Masked Deep Face Recognition using ArcFace and Ensemble Learning

Arun Kumar R
Department of IT,
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India

V Anoosh Solayappan
Department of IT,
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India

Sree Sharmila T
Department of IT,
Sri Sivasubramaniya Nadar College of Engineering,
Chennai, India

Ram Prasad K
Computer Science and Engineering,
Shiv Nadar University,
Chennai, India

***Abstract*— With advancements in technology, human biometrics, especially face recognition, has witnessed a tremendous increase in usage prominently in the field of security. Face recognition proves to be a convenient, coherent, and efficient way to identify a person uniquely. Face recognition systems are trained generally on human faces sans masks. With the ubiquitous use of face masks due to the ongoing COVID-19 pandemic, face recognition becomes a daunting challenge. In this paper, the deep learning architectures, namely MobileNetV2, DenseNet201, ResNet50V2 and VGG16 with the ArcFace loss function, were trained on the newly created dataset called "MaFaR", which consists of a mixture of masked and unmasked images of 75 distinct individuals, and ensemble learning techniques have been used to improve the performance, achieving an accuracy 93.65%.**

Keywords - ArcFace, Deep Learning, Face Recognition, MaFaR dataset, Soft voting

I. INTRODUCTION

Traditionally, humans safeguard and secure their devices and property using IDs and passwords, passphrases, etc., to protect them from crime, sabotage or attack, espionage, and privacy invasion. Face recognition has become a popular form of biometric security owing to advancements in technology. Further, an increasing number of countries have begun to view facial recognition technology as a law enforcement solution and utilise it to crack down on crime [1]. Deep learning architectures like DeepFace [2], FaceNet [3] and VGGFace [4] have been able to achieve high accuracy in establishing the identity of individuals. The deep learning algorithms extract features from the face to uniquely recognise the individual.

The COVID-19 pandemic has affected around 165 million people and killed more than 3 million people [5]. It has exacted a toll on the public health systems and has permanently changed the world in which we live. The COVID-19 virus spreads primarily through the respiratory droplets expelled by an infected person while sneezing, coughing and talking. These droplets can transmit the disease when landed in the noses or mouths of other persons. Studies have shown that masks when worn correctly, reduce the spray of droplets and thus the risk of infection. Wearing a mask has now become a norm and, in many cases, even mandatory. Many face recognition systems rely on deep learning techniques to recognise individuals. Face recognition systems are usually trained on unmasked images of the person, which performs poorly on identifying the masked images of the same individual. Face recognition meets a new challenge due to the ubiquitous usage of masks, as the mask worn by the person makes their nose, mouth, and the region around it occluded.

This paper proposes an approach to tackle the problem of masked facial recognition. In this work, the experiments were performed on MaFaR (**Ma**sked **Fa**ce **R**ecognition) dataset. This work uses four deep learning architectures, namely MobileNetV2 [6], DenseNet201 [7], VGG16 [8] and ResNet50V2 [9], all of which were instantiated with pre-trained ImageNet weights. Additionally, Additive Angular Margin Loss (ArcFace) [10] was used as the loss function for each deep learning architecture. The additive angular margin loss obtains highly discriminative features for face recognition [10], and it helps in stabilising the training process. All four deep learning models were trained individually with the same training set. Soft voting, a type of ensemble technique, was used on the class probabilities obtained from the prediction of each trained deep learning model on the test set to achieve

better performance. Thus, our contribution includes: (i) Fabricating a new dataset, MaFaR, which consists of masked and unmasked images, (ii) Performing facial recognition by training deep learning architectures with ArcFace loss function and (iii) Improving performance through soft voting.

II. RELATED WORKS

Computer vision is an area of high research interest because of its broad spectrum of applications. The field of computer vision has achieved major strides in recent years, particularly owing to the development and evolution of convolution neural networks. In 2012, the AlexNet architecture [11] achieved an error rate of 15.3% on the ImageNet dataset [12], establishing a major benchmark. In 2015, ResNet [13] outperformed AlexNet by achieving an error rate of 3.57% on the same ImageNet dataset. One popular problem faced in the domain of computer vision is face recognition. Deep learning architectures and techniques facilitated several advancements in the field of face recognition. FaceNet and VGGFace obtained 99.63% and 98.95% accuracy, respectively, on the LFW dataset [14].

One major challenge involved in face recognition is identifying the individual despite the occlusions present in the image. The scope of many simple face recognition algorithms becomes parochial when its objective is to identify an individual wearing a hat, sunglasses, masks, etc., which could act as a possible occlusion [15]. Earlier research works involving the identification of faces with occlusion primarily used two approaches to tackle the problem: one involves restoration while the other involves the removal of occlusion [15].

In the former approach, the occluded regions of images are restored with the help of images present in the training batch. Bagchi et al. [16] used the iterative closest point algorithm to register the input 3D face images. The occlusions are extracted by thresholding the depth map value of the 3D image, after which Principal Component Analysis (PCA) was used for restoration. Drira et al. [17] used a statistical approach to estimate and predict occlusions, and PCA was used to restore the regions of occlusions.

The discard occlusion-based approach detects and removes the regions of the face image that are found to be occluded so as to prevent improper reconstruction. The feature extraction and classification steps utilise only the non-occluded regions of the image. Priya et al. [18] split the face image into local patches. The support vector machine classifier was used to detect and remove the occluded region of the image. Following this, face recognition was performed using a mean based weight matrix on the non-occluded parts of the image. Alyuz et al. [19] used both occlusion removal and restoration techniques. Global mask projection was used to remove the occlusions, after which restoration was done using Gappy PCA.

The authors of the above papers focused on face recognition with occlusion in general, while Hariri [15] and Biswas et al. [20] primarily focused on face recognition with masks as occlusion. Hariri [15] performed masked face recognition by cropping out the masked face region, after which the feature maps of each face were obtained using Bag-of-Features (BoF) in the last convolutional layer of the pre-trained VGG-16, after which, Multilayer Perceptron (MLP) is applied for classification. Biswas et al. [20] utilised a ResNet50 architecture pre-trained on VGGFace2 to train the model on unmasked images and test it on masked images with the help of transfer learning.

Deng et al. [10] proposed Additive Angular Margin Loss (ArcFace) loss function. This loss function was able to obtain the discriminative features for facial recognition and consistently performs better than the other state-of-the-art loss functions. Montero et al. [21] implemented an ArcFace based face recognition system to identify masked images. They used a dataset consisting of unmasked images and generated masked images, and used ResNet50 as the backbone with multi-task ArcFace loss, a combination of ArcFace loss and mask-usage classification loss.

The objective of this paper is analogous with the goals of the research works mentioned above, which is to identify masked face images successfully. In this work, a **Ma**sked **Fa**ce **R**ecognition (MaFaR) dataset consisting of masked and unmasked images is fabricated and used to train an ensemble of deep learning models, each using the ArcFace loss function

and then performing soft voting to improve performance.

III. MASKED DEEP FACE RECOGNITION

The process of collecting and fabricating the masked and unmasked videos subsequently till the splitting of the 'MaFaR dataset' to training, validation and test set has been explained in Section 4.1. Figure 1 represents the pipeline of this work. The images present in the training, validation and test set are of RGB format with size *(224X224)*.

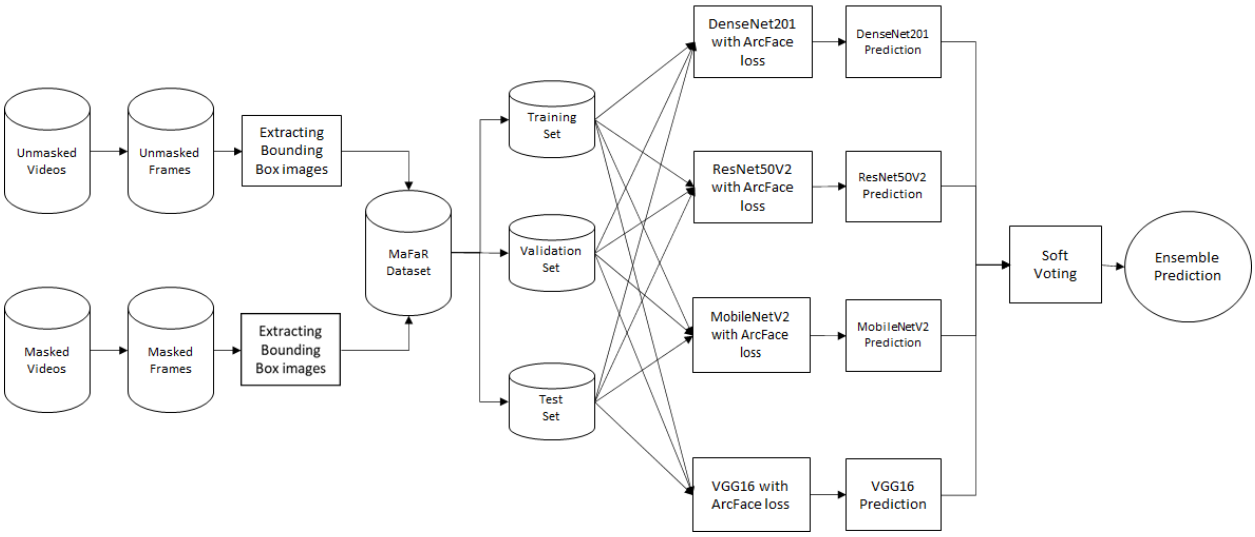


Fig.1: Model pipeline

This experiment used four deep learning architectures, namely VGG16, DenseNet201, ResNet50V2 and MobileNetV2. ResNet50V2 is a modified version of the ResNet50 architecture. In ResNet50V2, a modification was made in the propagation formulation of the connections between blocks. ResNet50V2 performs better than ResNet50 [13] and ResNet101 [13] on the ImageNet dataset. MobileNetV2 is an improvement of the MoblieNetV1 architecture [22], and it introduces two new features, which include linear bottlenecks between the layers and shortcut connections between the bottlenecks. The architectures mentioned above were instantiated with ImageNet weights and used the ArcFace loss function with the feature scale parameter *s* set to 30 and the angular margin parameter *m* set to 0.3. The last convolutional layer of each of these architectures was followed by batch normalization [23]–dropout [24]-fully connected-batch normalization structure as used in [10] to get a final 512-D embedding feature.

The SGD optimiser was used with a learning rate of 0.1, momentum of 0.9 and a weight decay of $5e^{-4}$. The class probabilities of each image given by the trained deep learning models are then summed, and the image is grouped to the class having the highest summed probability.

A. ArcFace

An additive angular margin penalty into the softmax loss by ArcFace [10] as mentioned in Equation (1).

$$L_{ArcFace} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\Theta_{y_i}+m)}}{e^{s\cos(\Theta_{y_i}+m)}+\sum_{j=1,j\neq y_i}^{N}e^{s\cos(\Theta_j)}} \quad (1)$$

Where N denotes the total number of classes, m denotes the angular margin parameter, and s denotes the feature re-scale parameter, The angle between the embedding feature x_i of the i-th face sample and the j-th class center W_j is denoted by Θ_j. If y_i is the class label of x_i, then Θ_{y_i} denotes the angle between x_i and the ground-truth center W_{y_i}.

IV. EXPERIMENTAL DISCUSSION AND RESULTS

This section discusses how the customised dataset "MaFaR" is created and the accuracy of the deep learning models using the proposed masked deep face recognition model.

A. Dataset Creation

The process of creating a customized dataset, MaFaR, involved the collection of two videos from each

subject, totalling 75 subjects, one video with the person wearing a mask and another video without wearing a mask. Each person who volunteered has been instructed to gently turn their head on both sides so that the centre, right and left sides of their faces are visible in the video. On average, these videos were 6-7 seconds long. The frames were extracted from the collected videos and were stored separately. The system extracted 18,684 frames from the videos containing masked individuals and 18,013 frames from the videos containing individuals not wearing masks. In total, 36,697 frames were extracted.

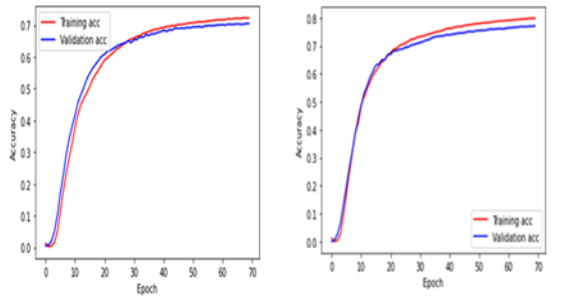
Fig. 2: Frames extracted from masked and unmasked videos

The facial bounding box was generated for all the frames extracted, and the region inside the bounding box was selected and stored. The OpenCV DNN module with Caffe framework [25] was used to identify the region of the image containing only the face for the masked and unmasked frames, which was then enclosed in a bounding box. The bounding box generated for certain images were improper, and they were manually discarded. The resulting bounding box images were used to create the MaFaR dataset. The MaFaR dataset consists of 22,500 images of both masked and unmasked images belonging to 75 different individuals, with each individual having 150 masked and unmasked images each. The images in the MaFaR dataset were randomly split into train, validation and test set. This work uses 15,000 images for training, 3,750 images each for validating and testing. The training set consists of 75 different individuals with 200 images per individual. The validation and test set each consist of 75 different individuals with 50 images per individual. A nearly equal number of masked and unmasked images were present for each person in the training, validation, and test set.

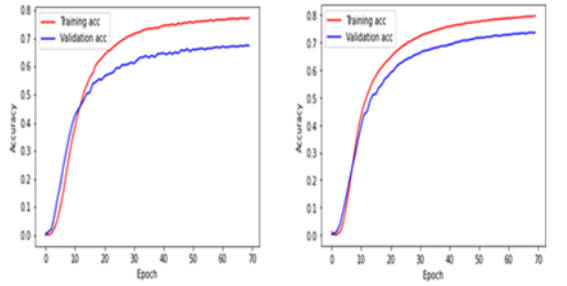
Fig. 3: Images obtained after generating the bounding box

A. Performance of masked deep face recognition

The deep learning models were set up as mentioned in model architecture. Each architecture was trained individually for 70 epochs with a batch size of 32. Training, validation and test accuracy obtained by each model is mentioned in Table 1. It is evident that ResNet50V2 has obtained the highest accuracy of 71.65% on the test set and has the least training loss compared to the other models. MobileNetV2 has obtained the highest training accuracy of 79.39% and the highest validation accuracy of 77.18% and also has the least validation loss. The accuracy curve of all the four deep learning models reaches a maximum and then becomes stable, providing a perfect accuracy curve, as seen in Figure 4. The loss curve of all the four deep learning models reaches a minimum and becomes stable, as observed in Figure 5, providing a perfect loss curve indicating that the models are learning from the data. Further, soft voting was implemented to enhance the performance of the deep learning models. The class probabilities for each image in the test set is obtained from the four trained deep learning models. The class probabilities of each image are then summed, and the image is grouped to the class with the highest summed probability.

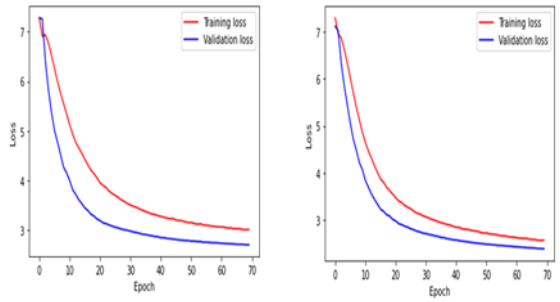


(i) DenseNet201 (ii) MobileNetV2

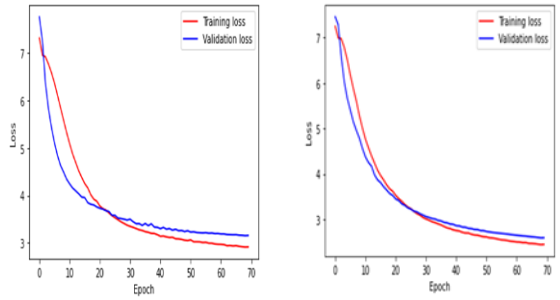


(iii) VGG16 (iv) ResNet50V2

Fig. 4: Accuracy Vs Epoch



(i) DenseNet201 (ii) MobileNetV2



(iii) VGG16 (iv) ResNet50V2

Fig. 5: Loss Vs Epoch

Finally, a test accuracy of 93.65% was achieved using this ensemble technique, thereby vastly improving the performance on the test set.

Tabe 1: Train, Validation and Test Accuracy of the Deep Learning Models

Architecture	Train Accuracy	Validation Accuracy	Test Accuracy
DenseNet201	71.92%	70.50%	64.02%
MobileNetV2	79.39%	77.18%	69.70%
VGG16	77.40%	67.40%	62.42%
ResNet50V2	79.01%	73.55%	71.65%

This ensemble approach has significantly improved the accuracy as compared to any individual model performance.

V. CONCLUSION

Face recognition systems are generally trained on unmasked images; therefore, it becomes challenging for such systems to recognise masked individuals. In this work, deep learning architectures were trained on the MaFaR dataset, and each of the architectures used the ArcFace loss function. The proposed ensemble-based ArcFace loss function, along with soft voting, helped to achieve an accuracy of 93.65%. This approach outperforms the accuracy of deep learning models evaluated individually.

REFERENCES

[1] nytimes.com/2020/01/12/technology/facial-recognition-police.html
[2] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," 2014 IEEE Conference on Computer Vision andattern Recognition, 2014, pp. 1701-1708, doi: 10.1109/CVPR.2014.220.
[3] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815-823. 2015.
[4] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.
[5] https://covid19.who.int/
[6] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520. 2018.
[7] G. Huang, Z. Liu, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in Proc. CVPR, Jun. 2016, pp. 4700–4708.
[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
[9] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In European conference on computer vision, pages 630–645. Springer, 2016.
[10] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. page 9.
[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. page 8.
[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
[14] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. page 11.
[15] Hariri Walid. Efficient masked face recognition method during the covid19 pandemic, 07 2020.
[16] Parama Bagchi, Debotosh Bhattacharjee, and Mita Nasipuri. Robust 3d face recognition in presence of pose and partial occlusions or missing parts, 2014.
[17] Hassen Drira, Boulbaba Ben Amor, Anuj Srivastava, Mohamed Daoudi, and Rim Slama. 3D Face Recognition Under Expressions,Occlusions and Pose Variations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2270–83, September 2013
[18] G Nirmala Priya and R S D Wahida Banu. Occlusion invariant face recognition using mean based weight matrix and support vector machine. Sadhana, 39(2):303–315, April 2014.
[19] N. Alyuz, B. Gokberk, and L. Akarun. 3-d face recognition under occlusion using masked projection. IEEE Transactions on Information Forensics and Security, 8(5):789–802, 2013.
[20] Mandal, Bishwas, Adaeze Okeukwu, and Yihong Theis. "Masked Face Recognition using ResNet-50." *arXiv preprint arXiv:2104.08997* (2021).
[21] Montero, David, Marcos Nieto, Peter Leskovsky, and Naiara Aginako. "Boosting Masked Face Recognition with Multi-Task ArcFace." *arXiv preprint arXiv:2104.09874* (2021).
[22] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).
[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015.
[24] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. JML, 2014.
[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia (MM '14). Association for Computing Machinery, New York, NY, USA, 675–678. DOI:https://doi.org/10.1145/2647868.2654889