ABSTRACT

This report discusses the application of deep learning algorithms to develop a face mask detection and authentication system. Due to the air borne nature of the novel Corona Virus it is necessary to wear a face mask. People working at places using a facial recognition attendance system have to remove the mask for it to work which can be dangerous. Therefore we propose an automated system which detects whether the person is wearing a mask and to perform the authentication without removing the mask. The mask detection module is a two stage model. The first stage is a face detector model which uses single shot object detector to detect faces from an image. The second stage is a classifier model to classify the images into 2 classes masked and unmasked. The face recognition module uses a siamese network approach and encodes the test image using a deep CNN to compare it with the encoding of the candidate images.

Contents

1 Introduction

The COVID19 pandemic has brought may changes in the world and arisen many different problems in the daily routine. Face masks are necessary for protection in this pandemic and have become part of the new normal. The topic of this project is "Face Mask detection and recognition" inspired by various problems face masks in our daily life bring. Face mask detection is becoming a very important aspect in surveillance and security monitoring. Face recognition with face masks is a wide-range challenge from face unlock systems in various devices to high end tracking and monitoring services used by all nations for security purposes. These problems have provided us with an opportunity to solve on current real-life problems.

These problems that have arisen are very time sensitive which require solution as soon as they originate. These problems are especially challenging because of unavailability of original images in a dataset. Many datasets contain morphed images of faces with masks which harms the accuracy of the model. Also the absence of large datasets makes this task more challenging. Currently many masks with various types of colors and designs along with animations with facial features are available which are special cases and need to be considered. However, lack of such data makes it difficult to incorporate these factors in the model.

This project is divided into two modules: Face mask detector and Face recognition with mask respectively. We have used a class balanced dataset of 3835 images by compiling from sources like Kaggle datasets, RMFD(Real-World Masked Face dataset) dataset. It consists of two classes: without_mask and with_mask. Both contain about same number of images 1916 and 1919 respectively. The with_mask class of the dataset contains only original images and hence no morphed images were used to create a large dataset. This has been very beneficial for the accuracy of the model.

2 Literature Survey

The exploration led so far for question discovery and following articles in Face Mask detection are discussed in this area. The main challenge to overcome in the Face mask detector models is the unavailability of sufficient data and improve the process of training, validation and testing of model. The feature extraction is another main aspect of this problem which have been explored thoroughly in the following Models. Lastly to reduce the computational load due to the feature extraction deep learning models (computational intensive), usually a classical machine learning classifier is used for the classification task.

2.1 Hybrid deep transfer learning Model

: Hybrid model utilizing deep learning and machine learning for face mask identification was introduced. The proposed model comprises of two segments. The principal segment is intended for feature extraction utilizing Resnet50. While the subsequent segment is intended for the classification of face masks utilizing decision trees, Support Vector Machine (SVM), and ensemble methods. Three face masked datasets have been chosen for examination. The Three datasets are the Real-World Masked Face Dataset (RMFD), the Simulated Masked Face Dataset (SMFD), and the Labeled Faces in the Wild (LFW).

The SVM classifier accomplished 99.64% testing precision in RMFD. In SMFD, it accomplished 99.49%, while in LFW, it accomplished 100% testing exactness.[Loe+20a]

Mainly ResNet50 is used for feature extraction phase, while machine learning approach like SVM is used for the the training, validation and testing phase. Residual Network(ResNet) is a kind of deep transfer learning based on residual learning. ResNet50 used here has 50 layers that start with a convolutional layer and end with a full connected layer. The proposed structure of ResNet50 is shown in figure. For the task of classification the last layer of ResNet50 is replaced by classic Machine learning classification algorithms like Support vector Machine, decision tree and ensemble methods. [Loe+20a]

2.2 Transfer Learning of InceptionV3:

A transfer learning model is proposed for recognizing the individuals who are not wearing masks. The proposed model is worked by calibrating the pre-prepared cutting edge deep learning model, InceptionV3. The proposed model is prepared and tried on the Simulated Masked Face Dataset (SMFD). Image augmentation method is received to address the restricted availability of data for better training and testing of the model. The model beat the other as of late proposed approaches by accomplishing a precision of 99.9% during preparing and 100% during testing. [Cho+20]

The first phase of the model is responsible for Image augmentation of the training data to deal with over-sampling. Image augmentation is a procedure used to build the size of the preparation dataset by misleadingly altering pictures in the dataset. In this research, the preparing pictures are increased with eight particular activities to be specific shearing, contrasting, flipping on a level plane, rotating, zooming, obscuring. The created dataset is then rescaled to 224 x 224 pixels, and changed over to a solitary channel greyscale portrayal. [Cho+20].

A transfer learning based methodology is proposed that uses the InceptionV3 pre-trained model for ordering the individuals who are not wearing face mask. For this work, the last layer of the InceptionV3 is taken out and is finetuned by adding 5 additional layers to the organization. The 5 layers that are added are an normal pooling layer with a pool size equivalent to 5 x 5, a leveling layer, followed by a thick layer of 128 neurons with ReLU enactment capacity and dropout rate of 0.5, lastly a conclusive thick layer with two neurons and softmax activation function is added to classify whether an individual is wearing mask. This transfer learning model is prepared for 80 epoch with every epochs having 42 steps. The architecture of the proposed model is shown in the figure. [Cho+20]

2.3 Deep learning model based on YOLO-v2 with ResNet50

: The proposed model comprises of two parts. The principal part is intended for the feature extraction process dependent on the ResNet-50 deep transfer learning model. While the subsequent segment is intended for the detection of clinical face masks dependent on YOLO v2. Two clinical face mask datasets have been consolidated in one dataset for this research. To improve the object detection , mean IoU has been utilized to gauge the best number of anchor boxes. The accomplished outcomes inferred that the adam streamlining agent accomplished the most noteworthy normal exactness level of 81% as an identifier.

The model consists of three main components. The first component is responsible for estimation of the number of anchor boxes. The second component is the data augmentation part while the third component is the object detector part. The proposed model utilizes mean Intersection over Union (IoU) which is given by the dividing Overlap Area between bounding box of target and predicted by Combined area between bounding box of target and predicted. IoU in object detection is a technique to figure the distance of similarity between the bounding box of target and predicted output. In the training data, the mean IoU ensure that the anchor boxes overlap with the boxes. The best number of anchors is 23 (mean IoU = 0.8634) for maximum detector perfomance according to this research. [Loe+20b]

Image augmentation is a procedure used to build the size of the preparation dataset by misleadingly altering pictures in the dataset. To increase the masked face dataset the data and their box labels were flipped horizontally. In the proposed model, ResNet-50 utilized as a deep transfer model for feature extraction. A residual neural network (ResNet) is a class of deep transfer learning dependent on a residual network . ResNet-50 has 16 lingering bottleneck obstructs each square has convolution size 1x1, 3x3 and 1x1 with feature maps (64, 128, 256, 512, 1024). The object detection network (YOLO v2) is a convolutional neural organization contain not many convolutional layers, transform layer, lastly output layer. The transform layer removes initiations of convolutional layer and improves the dauntlessness of the deep neural network. The transform layer changes over the bounding box forecast to be in the outlines of the target box. The areas of pure bounding box of the target is created by the output layer. [Loe+20b]

3 Related Work

3.1 Single Shot Detector (SSD)

Single Shot Detector is faster compared to regional proposal network [Ren+15] based approaches as unlike those approaches, in this approach a single shot is needed to detect multiple objects within the image. This approach named SSD, discretizes the output space of bounding boxes to a bunch of default boxes over various perspective aspect ratios and scales per feature map location. At prediction time, the network creates scores for the presence of each object classification in each default box and delivers acclimations to the box to all the more likely match the article shape. Furthermore, the network consolidates predictions from various feature maps with various resolutions to normally deal with objects of different sizes. SSD is basic comparative with strategies that require object proposal since it totally takes out proposal generation and respective pixel or feature resampling stages and encapsulates all calculation in a single network. This makes SSD simple to train and clear to coordinate into frameworks that require a detector component. [Liu+16]

The SSD approach depends on a feed-forward convolutional network that produces a fixed-size assortment of bounding boxes and scores for the presence of object class cases in those boxes, trailed by a non-maximum suppression step to create the final detections. The early network layers depend on a standard design utilized for high quality image classifications (shortened before any classifications layers [Ren+15]), which is known as the

base network. [Liu+16] The auxillary structure is added to the network which produces detections with the features like:

- **Multi-scale feature maps for detection:** The convolutional feature layers are added to the truncated base network [Ren+15]. These layers allow predictions of detections at multiple scales as they decrease in size progressively. This model responsible for predicting detections is different for each feature layer.

- **Convolutional predictors for detection:** Utilizing a set of convolutional filters, a fixed set of detection predictions are made by each added feature layer. The basic element for predicting parameters of a potential detection for a feature layer of size m x n with p channels is a *3* x *3* x p small kernel that can produce score for a class or a shape offset relative to the default box coordinates.The kernel applied at every one of these m x n locations produces an output value. Bounding box offset output values are measured relative to the default bounding box coordinates relative to each feature map location [Liu+16].

- **Default boxes and aspect ratios:** A set of default bounding boxes are associated with each feature map cell for multiple feature maps at the top of the network. The default boxes tile the feature map in a convolutional way, so that the location of each box is fixed in relation to its corresponding cell. We predict the offsets relative to the default box shapes in the cell for each feature map cell, as well as the per-class scores that display the existence of a class instance in each of those boxes. Specifically, we measure c class scores and the 4 offsets for each box out of k at a given position relative to the original default box form. This results in a total of filters of (c + 4) k that are applied around each place in the map of features, yielding (c + 4)kmn outputs for a map of m x n features. This default boxes are applied to several feature maps of different resolutions which allows different default box shapes in several feature maps and efficiently discretizes the space of possible output box shapes.

3.2 Xception Network

A novel deep convolutional neural network is the proposed model, Inception-inspired architecture, where Inception modules have been replaced with convolutions that are separable in depth. In the analysis [Cho17], this design, called Xception, marginally outperforms Inception V3 on the ImageNet dataset (for which Inception V3 was intended), and significantly outperforms Inception V3 on a larger dataset of 350 million images and 17,000 classes of image classification. The efficiency gains are not due to increased capability but rather to a more effective use of model parameters, because the Xception architecture has the same number of parameters as Inception V3. [Cho17]

Hypothesis of Xception architecture is: that it is possible to map cross-channel correlations and spatial correlations in convolutional neural network function maps. Decoupled completely. This proposed architecture is called Xception, which stands for "Extreme Inception," since this hypothesis is a stronger version of the hypothesis underlying the Inception architecture. There are 36 convolutional layers in the Xception architecture that form the network's feature extraction foundation. This exclusively investigates image classification in the experimental assessment and therefore a logistic regression layer

will follow our fully convolutional base [Cho17]. Optionally, fully connected layers may be added before the logistic regression layer. The 36 convolutional layers are organized into 14 modules, all of which, except for the first and last modules, have linear residual connections around them.

The architecture of Xception is a linear stack with residual connections of depthwise separable convolution layers. Unlike architectures such as Inception V2 or V3 that are much more difficult to describe, this makes the architecture very simple to define and alter. As part of Keras Applications Module2, an open-source implementation of Xception using Keras and TensorFlow is provided. In compliance with the MIT license. [Cho17]

4 Proposed Models

The proposed models is divided into two modules: Face mask detector and Face recognition with mask are described further:

4.1 Face Mask Detection

The task of this module is to detect whether the people in the image are wearing a face mask or not.We have developed a two stage model for this task. The first stage of the model detects the faces of people in the image, and the second stage uses the cropped image from the first stage to detect whether the person is wearing a mask or not.

In the first stage we use the Single shot detector which is an object detection model to detect and localise the faces from the image. Here we have used a pretrained SSD which is trained on 140000 iterations. The SSD model uses ResNet10 as the backbone architecture to extract the features. The model extracts features from different earlier layers of ResNet as well as it applies come convolution layers on the top of it.It applies the detector and classifier model using those features as the input. Doing this helps us in detecting all the faces in the image regardless of its size in the image because the features taken from earlier layers are not that aggregated and help in detecting the smaller faces and the features from the layers applied on top of ResNet are well aggregated and help detecting bigger faces.

(Architecture Image here)

In the second stage we use a classifier model to detect if the cropped face image obtained from the first stage is masked or unmasked. We used a dataset of around 4000 images collected from the internet for training the model. The dataset is class balanced that is, it contains equal number of masked and unmasked class of images.We used an Image data generator for data augmentation. In augmentation we used rotation range, zoom range, width shift, height shift and horizontal flip. We have used transfer learning from Xception model trained on Imagenet dataset. We stacked an Average Pooling layer, and two fully connnected layers on top of it. We achieved an accuracy of 0.99 by training the model on 255 epochs with a learning rate of 0.0001.

(Accuracy Graph here)

4.2 Face Recognition with Mask

In this module we added a feature of authentication using face recognition with the mask on. We have used the siamese network approach for the task. We pass the cropped image of the face through a deep CNN model to extract its features. It encodes the image into a 128 dimensional encoding. Then the encodings are compared with the encodings of the faces in our databse with which we want to check. If the distance of the test encodings from a candidate encoding is below a certain threshold then it is considered as a match and we write the title of the candidate on the top of the face bounding box generated on the test image.

5 Results

Result Images

6 Conclusion

We conclude that an accurate automated mask detector system can be constructed using a SSD model for face detection and CNN for classification. And facial recognition system using a siamese network approach can function correctly even with a face mask on.

Bibliography

[Ren+15] Shaoqing Ren et al. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[Liu+16] Wei Liu et al. "Ssd: Single shot multibox detector". In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[Cho17] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[Cho+20] G Jignesh Chowdary et al. "Face Mask Detection using Transfer Learning of InceptionV3". In: *arXiv preprint arXiv:2009.08369* (2020).

[Loe+20a] Mohamed Loey et al. "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic". In: *Measurement* 167 (2020), p. 108288.

[Loe+20b] Mohamed Loey et al. "Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection". In: *Sustainable Cities and Society* (2020), p. 102600.

List of Useful Websites:

- https://github.com/chandrikadeb7/Face-Mask-Detection
- https://github.com/X-zhangyang/Real-World-Masked-Face-Dataset
- https://maelfabien.github.io/deeplearning/xception/#