Face Mask Detection in Real-Time using MobileNetv2

Mohamed Almghraby, Abdelrady Okasha Elnady

***Abstract:** Face mask detection has made considerable progress in the field of computer vision since the start of the Covid-19 epidemic. Many efforts are being made to develop software that can detect whether or not someone is wearing a mask. Many methods and strategies have been used to construct face detection models. A created model for detecting face masks is described in this paper, which uses "deep learning", "TensorFlow", "Keras", and "OpenCV". The MobilenetV2 architecture is used as a foundation for the classifier to perform real-time mask identification. The present model dedicates 80 percent of the training dataset to training and 20% to testing, and splits the training dataset into 80% training and 20% validation, resulting in a final model with 65 percent of the dataset for training, 15 percent for validation, and 20% for testing. The optimization approach used in this experiment is "stochastic gradient descent" with momentum ("SGD"), with a learning rate of 0.001 and momentum of 0.85. The training and validation accuracy rose until they reached their maximal peak at epoch 12, with 99% training accuracy and 98% validation accuracy. The model's training and validation losses both reduced until they reached their lowest at epoch 12, with a validation loss of 0.050% and a training loss of less than 0.025%. This system allows for real-time detection of someone is missing the appropriate face mask. This model is particularly resource-efficient when it comes to deployment, thus it can be employed for safety. So, this technique can be merged with embedded application systems at public places and public services places as airports, trains stations, workplaces, and schools to ensure subordination to the guidelines for public safety. The current version is compatible with both IP and non-IP cameras. Web and desktop apps can use the live video feed for detection. The program can also be linked to the entrance gates, allowing only those who are wearing masks to enter. It can also be used in shopping malls and universities.*

***Keywords:** Coronavirus, Computer vision, Face Mask Detection, CoVid-19, MobileNetV2*

I. INTRODUCTION

The World Health Organization (WHO) has proclaimed Covid-19 a public health emergency of worldwide concern. According to a WHO study, information about unexplained aetiology (unknown cause) pneumonia cases in Wuhan Metropolis, Hubei State, China was released for the first time in the last week of December 2019 [1].

Since that time, the World Health Organization has announced precautionary measures that must be followed, the most important of which are social distancing and wearing a protective mask.

Artificial intelligence supports image and video-based detection methods which can detects an object with high accuracy and classify whether the human is wearing or not wearing a mask. Face mask identification can be done using deep learning and machine learning methods such as "Support Vector Machines" and "Decision Trees", but with different datasets. [2].

The method of recognizing whether or not someone is wearing a mask and where his face is located is known as face mask detection. The issue is intertwined with general object detection, which identifies object classes. Face detection is the detection of a specific type of object. Object and face recognition has a Variety of applications. The related work in applying face mask detection is huge, not just using a neural network but also using machine learning algorithms, and applying various classifiers to the system.

"Nieto-Rodríguez, A., Mucientes, M., Brea, V.M" [4] trained the model on Celeb dataset and used a combination of LogitBoost and AdaBoost to detect and classify whether medical personnel is wearing masks. To decrease the number of false positive face detections, the system obtained a 95% accuracy rate. "Ejaz, Md.S., Islam, Md.R., Sifatullah, M., Sarker," on the other hand, employed "Principal Component Analysis" (PCA) to extract facial features and Viola-Jones to recognize faces [5]. Wearing masks reduces the accuracy of face resonation utilizing the PCA, with accuracy dropping to less than 70%. Using MFDD, RMFRD, and SMFRD, Wang, Z., Wang, G., and Huang, B. [6] created three face-mask datasets of artificially generated masked face images. Although face recognition has a low accuracy, it reduces below 50% when wearing masks, the datasets created are routinely used in industry and academics in various studies. A RetinaFaceMask detection architecture has been presented by Jiang, M., and Fan, X. The backbone is "ResNet" and "MobileNet", the neck is FPN, and the heads are context attention modules, with the heads being initialized using kaiming's technique that concentrates on the face and mask features [7]. In France, certain proposed AI-based approaches and software solutions have recently been combined with security cameras to block the spread of "COVID-19" by recognizing people without face masks ("Paris Metro system"). There are many object identification approaches based on deep learning. Only a few of these algorithms are employed in face mask detection "Li, C., Wang, R., Li, J., Fei, L." [3].

Manuscript received on August 06, 2021.
Revised Manuscript received on August 17, 2021.
Manuscript published on August 30, 2021.
* Correspondence Author

Mohamed Almghraby, UG Student, Department of Mechatronics, Faculty of Engineering October 6 University, Egypt

Abdelrady Okasha Elnady*, Head, Department of Mechatronics, Faculty of Engineering October 6 University, Egypt

One of these algorithms is "YOLOv3". "You Only Look Once", version 3 (YOLOv3) is a real-time object detection system that recognizes specific things in videos, live feeds, and images Artificial Intelligence (AI) is utilized in object categorization systems programs to perceive certain objects in a class as subjects of interest. Objects in images are categorized into groups, with comparable qualities being grouped together and others being ignored until otherwise told.

In this research, a face mask detector model is introduced. This model uses (MobileNetV2) Architecture to train the model. The model was trained and tested using two independent datasets of face masks. To recognize faces from video streams, a pretrained model of MobileNetV2 architecture was used. A variety of packages of "machine/deep" learning methodologies and "image processing techniques" were used in addition to the OpenCV framework. To provide effective monitoring and enable proactively, different steps were also used, including "data augmentation, loading the classifier (MobileNetV2), building the fully-connected (FC) layer, pre-processing and loading the image data, applying classifier, training phase, validation, and testing phase".

II. METHODOLOGY

A. Network Architecture

MobileNetV2 is an architecture of bottleneck depth-separable convolution building of basic blocks with residuals [8]. It has two types of blocks as shown in Fig. 1. Both blocks have three layers. The first one is 1x1 convolutions with "ReLU6". The second layer contains depth-wise "convolution," and the third layer contains a 1x1 "convolution" with no non-linearity. The first layer is a one-stride residual block. As seen in Table 1, the second layer is also a residual block with stride 2 and is used for shrinking.
The methodology of object detection is to make a classification to determine the input class, regression to adjust the bounding box. With the exception of the last completely connected layers, most backbone networks for detection are networks for classification tasks. The backbone network serves as a simple feature extractor for object detection tasks, taking images as input and producing feature maps for each input image.

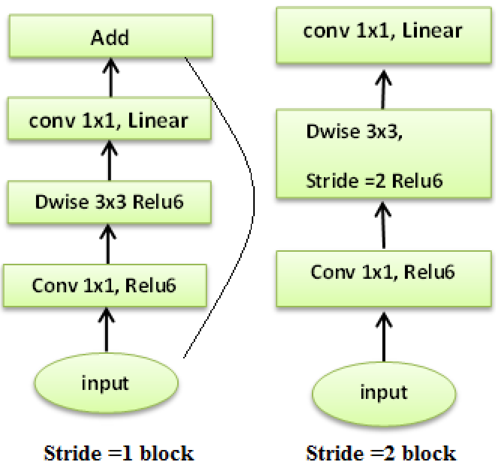

Fig. 1: MobileNetV2 blocks

Table I: Residual block in the k to k' channel

Input	Operator	Output
"h x w x k"	1 x 1 "conv2d", "ReLU6"	"h x w x tk"
"h x w x tk"	3 x 3 "dwise" S=s, "ReLU6"	$\frac{h}{s}$ x $\frac{w}{s}$ x tk
$\frac{h}{s}$ x $\frac{w}{s}$ x tk	Linear 1 x 1 conv2d	$\frac{h}{s}$ x $\frac{w}{s}$ x k′

The predefined trained techniques are usually used to extract feature maps with high-quality classification problems. This part of the model is called the base model. The base model is the MobileNetV2 network which is used the "image net" weights. ImageNet is an image database that has been trained on hundreds of thousands of images, and as a result, it is extremely useful for image categorization. The evaluated "bounding boxes" are compared to the "ground truth boxes" during training, and the trainable parameters are changed as needed during backpropagation. A kernel is utilized in each feature space to produce outcomes that show corresponding scores for each pixel, whether or not an item exists, as well as the appropriate bounding box dimensions. The MobileNet design is made up of two parts: a base model and a classifier. In this model, the base model is reused, the head is trimmed, and two fully connected layers are employed [20].

B. Transfer Learning

Deep neural network training is costly and time-consuming since it necessitates a lot of processing power and other resources. Transfer learning based on deep learning has evolved to speed up the network train. Transfer learning is essentially the process of training and predicting on a new dataset using a pre-trained model that has been learned on a previous dataset. The Imagenet dataset is used in most computer vision applications as a source of pre-trained weights [9]. This arrangement is used as a baseline. The network employs pre-trained a set of weights derived from a similar task–image classification– and has been trained on a certain collection of data like mask face detection in the ablation study of transfer learning. Using the MobileNet backbone, the results reveal that "transfer learning" can improve "detection performance" by 34% in both cases. One possible explanation is that using pre-trained weights from a closely related job improves feature extraction capabilities. "InceptionV3, Xception, MobileNet, MobileNetV2, VGG16, ResNet50", and other pre-trained models [12, 13, 14, 15, 16, and 17] have been trained using the "ImageNet dataset", which contains 14 million images.

The last layer of MobileNetV2 is removed in this effort, and the network is fine-tuned by adding four more levels. Finally, to classify whether a person is wearing a mask, a key thick layer with two neurons and a softmax activation function is added. Figure 2 shows a schematic representation of the proposed methodology.

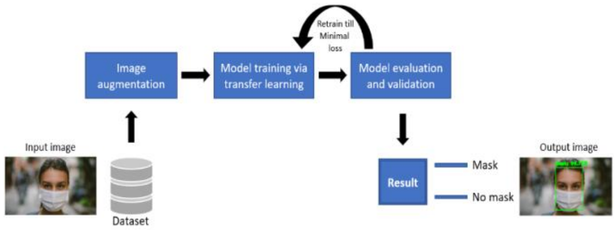
Retrain till Minimal loss
Image augmentation
Model training via transfer learning
Model evaluation and validation
Input image
Dataset
Result
Mask
No mask
Output image

Fig. 2: Schematic diagram of the proposed system.

C. Data augmentation

Because the training data set is limited in size, data augmentation is employed to extend the amount of the dataset for training by manipulating images in the dataset artificially. “Shearing, contrasting, horizontally flipping, rotating, zooming, and blurring” are all used to enhance the training images. By rescaling the input image by 224*224 and converting it to a single channel, the size of the current model can be reduced.

III. EXPERIMENT RESULTS

A. Dataset

Face Mask Dataset contains 1800 images with and without a mask. From this data set, 1000 photos were used for training and 800 photos were used for testing. All those images are actual images extracted from the Kaggle dataset. There is no class imbalance because the ratio of masked images to non-masked images is equal. A dataset is divided into three parts after it is created: a training “dataset”, a testing “dataset”, and a validation “dataset”. The goal of data splitting is to prevent overfitting.

The goal is to achieve generalization, which is required for this model to perform well on test data. The model is trained using the training set. This data is used to train the model, which then optimizes its parameters. The validation dataset is used to fine-tune the hyperparameters (number of hidden layers, learning rate, regularization parameters). The learning can be halted using a training dataset if the model has performed well enough on the validation dataset. The test set is the final set of data used to objectively assess a final model fit on the training dataset. The split of the dataset is highly dependent on the model and the dataset itself; for example, if you have a lot of training data, you'll need a larger model to change that many hyperparameters, in addition to a sizable validation dataset. If the model includes a small number of datasets, it's important to avoid underfitting and keep the number of validation datasets to a minimum. The present model dedicates 80 percent of the training dataset to training and 20% to testing, and splits the training dataset into 80% training and 20% validation, resulting in a final model with 65 percent of the dataset for training, 15 percent for validation, and 20% for testing.

B. Evaluation Metrics

To evaluate the performance of the proposed model, many performance measures are required. “Accuracy, precision, F1 score, and recall” are the performance measurement parameters, and they may be found in equations (1-4) below [18]. “True Positive” is “TP”, “True Negative” is “TN”, “False Positive” is “FP”, and “False Negative” is “FN”.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$"Recall = \frac{TP}{TP+FN}" \quad (2)$$

$$"Accuracy = \frac{(TP+TN)}{[(TP+FP)+(TN+FN)]}" \quad (3)$$

$$Fl\ score = 2\,\frac{Precision*Recall}{(Precision+Recall)} \quad (4)$$

C. Training

The “bounding boxes” with varied sizes and aspect ratios are compared to ground truth at training time for each pixel, and the best fitting box is selected using the Union and Intersection method. The measure of Union and Intersection is used to determine how accurate an item detector is on a particular dataset. Union and Intersection is commonly used to evaluate the performance of the MobilNetV2 architecture by using transfer learning from the base network, removing the last layer, and adding two FC layers and softmax to the output layer. Any technique that produces anticipated bounding boxes as an output can be evaluated using IoU. Two parameters from the current model are required to evaluate an item detector employing Union and Intersection: the “ground-truth bounding” and the anticipated “bounding boxes”. As long as the two sets of bounding boxes are compatible, Union and Intersection [22] can be employed. In Fig. 3 below [19], a visual example of a “ground-truth bounding box” vs. a predicted “bounding box” is shown.

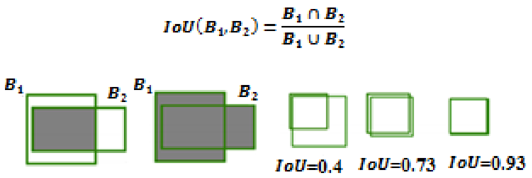

Fig. 3: Visual explanation of IoU

IV. RESULT AND ANALYSIS

The optimization approach used in this experiment is “stochastic gradient descent” with momentum (“SGD”), with a learning rate of α = 0.001 and momentum of β = 0.85. An NVIDIA GeForce 940 MX was used to train the models. The data was separated into three parts: a “train set”, a “validation set”, and a “test set”, each comprising 6000, 1500, and 2000 images. The algorithm was created using the TensorFlow deep learning framework, and it is implemented using MobileNet. There are 12 epochs in each experiment. The input picture size for the MobileNet backbone is 224x224 with a batch size of 32.

The “training and validation” accuracy rose until they reached their maximal peak at epoch 12, with 99% training accuracy and 98% validation accuracy as shown in Fig. 4. The model's training and validation losses both reduced until they reached their lowest at epoch 12, with a validation loss of 0.050% and a training loss of less than 0.025% as shown in Fig. 5.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

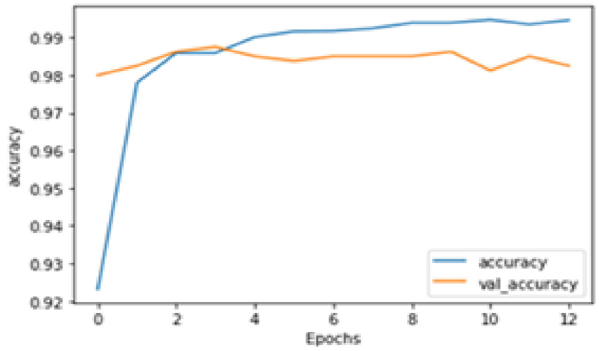

Fig.4: Accuracy in training and validation

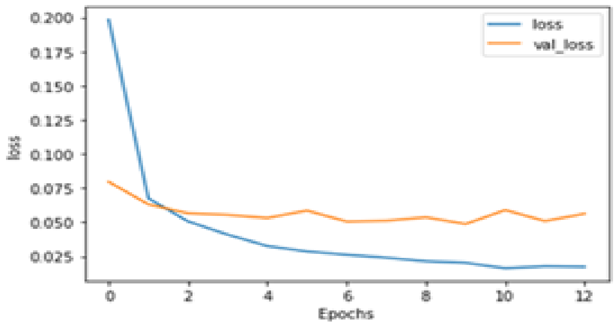

Fig.5: Loss in training and validation

A. Limitation

There are some hardware challenges: my GPU takes a long time and consumes a lot of power to train the model, so I train the model on 224*224 images. Inference after freezing the backbone weight, there is a little bit of lag in live streaming, so I tried to use Google Colab, but there was a problem with live streaming on Google Colab, so I tried to use my phone's camera to stream, but Colab refused to accept it.

B. Future work

It is planned to improve Face Mask Detection to identify people who aren't wearing a mask and apply it on embedded devices like raspberry pi or jetson Nano. An alarm system can also be implemented to make a sound when someone without a mask enters the area or laser detectors that mark the individual who is not wearing a mask.

V. CONCLUSIONS

A face mask detector is proposed in this research, which can help with public healthcare. The backbone of the MobileNet face detection architecture is MobileNet, with dense as the last layer. MobileNet is a light backbone that can handle both high and low processing workloads. Transfer learning is used to adopt weights from a related job, face detection, which is learned on a big dataset, in order to extract more robust features. On a public face mask dataset, the suggested method yields state-of-the-art results. The current version is compatible with both IP and non-IP cameras. Web and desktop apps can use the live video feed for detection. The program can also be linked to the entrance gates, allowing only those who are wearing masks to enter. It can also be used in shopping malls and universities.

ACKNOWLEDGMENT

I'd like to express my thankfulness and appreciation for Assoc. Prof. *Abdelrady Okasha Elnady*, whose advice, encouragement, and support have been vital throughout this project. I admire how compassionate and instructive you are. Thank you for everything you taught us, in my academic life even though in my daily life. These will stay with me for the rest of my life.

REFERENCES

1. Asadi, Sima, et al. "Aerosol emission and superemission during human speech increase with voice loudness." Scientific reports 9.1 (2019): 1-10. https://doi.org/10.1038/s41598-019-38808-zhttps://doi.org/10.1038/s41598-019-38808-z
2. Loey, Mohamed, et al. "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic." Measurement 167 (2021): 108288. https://doi.org/10.1016/j.measurement.2020.108288https://doi.org/10.1016/j.measurement.2020.108288
3. Li, Chong, et al. "Face detection based on YOLOv3." Recent Trends in Intelligent Computing, Communication and Devices. Springer, Singapore, 2020. 277-284. https://doi.org/10.1007/978-981-13-9406-5_34.
4. Nieto-Rodríguez, A., Mucientes, M., Brea, V.M.: System for medical mask detection in the operating room through facial attributes. In: Paredes, R., Cardoso, J.S., and Pardo, X.M. (eds.) Pattern recognition and image analysis. Pp. 138–145. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-19390-8_16.
5. Ejaz, Md.S., Islam, Md.R., Sifatullah, M., Sarker, A.: Implementation of principal component analysis on masked and non-masked face recognition. In: 2019 1st international conference on advances in science, engineering and robotics technology (ICASERT). Pp. 1–5 (2019). https://doi.org/10.1109/ICASERT.2019.8934543.
6. Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei,Y., Chen, H., Miao, Y., Huang, Z., Liang, J.: Masked face recognition dataset and application.arXiv:2003.09093 [cs]. (2020).
7. Jiang, M., Fan, X., Yan, H.: RetinaMask: a face mask detector. arXiv:2005.03950 [cs]. (2020)
8. M. Sandler, 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(CVPR), 2018
9. Imagenet Dataset https://image-net.org/download.php
10. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CoRR abs/1610.02357 (2016). http://arxiv.org/abs/1610.02357
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385.
12. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017). http://arxiv.org/abs/1704.04861
13. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). pp. 730–734 (2015)
14. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR abs/1801.04381 (2018). http://arxiv.org/abs/1801.04381
15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR abs/1512.00567 (2015).http://arxiv.org/abs/1512.00567

16. Faisal Bashir, Fatih Porikli: Performance Evaluation of Object Detection and Tracking Systems. https://www.researchgate.net/publication/237749648
17. Hamid Rezatofighi, Nathan Tsoi, Jun Young Gwak, Amir Sadeghian, Ian Reid, Silvio Savarese. https://arxiv.org/abs/1902.09630
18. MobileNetV2: Inverted Residuals and Linear Bottlenecks, Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. arXiv preprint. arXiv:1801.04381, 2018.
19. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, Howard AG, Zhu M, ChenB, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H, arXiv:1704.04861, 2017.
20. 20. Haddad, J., 2020. How I Built A Face Mask Detector For COVID-19 Using Pytorch Lightning. [online]Medium. Available at:https://towardsdatascience.com/how-i-built-a-face-mask-detector-for-covid-19-using-pytorch-lightning-67eb3752fd61.
21. Rosebrock, A., 2020. COVID-19: Face Mask Detector With Opencv, Keras/Tensorflow, And Deep Learning-Pyimage search. [online] PyImageSearch. Available at: https://www.pyimagesearch.com/2020/05/04/covid-19-face-mask-detector-with-opencv-keras-tensorflow-and-deep-learning/.
22. Review about Intersection over union. https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/

AUTHORS PROFILE

Mohamed Khalid Almghraby, UG Student, Mechatronics Engineering, October 6 University. Main Interests: Automatic Machine Learning, Deep Learning, Hyper Parameter Optimization, Computer Vision. Mail Id: mohamed.k.almghraby@gmail.com

Assoc. Prof. Abdelrady Okasha Elnady, Head of Mechatronics Department, Faculty of Engineering, October 6 University. Main Interests: Robotics, Drones, UAV, Image Processing. Mail ID: rady_nady.eng@o6u.edu.eg.