EasyChair Preprint

№ 9656

DCNN-Based Transfer Learning Approaches for Gender Recognition

Md Shahzeb, Sunita Dhavale, D Srikanth and Suresh Kumar

February 3, 2023

DCNN-based Transfer Learning approaches for Gender Recognition

Md Shahzeb, Sunita Dhavale, D Srikanth, Suresh Kumar

Md Shahzeb, Defence Institute of Advanced Technology (DIAT), Pune, India (e-mail: mdshahzeb034@gmail.com)
Sunita Dhavale*, Defence Institute of Advanced Technology (DIAT), Pune, India (e-mail: sunitadhavale@gmail.com)
D Srikanth, Defence Institute of Advanced Technology (DIAT), Pune, India (e-mail:dasarisrikanth@diat.ac.in)
Suresh Kumar, Defence Institute of Psychological Research (DIPR), Delhi, India (e-mail: skdipr@gmail.com)
*Corresponding Author: Sunita Dhavale, (e-mail: sunitadhavale@gmail.com)

Abstract:

Gender recognition becomes a very critical task for security agencies while assessing protest activities. At present, with the advent of GPUs, high computing machines, and Deep Convolution Neural Networks (DCCN), automated gender recognition is possible. In this research work, we explore the performance of various DCNN architectures using transfer learning approaches for gender recognition. We performed a detailed ablation study on different input sizes and on different architectures to see the trade-off between latency and the accuracy of the classification. The performance of models tested against standard dataset WIKI, UTKFace, and Adience. We explored VGG-16 and MobileNetV3 architectures for comparison against accuracy and latency parameters in order to select a model suitable for the embedded device considering their low processing and less storage capacity. Experiments conducted using standard architecture against the standard dataset by changing the resolution and fine-tuning it.

Keywords: Transfer Learning, Deep Learning, Real Time Gender Recognition, Video Surveillance, Convolution Neural Network.

1. Introduction:

Gender Recognition is very crucial at the protest site, crowded places, etc. Based on gender distribution, an authority can make preparatory steps in that place for e.g. the number of women personnel and male personnel required in case of the crowd getting violent and damaging public/private properties. This makes gender recognition tasks important in the case of video surveillance. Further, to automate many things in video surveillance, it is required to install smart cameras or embedded devices which are capable of doing gender recognition. Also in many cases, the system should work reliably in "the wild environment" providing real-time performances. Considering latency and non-availability of the internet in many places, gender recognition algorithm may require running directly on embedded devices like smart cameras or edge devices providing both accurate and real-time performance. These embedded devices generally suffer from low storage and low processing capabilities.

In this work, we use the VGG16 architecture as the benchmark for comparing the accuracy of image/video based gender recognition task as VGG-16 architecture has already proven its state-of-the-art accuracy in many image classification problems [1]. Further, we experimented with MobileNetV3 [2], light weight CNN architecture with depth-wise separable convolutions to get faster performance compared to same depth architectures. To get more generalization capability specially in the case of the Indian context, we train our algorithm with a very large dataset [3, 4, 5] using transfer learning approaches [6] that have a significant variation in faces. We compared both architectures by training them with the same datasets.

In [7], the authors proposed a new architecture with 160x160x3 resolution, a width multiplier of 0.75, and a depth of 17 layers which is based on MobileNetV2 as a reference architecture. The authors stated the accuracy of the model is 95.78% for the WIKI dataset and 84.45% for the Adience dataset [7]. However, in the case of ImageNet classification tasks, MobileNetV3 is 3.2% more accurate compared to MobileNetV2 and also has 20% reduced latency compared to MobileNetV2 [2]. Hence, we selected MobileNetV3 for training on the same standard datasets. The number of parameters in inference using MobileNetV3 is less compared to other architectures used for gender recognition; this means less latency.

The paper is organized as follows: Section 2 describes the Datasets; Section 3 describes Transfer Learning; Section 4 describes the Evaluation; Section 5 describes experimental results, and section 6 consists of the Conclusion.

2. DATASETS:

A)**WIKI:** The WIKI datasets [4] include 40,216 images collected from Wikipedia. The sample images are shown in Figure 1.

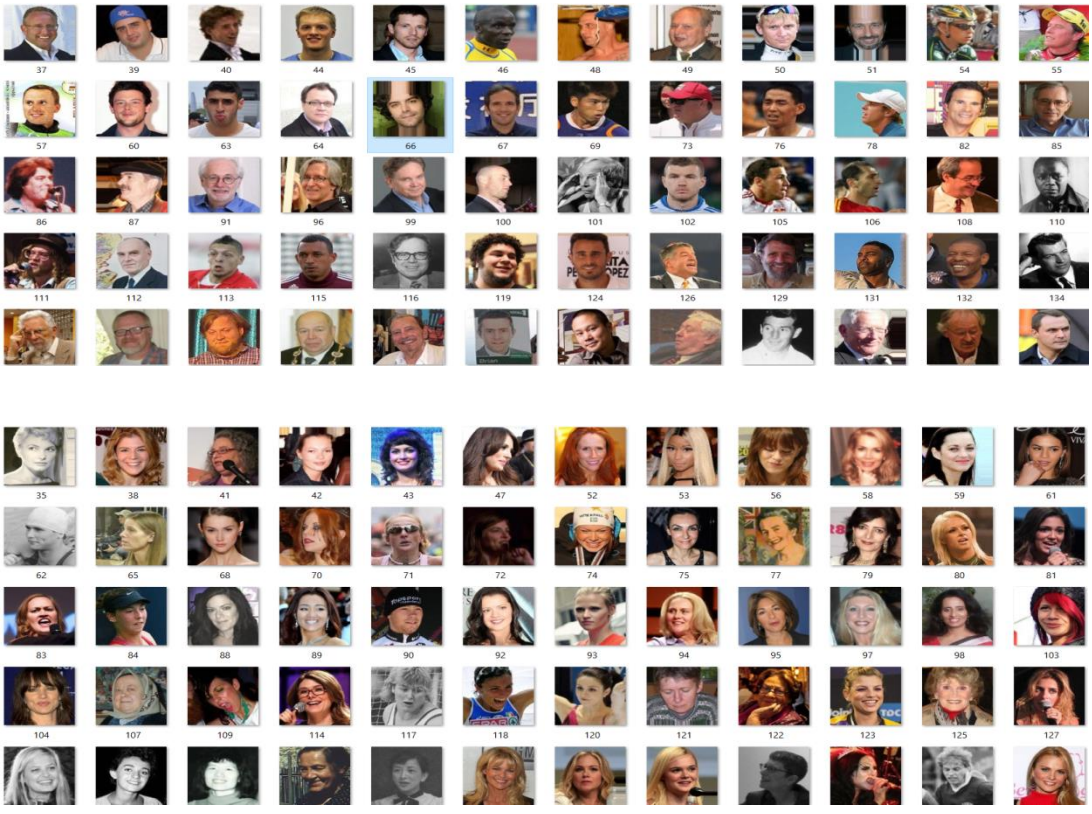


Figure 1: WIKI Dataset Sample Images

The wiki dataset is annotated with age and gender which is prepared by taking images from Wikipedia of celebrities. In the experiment, the dataset divided into 70:30 ratios for training and validation purposes.

B) Adience [5]: The Adience dataset consists of 26,580 images of 2,284 identities with gender and age labels. The images belongs to a wide variety of appearance, pose lightning condition, and image quality. The sample images are shown in Figure 2:

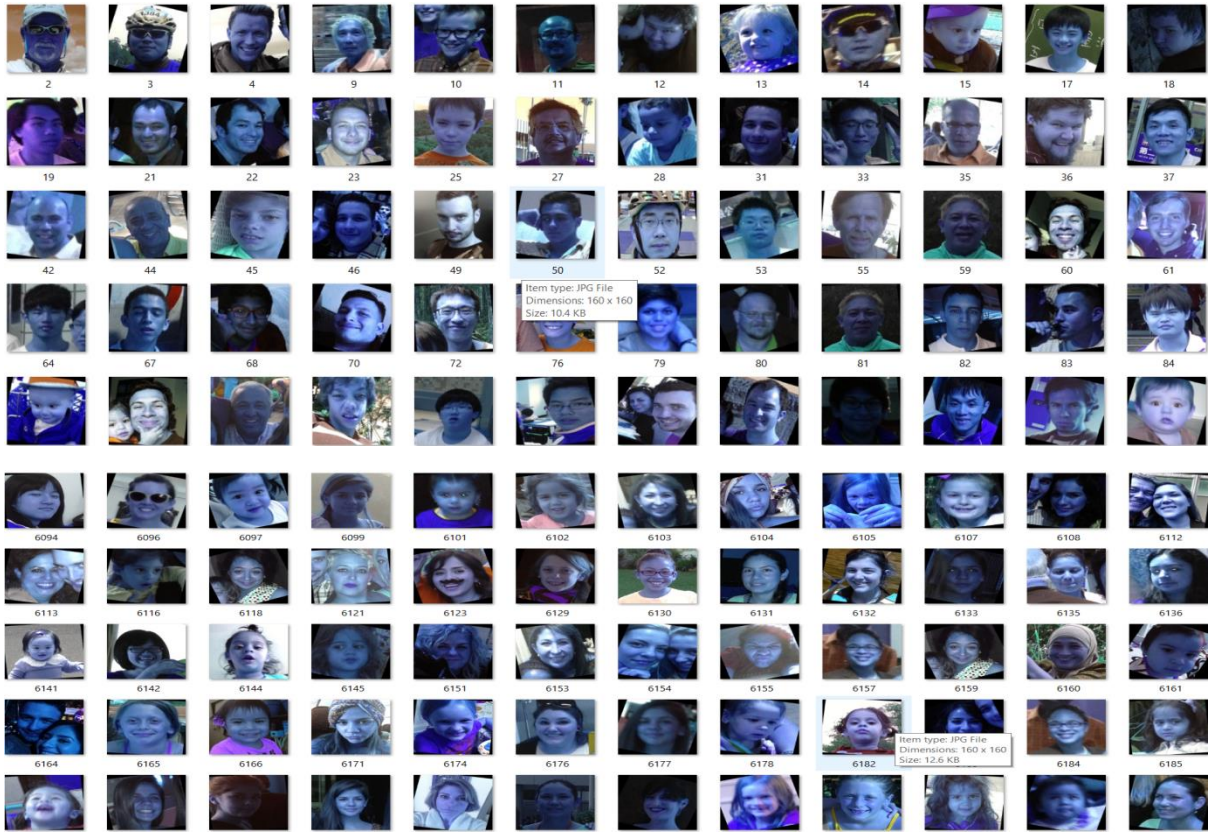
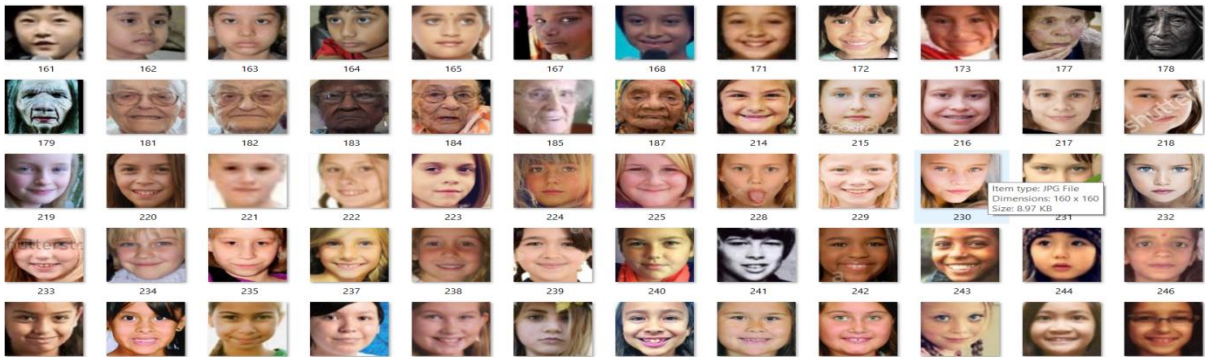


Figure 2: Adience dataset sample image

C)UTKFace [3]: UTKFace dataset consists of 20,000 face images with annotations of age, gender, and ethnicity. It covers the age range of 0 to 116. The image is captured in a large variety of poses, resolutions, and illuminations so that it can be used in the real world. The sample image is shown in Figure 3:



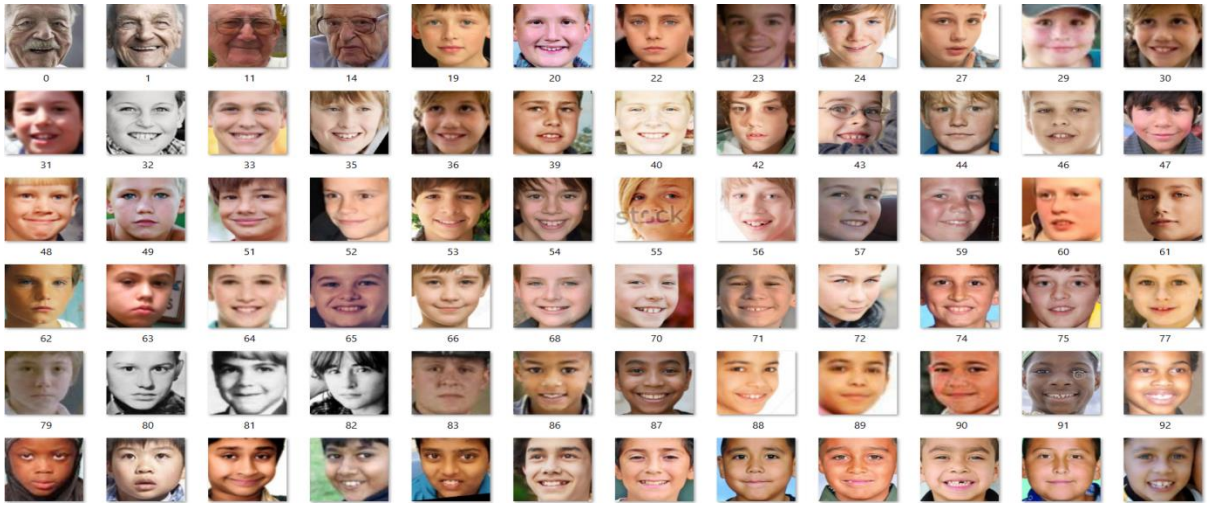


Figure 3: UTKFace Dataset Sample Images

Table 1. Dataset Distribution between Male and Female classes.

Dataset	Male	Female	Remarks
WIKI	29727	10489	Imbalanced Dataset
Adience	8491	10913	Almost Balanced Dataset, low resolution, wide real-time scenarios, low-quality images, variation in pose, appearance, illumination, image quality
UTKFace	12391	11317	Almost Balanced Dataset, variation in pose, facial expression, illumination, occlusion, resolution

3: TRANSFER LEARNING

Training any deep neural network architecture from scratch is expensive and time-consuming and the solution to this problem is a technique called Transfer Learning [6]. In transfer learning, weights and filters that are used for a specific task are used can be reused for different tasks by retaining some proportion. In this process, we load the architecture with pre-trained weights and then modify that architecture slightly according to the task we have to perform. In this paper, we have used the pre-trained weight of the ImageNet with architecture VGG-16 and MobileNetV3.

VGG16 [1]:

VGG16 is Deep Neural Network architecture, considered a very useful architecture for image classification. VGG16 contains a total of 21 layers which include 13 convolution layers, 5 max-pooling layers, and 3 Dense layers.

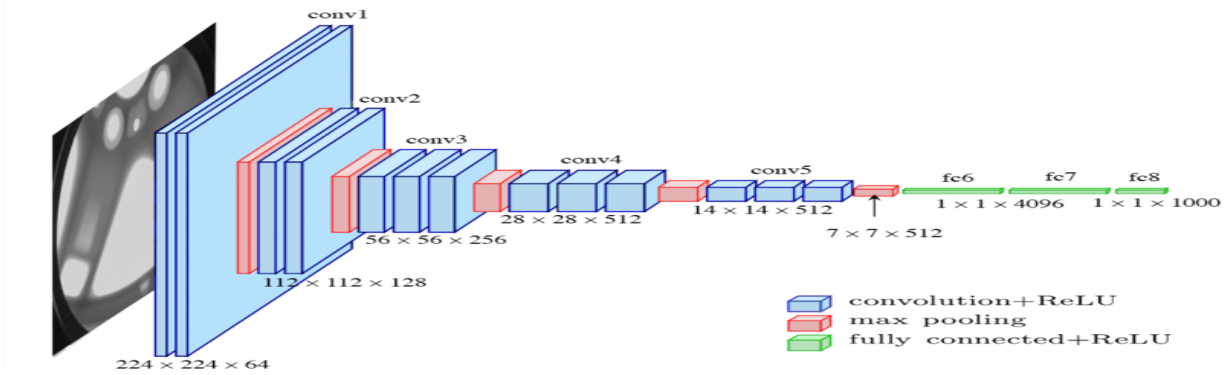


Figure 4: The original VGG16 architecture (width multiplier = 0.5, input size=224)

MobileNetV3:

MobileNetV3 [2] is a deep convolution neural network that is specifically designed for mobile phone CPUs.This CNN design includes the use of hard swish activation and squeeze-and-excitation modules in the MBConv blocks.

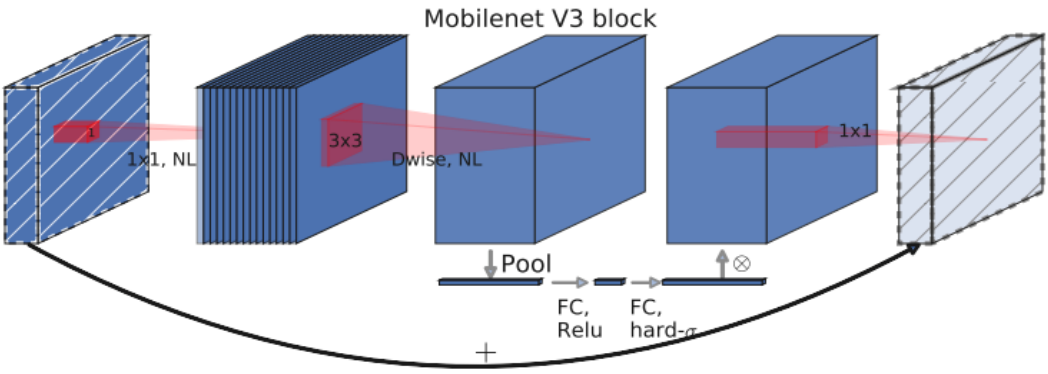


Figure 5:MobileNetV2+Squeeze-and-Excite

Training and Testing:

Since the data set images were already aligned, so by changing the resolution of the image and keeping it in the RGB channels were passed for the training in network architecture. For gender classification, we used the cross-entropy function as a loss function [8]. The optimizer used is the AMSGrad variant of Adam [8] with an initial learning rate of 0.0001 which is 0.00001 in the fine-tuning process.

Each model was first trained for 10 epochs and then fine tunning is done for 40 epochs. In the first 10 epochs training is done on the newly added layer in the original architecture taken by removing the top layers.

Comparison of VGG16_80 accuracy on Adience dataset with MobileNetV3 shown in Fig.6 .That clearly shows that the MobileNetV3_80 accuracy after transfer learning using imagenet weight is 68.23% and MobileNetV3 on the same place is 67.8% which is approx same. And after fine-tuning, the VGG16_80 and MobileNetV3_80 for forty epochs accuracy increased to 98.33% for VGG16_80 and MobileNetV3_80 accuracy is 89.24%.

In VGG16, the number of layers unfreezes is 9 out of 19 layers, and in MobileNetV3, the number of layers unfreeze is 115 out of 235 layers. To see the effect on accuracy by changing the number of the trainable layer in the MobileNetV3, we have taken one configuration, 80×80×3, and done the experimentation.

In Figure 6, the following convention is used during the ablation study. Here, 'X' and 'Y' represent the resolution and number of layers in CNN. Trained VGG16_X_Y and MobileNetV3_X_Y are the corresponding fine-tuned architectures using 'X' and 'Y' settings.

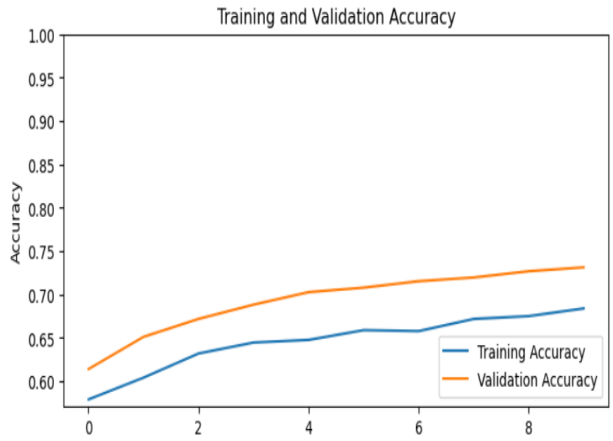
VGG16_80_9	MobileNetV3_80_115
 (a)	 (b)
 (c)	 (d)

Figure 6.Training plots depicting training and validation accuracy, training, and validation loss.

MobileNetV3_80_175	MobileNetV3_80_195

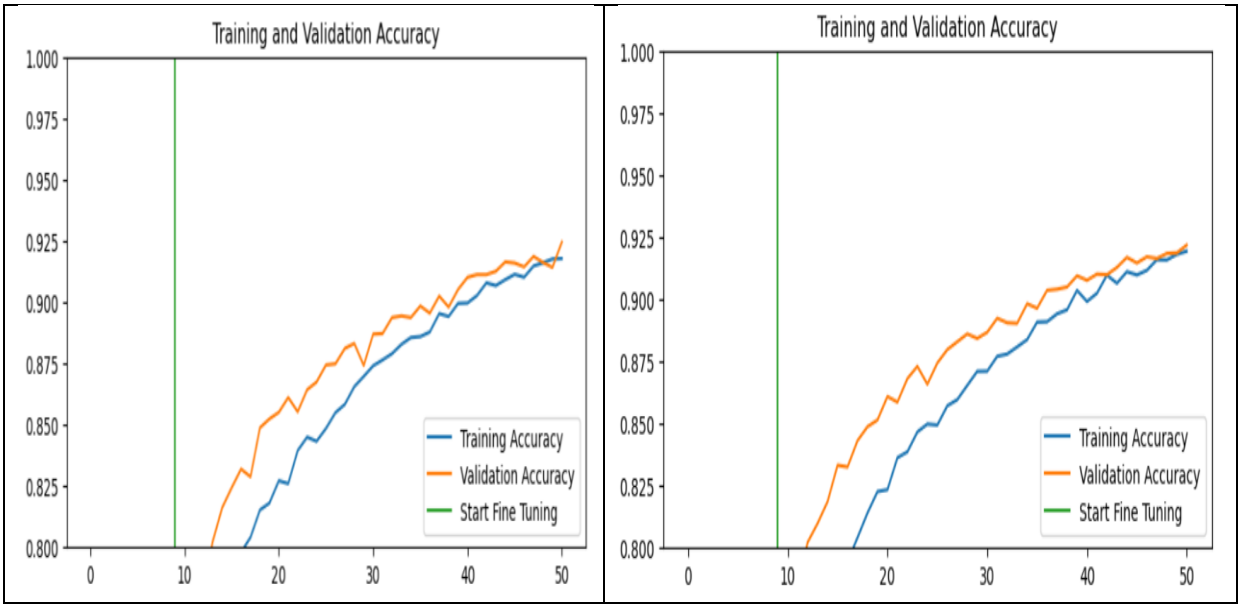


Figure 7:Training plots depicting accuracy by changing the number of trainable layers.

4: Evaluation

For gender classification, we used the evaluation metric: accuracy which is given as the fraction of gender images which are classified correctly over the total number of gender images.

5: Experimentation and Results

The experiment is conducted against various standard datasets with VGG16 and MobileNetV3 architecture with varying the input size of the image. The input size of the images chosen is 40,80 and 120. The pre-trained architecture of VGG16 and MobileNetV3 are loaded by dropping the top layer and then adding the global average pooling layer and dense layer with one output. Drop out in experimentation is 20%. The first extra added layer is trained and all other layers are freezed thereafter layers of standard architecture are unfreeze and training is done on that unfreeze layer, this process is called fine-tuning.

Table 2. Evaluation of different architecture on different datasets. This table contains the accuracy of each dataset.

	Accuracy(%)		
Model	WIKI	Adience	UTKFace
VGG16_40	82.99	95.60	87.80
VGG16_80	89.35	98.33	90.70
VGG16_120	90.04	99.37	91.94
MobileNetV3_40	74.37	74.94	76.37
MobileNetV3_80	75.51	89.24	87.01
MobileNetV3_120	76.37	94.49	89.92

As we got an accuracy of 89.24% on the audience dataset and configuration of 80×80×3 with MobileNetV3, the number of trainable layers is 115. So we experimented by changing the number of trainable layers, results are summarised in table 3.

Table 3. Evaluation of MobileNetV3_80 with varying the number of trainable layers.

Number of Trainable layers	Accuracy (%)
MobileNetV3_80_115	89.24
MobileNetV3_80_135	90.28
MobileNetV3_80_155	91.15
MobileNetV3_80_175	91.76
MobileNetV3_80_195	91.90

As from table 3, it is evident that by increasing the number of the trainable layer the accuracy got increased but after a certain point that got saturated so we stopped further increasing the number of the trainable layer. The accuracy we got from that saturation point is 91.90%.

6. Conclusion:

In this work after carrying out ablation study, we observed that the MobileNetV3_80 model provides 91.90% accuracy, MobileNetV3_120 provides 94.49% accuracy and VGG16_120 provides 99.37% accuracy on the Adience dataset respectively. Here, the MobileNetV3 model has very low latency because of less number of parameters (1,530,993) as compared to VGG16 (14,715,201), and hence we selected the MobileNetV3_80 in order to implement the model on embedded devices. Choosing image resolution of 80×80×3 will also help to get fast inference with good accuracy in the case of embedded devices.

ACKNOWLEDGEMENT

This research is supported by the Life Sciences Research Board (LSRB) in association with Defence Institute of Psychological Research (DIPR) sanction letter no. LSRB/o1/15001/M/LSRB-381/PEE& BS/2020, dated 15.03.2021. The authors want to thank NVIDIA for the academic GPU research grant.

References:

[1] Walid Hariri. Efficient Masked Face Recognition Method during the COVID-19 Pandemic, 07 July 2020, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-39289/v1]

[2] A. Howard et al., "Searching for MobileNetV3," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314-1324, doi: 10.1109/ICCV.2019.00140.

[3] UTKFace. (n.d.). Retrieved Aug 20, 2022, from http://aicip.eecs.utk.edu/wiki/UTKFace

[4] IMDB-WIKI – 500k+ face images with age and gender labels. (n.d.). Retrieved July 14, 2020, from https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/

[5] The OUI-Adience Face Image Project (n.d.) Retrieved Aug 21,2022,from https://talhassner.github.io/home/projects/Adience/Adience-data.html

[6] F. Zhuang et al., "A Comprehensive Survey on Transfer Learning," in Proceedings of the IEEE, vol. 109, no. 1, pp. 43-76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[7] A. Greco, A. Saggese, M. Vento and V. Vigilante, "A Convolutional Neural Network for Gender Recognition Optimizing the Accuracy/Speed Tradeoff," in IEEE Access, vol. 8, pp. 130771-130781, 2020, doi: 10.1109/ACCESS.2020.3008793.

[8] Reddi, S. J., Kale, S., & Kumar, S. (2018). On the convergence of Adam and beyond. ICLR 2018

[9] Sheoran, V., Joshi, S., Bhayani, T.R. (2021). Age and Gender Prediction Using Deep CNNs and Transfer Learning. In: Singh, S.K., Roy, P., Raman, B., Nagabhushan, P. (eds) Computer Vision and Image Processing. CVIP 2020. Communications in Computer and Information Science, vol 1377. Springer, Singapore. https://doi.org/10.1007/978-981-16-1092-9_25