Determining correct face mask usage with Inception Resnet and MaskedFace-Net dataset

Pratyush Juneja[a] and **Eric Kang**[b] and **Elizabeth Kim**[c]
[a] pjuneja@ucsd.edu
[b] ekang@ucsd.ed
[c] mek017@ucsd.edu

Abstract

Many models and algorithms in artificial intelligence are considered "black box models", or models that do not provide transparency into how it reached a certain conclusion; ultimately, although they may be accurate, they are uninterpretable by humans and lack trust and transparency. We aim to provide this transparency through explainable artificial intelligence. First and foremost, we aim to present a model that can determine if individuals are properly wearing a mask, improperly wearing a mask, or are not wearing a mask at all in the light of the Covid-19 pandemic. Especially as this is a high stakes situation, with businesses and individuals at risk, transparency is key. Secondly, we aim to provide this transparency through GradCAM, which will highlight how our model came to its decision. The method uses an untrained Inception Resnet V1 in order to determine the mask usage in a given image. Gradient Descent is used for training and Cross Entropy as a loss function. Finally, GradCAM is applied to the images and outputs a coarse heatmap from the last layers of our convolutional neural network that shows exactly what our model is looking at. Currently, we are able to reach a model with 96% accuracy.

1 Introduction

An MIT article, published in late 2019 by Rudin and Radin, explains the unnecessary supply of black box models, or models which can be viewed in terms of its "inputs and outputs, without any knowledge of its internal workings".[1] Rudin and Radin explain how the advances in deep learning for computer vision have led to the trend of "inherently uninterpretable and complicated"[2] models, in which there is a lack of transparency to humans in being able to understand how a certain model came to a certain conclusion. Stemming from the historical use of machine learning in society, such as "online advertising and web search,"[2] black box machine learning models initially proved effective; however, these initial models did not necessarily deeply impact human lives. With the growing use and power of machine learning that now has the power to do so, it is clear that explainable artificial intelligence is a necessity, that "the belief that accuracy must be sacrificed for interpretability is inaccurate"[2], and that we should aim to provide transparency and understanding to any model, especially for high stakes decisions. Our aim was to apply transparency to face mask detection, in light of the Coronavirus pandemic. As of early February, Coronavirus (known as Covid-19) has infected over 100 million individuals throughout the world, and has killed over 2.3 million, nearly 500,000 of those being Americans[3]. According to the CDC, Covid-19 spreads most commonly through close contact, primarily when individuals with COVID-19 "cough, sneeze, sing, talk, or breathe"[4] as the virus is transmitted through respiratory droplets. Moreover, these viruses may infect others not in close contact through airborne transmission. According to John Brooks, a medical epidemiologist at the CDC in Atlanta, "masks bring

[1] Kenton, Will. "Black Box Model." Investopedia, 25 Aug. 2020

[2] Rudin, Cynthia, and Joanna Radin. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition · Issue 1.2, Fall 2019." Harvard Data Science Review, PubPub, 22 Nov. 2019

[3] "Coronavirus Cases:" Worldometer, 6 Feb. 2021,

[4] "COVID-19: Considerations for Wearing Masks." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 18 Dec. 2020

down the community viral load"[5] and are used to protect others, rather than oneself. Masks are imperative to helping prevent the spread of coronavirus; however, they must also be worn properly. However, implementing this on a grand scale is difficult. Establishments, such as schools and businesses, would need a scalable system to ensure correct mask usage in order to keep their brand image, as they could be shunned or shut down for not following mask protocols, and more importantly, keep their employees and customers safe. We aim to provide this solution through a face mask image recognition algorithm using a subsection of the MaskedFace-Net dataset[6], which contains images of properly and improperly worn masks. Moreover, we aim to provide transparency through GradCAM, in order to ensure that our model is coming to the right conclusions for the right reasons, especially in such an environment concerning public health. Ultimately, our hope is to create an algorithm that ensures transparency and trust, that has the potential to help human lives.

2 Literature Review

There have been several other advances when it comes to face mask detection in light of the coronavirus. Nagrath, Jain, Madan, Arora, Kataria, and Hemanth trained a real time DNN-based face mask detection system, and utilized deep learning, TensorFlow, Keras, and OpenCV[7]. Utilizing a dataset containing "with mask" and "without mask" labels of 5521 images each, they were able to utilize MobileNetV2, a Deep Neural Network deployed for classification, and utilized the pretrained weights of ImageNet. Moreover, although the dataset was smaller than our FaceMasked-Net, it held a combination of real and artificial images. Their approach was the SSDMNV2 approach, which utilized Single Shot Multibox as a face detector and MobilenetV2 as a framework. After 100 epochs, they received a training accuracy of 92.64% and an F1 score of 93%. As reference, AlexNet and LeNet-5 were also two architectures used, and received a lower accuracy of 89.2% and 84.6%, respectively. The work of SSDMNV2: A real time DNN-based face mask detection system, had incredibly high accuracy and speed, as it is a real time detector. Similarly, A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3 had a high accuracy rate of 96%, trained over 4000 epochs and had real time detection, with an average output of 17 fps [8]. Their dataset consisted of a web-scraping tool to collect 650 images of masked and unmasked, each. This dataset consisted of natural face mask images and was not artificial, unlike ours. It also contained several various types of face masks of different materials. There are a myriad of other algorithms done on face mask detection, with various CNN architectures; however, it should be noted that many of the current algorithms do not necessarily go in depth to incorrectly worn masks, which is a large problem in the context of the pandemic. This is a gap we aim to solve with our FaceMasked-Net dataset.

3 Dataset

We utilized a subsection of the MaskedFace-Net dataset[9]. The dataset itself is a cumulation of 133,783 images based on the dataset Flickr-Faces-HQ, and consists of a blue medical face mask photoshopped over various faces, either correctly or incorrectly. Of the incorrectly worn masks, there are three categories: Uncovered chin, uncovered noise, and uncovered nose and mouth, as wearing a mask properly ensures its effectiveness. The dataset for our model was taken off Sheldon Sebastian's MaskedFace-Net Kaggle[10] dataset, which contains 58,582 images total. This dataset contained images as follows: for covered faces, 12362 images in the train folder, 3863 images in the holdout folder, and 3090 images in the validation folder; for incorrect faces, 12338 images in the train folder, 3856 images in the holdout folder, 3084 images in the validation folder; for completely uncovered faces, 12800 images in the train folder, 3090 images in the holdout folder, and 3199 images in the validation folder. Of the incorrectly worn mask labels, 10184

[5]Goodman, Brenda. "How Much Does Wearing a Mask Protect You?" WebMD, WebMD, 19 Nov. 2020

[6]Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19", Smart Health, ISSN 2352-6483, Elsevier, 2020,https://github.com/cabani/MaskedFace-Net

[7]Preeti Nagrath, Rachna Jain, Agam Madan, Rohan Arora, Piyush Kataria, Jude Hemanth, SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2, Sustainable Cities and Society, Volume 66,2021,102692,ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.102692.

[8]M. R. Bhuiyan, S. A. Khushbu and M. S. Islam, "A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225384.

[9]Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19", Smart Health, ISSN 2352-6483, Elsevier, 2020, DOI:10.1016/j.smhl.2020.100144 https://github.com/cabani/MaskedFace-Net

[10]Sebastian, Sheldon MaskedFace-Net , MaskedFace-Net dataset along with Flicker Faces dataset https://www.kaggle.com/sheldonsebastian/maskednet-flicker-faces

of the images for Mouth/Chin were in the train folder, 3192 in the holdout folder, and 2580 in the validation folder. 1154 of the images for Nose/Mouth were in the train folder, 352 in the holdout folder, and 274 in the validation folder. 1000 of the images for Chin were in the train folder, 230 images in the holdout folder, and 230 in the validation folder. The majority of the incorrect labels were a correctly covered Mouth and Chin but uncovered nose, making the dataset of incorrectly worn masks slightly unbalanced. The dataset contains images of faces, including a variety of ethnicity, genders, and ages, which ensures that our model will be applicable to a variety of individuals.

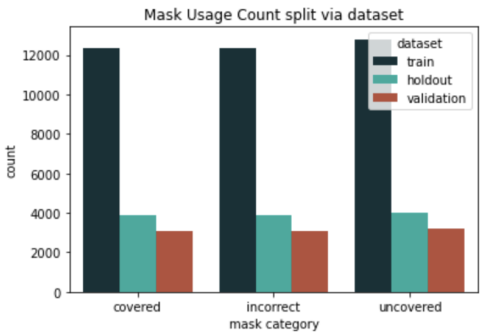

Figure 1: Number of correctly worn masks, incorrectly worn masks, and uncovered faces

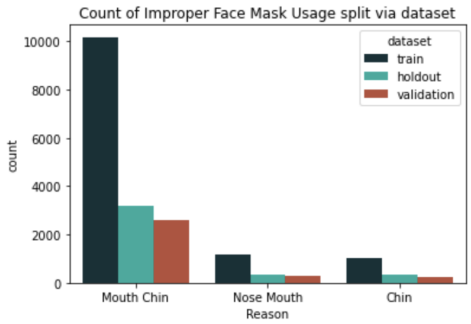

Figure 2: Count of images per reason for incorrectly worn mask

Moreover, as the coronavirus pandemic began in late 2019 to early 2020, there were few public datasets with extensive mask-wearing; although the masks in this dataset are photoshopped on rather than naturally worn, the dataset was one of the few that had a large number of images with the granularity of why a mask was worn improperly and we found as appropriate for our problem. However, it should be noted that there are some disadvantages to this dataset, which are outlined in our discussion section below.

Figure 3: A random image taken from the MaskedFace-Net dataset of an individual with a properly worn mask. The mask is photoshopped.

4 Methodology

Inception ResNet V1 is a variant of the original FaceNet Model and was the architecture we utilized in pytorch; although it is pre-trained on both the VGGFACE2 and Casia-Webface datasets, we utilized an untrained model with 3 classes. The Inception ResNet model is unique in that typical CNNs have stacked convolutional layers followed by one or more fully connected layers; Inception on the other hand, all filters in the Inception model are learned and repeated many times.The main idea of Inception is "how an optimal local sparse structure in a convolutional vision network can be approximated and covered by readily available dense components[11] Inception Resnet finds the optimal local construction and repeats it spatially; this was done through a layer by layer construction in which correlation statistics of the previous layer are clustered in to groups, which form the units of the next layer. In the lower layers, various clusters would be concentrated in a single location and covered by a layer of 1x1 convolutions; however, as features of higher abstraction are captured by higher layers, this concentration should decrease, and therefore increase in 3x3 and 5x5 convolutions, as Inception Resnet is limited to 1x1, 3x3, and 5x5 filters. However, as these are computationally expensive, Inception Resnet applies dimensionality reduction wherever it is too computationally expensive. In terms of our model classification, 3 classes were utilized - incorrectly worn face masks, correctly worn face masks, and no face masks

[11] https://arxiv.org/pdf/1409.4842v1.pdf

detected/uncovered face. In terms of labelling, this dataset did not contain any bounding boxes, but rather, had images in folders with the folder name having the respective label. For image preprocessing, the images were resized to 256 x 256 and transformed into tensors; as Inception ResNet V1 works best with images that are cropped to the face, our dataset proved helpful as it was precropped and no other preprocessing steps took place. Cross Entropy was the loss function utilized, and the Adam algorithm was utilized as the optimizer object, which holds the current state and updates parameters based on computed gradients with a learning rate of 0.009 as a hyperparameter; 2 epochs were used to train the model.

GradCAM, or Gradient weighted Class Activation Mapping, was utilized in order to provide transparency. GradCam uses gradient information from a target concept that is flowing into the last convolutional layer of the CNN. Ultimately, it results in a coarse heat map that essentially show which parts of the image helped the model lead to it's final decision - essentially, the most important parts of the image. GradCAM was applied to our model in order to ensure that it was looking in the correct areas.

5 Results and Discussion

The result of our model was an accuracy of 96% in being able to classify between the three classes: improper face mask usage, no mask, and proper face mask usage. In terms of Grad-CAM, the implementation was successful in building trust and transparency within our model: the model was looking at the correct areas to determine the face mask usage.

Figure 4: A female with proper face mask usage

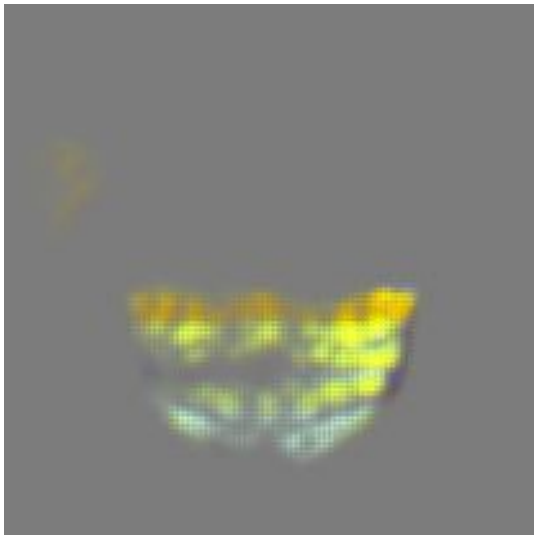

Figure 5: The coarse localization map of GradCAM

Figure 6: A GradCAM heatmap overlayed onto the original image, which pinpoints the important parts of the image that led to the result

Figures 4 to 6 show the GradCAM heatmap on an individual who wore a face mask properly.

Figures 7 to 9 show an individual with improper usage as it only covers his chin; this is seen in our GradCAM output as well - GradCAM shows that our model was looking heavily at the chin area. Overall, our model was fairly successful, and was able to reach a level of trust. However, there were several weaknesses in our model. To begin with: the dataset. The dataset only contains blue medical face masks - it does not contain cloth masks, KN95, or N-95 masks, which are also commonly used; therefore, we are not sure how our model may perform given another face mask. More-

Figure 7: A male with improper face mask usage; his mask only covers his chin

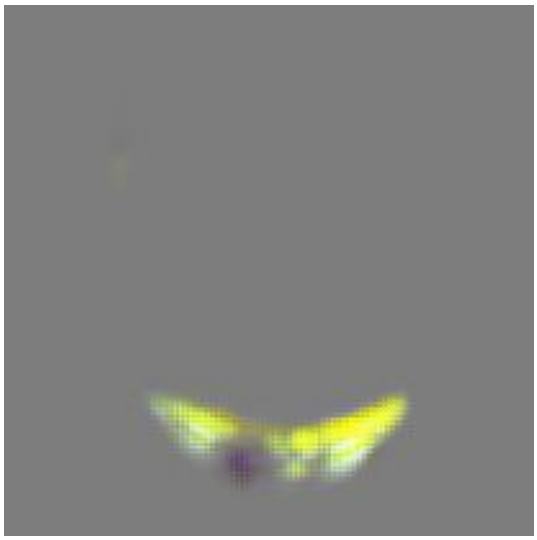

Figure 8: The coarse localization map of GradCAM

over, some face masks are less effective than others, such as a cloth mask vs an N-95 mask, which is less available to the public. This weakness can be attributed to the lack of datasets available on face masks - there are several datasets with face masks in real images; however, many did not include improper face mask usage (only mask or no mask) as well as the reason behind why it may have been deemed as "improper." Furthermore, the dataset consisted of photoshopped face masks on cropped faces: ultimately, this is an artificially made dataset. Another weakness would be the accuracy of 96%; although it is fairly high, considering the high stakes situation, having 4% inaccurate classifications could have detrimental effects (such as

Figure 9: A GradCAM heatmap overlayed onto the original image, which pinpoints the important parts of the image that led to the result

if someone who was Covid-19 positive did not wear a mask properly, and it was deemed proper usage by our model). Finally, due to computational and time constraints, we were not able to implement our model in real time, which would be necessary if one were to use our model to scale it for a business or school. However, regardless of various weaknesses, we were able to complete our goal of having a model that users could trust. The outputs of our project included a front-end static website for general audiences, a recorded presentation, and a GitHub repo. Links can be found in the Appendix section.

1) Kenton, Will. “Black Box Model.” Investopedia, 25 Aug. 2020

2) Rudin, Cynthia, and Joanna Radin. “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition · Issue 1.2, Fall 2019.” Harvard Data Science Review, PubPub, 22 Nov. 2019

3) “Coronavirus Cases:” Worldometer, 6 Feb. 2021,

4) Goodman, Brenda. “How Much Does Wearing a Mask Protect You?” WebMD, WebMD, 19 Nov. 2020

5)Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19", Smart Health, ISSN 2352-6483, Elsevier, 2020,https://github.com/cabani/MaskedFace-Net

6) Preeti Nagrath, Rachna Jain, Agam Madan, Rohan Arora, Piyush Kataria, Jude Hemanth, SS-

DMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2, Sustainable Cities and Society, Volume 66,2021,102692,ISSN 2210-6707, https://doi.org/10.1016/j.scs.2020.102692.

7) M. R. Bhuiyan, S. A. Khushbu and M. S. Islam, "A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225384.

8) Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19", Smart Health, ISSN 2352-6483, Elsevier, 2020, DOI:10.1016/j.smhl.2020.100144 https://github.com/cabani/MaskedFace-Net

9)Sebastian, Sheldon MaskedFace-Net , MaskedFace-Net dataset along with Flicker Faces dataset https://www.kaggle.com/sheldonsebastian/maskednet-flicker-faces

References

6 Appendix

Github for Front End: https://elizabethmkim.github.io/FaceMaskDetection/

Github for Model: https://github.com/eric-h-kang/FaceMaskDetection

Project Proposal Statement: A lot of neural networks are a “black box” which does not provide understanding or insight into how a model or algorithm reached a conclusion; this means a certain model could reach a conclusion, although a correct conclusion, for the wrong reasons and we would not be aware. Transparency is needed, first to recognize and pinpoint any failures a model can have, but also second, if a model works, this transparency builds trust to those utilizing an algorithm. We propose to bring this transparency into a mask recognition model, which would be able to identify whether or not an individual is properly wearing a mask. We aim to train a model which can determine whether or not an individual is wearing a mask properly or improperly and which concerns public safety - with high stakes such as these, trust and transparency into our model is a must. As Covid-19 has killed over 290,000 Americans as of December, proper mask usage is increasingly important, as it helps prevent the spread of Coronavirus. People are either outright refusing to wear masks, or sometimes cheating the system by wearing masks incorrectly or unknowingly. By this, we refer to those that wear masks that do not tightly cover the mouth, nose, and chin. As a result, the efficacy of the mask is nearly zero, posing both a large public health risk as well as also potentially being detrimental to brand imaging. Establishments risk being shut down if they do not abide by CDC guidelines, and furthermore people simply won’t want to go to these establishments if they are a large safety risk. Clearly, businesses are motivated to ensure the people in their establishments are wearing masks properly, and current solutions are inefficient when scaled. Our solution scales well and is much more cost effective than hiring manual workers to make sure everyone is wearing a mask properly. Real-time convolutional neural networks are getting significantly better with the advent of YOLOv4. Even if people only appeared in the frame of the camera for a short amount of time, we would be able to build a network to detect them. Building Grad CAM into YOLOv4 is relatively simple as well, since with the open source Darknet version, all we would need to do is implement Grad CAM into the convolutional and pooling layers to build our explanation. Using Grad CAM simplifies a lot of the possible issues that may occur with this model. By taking in images that are trained on our Neural Network, it will return a density of areas that displays the “important areas” that a facemask model should look at to identify whether or not the mask is being worn currently. For example this model should at least look at our nose, chin and general face features to see if the mask is covering a certain targeted area which can help the model then make its predictions more understandable to us. Feasibility otherwise has been historically demonstrated through the widespread popularity of convolutional neural networks, as well as the current speed of real-time object detection. We propose to create a website which showcases not only our model, but highlights the necessary transparency that will ensure further trust in our work. Ultimately, we would hope to allow individuals to submit photographs of themselves wearing a mask, and have our model be able to recognize whether or not they are wearing it properly. However, other aspects of the website would include going further in depth to our model, any discussions and obstacles in our model, and ultimately highlight any successes our model achieves. In short, to be considered a successful model, the outputs should correspond with the correct wearing of masks and the usage of Grad CAM should serve as some sort of evidence that the model is looking at the right areas. Currently, there are several models which look at whether faces are correctly or incorrectly wearing a mask. Although there exists

large datasets with mask face images; there are few that include whether the mask is actually worn properly except for a few. One is MaskedFace-Net, which includes 67,193 images with masks that are worn correctly, and 66,900 images of masks work incorrectly. Moreover, this dataset also breaks down the incorrectly worn masks into why they were deemed "improper," such as uncovered nose, mouth, chin, etc and contains the information needed. As we aim to provide transparency into our model, these specificities will allow us to ensure our model comes to a conclusion for the correct reasons. Another dataset which includes improperly worn masks is the Face Mask Detection with 3 classes dataset on Kaggle, which has 853 images with bounding boxes in PASCAL VOC format. However, this dataset is far smaller than the MaskedFace-Net dataset. There are multiple data sources out there which contain the presence of a mask, such as PyImage Search Reader's covid-19 face mask detector and NVIDIA's face mask detection; however, both do not go in depth into whether a mask is worn properly or improperly, which is crucial to the efficacy of a mask during the covid-19 pandemic. We will most likely utilize the MaskedFace-Net dataset or the Kaggle dataset.