IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 15, Issue 3, Ser. I (May - June 2020), PP 37-46
www.iosrjournals.org

A Convolutional Neural Network for Real-time Face Detection and Emotion & Gender Classification

Md. Jashim Uddin[1], Dr. Paresh Chandra Barman[2], Khandaker Takdir Ahmed[3], S.M. Abdur Rahim[4], Abu Rumman Refat[5], Md Abdullah-Al-Imran[6]

[1,3]*Assistant Professor Dept. of ICT, Islamic University, Kushtia-7003, Bangladesh,*
[2]*Professor Dept. of ICT, Islamic University, Kushtia-7003, Bangladesh,*
[4]*Assistant Professor Dept. of EEE, Islamic University, Kushtia-7003, Bangladesh,*
[5,6]*Student (M.Sc) Dept. of ICT, Islamic University, Kushtia-7003, Bangladesh.*

Abstract: *We implement a general Convolutional Neural Network (CNN) to design a real-time model and validates our model by creating a real-time vision that accomplishes the task of face detection, gender, and emotion classification simultaneously. We got accuracies of 95% in the IMDB-WIKI age and gender dataset and 66% in the FER emotion recognition dataset. We have used a real-time guided back-propagation technique to visualize the weighed of real-time CNN that uncovered the dynamic weight change and evaluate the learning feature. We think in the modern CNN architecture regularization and visualization of previously hidden layer features are necessary to reduce the gap between slow performances and real-time architecture.*

Keywords: *Back Propagation, Convolutional Neural Network (CNN), Computer Vision (CV), Emotion Detection, Face Detection, Gender Classification.*

Date of Submission: 14-05-2020 Date of Acceptance: 29-05-2020

I. Introduction

Over the last decade, the rate of image uploads to the internet has grown at a nearly exponential rate. This new-found wealth of data has empowered computer scientists to tackle problems in computer vision that was previously either irrelevant of intractable. Consequently, we have witnessed the dawn of highly accurate and efficient facial detection e.g. identify the emotional state or deduce gender. Interpreting correctly any of these elements using machine learning (ML) technique has proven to be complicated due to the high variability of the sample within each task [4]. This leads to models with millions of parameters trained under thousands of samples [5]. Furthermore, the human accuracy for classifying an image face in one of seven different emotions is 65 ± 5 %.

Moreover, the state-of-the-art functions in images related jobs such as image classification [2] and object detection are all based on the Convolutional neural network (CNN). These jobs require CNN architecture with millions of parameters, therefore their real-time systems become unfeasible. Because of this, we proposed and designed a general CNN building for designing real-time CNNs that implementation has been validated in a real-time facial expression system that availabels for face detection, gender classification and that achieves human-level performance when classifying emotion.

Now day computer vision research is more relaxed in the practical application in ranging object detection, image classification. Gender classification and Facial emotion recognition have dawned more interest in practical because of many applications ranging from human behavior understanding, mental disorders detection, synthetic human expressions, etc. That being said, this problem is also difficult. This problem is usually split into different sub-problem to make easier to with mainly face detection in an image that can be performed some tasks in between such as frontolysis face or extracting additional from an image.

Our main target is to provide a robust system that can perform some of the following tasks like a human performance like facial emotion recognition and gender classification that capable of working with any kind of images and real-time scenario with a human face. Finally, we have to establish a benchmark for the task based on state-of-the-art network architectures and show that chaining the prediction. of gender with that of emotion can improve overall accuracy. We are working with the back-propagation algorithm and soft-max activation function and ReLUs.

Figure 1.1 shows a general approach of facial emotion recognition and gender classification. The human face is first given to the system as image input. The input image is the first processing to detect face and remove noise. This is done by many filters and data augmentation. After taking processing human face from our real-time input image feature is extracting from the image face to differentiate it with others that are done by the classification part of our architecture.

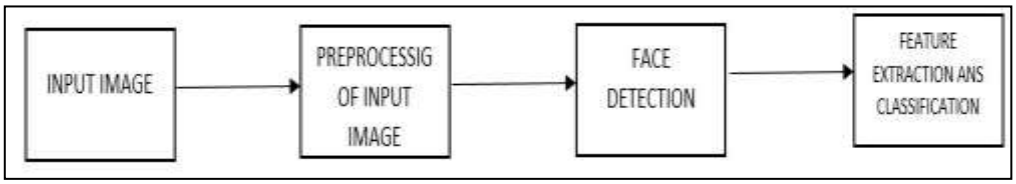

Fig.1.1. A general approach for real-time face detection, emotion recognition and gender classification.

The rest of this paper is organized as follows. Related works are described in section II. A basic face detection idea is given in section III. Basic Operation and training process of CNN are described in section VI. The proposed method is described in section VII. Experimental results are described in section VIII. Finally, conclusions are given in Section IX.

II. Related Works

Several works have been done so far for real-time face detection, facial emotion recognition, and gender-age classification [1,4,5]. For this project, we reviewed the current literature on convolutional face detection and gender and classification and facial emotion recognition [1,2,3,4,5,6,7]. We found that convolutional face detection and gender & emotion classification is still evolving as a technology, despite outranking other face detection and gender classification methods For the free availability of datasets and pre-trained networks, it is possible to make a functional implementation of a deep neural network without permition to specialist hardware. Pretrained networks can also be used as a starting point for training new networks, decreasing costly training time such as vgg16 [9] and inception v3 [10].

III. Face Detection

The Facial appearance detection and recognition system perform the three learning stages in just one Convolution neural network (CNN). The proposed function operates in two main phases: training and test. During training, the system receives a training data comprising gray-scale images of faces with their respective expression id and eye center locations and learns a set of weights for the network. To ensure that the training performance is not expected by the order of presentation of the examples, a few images are separated as validation and are used to choose the best set of weights out of a set of training performed with samples presented in different orders. During the test, the system receives a gray-scale image of a face along with its respective eye center locations and outputs the predicted expression by using the neural network weights learned during training.

Face Detection is a technology that used in different applications that detect human faces in digital images. Face detection also used for the psychological process by which humans locate and attend to faces in a visual scene We used OpenCV to catch the live image. Here, for detection the human faces, we used the Haar-Cascade image processing method. We saw that there was a situation where it didn't detect the human faces in the live images for the lack of contrast. So, we used histogram equalization to improve detection by increasing contrast. Haar-cascade: Face detection using Haar-cascade is based upon the training of a Binary classifier system using the number of positive images that represent the object to be recognized (such as faces of different peoples at the different scene) and even large number of negative images that indicate objects or feature not to be detected (images that are not human faces but can be anything else like a table, chair, wall, etc.) Actual Image Extracted human face.

IV. Gender Classification

In Gender Classification two of the key facial functions are age and gender, play a very vital role in social interactions, making age and gender estimation from a single human face image an important job in intelligent applications, such as human-computer interaction, access control, marketing intelligence, law enforcement, visual surveillance, etc. A preprocessing method which can collect facial and other physical characteristics from the image, a neural network which can classify the gender from the ensemble, an algorithm which can integrate the part-based information and ensemble, based on the database that connects the peculiarity of these physical features for females and males should work.

V. Emotion Classification

Emotion Classification Images can both express and affect people's emotions. It is interesting and essential to understanding what emotions are conveyed and how they are implied by the visual content of images. Inspired by the recent success of deep convolutional neural networks (CNN) in visual recognition, there explore simple, yet effective deep learning-based methods for image emotion analysis. we extract attributes using the fine-tuned CNN at the different addresses at multiple levels to capture both the global and local

information. The features at different locations are aggregated using the Fisher Vector for each level and concatenated to form a compact representation.

VI. Basic Operations and training process on CNN

Primary Operations on CNN Convolutional Neural Network (CNN or ConvNet) is a class of deep artificial neural networks that have successfully been functional in analyzing visual imagery. Simple ConvNet for Emotion & Gender classification could have the architecture [INPUT - CONV – ReLU - POOL - FC] [13]. There are four main operations in the ConvNet. Figure 6.1 shown the basic CNN architecture for classification where the first portion describes as the feature extraction part and the next portion describe as the classification part.

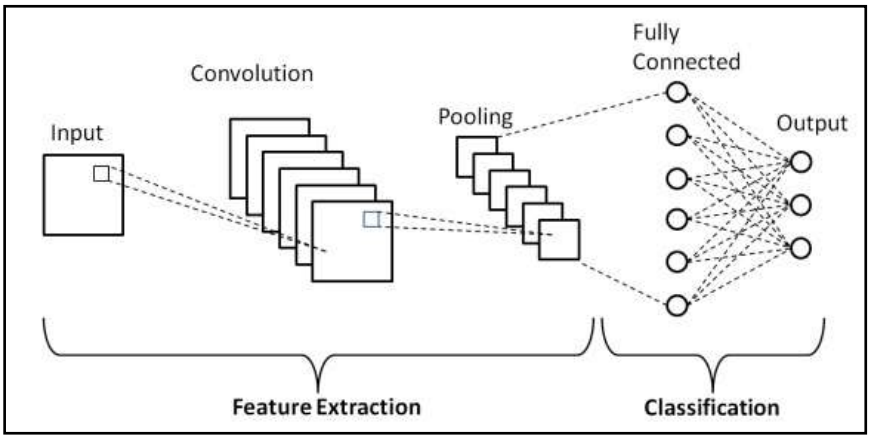

Fig. 6.1: Basic CNN Architecture [13].

A. Steps in the training process of CNN

Step 1: Initialize all filters and attributes/weights with random values
Step 2: The method takes a training image as input, goes through the forward propagation step (convolution, ReLU, and pooling operations with forwarding propagation in the completely connected layer and searches the output probabilities for each class.
Step 3: Calculate the total error at the output stage
Total error = $\sum 1/2$ (Target Probability – Outcome Probability)2
Step 4: Use Back-propagation to evaluate the gradients of the error concerning all weights in the method and use gradient descent to update all filter (values)/weights and parameter values to minimize the output error.
Step5: again steps 2-4 with all images in the training set.

VII. Proposed Method

Due to the classification of gender and emotion from each image, a group of steps needs to take. Such as dataset preparation, preprocessing, and powerful classification model. Each of the step's performance causes an effect on the total classification accuracy.

Work Flow of Real-time face detection and gender & emotion recognition is a robust complex problem in computer vision because of the real-time image frame. At first, we have to take a real-time video frame then convert it as an image and we have extract face from image to detect a human face. After extracting face, we consider each face part of the image as a full image for further process. Each extracting face image is then providing as input to preprocess step of classification model and each preprocessing step takes some operation on its input to resize as model input and data augmentation as input to our proposed convolutional neural network (CNN) model for classification of the emotion and gender. The resulting label that is the output of the CNN is then used for making a description of gender {"man" or "women"} and facial emotion classification {"angry", "disgust", "fear", "happy", "sad", "neutral"}. The working diagram is given in figure 7.1.

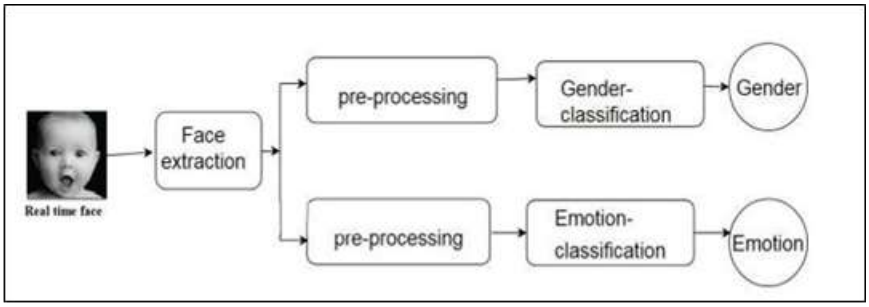

Fig. 7.1: Working principle diagram of our proposed approach.

Feature collection of each face is the most hard part for any face due to its face shape and structure in real-time. To solve this problem, we apply a Convolutional neural network that doesn't require any predefined feature for the classification of the specific human image.

Proposed CNN Architecture We propose two models that we evaluated following their test accuracy and number of parameters. Both proposed models were designed with the idea of creating the best accuracy over many parameters ratio. Reducing the number of parameters that help us to overcome two important problems. First, the use of small CNN's alleviates us from slow performances in hardware constrained systems. And second, the reduction of parameters gives a better generalization under Occam's razor framework. Our first method relies on the idea of demolishing the fully connected layers. The second architecture adds the inclusion of the combined depth-wise separable convolutions and residual modules and the deletion of the fully connected layer. These methods were trained with the ADAM optimizer [8]. Following the previous method schemas, our primary architecture used Global Average Pooling to completely remove any fully connected layers. This was earned by having in the last convolutional layer the same number of feature maps as the number of classes and applying a softmax activation function to each reduced feature map. Our primary proposed method is a standard fully-convolutional neural network composed of 9 convolution layers, ReLUs [10], Batch Normalization [9], and Global Average Pooling.

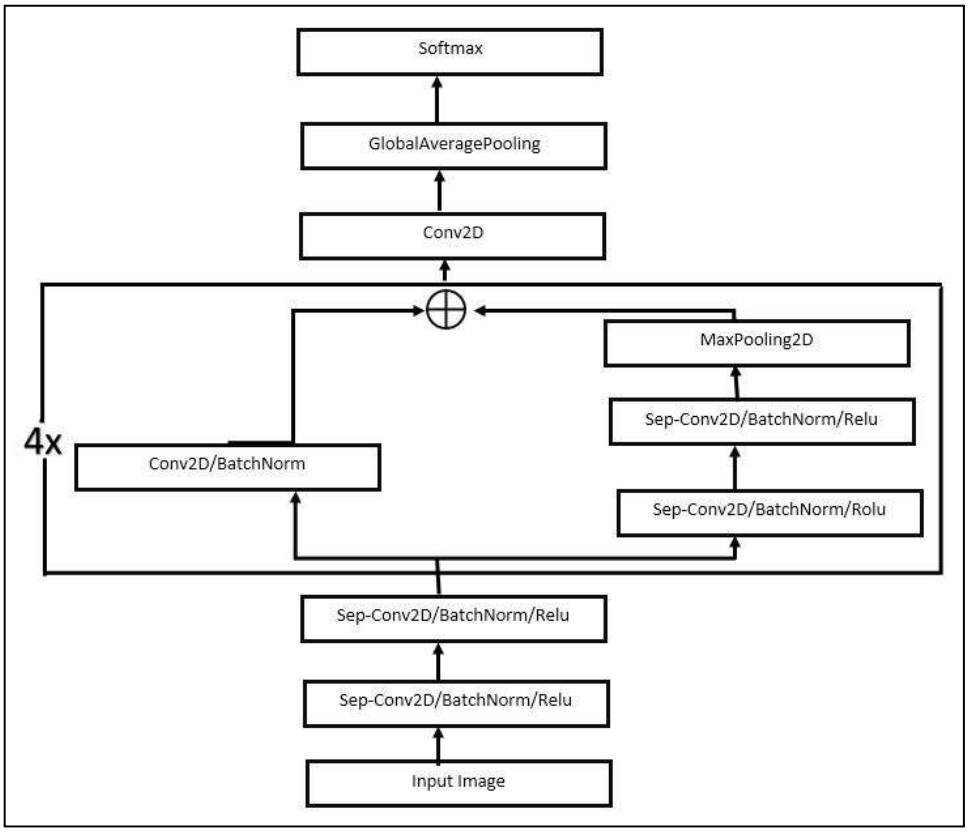

Fig. 7.2: Our proposed model for real-time classification.

This model accommodates approximately 600,000 frameworks. It had been trained on the IMDB gender data set, which contains 460,723 RGB images where each image divides into the class "woman" or "man", and it earns an accuracy of 95% in this data set. We also validated this model in the FER-2013 dataset. This data set accommodates 35,887 grayscale images and each image relate to one of the following classes {"angry", "disgust", "fear", "happy", "sad", "surprise", "neutral"}. Our initial model earned an accuracy of 66% in this data set. We will be recommended for this model as "sequential fully-CNN".

Our second method is inspired by the Xception [3] demo. This demo adds the use of residual modules [11] and depth-wise separable convolutions [12]. Residual modules revise the desired mapping between two subsequent layers so that the learned features become the difference of the desired features and the original feature map. Consequently, the desired features H(x) are modified to solve an easier learning problem F (X) such that:

$$H(x) = F(x) + x \quad \ldots\ldots (1)$$

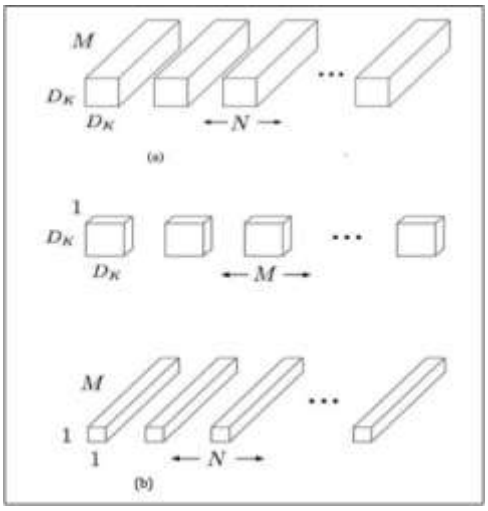

Fig. 7.3: Difference between (a) standard convolutions and (b) depth-wise separable convolutions [12].

We reduced further the number of attributes by eliminating them now from the convolutional layers. This method was done by the use of depth-wise separable convolutions. Depth-wise defferential convolutions are cosists of two different layers: depth-wise convolutions and pointwise convolutions. The vital aim of these layers is to separate the spatial cross-correlations from the channel cross-correlations [3]. They do this by first applying a D × D filter on every M input channel and then applying N 1 × 1 × M convolution filters to combine the M input channels into N output channels. Without considering their spatial relationships within the channel applying 1 × 1 × M convolutions combines each value in the feature map. Depth-wise separable convolutions reduce the computation concerning the standard convolutions by a factor of N1 +1D 2 [12]. A visualization of the distinction between a normal Convolution layer and a depth-wise separable convolution can be observed in Figure 7.3.

Our final method is a totally convolutional neural network that contains 4 residual depth-wise separable convolutions where each convolution is followed by a batch normalization operation and a ReLU activation function. The last layer applies a global average pooling and a soft-max activation function to create a prediction. This architecture has approximately 60; 000 parameters; which corresponds to a reduction of 10× when compared to our initial naïve implementation, and 80× when compared to the original CNN. Figure 3 displays our complete final architecture which we refer to as miniXception. This architecture obtains an accuracy of 95% in the gender classification task. Which corresponds to a reduction of one percent for our initial implementation.

Furthermore, we tested this method in the FER-2013 dataset and we gained the same accuracy of 66% for the emotion classification task. Our final method weights can be stored in an 855 kilobytes file. We are now able to join both models and use them consecutively in the same image without any serious time reduction, by reducing our architecture's computational cost. Our complete pipeline including the OpenCV face detection module, the gender classification, and the emotion classification takes 0:22 ± 0:0003 ms on an i55200U CPU. When compared to the original architecture of Tang this corresponds to a speedup of 1:5×. We also combined to our implementation a real-time guided back-propagation visualization to observe which pixels in the image activate an element of a higher-level feature map. Given a CNN model with only ReLUs as activation functions for the intermediate layers, guided-back propagation takes the derivative of every element (x; y) of the input

image I concerning an element (i; j) of the feature map fL in layer L. The reconstructed picture R filters all the negative gradients; consequently, the remaining gradients are chosen such that they only increase the value of the chosen element of the feature map. Following [10], a fully ReLU CNN reconstructed image in layer l is given by Ri;

$$j = (Ri; j\ l+1 > 0) * Ri; j\ l+1 \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (2)$$

VIII. Experimental Result and Discussions

In the future, we will further improve the algorithm. Blind texts it is an almost unorthographical life One day however a small line of blind text by the name of Lorem Ipsum decided to leave for the far World of Grammar by adding these effects.

A. Dataset

In this project, we used the FER-2013 emotion dataset [4] for emotion classification task and we used the IMDB-WIKI age and gender dataset [5] for gender classification tasks simultaneously. The FER-2013 dataset in figure 8.1 consists of 48x48 pixel gray-scale 35886 images of faces. The face is more or less centered and occupies about the same amount of space in each image so that the faces have been automatically registered. The task is to categorize each face based on the emotion shown in the facial expression into one of seven categories with the following classes {"0-angry", "1-disgust", "3-fear", "4-happy", "5-sad", "6-neutral"}

Fig. 8.1: Samples of the FER-2013 emotion dataset [4].

Fig.8.2: Samples of the IMDB dataset [5].

In IMDB dataset in figure-8.2 consists of a total of 460,723 face images from 20,284 celebrities. We used only face image for our training purpose of gender classification task of two classes as following {"0-Women" and "1-Man"}.

B. Experiment After training of the network

We use 10% images in both datasets [2,5] as a test image for the task of recognition of emotional expression and gender. Our testing result can be observed as figure 8.3 where our proposed model can classify gender as two-class properly but all kind of our training images is such as western actors, models, politician, etc. our emotion recognition performed as like human where our proposed method can understand human behavior of seven class such angry, happy, sad, fear, surprise, neutral, etc.

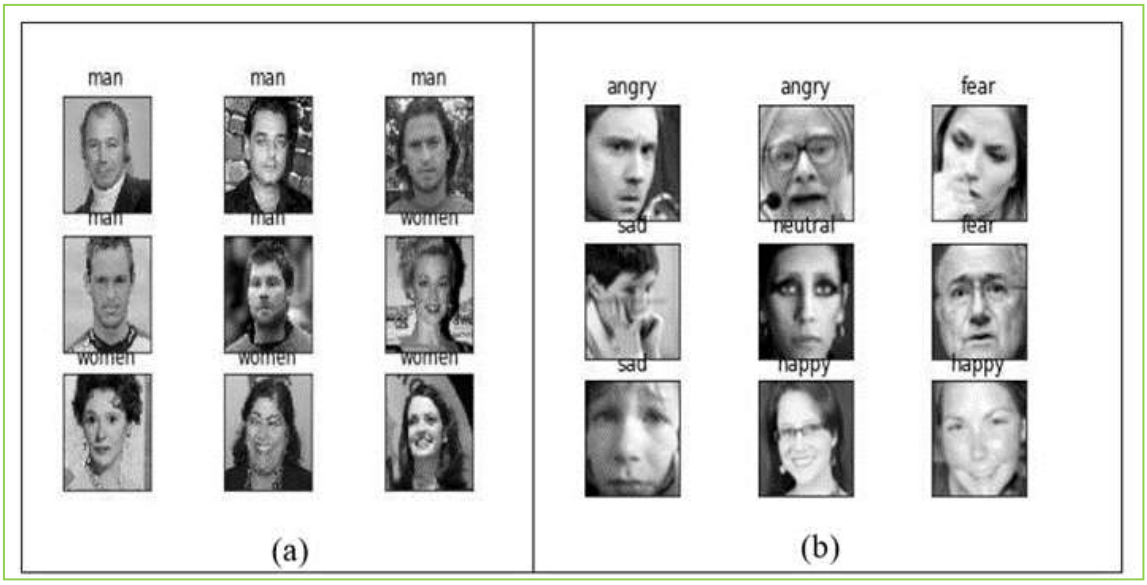

Fig.8.3: Predicted results of the emotion and gender classification on testing data. (a) Result of gender classification on the IMDB dataset [5]. (b) result of emotion recognition on the FER-2013 dataset [2].

C. Experiment Result Experiment results

Experiment results of the real-time emotion classification task in unseen faces can be observed in Figure 8.4. Our complete real-time pipeline including face detection, emotion, and gender classification has been fully integrated with our Intel Core-i5 5200U processor.

Fig.8.4: Results of the real-time emotion classification

An example of our complete pipeline can be seen in Figure 8.5 in which we provide emotion and gender classification

Fig.8.5: Results of the combined gender and emotion recognition. The color blue represents the assigned class man and red the class woman.

In Figure 8.6 we provide the confusion matrix results of our emotion classification of our proposed mini-Xception model.

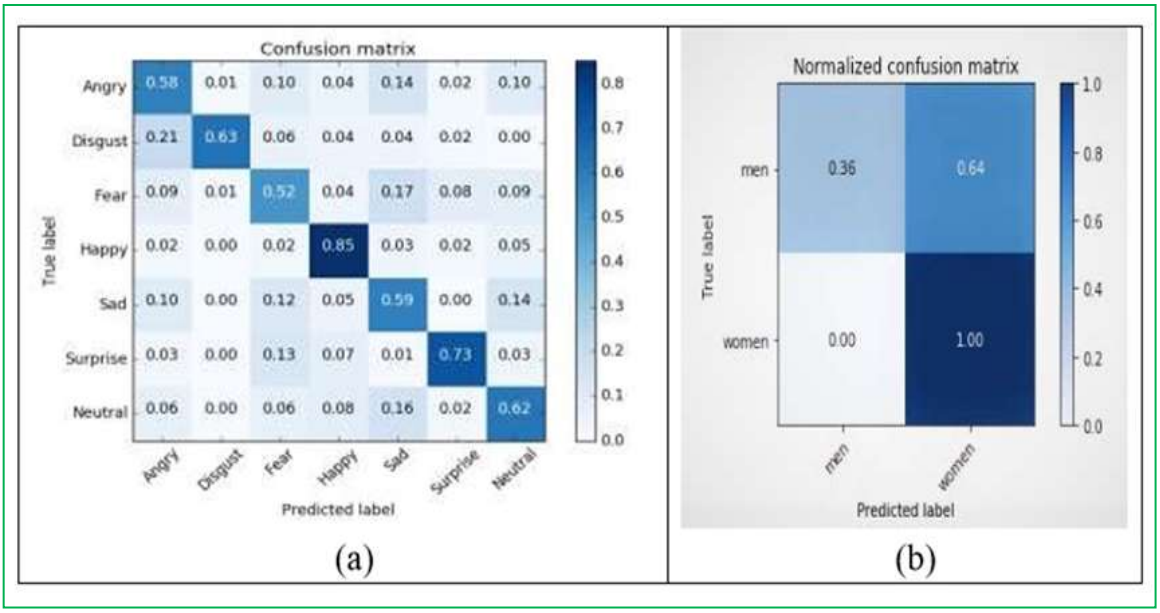

Fig.8.6: Normalized confusion matrix of our mini-Xception network (a) confusion matrix for facial emotion recognition (b) confusion matrix for gender recognition.

We can observe several common misclassifications such as predicting "sad" instead of "fear" and predicting "angry" instead of "disgust".

A comparison of the learned features between several emotions and both of our proposed models can be observed in Figure 8.7. The white areas in figure 8b correspond to the pixel values that activate a selected neuron in our last convolution layer. The selected neuron was always selected for the highest activation.

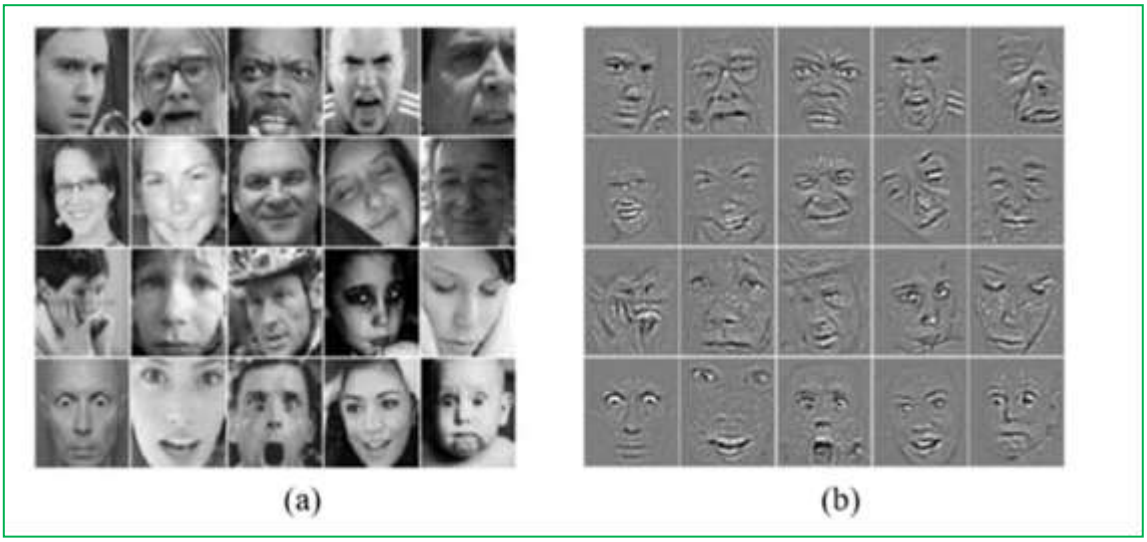

Fig. 8**.7:** All sub-figures contain the same images in the same order. Every row starting from the top corresponds respectively to the emotions: angry, happy, sad, and surprise (a) Samples from the FER-2013 dataset (b) Guided back-propagation visualization of our proposed mini-Xception model.

By considering features such as the frown, the teeth, the eyebrows and the widening of one's eyes, and that each feature remains constant within the same class. , we can find that CNN learned to get activated. These examined results reassure that CNN learned to interpret understandable human-like characteristics that provide generalizable elements. These interpretable results have helped us understand several common misclassifications such as persons with glasses being classified as "angry" in figure 8.8.

When it believes a person is frowning and frowning features get confused with darker glass frames, this happens since the label "angry" is highly activated. Moreover, we can also observe that the features learned in our proposed mini-Xception model are more interpretable than the ones learned from sequential fully-CNN. Consequently, the use of more parameters in our naive implementations leads to less robust features.

Fig. 8.8: Results of the misclassification in real-time emotion and gender recognition.

IX. Conclusion

We have proposed and tested general building designs for creating real-time CNNs. Our proposed architectures have been systematically built to reduce the number of parameters. We began by eliminating the fully connected layers and by reducing the number of parameters in the remaining convolutional layers via depth-wise separable convolutions. We have shown that our proposed models can be stacked for multi-class classifications while maintaining real-time inferences. Specifically, our vision system can perform face

detection, gender classification, and emotion classification in a single integrated module. We have achieved human-level performance in our classification tasks using a single CNN that leverages modern architecture constructs. Our architecture reduces the number of parameters 80× while obtaining favorable results.

Finally, we developed a visualization of the learned features in CNN using the guided back-propagation visualization. This visualization technique can show us the high-level features learned by our models and discuss their interpretability.

Acknowledgments

This work was supported by Dept. of ICT, Islamic University, Kushtia-7003, Bangladesh.

.

References

[1]. Arriaga, Octavio et al. "Real-time Convolutional Neural Networks for Emotion and Gender Classification." CoRR abs/1710.07557 (2017): n. pag.
[2]. Goodfellow, Ian J., et al. "Challenges in Representation Learning: A report on three machine learning contests." Neural networks: the official journal of the International Neural Network Society 64 (2013): 59-63.
[3]. Chollet, François. "Xception: Deep Learning with Depthwise Separable Convolutions." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1800-1807
[4]. Levi, Gil, and Tal Hassner. "Age and gender classification using convolutional neural networks." 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2015): 34-42.
[5]. Rothe, Rasmus et al. "DEX: Deep Expectation of Apparent Age from a Single Image." 2015 IEEE International Conference on Computer Vision Workshop (ICCVW) (2015): 252-257.
[6]. Gurnani, Ayesha et al. "VEGAC: Visual Saliency-based Age, Gender, and Facial Expression Classification Using Convolutional Neural Networks." CoRR abs/1803.05719 (2018): n. pag.
[7]. Wang, X., Guo, R., Kambhamettu, C.: Deeply-learned feature for age estimation. In: WACV. (2015)
[8]. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014
[9]. Ioffe, Sergey, and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." ICML (2015).
[10]. Glorot, Xavier et al."Deep Sparse Rectifier Neural Networks." AISTATS (2011).
[11]. He, Kaiming et al."Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.
[12]. Howard, Andrew G. et al. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." CoRR abs/1704.04861 (2017): n. pag.
[13]. Course note of cs 231 at Stanford University: [http://cs231n.github.io/convolutional-networks/]
[14]. Nair, Vinod, and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines." ICML (2010).
[15]. Krizhevsky, Alex et al. "ImageNet Classification with Deep Convolutional Neural Networks." Commun. ACM 60 (2012): 84-90.
[16] Mao, Junhua et al. "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)." CoRR abs/1412.6632 (2015): n. pag.
[17] Shankar, Sukrit et al. "DEEP-CARVING: Discovering visual attributes by carving deep neural nets." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015): 3403-3412.

Khandaker Takdir Ahmed, et. al. "A Convolutional Neural Network for Real-time Face Detection and Emotion & Gender Classification." *IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)* 15(3), (2020): 37-46.