University of Science and Technology at Zewail City

Communications and Information Engineering Program

CIE 555: Neural Networks and Deep Learning

Project Report

Team Information

Asmaa Mohammed Ibrahim	ID: 201701056
Elsayed Mohammed Elsayed Mostafa	ID: 201700316
Muhammed Magdy Alasmar	ID: 201700038
Salma Hasan Elbess	ID: 201601152

Table of Contents

Problem definition and motivation

The whole world is now fighting Covid-19. There are more than one hundred million detected cases across the globe since the beginning of the crisis. [1] According to the world health organization (WHO) [2], it is essential for individuals to wear face masks upon interacting with others. With the aid of visual systems, it can be detected whether a person is wearing a face mask or not. This will help researchers to incorporate the factor of wearing masks in their research projects about Covid-. In addition, it will help authorities to monitor individuals and take action upon bare face detection.

The general problem is an object detection problem [3] which is detecting people with and without masks in offline or live videos. In this project, we tackle part of the problem; We tackle the image classification part [4]. Given an image with one person as input, the problem is to classify the image whether it is an image of a masked person or non-masked person.

Dataset

In this project, we are using a popular public domain dataset for face masks detection [5]. This dataset consists of 12000 images collected through scrapping on google search as well as the CelebFace dataset created by Jessica Li [6]. The dataset is primarily separated into 3 folders: train, validation and test. Each folder has 2 classes folders named 'WithMask' and 'WithoutMask'. After investigating the dataset, it is extremely balanced with 5000 masked faces images and 5000 non-masked faces images in the train directory. The same happens in the validation directory which has 400 masked faces images and 00 non-masked faces images. For the test folder, it contains 483 masked faces images and 509 non-masked faces images. Hence, the dataset is suitable for unbiased classification. Moreover, the dataset is highly generic and representative of various ethnicities and mask shapes, so it is possible to have generic unbiased classificer out of it. According to the transfer learning model used for classification, data augmentation will be performed in order to make the best use of the pre-trained weights.

For testing, a whole different dataset is used. This testing dataset is called Face Mask Detection Dataset [7]. The dataset contains a total of 7553 images of 3 channels (RGB) , 3725 of them are labeled as masked faces images while 3828 of them are labeled as non-masked faces images.

Literature Review

Due to the recent spread of the problem and its related datasets, most of the similar work on such classification problems used hybrid datasets. In each research, there were different datasets used to detect the existence of the face mask. Moreover, a lot of papers encounter the more sophisticated problem of face recognition while wearing masks. Hence, this section includes general work on the problem using the chosen dataset and other similar datasets.

In [8] a mask existence detection system is proposed using a traditional classification method based on color processing technique in computer vision. After training phase and test phase on multiple datasets, the proposed system achieved a true positive rate of about 95% with 5% false positive rate. These rates were achieved without using deep learning or machine learning models due to the limitation on the dataset used; the images were restricted to the medical staff in operation rooms. More generic systems are still preferable to have.

In [9], the authors proposed a face-mask condition detection model using a two-stage hybrid network (SRCNet). SRCNet consists of a super-resolution (SR) network for image processing followed by a classification network. SR network has proved credibility in achieving good accuracy after using it for image processing. In the work introduced in [9], three categories classification achieved 98.70% accuracy using deep learning and image processing in MATLAB. The three classes are no face-mask wearing (NFW), incorrect face mask-wearing (IFW) and correct face mask-wearing (CFW). The high classification accuracy achieved encourages the usage of deep learning models in similar projects more than traditional machine learning models.

In [10], a relatively small dataset is used for PCA algorithms to recognize the existence/non-existence of a mask. The result of this work shows that PCA gives a poor recognition rate for masked face images rather than non-masked faces.

In [11], a hybrid model of deep learning and machine learning is introduced to detect the face mask existence on three different datasets. In this model, deep learning is used for feature extraction. Specifically, the ResNet50 network is used to extract representative features from the images before using decision trees, SVM and ensemble learning models for classification. Testing on the Real-World Masked Face

Dataset (RMFD) achieved 99.54% testing accuracy with the SVM classifier [12]. The very high testing accuracy is generally related to the usage of deep learning as feature extraction before introducing the classification stage.

In [13], the same problem is tackled in a real-time mask detection system. The proposed system applies data augmentation to overcome the problem of unbiased dataset. After data augmentation, the model uses ready-made face detection in open-CV with DNN and then uses MobileNetV2 as the base classifier network to classify whether the person is wearing a mask or not. The system also adds new layers for tuning the accuracy and the performance of the model. This system achieves an accuracy score of 0.9264 and an F1 score of 0.93 from different datasets.

Using transfer learning with state of the art pretrained model of InceptionV3 to be trained on Simulated Mask Face Dataset (SMFD). In [14], as the proposed dataset contains only 1570 images, data augmentation has been applied before using the model. InceptionV3 is a 48 layered convolutional neural network architecture developed by google. In the proposed architecture, the last layer of InceptionV3 is removed, and five layers are added to the network structure. Those layers are an average pooling layer with a pool size of 5x5, flattening layer followed by a dense layer of 128 neurons with Relu activation function and dropout rate of 0.5, and the last layer is a decision making layer of two neurons with softmax activation function to determine whether a person is wearing a mask or not. Model is trained for 80 epochs having 42 steps. The proposed model achieved a precision and specificity of 99.92%,99.9% for training and 100% for both metrics while testing. The implementation could be better implemented by imposing a larger dataset and appending a face recognition system.

Using the architecture of YOLO V3 developed by Josef Redmon and Ali Farhadi ,and faster RCNN for deployment on face mask detection tasks [15]. The architecture of YOLO V3 contains 53 convolutional layers. In the proposed structure, an additional 53 layers are stacked into it. During training the model, all model layers were freezed except the last 3 layers that were trained. Initially, this resulted in insufficient accuracy. This challenge may be due to the fact that available images to be trained on were limited. To better improve accuracy on images, all layers were not freezed and all of them got trained for more than 70 epochs. YOLO V3 resulted in an average precision of 55 %. What makes it unique is that it could be implemented easily on mobile phones, surveillance systems as it is sensitive to camera angle orientation.

As a part of a smart city system, a straightforward deep CNN architecture is proposed and used in [16]. The proposed architecture is simply 2 layers of convolutional layers with kernel size 3x3 followed by a 2x2 max pooling layer and this sequence is repeated 3 times. Three dense layers, each one followed by a dropout layer, are then used after flattening the output of the latest convolutional layer. A 2 units layer is then added as an output layer for the architecture. After training for 100 epochs, the proposed model achieved 98.4% accuracy on unseen test data in addition to 0.985 AUC value which indicates high true positive rate and low false positive rate. Although achieving high accuracy, the model was trained on a total 1231 images which is relatively small. Also, there was no guarantee about the diversity of the images such as the mask color or type in addition to the human race or color in the used images.

In [17], an IoT-based system using real-time deep learning models for face mask detection and classification is deployed. The classification part is done with different transfer learning models such as VVGG-16, MobileNetV2, Inception V3, ResNet-50 and traditional CNN model. With each pretrained model, the authors freezed its weights and just added output layer or dropout layers in some cases in addition to performing data augmentation to avoid overfitting. The classification was done into 3 classes which are masked, non-masked and improperly masked. By using the VGG-16, MobileNetV2, Inception V3, ResNet-50, and CNN models, they achieved 99.8%, 99.6%, 99.4%, 99.2%, and 99.0% accuracies, respectively, on their test set. The VGG-16 model achieved a recall value of about 0.993 and a precision value of about 0.9967 as an average across the three classes as the highest model. The MobileNetV2 came second with 0.993 and 0.986 as average precision and recall respectively.

Our mission is mainly to achieve similar high performance with different tuned deep learning architectures. Generally, the transfer learning concept is used to get a pretrained model and tune multiple layers afterwards. The main accuracy metric to use is F1-score to consider both accuracy and recall. Specifically, the recall metric is more important than precision in such an application. This is because ,due to the dataset organization, the non-maksed class is the true (positive) class and we're concerned with avoiding classifying non-masked as masked. Hence, we're more strict against false negative class. So, the recall metric is more important than precision.

Objectives

The literature review shows that good results were already obtained using transfer learning. In this project our objective will be to generate a generalizable model that detects masks with accuracy more than 97% regardless of the brightness and poses of faces. After the desired accuracy is achieved on the dataset described above. We will test and generalize our model to work on the second dataset [7]. After fitting and testing multiple transfer learning models, they are compared and the best model based on its recall score, generalization and efficiency is stated. Our models are to be tested on the test data from different datasets, as well as random images. This generalization will make sure that the models, especially the best one, can be used to detect non masked faces in hotels, organizations and different public areas in order to give warning or invoke penalties.

Methodology

The main approach is to train two transfer learning models on the proposed dataset, evaluating the recall and precision of each model and then testing each model on different random images and dataset. The results of each model are then compared and analysed to choose the best model to use as a final product.

I. VGG16

VGG16 layers are all convolutional layers with 3x3 kernels and 2x2 max pooling layers until the 3 final dense layers. Hence, we think that VGG16 architecture is suitable for capturing small features variations in the input layer and wider variations in the consecutive layers of the features space. This practically seems useful as the deeper we go in the images in this project the less important the details are. Also, the input layer for VGG16 is RGB images which is totally important for this project to capture the mask color variations. At first, the original VGG16 pretrained weights are used except for the first 2 layers which we choose to fine tune to suit our dataset. Also, 2 dense layers are added to the output of the pretrained model with 64 and 2 units respectively. This initial model behaves well for the 5 epochs we trained it with. However, it seemed a little bit slow so the learning rate of the used stochastic gradient descent is increased 10 times and the model is retrained. The modified model became much faster with the updated learning rate as it achieved 89% validation accuracy in 5 epochs. For more model trails and improvements, Adam optimizer is used and an additional layer with 256 units is used for more robust noise and avoiding overfitting. After the initial 5 epochs, the model performance is boosted compared to the previous trials. Hence, the later model is retrained with a higher number of epochs as the final model with VGG16. Next, the model performance is tested and evaluated using the recall and precision metrics on the test set as well as testing on random images and different dataset.

II. EfficientNet

EffecientNets is a series of neural networks. They are also mobile size like MobileNet. The were introduced in 2019 [18], that's when they reached state-of-art accuracies (and up to 8.4x parameter reduction and up to 16x FLOPS reduction) in ImageNet, which the series was trained on, as well as other famous dataset such as CIFAR-100. EffecientNets has 7 scaled versions of the same model, EfficientNetB0 - B7. A method of compound scaling was also introduced in the same paper. The three parameters (length, width and resolution) are scaled with a dependent coefficient.

$$
\begin{aligned}
\text{depth: } & d = \alpha^{\phi} \\
\text{width: } & w = \beta^{\phi} \\
\text{resolution: } & r = \gamma^{\phi} \\
\text{s.t. } & \alpha \cdot \beta^{2} \cdot \gamma^{2} \approx 2 \\
& \alpha \geq 1, \beta \geq 1, \gamma \geq 1
\end{aligned}
$$

However, we used the smallest version (EffecientNetB0). EffecientNetB0 resolution should be 224x224 while our images resolution is 128x128. We tried removing the top layer, adding it and we tried adjusting input size.
We have also used EffecientNetB1 for experimentation.

Results and Discussion

In this section, the results of each model fitting and testing are summarized and discussed. Generally, the testing is done through three steps. The first one is to get the recall, precision, f1-score metrics of the model using test images provided in the dataset in addition to the evaluation accuracy on them. As mentioned before, the recall metric is more important than precision due to the nature of the project and the dataset labeling. The second step is to test the model prediction on random unseen images collected from the internet and our own images with a variety of poses and backgrounds. Tricky images are also used such as nose masks worn by some footballers. The third step is to test the model performance over the previously describer testing dataset to investigate the model generalization.

I. VGG16

As expected, the VGG16 model indeed behaved very well with 99.5% and 99.6% as a training accuracy and validation accuracy respectively. Figure (1) shows the final accuracy and validation accuracy curves.

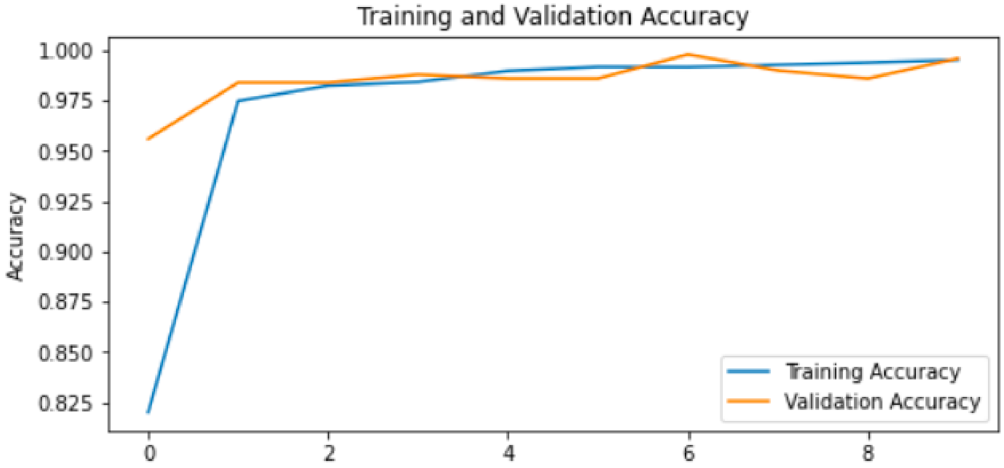


Figure 1. Accuracy and validation accuracy for trained VGG16 model

As an indication of the great performance of the model, the average recall metric over the two categories is 1 which is the optimal value we can get for the recall. Moreover, the average precision over the two classes is also 1, hence the overall F1-score is also 1 which indicates optimal performance for this model on the test data. This model is also tested on random images selected from the internet in addition to our own images and got very accurate results as shown in figure (1). However the good accuracy and recall obtained, two major points can fail the model. The first one is that it behaves very well on the images where only the face is focused on. In general images where the images are not cropped around the face, it makes more mistakes. The second one is that it can be perceived easily with inaccurately worn masks such as having a mask over half of the face only as shown in one of the test cases in figure (2). This deception can be also done with other mouth coverages such as hands or american football masks as shown in one of the test cases in figure (2). This point is generally due to the lack of training images regarding this sort of non-masked cases.

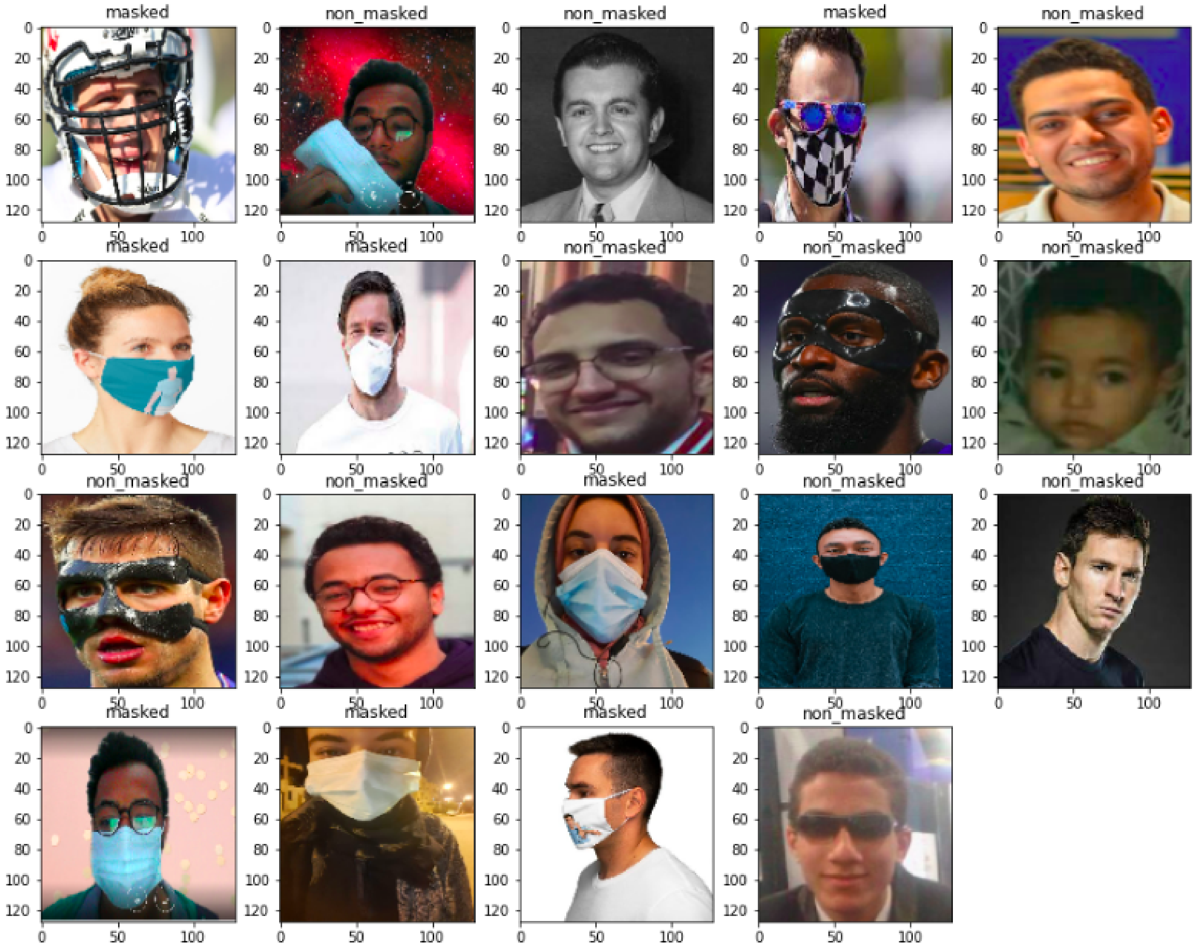
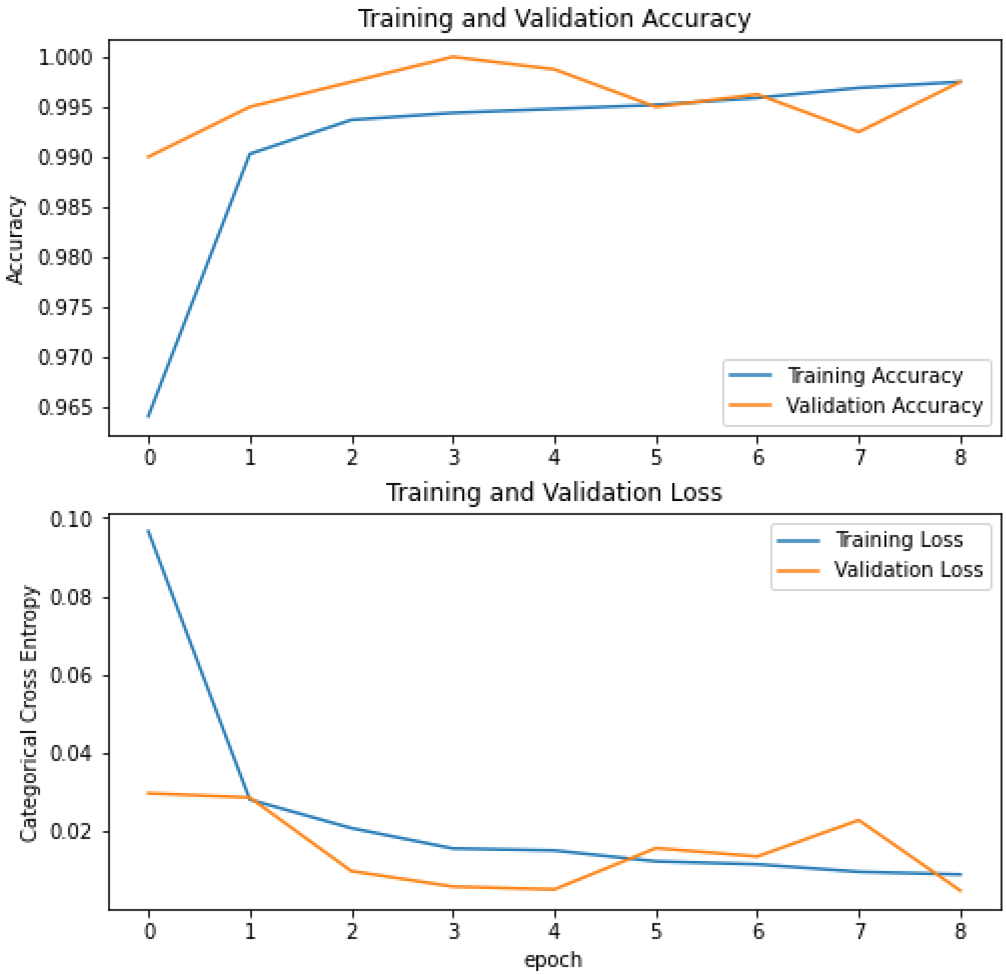


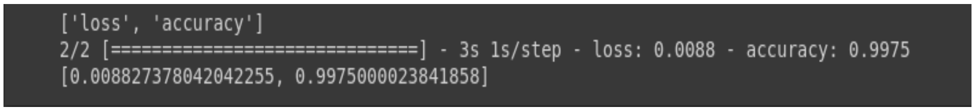
Figure 2. Random test images for VGG16 model

II. EfficientNet

EfficientNetB0 produced very promising results that were expected since the first epoch in the training. It took 5-10 epochs or even less to reach its potential on our dataset. The training and validation accuracy/loss were plotted, and are as shown:

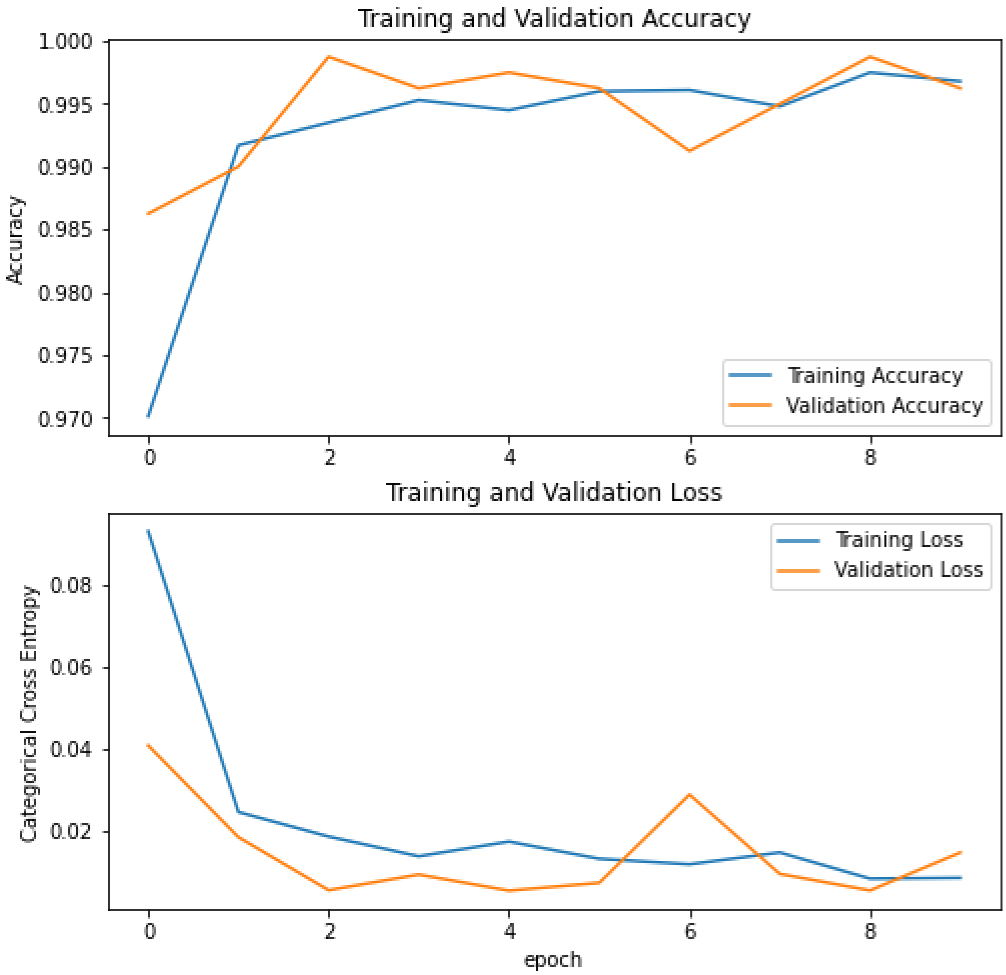


At evaluating the model on the test portion of our dataset the following results were obtained



```
['loss', 'accuracy']
2/2 [==============================] - 3s 1s/step - loss: 0.0088 - accuracy: 0.9975
[0.008827378042042255, 0.9975000023841858]
```

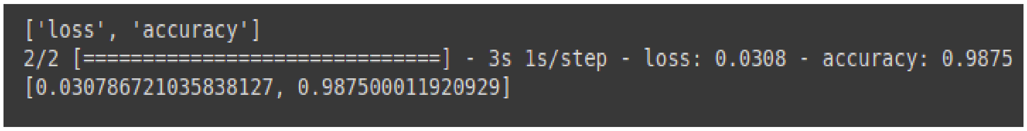
With accuracy 99.75% and loss 0.0088
Upon trial of EfficientNetB1, the results were as follows:



Training and Validation Accuracy

Training and Validation Loss

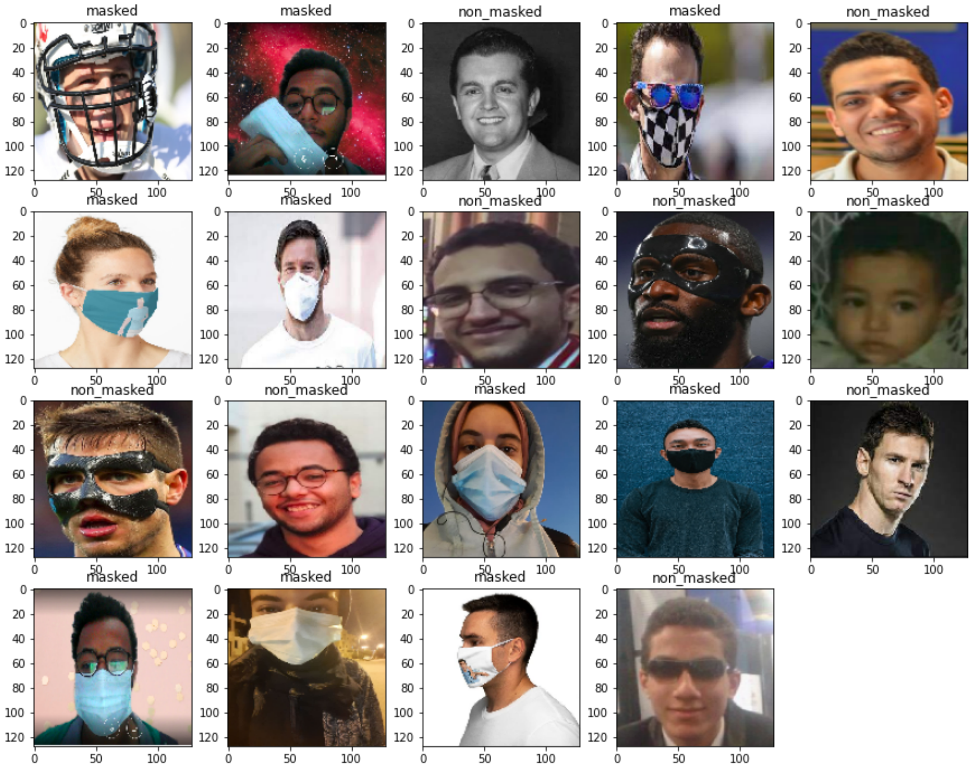
The accuracy obtained on the test portion:



```
['loss', 'accuracy']
2/2 [==============================] - 3s 1s/step - loss: 0.0308 - accuracy: 0.9875
[0.030786721035838127, 0.987500011920929]
```

We ended up choosing EfficientNetB0 over EfficientNetB1 because it gets better results and uses less parameters 5.3M compared to 7.8M in EfficientNetB1

Running on the micro dataset we collected, we get the following results:



To test the generalization of our model we imported another dataset containing masked and unmasked faces.

Using the VGG16 model on the test dataset resulted in an accuracy of 0.89. Through visualising some of the model's classified images, it appears that out of focus images or if they are not centred or zoomed out, that represents a potential mistake for the model. Attaching a high quality face recognition system before passing images to the model would boost the model's reliability. Rather,using the efficient network model on the same test dataset resulted in higher accuracy of 0.94. Efficient net was more robust against input variation in images, although it mistaken some images to be masked if there are more than one face as an input to the model. This encourages the idea of attaching a preprocessing face recognition system to gain best results out of the model.

III. Final Comparison

As a final comparison to choose the best model to go with in any needed application, we compared the two models by their number of parameters, floating point operations and recall metric. The following table shows the comparison.

MODEL	PARAMETERS	FLOPS	RECALL
VGG16	16.8 million	15.3 billion	1
MobileNetV2		0.289 billion	-
EfficientNetB0	5.3 Million	0.39 billion	1

According to the shown results we decided to choose ***EfficientNet.***

The results obtained over the test data for both models are pretty much comparable or even better than those mentioned in the literature review with the same models, especially the VGG16 model. For the EfficientNet model, as a state-of-the-art model, there was no comparing reference; however the model efficiency itself is pretty much optimum with a recall value of 1.

References

[1]"COVID Live Update: 158,564,907 Cases and 3,300,353 Deaths from the Coronavirus - Worldometer", Worldometers.info, 2021. [Online]. Available: https://www.worldometers.info/coronavirus/ . [Accessed: 09- May- 2021].
[2]"When and how to use masks", Who.int, 2021. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/when-and-how-to-use-masks. [Accessed: 09- May- 2021].
[3]"Object detection | TensorFlow Lite", TensorFlow, 2021. [Online]. Available: https://www.tensorflow.org/lite/examples/object_detection/overview. [Accessed: 09- May- 2021].
[4]"Image classification | TensorFlow Lite", TensorFlow, 2021. [Online]. Available: https://www.tensorflow.org/lite/examples/image_classification/overview. [Accessed: 09- May- 2021].
[5] Ashish Jangra, *Face Mask Detection ~12K Images Dataset* Available: https://www.kaggle.com/ashishjangra27/face-mask-12k-images-dataset [Accessed 29 May 2021].
[6] Jessica Li, *CelebFace dataset*, Available: https://www.kaggle.com/jessicali9530
[7] Omkar Gurav, *Face Mask Detection Dataset*. Available: https://www.kaggle.com/omkargurav/face-mask-dataset [Accessed 30 May 2021].

[8] A. Nieto-Rodríguez, M. Mucientes and V. Brea, "System for Medical Mask Detection in the Operating Room Through Facial Attributes", Pattern Recognition and Image Analysis, pp. 138-145, 2015. Available: 10.1007/978-3-319-19390-8_16 [Accessed 9 May 2021].
[9] B. Qin and D. Li, "Identifying Facemask-Wearing Condition Using Image Super-Resolution with Classification Network to Prevent COVID-19", Sensors, vol. 20, no. 18, p. 5236, 2020. Available: 10.3390/s20185236 [Accessed 9 May 2021].

[10] M. Ejaz, M. Islam, M. Sifatullah and A. Sarker, "Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition", 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 2019. Available: 10.1109/icasert.2019.8934543 [Accessed 9 May 2021].
[11] M. Loey, G. Manogaran, M. Taha and N. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic", Measurement, vol. 167, p. 108288, 2021. Available: 10.1016/j.measurement.2020.108288 [Accessed 9 May 2021].
[12] Z. Wang et al., "Masked Face Recognition Dataset and Application", arXiv.org, 2021[Online]. Available: https://arxiv.org/abs/2003.09093v2 [Accessed: 09- May- 2021].
[13] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria and J. Hemanth, "SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2", Sustainable Cities and Society, vol. 66, p. 102692, 2021. Available at: 10.1016/j.scs.2020.102692 [Accessed 9 May 2021].
[14] Jignesh Chowdary, G., Punn, N., Sonbhadra, S. and Agarwal, S., 2020. Face Mask Detection Using Transfer Learning of InceptionV3. Big Data Analytics, pp.81-90.
[15] Li, C., Wang, R., Li, J. and Fei, L., 2019. Face Detection Based on YOLOv3. Recent Trends in Intelligent Computing, Communication and Devices, pp.277-284.
[16] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud and J. -H. Kim, "An Automated System to Limit COVID-19 Using Facial Mask Detection in Smart City Network," 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2020, pp. 1-5, doi: 10.1109/IEMTRONICS51293.2020.9216386.
[17] S. Hussain et al., "IoT and Deep Learning Based Approach for Rapid Screening and Face Mask Detection for Infection Spread Control of COVID-19", Applied Sciences, vol. 11, no. 8, p. 3495, 2021. Available: https://www.mdpi.com/2076-3417/11/8/3495. [Accessed 16 June 2021].
[18] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", International Conference on Machine Learning, 2019, 2021. Available: https://arxiv.org/abs/1905.11946. [Accessed 1 June 2021].