International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

EMOTION AND GENDER CLASSIFICATION IN REAL-TIME

Utkarsh Agrawal[1], Anjnesh Sharma[1], K. Senthil Kumar[2]

[1]*Student, SRM Institute of Science and Technology, Chennai*
[2]*CSE, Assistant Professor, SRM Institute of Science and Technology, Chennai*

Abstract - We propose a CNN for designing a real time classification system for face detection, gender classification and emotion recognition. Each of the three tasks will have a separate model trained for the specific problem of either face detection, gender classification or emotional classification. The three models will then be pipelined so get one common output for all the three tasks. The architecture of our model is designed so that it can give acceptable performance even on low end systems which lack powerful hardware.

***Key Words*: Convolutional Neural Networks, Face Recognition, Gender Classification, Emotion Classification, Face Detection**

1. INTRODUCTION

Facial expressions are an important aspect of human emotion recognition. Facial Expressions were introduced as a research field by Darwin. Automatic emotion recognition can be useful in areas such as computer-human interaction, security systems and many others services-based systems.

Identifying one's is a task we perform on a daily basis without much effort. Doing that automatically using computers is a complicated task. Accurately identifying one's expression using ML algorithms is a complicated task since there are a huge number of variable states possible. To achieve decent accuracy, the model will have to be trained using hundreds of thousands of training samples. Humans themselves have an accuracy of just 65% in identifying these expressions. One can see the complexity of the task by trying to identify the different emotions in the FER-2013 dataset **(Figure 1).**

Fig -1: FER-2013 Dataset

2. Related Work

The CNNs that are commonly used for computer vision applications typically include some fully connected dense layers in the final stages of the model. Most of the parameters in a model are present in these dense layers. For example, in VGG16 [11], the last fully connected dense layers contains approximately 90% of all the parameters in the model. Recent architectures such as Inception V3 [13], decrease the number of parameters at the end of model by including a Global Average Pooling operation. Global Average Pooling takes the average over all elements of the feature map and reduces it into a scalar value. This forces the neural net to extract the global features from an input image. Recent architectures like the Xception [1] utilize a combination of the 2 most successful assumptions in CNNs: the use of depth-wise separable convolutions [2] and the residual modules [8]. Depth-wise separable convolutions separate the tasks of feature extraction and combination within the same convolutional layer to further reduce the number of parameters.

The current best performing model on the FER-2013 dataset utilizes a CNN trained with square hinged loss [12]. This model has an accuracy of 70% [6] and contains around 5 million parameters. This architecture contains 98% of all the parameters of the model in the dense layers located at the end of the model.

Other models presented in [6] obtained an accuracy of 66% by using an ensemble of CNNs.

3. Model

We have designed two models which will be evaluated on the basis of accuracy and the number of parameters in the model. The models have been designed to achieve the highest accuracy to number of parameters ratio. This reduces the size of CNNs which greatly enhances the performance in systems with hardware- constrains like robots. Also, decreasing of number of parameters will lead to better generalization. The initial model completely re- moves the fully connected dense layers. The final model also includes depth wise separable convolutions with combined with residual modules. An ADAM optimizer was used in training both models [14].

Our initial architecture uses Global Average Pooling to eliminate the need of any fully connected dense layers. The number of feature maps is made equal to the number of classes in the last convolutional layer and then a soft max

activation function is used on all reduced feature map. First architecture is a fully connected CNN having 9 layers, ReLU activation functions [7], Batch Normalization [4] and Global Average Pooling. The model has nearly 700,000 parameters. IMDB gender data set with 460,723 RGB images, each classified as "woman" or "man" was used as training data, and a 96 percent accuracy in this data set was achieved. We validated this model using the FER-2013 data set which consists of 35,887 grayscale images classified as “angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, “neutral”. This model had an accuracy of 67%.

We have also designed another model based on the Xception [1] architecture. This architecture utilizes the combination of depthwise separable convolutions [2] and residual modules [8]. Residual modules change the mapping between two adjacent layers such that the difference between desired features and the original feature map be- comes the learned feature. The desired features D(x) are therefore changed to solve an easier learning problem L(X) such that:

$$D(x) = L(x) + x \quad (1)$$

Since our first architecture removed the fully connected dense layer, we now have further reduced the number of parameters by deleting many of the parameters from the convolutional layers. This was accomplished by using depth-wise separable convolutions. There are 2 different layers of deep-separable convolutions: point-wise convolutions and depth-wise convolutions. These layers’ main purpose is to disassociate the channel cross correlations from the spatial cross correlations [1]. This is done by using one D X D filter on each of the M input channels and then applying N 1 X 1 X M convolution filters to combine the input channels into output channels. Applying 1 X 1 X M convolution will combine each value in the feature map, ignoring the channel’s spatial relationships. When compared with the standard convolutions, depth-wise separable convolutions reduce the computation by a factor of 1/N + 1/D2 [2].

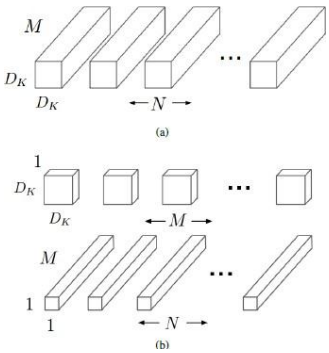

Fig.2. Differences between (a) standard convolutional layer and (b) depthwise separable convolutional layer.

Figure 2 shows the difference between a separable depth-wise convolution and a normal Convolution layer.

The architecture that we have finalized is a fully-convolutional neural network consisting of four residual depth-wise separable convolutions and a batch normalization operation and a ReLU activation function follow each convolution. A global average pooling and a soft-max activation function are applied in the last layer to produce the output prediction. This architecture contains approximately 60,000 parameters; this is 10x less than our initial naive implementation, and 80X less than the base CNN. Figure 3 shows mini-Xception, our final architecture. This architecture achieved an accuracy of 94% in the task of gender classification. This corresponds to a decrease of just 1% compared with our initial implementation. Upon testing this model on the FER-2013 data set for the emotion classification task, the model achieved the same accuracy of 66 percent. By reducing the computational cost of our architecture, we can now join the two models and consecutively use them on an image without any significant increase in computational time. The complete pipeline which includes the open CV module for face detection, the modules for emotion classification and gender classification takes about 0:15 ms on an i7-8850H CPU.

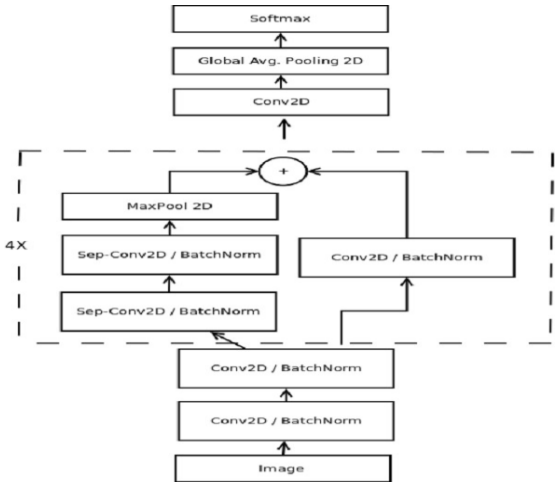

Fig.3. Proposed mini-Xception Architecture

Compared to Tang’s original architecture, this corresponds to a speedup of 1.5x.

4. Result

The algorithm classifies faces by looking at characteristics such as teeth, the frown, the eyebrows, the widening of eyes, etc. and within the same class, each feature remains constant. The results show that our CNN can interpret human features and gives a generalized understanding of it. The results help computers understand emotions with human like accuracy.

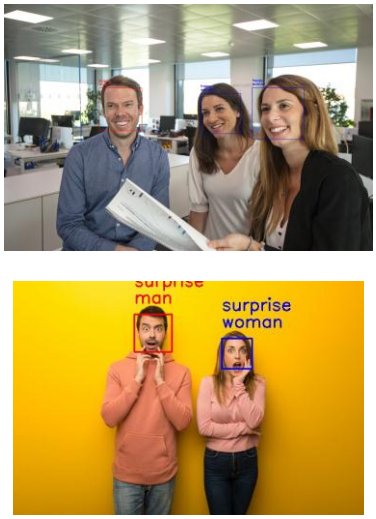

Fig.4. Results of the classification model. Blue colour represents the assigned class woman and red the class man.

5. CONCLUSION

We propose a CNN for real time facial expression classification. Our proposed model has been purpose built to lessen the parameters required. This was done by ruling out the connected layers. We also reduced the number of parameters in the other layers with the help of depthwise separable convolutions. Doing this decreases the number of parameters by 80 times. Our model does three tasks in real time, i.e. facial recognition, emotion classification and gender classification. The system is very fast and achieves human level accuracy.

REFERENCES

[1] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258).

[2] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

[3] Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, J. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning (pp. 173-182).

[4] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

[5] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[6] Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., ... & Zhou, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59-63.

[7] Glorot, X., Bordes, A., & Bengio, Y. (2011, June). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323).

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[9] Rothe, R., Timofte, R., & Van Gool, L. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2-4), 144-157.

[10] Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

[11] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[12] Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239.

[13] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[14] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.