Facemask Detection using Inception V3 Model and Effect on Accuracy of Data Preprocessing Methods

Yongyuan Li
University of Nevada, Las Vegas, USA
*Corresponding author’s e-mail: Yongyuan.Li.Sae@gmail.com

Abstract. Nowadays, image classification done by Machine Learning can classify images within an instant after an efficient model is built. Such techniques can help identify whether a person correctly puts a mask on. During the Covid situation, it is important to ensure the people in public areas put on a mask correctly so it can cut off the route of mass infection. In this paper, three classes of image classification were tested: with mask, without mask, and mask-wore-d-incorrectly. Based on the Google Colaboratory platform and Inception V3 model, a three-classes-detect-model with 94.52% testing accuracy was built. In addition to building the models, the effectiveness of data preprocessing has also been tested. After applying different methods for data preprocessing, the testing accuracy for the two-classes-detector model improved to 97.11% but the three-classes-detector model decreased to 94.26%.

1. Introduction
Computer versions can be used for image classification; one of the advantages is it can classify images within a short period of time and with high accuracy if the built model is efficient. One of the requirements for building an efficient model is the data need to have a sufficient size and with high quality. Data preprocessing can be used to both increase the data size and data quality.

Data preprocessing is a useful data mining technique that can edit the raw data to an appropriate and efficient format [1]. With the correct technique, a raw and non-processable data set can be processed to a processable and appropriate data set. In this paper, the cropping images technique was used. This technique can help to crop the specific areas of the image so the noise would be cropped out and keep the useful information. In this way, the quality of the data can be increased. Furthermore, image cropping allows cropping out different objects from a single data, so the data size increases.

In this paper, the Inception V3 model was used. Inception V3 Model is a widely used image classification model which successfully combined ideas from multiple researchers and the building process took years [2]. A high-level diagram of the Inception V3 model is shown below as Figure 1.
As shown in Figure 1, the architecture of the Inception V3 Model can be divided into several inception blocks. Each inception block can contain a different combination of Convolution layer, AvgPool layer, MaxPool layer, Concat layer, Dropout layer, Fully Connected layer, and Softmax output. As the graph shows, the Inception V3 Model could have some softmax outputs during the learning process. Those softmax outputs are built for obtaining loss and accuracy during the learning process. This model allows inception work with Factorizing Convolutions which could not only effectively speed up the training process by reducing the number of connections, but also prevent overfitting by reducing the parameters to learn [2]. To introduce the advantages of the Inception V3 model, it will be compared to one of the classical image classification models named AlexNet. Figure 2 shown below is the architecture for AlexNet.

As Figure 2 shows, the architecture of the AlexNet model is a combination of convolution layers, pooling layers, and fully connected layers. Unlike AlexNet, the advantage of the Inception V3 Model is it can allow different Convolution layers, MaxPool layers, AvgPool layers to exist in the same inception block and apply a Concat layer to concatenate the outputs coming from those different layers and feed them to the next inception block. In this way, the Inception V3 Model can contain more layers but with fewer connections. Additionally, the combination of a few convolution layers can
replace a large convolution layer shown in AlexNet. In this way, the Inception V3 Model can reduce the learning parameters which leads to less possibility of overfitting [2].

In this paper, two data sets were used. Data set 1 contains 853 images. Most of the images contain multiple objects inside one image. Besides images, this data set also contains annotations for each image. For this paper, each annotation contains the information of filename, file size, objects’ class, post, truncated, occluded, difficult, and boundary box. For the boundary box, the annotations provide four data points, which are xmin, xmax, ymin, and ymax. Additionally, in order to find out the effect of the face cropping and other possible factors, a control group was needed. Therefore, another data set was used and trained with the same Inception V3 model. This data set contains 700 images belonging to two classes, with half of the images labeled as “with mask” and half of the images labeled as “without mask”. Different from data set 1, this data set contains a single face inside each image and all images are labeled. Therefore, data set 2 can be trained without any preprocessing and its accuracy will be compared to the accuracy from data set 1 and preprocessed data sets.

2. Methodology:
For data set 1, due to the multiple objects inside one image, the image cropping technique was needed to separate objects and classify each object into different classes. To achieve that, a code needs to read the annotation file and get the four data points from the list under the object tree, which are xmin, xmax, ymin, and ymax. After getting those data, the faces were cropped out one by one. The Figure 3 shown below shows four parameters that provided the locations for image cropping.

![Figure 3. Face Cropping Based on Xmin, Xmax, Ymin, and Ymax](image)

Besides face cropping, two other preprocessing procedures were also applied to data set 1. The first one was deleting the “mask-/weared-incorrect” class and the other one was deleting the burry face images cropped by the cropping process.

By deleting the “mask-/weared-incorrect” class, data set 1 had the same classes as data set 2. In this way, the major difference between the two models is whether these data sets had been processed by image cropping.

The annotations from data set 1 marked every face in images without considering the image quality. Therefore, after image cropping, some data from data set 1 became burry. To determine whether the burry images damaged the accuracy for data set one, the face images that were less than 2 kilobytes were deleted. After this preprocessing, the effect of burry images was determined after comparing the new accuracy to the original accuracy.

To make the results more accurate and comparable, the training for different situations were repeated ten times and the average accuracy and loss were used for the below sections.
3. Result and Discussions:
For data set 1, after image cropping, the data size increased from 853 faces-and-environment images to 4072 face-only images. 79.4% of the data classified as “with mask,” 17.6% classified as “without mask,” and 3.0% classified as “mask-wore-incorrectly.”

For data set 1, after image cropping and the 30-epochs Inception V3 model training, the average testing accuracy for this data set was 94.52%. Without image cropping, the average testing accuracy for data set 2 was 96.68%. The accuracy graphs for these two data sets are shown below as Figure 4 and Figure 5.

![Figure 4. Accuracy for Data Set 1](image1.png)

![Figure 5. Accuracy for Data Set 2](image2.png)
As the results have shown, the average testing accuracy from data set 2 is 2.16% higher than the data set 1. However, the image cropping technique should have increased the accuracy of this data set since it cropped out the useful information and left the noise. To obtain the reasons for the low accuracy, the effect of the two main differences between the two data sets were focused on. The first difference is the data set 1 has a class named “mask-wereed-incorrectly” and the data set 2 does not. The other difference is that the data set 1 contains some low-quality face images but the data set 2 contains high-quality images only.

To obtain the relationship between the two main differences and the low accuracy of data set 1, two more training models were created. By deleting the whole “mask-wereed-incorrectly” class and training data set 1 again, the testing accuracy increased from 94.52% to 97.11%. However, by deleting all the face images that are smaller than 2 kilobytes and training the data set again, the testing accuracy decreased from 94.52% to 94.26%. The accuracy graphs for both trainings are shown below as Figure 6 and Figure 7.

![Figure 6. Accuracy for Two Classes Model](image1)

![Figure 7. Accuracy for High Quality Images](image2)

One possible reason the two-class model’s accuracy increased is the images in “mask-wereed-incorrectly” are hard to detect and classify. The data in this class is comparable to both “with mask”
and “without mask” depending on how large the mouth and nose area is exposed. If the mask is only pulled down a bit and the nose exposed, this data will be similar to data in the “with mask” class. If the mask is pulled down a lot more and both the nose and mouth are exposed, this data will be comparable to the data in the “without mask” class. Another possible reason is the data size for “mask-wearred-incorrectly” is not large enough. Therefore, the model did not obtain enough data for this class during the training process, so the accuracy was lower for the three-class model. This result shows for the “with mask” and “without mask” classification, the data set after image cropping obtained a higher accuracy model compared to the two classes model without applying the image cropping technique.

From the model of high-quality images, the accuracy decreased due to the overfitting of the model. The validation accuracy from Figure 7 was declining overall after 15 epochs. The loss graph below also indicates there was an overfit during the training process.

Figure 8. Loss Graph for High Quality Images

Compared to the accuracy graph, the validation loss from Figure 8 increased after 15 epochs, which also indicates overfit happened. The reason for the overfit is that after all the images less than 2 kilobytes were deleted, the data size decreased 34%. Therefore, the effect of the blurry images was not obtained since the decrease of accuracy can be caused by the overfitting. To prevent overfitting, after deleting the low-quality images, the exact amounts of high-quality images should feed in and fill the vacancy, so the low-quality images would be replaced by high-quality images and no data size will change.

4. Conclusion
In this paper, the Inception V3 Model showed a high performance by outputting the high accuracy models with up to 97.11% testing accuracy. The image cropping technique was proven to improve the accuracy of the two-class mask detection model. Due to lack of data and overfit, the relationship between accuracy and low-quality images was not found. In the end, two models were created, a three-classes detection model with 94.52% testing accuracy and a two-classes detection model with 97.11% testing accuracy.

5. Future works
Due to the lack of data, an overfitting occurred which led to the unclear effect of deleting the low-quality images. To figure out the effect of this, the data set needs to remain the same size while deleting the data smaller than two kilobytes. Therefore, one future experiment is filling the 34%
vacancy with high-quality face-only data for all classes, training the model again, and comparing the new accuracy to the original one.

For this paper, the image cropping was based on the parameters provided by the annotations. One of the disadvantages of that is some blurry images were created and those images might lead to lower accuracy. Therefore, another future project is applying a face detector to crop the faces and compare the two models’ accuracy.

The last future study for this paper is applying the models obtained from this project to OpenCV and creating a real-time mask detection program. The program should detect and crop the face from a camera and feed the image to the model. If the image is classified as “without mask” or “mask-wereed-incorrectly,” a warning message and a warning sound will be produced.

References
[1] Hutajulu, Oscar Daniel. (2019) “Basics of Data Preprocessing” https://medium.com/easyread/basics-of-data-preprocessing-71c314bc7188
[2] Patel, Khush. (2020) “Architecture comparison of AlexNet, VGGNet, ResNet, Inception, DenseNet” https://towardsdatascience.com/architecture-comparison-of-alexnet-vggnet-resnet-inception-densenet-beb8b116866d
[3] Denny Britz (2020) “Deep Learning’s Most Important Ideas” https://www.kdnuggets.com/2020/09/deep-learnings-most-important-ideas.html
[4] Chowdary G J , Punn N S , Sonbhadra S K , et al. Face Mask Detection using Transfer Learning of InceptionV3[J]. 2020.
[5] Alajlan A A , Saudagar A . Threat Detection in Social Media Images Using the Inception-v3 Model[M]. 2020.
[6] Tornow M A . The user behavior in social media networks using the example of Instagram.