The Training of Different CNN Architectures for Human Face Detection and Analysis

1st Shehab Ahmed Bassiouni
Faculty of Computer and Information Sciences
Computer Science dept.
Ain Shams University
Cairo, Egypt
swe.shehab.bassiouni@gmail.com

2nd Omar Essam Abdelhadi
Faculty of Computer and Information Sciences
Computer Science dept.
Ain Shams University
Cairo, Egypt

3rd Mohammed Mas'ad Saad
Faculty of Computer and Information Sciences
Computer Science dept.
Ain Shams University
Cairo, Egypt

4th Mohammed Adel Abdullhamid
Faculty of Computer and Information Sciences
Computer Science dept.
Ain Shams University
Cairo, Egypt

5th Ibrahim Abdelghani Mansour
Faculty of Computer and Information Sciences
Computer Science dept.
Ain Shams University
Cairo, Egypt

Under Supervision of:
Dr.Nermine Naguib, Dr.Assma Bahi
Faculty of Computer and Information Sciences
Computer Science dept. Ain Shams University
Cairo, Egypt

***Abstract*—This paper presents the training methods and results of different convolutional neural networks models on different datasets to build a system capable of detecting human faces and predicting age, gender, emotions, and ethnicity of the detected faces. The system achieved an accuracy of 71% for age classification, 81% for gender classification, 86% for Ethnicity classifications and 67% for emotions classifications. The method used to detect human face is haar cascade algorithm.**

***Index Terms*—convolutional neural networks, face recognition, classification**

I. INTRODUCTION

In recent years, the need for systems capable of detecting and analyzing the human face is increasing rapidly with the wide possible application of it in our life. Those systems can be used in many areas like marketing by analyzing customer emotions from facial image to measure the degree of satisfaction or analyzing the ages of most-visiting customers etc. it can also be integrated in surveillance systems to detect specific features. With the traditional machine learning techniques, the needs of hand-crafted features and domain experts presented many challenges due to the variety of human races and features. With the rabid development of machines and big datasets, applying deep learning techniques such as neural networks became possible. Proposing AlexNet Model by Alex, Ilya and Geoffrey in their paper "ImageNet Classification with Deep Convolutional Neural Networks" [1], promoted the development of more Deep CNN models that achieved state-of-art performance in many computer vision tasks. The aim of this paper is to improve, optimize and train CNN architectures on a variety of datasets to achieve responsible results in detecting and analyzing human face.

II. RELATED WORK

There are many approaches used to detect human faces. One of these methods is haar cascade algorithm proposed in [2]. The process involves sliding a window of varying sizes across the image, at multiple scales. At each window position, a set of classifiers is applied to evaluate whether the features within the window match the learned patterns of a face. Another state-of-art method is YOLO algorithm proposed in [3] that works by dividing the image into a grid of cells and scan the image and assign confidence score to each bounding box indicating the likelihood of it containing a face. Various CNN architecture proposed to analyze human face. Example of those architectures are models in Table 1.

TABLE I
SUMMARY OF THE CNN MODELS

REF.	***Task***	***Accuracy***
[4]	Gender Prediction	97.45%
[5]	Gender Prediction	90.35%
[6]	Emotions Prediction	55%
[7]	Emotions Prediction	65%
[5]	Age Prediction	80.1%
[8]]	Ethnicity Prediction	76%

III. Datasets

There are many available datasets to train CNN models on different tasks in computer vision. Most high-quality datasets with less noise and misclassifications are private datasets. Such datasets are hard to access. In our training we used free available datasets. In the training of Age model, UTK- FACE and Facial Age datasets were used, UTK-Face is a large-scale face dataset consisting of over 20,000 face images with annotations of age, gender, and ethnicity. The images cover large variation in pose, facial expression, illumination, occlusion, resolution. Facial Age is a dataset generated from WIKI-ART containing human faces with various ages split into 7 Age categories starting from age of 1 to 116. In the training of Gender Model, UTK-FACE and B3FD datasets were used. B3FD dataset is a cleaned version of both IMDB and WIKI datasets, it contains images of celebrities of different races. In the training of Ethnicity Model, both Ethnicity-Aware and Arab Celebrities Faces datasets were used. Ethnicity-Aware dataset is a large-scale dataset consist of 1.3M images from 28K celebrities from different races split into Caucasian, Indian, Asian and African. Arab Celebrities Faces Dataset consist of 7453 face images of Arab Celebrities. In the training of Emotions Model, FER-2013 dataset was used which containing 35887 gray scale images of human faces with different ages and genders expressing different emotions, the images are split into 7 categories representing the main human emotions which are anger, disgust, fear, happiness, sadness, surprises and Neutral.

IV. Preprocessing

All datasets were cleaned using haar cascade algorithm to remove any images that do not contain a human face. Data Augmentation techniques applied on the datasets to increase the datasets size and decrease overfitting during training by making the models more robust to changing in the data by applying 20% rotation, 0.2 width shift range, 0.2 height shift range, 0.2 shear range, 0.2 zoom range and horizontal flip with "nearest" fill mode. All images in the datasets have been normalized with output value in range [0,1] and mean-centering have been applied. For Age Model the datasets split into 7 categories each category groups the images of humans with high similarity, those categories are, ages from 1 to 2, 3 to 9, 10 to 20, 21 to 27, 28 to 45, 46 to 65 and 66 to 116, All the images have been converted to grayscale and resized to 128x128, the final dataset split to 70% for training and 30% for validation and testing. In Gender Model datasets images split into 2 categories male and female and images resized to 128x128x3, the final dataset split to 75% for training and 25% for validation and testing. In Emotions Model dataset, all images have been resized to 48x48x1, final dataset split to 70% for training and 30% for validation and testing. In Ethnicity Model datasets, the datasets split into 5 categories Caucasian, African, Indian, Asian and middle eastern and all images have been resized to 224x224x3, the final dataset split to 70% for training and 30% for validation and testing.

V. CNN ARCHETICTURES

A. Age and Gender Model

The CNN architecture proposed by Juel in [9] was used in age and gender models. In Age model, the last layer had softmax activation with 7 units. In gender model, the activation changed to sigmoid activation in the last layer.

B. Ethnicity Model

In ethnicity model a pre trained VGG-16 architecture with VGG-Face weights was used by fine tuning it. The architecture was fine-tuned by adding Global Average Pooling layer instead of the flatten layer, a dense layer with 1024 units and RelU activation, a dropout layer with probability 0.5 and a dense layer with 5 units and SoftMax activation.

C. Emotions Model

The following architecture in figure 1 was used for emotions model. Input dimension is 48x48x1, all convolution layers have kernel size of 3 and relu activation with "same" padding, Max pooling layers have pool size of 2, Dense layer have 256 units with relu activation and softmax layer have 7 units.

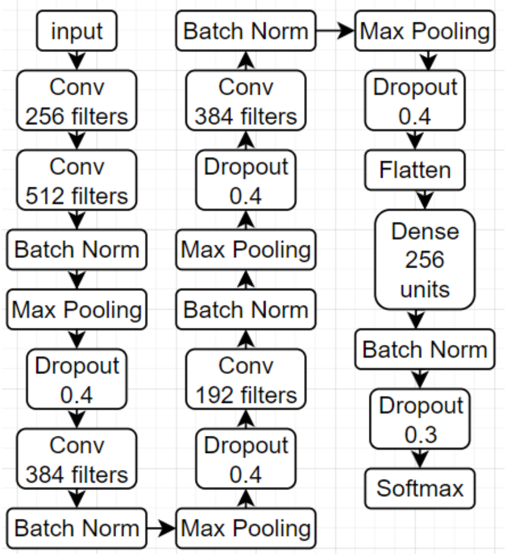

Fig. 1. Emotions Model Architecture.

VI. Training

TensorFlow was used to implement all models.

A. Loss and Validation

categorical cross entropy loss and accuracy evaluation metric were used. validation method used is cross validation.

B. Hyperparameters and Optimizer

Adam Optimizer from TensorFlow was used with the default hyperparameters and 0.001 learning rate.

C. Optimizations

early stopping and checkpoints methods from TensorFlow were implemented with 5 patience. Class weight balancing was applied to all models to avoid any biases. All models were Fitted with 100 epochs. batch sizes chosen were 32 for age and gender models and 128 for ethnicity and emotions models.

VII. Results

early stopping callback was triggered in all models, Age model stopped after 14 epochs, Gender model stopped after 20 epochs, Ethnicity model after 10 epochs, Emotions model after 69 epochs. the models. The models achieved the following results in Table 2 in the test data.

TABLE II
Test Results of the Models

Model	*Loss*	*Accuracy*
Age Model	0.75	71%
Gender Model	0.8463	81%
Ethnicity Model	0.3744	86%
Emotions Model	0.9411	67%

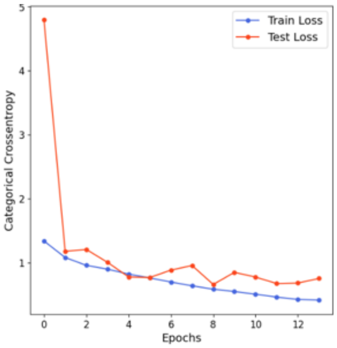

Fig. 2. Age Model Loss.

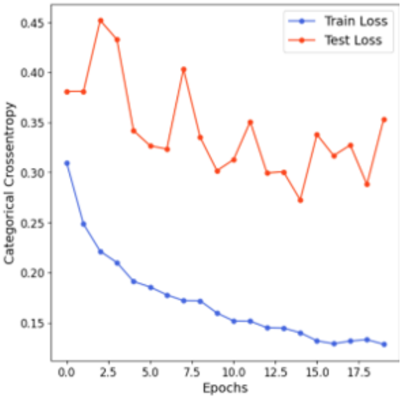

Fig. 3. Gender Model Loss.

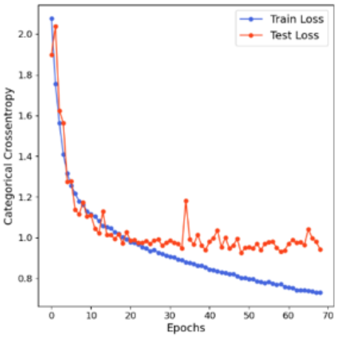

Fig. 4. Emotions Model Loss.

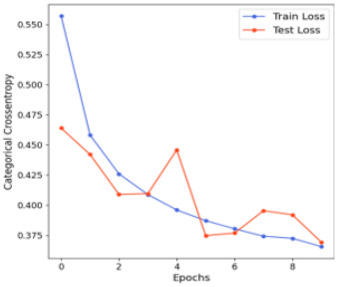

Fig. 5. Ethnicity Model Loss.

VIII. Confusion matrices

Age Model's Confusion Matrix showing some Misclassifications with the highest being 35% probably in the ages near the borders of each category because of the increased similarities. The Male class in Gender Model's Confusion matrix had the highest accurate classification with 87%. Ethnicity Model's highest Misclassifications is 10% between Indian and Middle eastern classes, Those two races have many common features which are sometimes hard for even humans to classify. Emotion Model also showed some Misclassifications with the highest being 28% between emotions that have similar patterns in the human face like angry emotion against disgust emotions and neutral emotion against sad emotions.

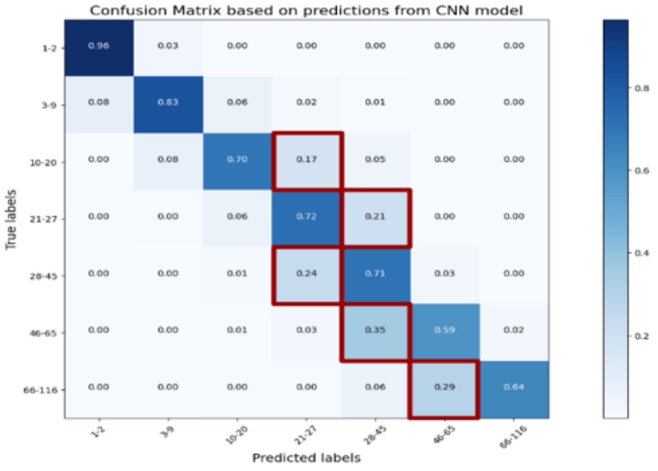

Fig. 6. Age Model Confusion Matrix.

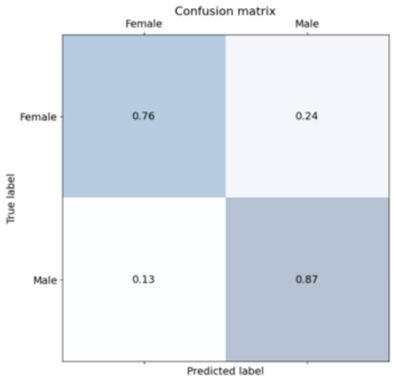

Fig. 7. Gender Model Confusion Matrix.

IX. Conclusions

The paper aimed to develop and train models capable of detecting and predicting human age, gender, ethnicity, and emotions by utilizing the power of CNN and optimizations techniques like Data Augmentation, Class weighs balancing and early stopping.The method used to detect human face is haar cascade algorithm [2]. The models achieved satisfactory results considering the high degree of noise and mislabeled data in the free datasets used and the limits of the machine that was used in the training. The models had the following accuracies:

- 71% in Age Classification.
- 81% in Gender Classification.
- 86% in Ethnicity Classification.
- 67% in Emotions Classification.

Those Models can be used in many fields, for example monitoring patients' mental health through their facial emotion, applying constraints for specific ages, and in security by integrating it into Surveillance systems. Further training will be conducted when better datasets are available.

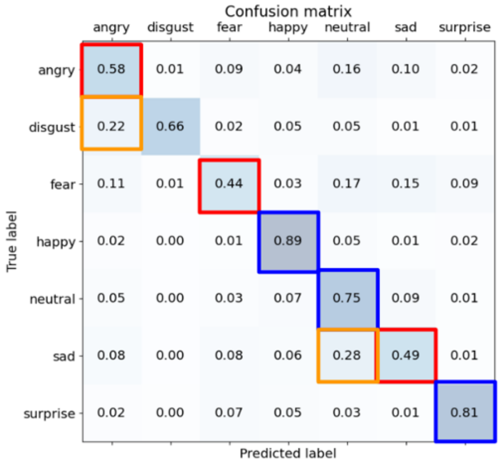

Fig. 8. Emotions Model Confusion Matrix.

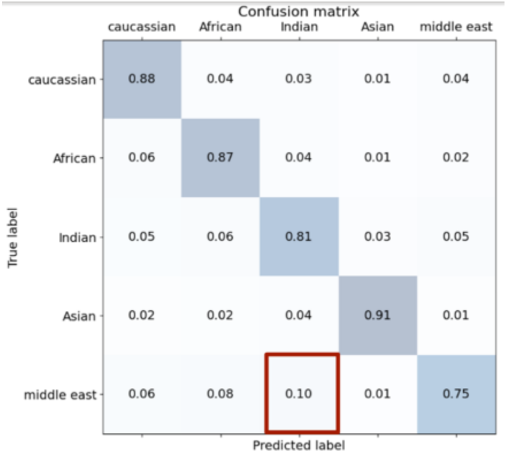

Fig. 9. Ethnicity Model Confusion Matrix.

References

[1] A. Krizhevsky, I. Sutskever,G. E. Hinton " ImageNet Classification with Deep Convolutional Neural Networks," University of Toronto.

[2] P. Viola, M. Jones, " Rapid Object Detection using a Boosted Cascade of Simple Features," CVPR conference, 2001.

[3] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE, 2016.

[4] Simanjuntak, Frans,G. Azzopardi, ""Fusion of cnn-and cosfire-based features with application to gender recognition from face images."," IEEE, 2019.

[5] Wang, Xiaofeng, A. M. Ali, P. Angelov, "Gender and age classification of human faces for automatic detection of anomalous human behaviour," IEEE, 2017.

[6] Mollahosseini, Ali, D. Chan,M. H. Mahoor, "going deeper in facial expression recognition using deep neural networks.," IEEE, 2016.

[7] Agrawal, Abhinav, N. Mittal, "Using CNN for facial expression recognition: a study of the effect of kernal size and the number of filters on the accuracy.," IEEE, 2020.

[8] Mohammad, A.S., Al-Ani, J.A., "Convolutional neural network for ethnicity classification using ocular region in mobile environment.," IEEE, 2018.

[9] J. S. Juel, "An Approach Based on Deep Learning for Recognizing Emotion, Gender and Age.," Rangamati Science and Technology University, 2022.