creating and training the network architecture with FER plus dataset. Moreover, I wrote the scripts to create confusion matrix. I worked on the d-lib libraries to generate the key-points for all the images in the dataset.

Kavin – I worked with generating key-points or landmarks with d-lib (based on an ensemble of regression trees) for the FER plus dataset. I focused on training the base model (fully connected layers) so that we have a reference model to compare with the GAP model and also contributed to the one-hot encoding which was required on the FER plus dataset and written most of the CVPR paper of this project.

Rahul - I have done a literature review of the various methods and techniques used for face detection, generating key points, extracting good correlation features as parameters from the key points and architecture for emotion classification based on the input features. Helped in writing the report and to get references. Done a literary survey of how this project can be a module to human computer interaction for affective computing for better human computer communication.

8. Future Work

Until now, we have used grayscale images in our datasets, both FER and FER PLUS for generating the facial landmarks and training the networks., In near future, we can also model the facial landmarks using 3D key points by creating the 3D models out of faces to extract more minuscule features and train the network on them to predict more accurate real-life emotions.

8. Acknowledgement

We would like to thank Octavio Arriaga,Paul G. Ploger, Matias Valdenegro for making their source code available on github and we also thank Barsoum, Emad and Zhang, Cha and Canton Ferrer, Cristian and Zhang, Zhengyou for making the FER plus data available in github.

9. Reference

[1] Octavio Arriaga,Paul G. Ploger, Matias Valdenegro. Real time Convolutional Neural Networks for Emotion and Gender classification. https://github.com/oarriaga/face_classification/blob/master/report.pdf

[2] Lin Bai, Student Member, IEEE, Yiming Zhao, and Xinming Huang, Senior Member, IEEE. A CNN Accelerator on FPGA Using Depthwise Separable Convolution. arXiv:1809.01536v2 [eess.SP] 6 Sep 2018

[3] O. Déniz, G. Bueno, J. Salido, F. De la Torre. Face recognition using Histograms of Oriented Gradients.

[4] Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357v3 [cs.CV] 4 Apr 2017

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. arXiv:1512.03385v1 [cs.CV] 10 Dec 2015

[6] Ma Xiaoxi, Lin Weisi, Huang Dongyan, Dong Minghui, Haizhou Li. Facial emotion recognition. 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)

[7] James Pao. Emotion Detection Through Facial Feature-Recognition. https://web.stanford.edu/class/ee368/Project_Autumn_1617/Reports/report_pao.pdf

[8] Min Lin, Qiang Chen, Shuicheng Yan. Network In Network. arXiv:1312.4400v3 [cs.NE] 4 Mar 2014

[8] Andrew G. Howard et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017

[9] R. Picard. Affective computing. MIT Press, 1997. M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321

[10] Barsoum, Emad and Zhang, Cha and Canton Ferrer, Cristian and Zhang, Zhengyou. Training Deep Networks for Facial Expression recognition with Crowd-Sourced Label Distribution. ACM International Conference on Multimodal Interaction (ICMI) 2016