Face Mask Detection
MTCNN_Xception

Abstract- The task of finding proportion of people wearing masks in a given image can be subdivided into two sub-tasks as first identifying the faces in the image then classifying whether a person is wearing a mask or not.
Recent studies show that deep learning approaches can achieve impressive performance on these two tasks.
For the purpose of face detection we use MTCNN model pretrained on WIDER FACE benchmark for face detection, while for Classification Xception model is used pretrained on Imagenet dataset.

1. Architecture

MTCNN :In this paper, authors propose a deep cascaded multi-task framework which exploits the inherent correlation between them to boost up their performance. In particular, our framework adopts a cascaded structure with three stages of carefully designed deep convolutional networks that predict face and landmark location in a coarse-to-fine manner.

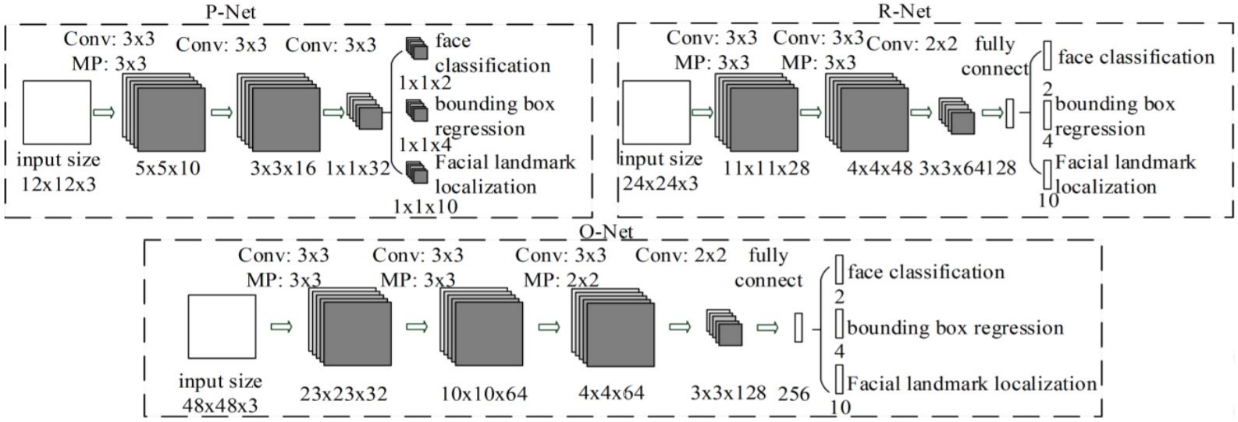


Fig. 1. The architectures of P-Net, R-Net, and O-Net, where "MP" means max pooling and   "Conv" means convolution.

P-Net : A fully convolutional network, to obtain the candidate windows and their bounding box regression vectors. Then uses the estimated bounding box regression vectors to calibrate the candidates. After that, employing NMS to merge highly overlapped candidates.
R-Net: All candidates are fed that further rejects a large number of false candidates, performs calibration with bounding box regression, and NMS candidate merge.
O-Net : This stage is similar to the second stage, but in this stage we aim to describe the face in more detail. In particular, the network will output five facial landmarks' positions.

After extracting face boxes each box is passed to the Xception classification model.

Xception :Authors present an interpretation of Inception modules in convolutional neural networks as being an intermediate step in-between regular convolution and the depthwise separable convolution operation In this light, a depthwise separable convolution can be understood as an Inception module with a maximally large number of towers. They propose a novel deep convolutional neural network architecture inspired by Inception, where Inception modules have been replaced with depthwise separable convolutions.

The Xception architecture has 36 convolutional layers forming the feature extractor base of the network. The 36 convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules. In short, the Xception architecture is a linear stack of depthwise separable convolution layers with residual connections

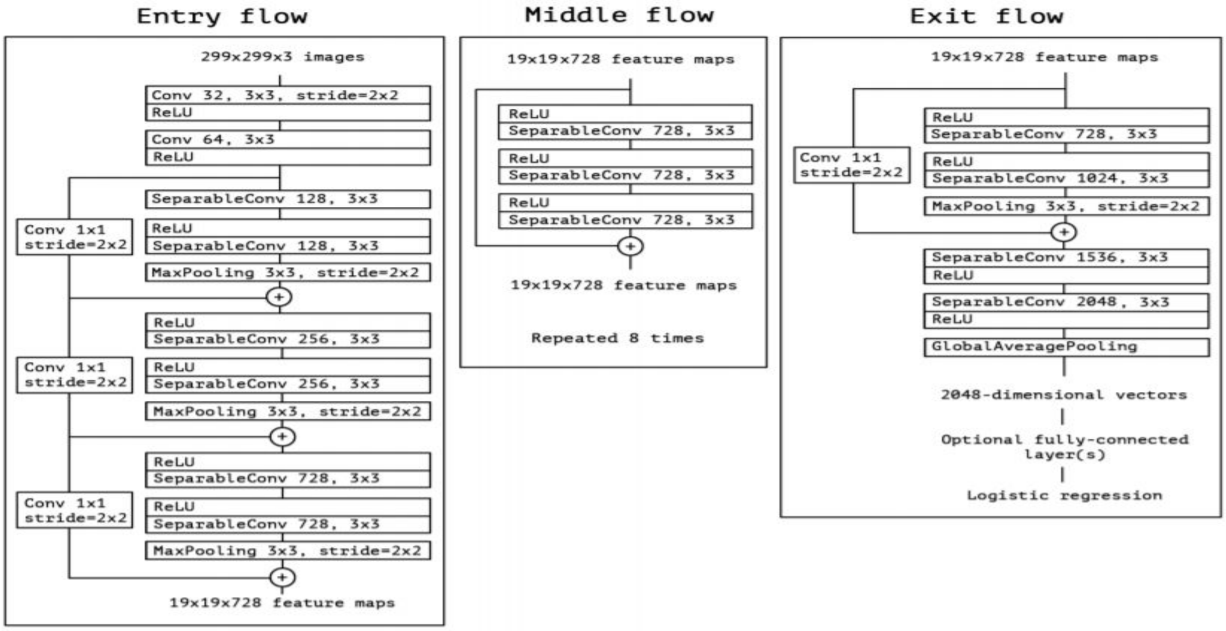


Figure 2. The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow. Note that all Convolution and SeparableConvolution layers are followed by batch normalization(not included in the diagram).

After getting feature extractor from “Xception” as base model, Global-AvgPooling -Dense Dropout-Dense are used for fine tuning on Mask/No-Mask classification.

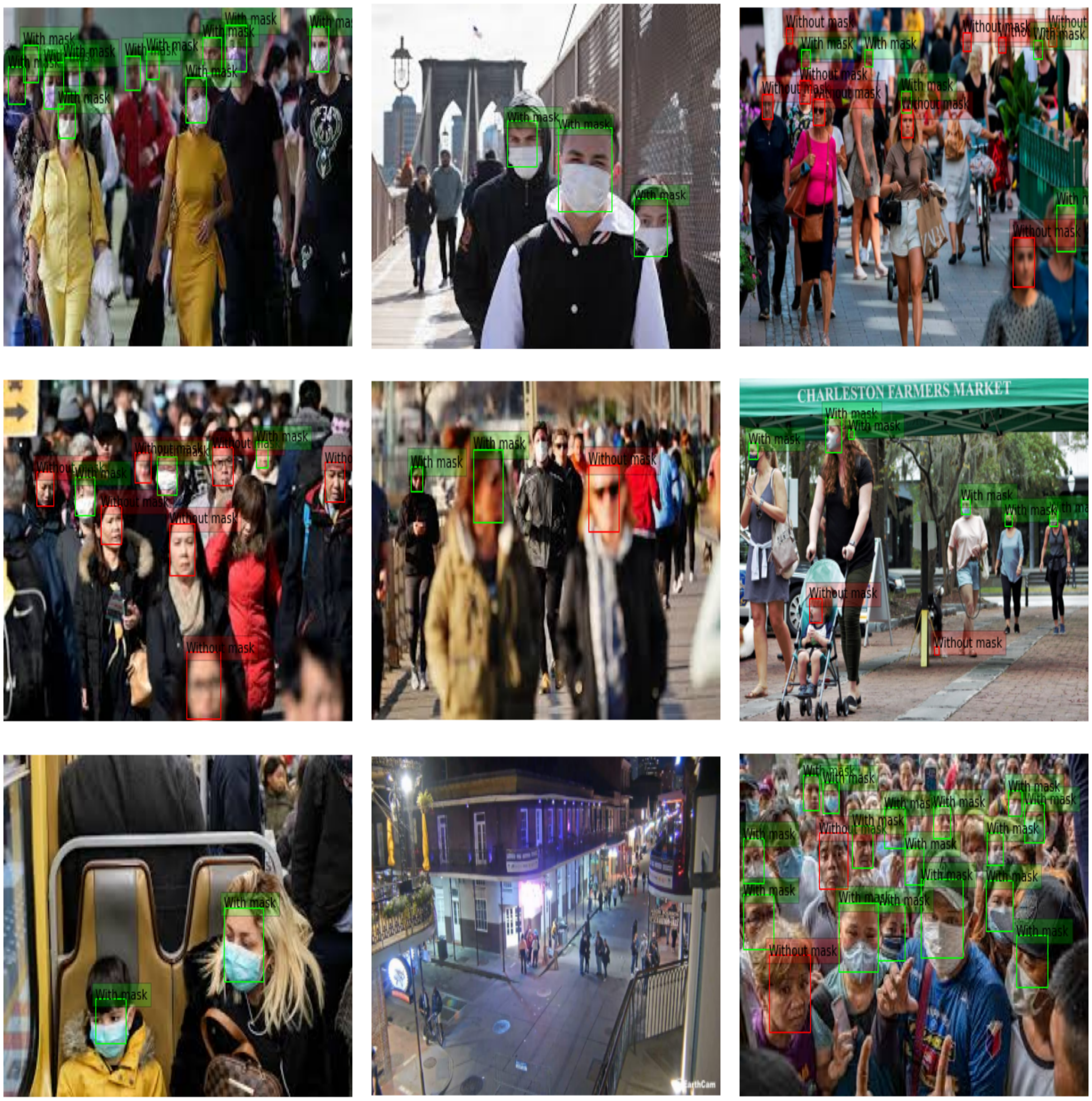
2. Training Specifications

The Real World Masked Face Detection (RMFD) dataset is used for classification training containing, With mask: 8072 images and Without mask: 8086 images.

Dense layer with 1024 units and Dropout of 0.5 is used for training model for 3 epochs with Adam as optimizer , learning rate of 10^{-4} and decay of 10^{-5} . Standard 80/20 train-test split is used. Accuracy of 0.997 on training set and 0.991 on test set is obtained.

For MTCNN (0.7,0.7,0.7) threshold for NMS, (0.6,0.7,0.8) threshold for MTCNN steps is used. Minimum face size used is 20 pixels.

3. Results


With mask
Without mask
CHARLESTON FARMERS MARKET
EarthCam