CIS 522 Spring 2021, Project Report

Mask Yourself Correctly: Identifying Improper Mask Usage Using Limited Data

Team Members

- **M**ohit Bagri; Email: `mobagri@seas.upenn.edu`
- **Y**oonduk Kim; Email: `yoonduk@wharton.upenn.edu`
- **C**atrina Hacker; Email: `cmhacker@pennmedicine.upenn.edu`

Abstract

The COVID-19 pandemic has demonstrated the importance of speed and efficiency in developing algorithmic solutions for novel problems. We compare four different approaches for building classifiers for proper mask usage using a limited training dataset of 240 images. We find that though SVM and CNN models work adequately, transfer learning with pretrained AlexNet eclipses other models in performance metrics. Additionally, we introduce a Symbolic Reasoning Model using predefined facial feature detectors with the potential to aid in building data-efficient models. While the performance of the Symbolic Reasoning Model is only on par with SVM and vanilla CNN, the method shows potential to develop into an explainable and flexible alternative for deep learning models.

1 Introduction

1.1 Motivation

As the COVID-19 pandemic continues to affect populations around the world, many public spaces require that masks be worn to prevent disease transmission. While wearing masks can be an effective measure for reducing airborne transmission of infectious diseases (Howard et al., 2021), incorrect use of masks diminishes their efficacy (World Health Organization, 2020). An automated system to detect incorrect mask usage can benefit public health by helping identify individuals at risk of spreading or becoming infected with the virus.

The convolutional neural net (CNN) architecture and its variations are extremely effective at supervised image classification tasks. However, obtaining a good dataset is still a challenging obstacle. First, the price of generating labeled data presents entry barriers for researchers and practitioners that are constrained on resources. Training a new deep learning network from scratch often requires thousands of training examples. Though methods such as crowdsourcing can mitigate the cost, concerns have been raised over crowd workers being underpaid (Fort et al., 2011) and the labels amplifying existing biases (Smith, 2019) or not being representative of the underlying concept (Crawford and Paglen, 2019). Second, use of facial images without the consent of the respective individuals may invade their privacy (Garvie, 2019). This is especially problematic if the data label can put the individuals in a negative light, in this case improperly wearing masks in public.

The task of identifying proper mask usage serves to emphasize the importance of data efficiency. Efficiency is often overlooked in favor of performance in the literature, as the research landscape of machine learning incentivizes competition on established benchmarks (Al-Jarrah et al., 2015). The researchers that develop state of the art models are oftentimes supported by ample funding from grants and industry backing. As deep learning opens up to researchers, engineers, and students over the globe, we face an increasing need to explore accessible methods for developing functioning models from limited data. Data efficiency can lower the entry barriers for model development and increase flexibility in responding to unanticipated emergencies such as the COVID-19 pandemic.

In this project, we aim to build data-efficient image classifiers that can identify whether a person is wearing face masks properly. Given a facial image, the model returns a ternary classification output of

whether the person is 1) wearing the mask properly, 2) wearing the mask improperly, or 3) not wearing a mask at all. The model is trained on a small dataset of 240 labeled facial images. We review and employ a variety of classification methods ranging from modern deep learning techniques such as transfer learning to the classic Symbolic Reasoning Model using predefined Haar-like facial features.

1.2 Literature Review

At the time of writing, the authors did not find a suitable benchmark or a generalizable model for ternary classification of face mask usage in real world facial images. State-of-the-art models for face mask recognition are primarily trained for binary classification and do not take into account whether the mask is worn properly (Nagrath et al., 2021; Mercaldo and Santone, 2021).

The absence of ternary classification models can be attributed to the scarcity of well-formatted labeled data on real world images of improperly masked faces. A dataset published on Kaggle by an independent researcher (Larxel, 2020) contains 853 images of one or more persons with varying degrees of mask usage, but only 123 out of 4072 (3%) faces in the dataset are labeled as worn incorrectly. MaskedFace-Net (Cabani et al., 2021) addresses the size issue by providing 60,000 simulated variations of the Flicker-Faces-HQ Dataset (Karras et al., 2019) in which virtual masks are superimposed on faces. However, the data was not applicable for classifying facial images in the real world. Training on the MaskedFace-Net data led models to overfit for detecting idiosyncrasies of the simulated mask overlays and did not generalize well to masks with varying shapes, textures, and colors. A pretest on the dataset using a standard AlexNet architecture yielded a classifier that had 98% accuracy on the validation set, but performed poorly on real world images.

1.3 Project Outline

In the following sections we will first describe our data, compare four model variations, and discuss the results. We measure the performance of each model on 2 metrics (Accuracy and F1 score). We first show that the CNN model outperforms SVM. The two advanced models were based on vastly different methods and yielded differing outcomes. Transfer learning from pretrained AlexNet was the best performing model, while the Symbolic Reasoning Model did not meet our expectations. The challenge nonetheless provided insights and directions for improvements. In our last section, we discuss the implications of our results and map out plans for the future.

2 Methods

2.1 Data

The goal of this project is to do a three-way classification of images depicting subjects wearing masks properly, improperly, or not at all. This is a supervised learning problem, so images are labeled with one of the three target classifications. As stated earlier, one of the key challenges in this project was finding real world images of people wearing masks improperly. We could not find any dataset which by itself had a balanced dataset of the three classes and hence we had to get creative and generate our own balanced dataset which we made by combining data from three different sources.

We built a dataset of 300 facial images that were evenly distributed across each label (Figure 1). We pulled images from several different sources available online. First, we randomly selected 100 images from the FFHQ face dataset (Karras et al., 2019) to serve as images without a mask. For our correctly worn mask images we randomly selected 100 images from the Kaggle dataset "Face Mask Detection" (Larxel, 2020). While the Kaggle dataset had a few incorrectly worn images, to get to 100 images in this condition we manually collected images from Google image search. This left us with 100 images in each class that contained a single face approximately centered in the image.

We note that sourcing each class of images from varying sources may generate biases due to the peculiarities of each datasource. We set up the following inductive biases and processed the images to reduce correlation between the incidental features and our main classification task. Our first inductive bias is that the characteristics of the person (race, gender, and age), the mask (color, shape, texture), and the environment (lighting, background) in the image have no bearing on whether or not the mask is being worn properly.

We manually curated the images to make sure that each class was represented by a diverse population of individuals, masks, and environments. The images were processed such that only one person was visible and the face was the main focus. Each image was cropped and/or resized to 256x256 and formatted as a 2d-array with normalized values for the 3 channels (RGB). The dataset was augmented via horizontal flip, shift, scale, and rotations. The train and test data were split 8:2, resulting in the training set of 240 images and the test set of 60. The batch size for the data loader was set to 4 to account for the small size of the dataset.

Figure 1: Left: Three example images from each of the image classes used for this project.

2.2 Non-Deep Learning Baseline: SVM

For our non deep learning baseline model we decided to use a Support Vector Machine with RBF kernel. Non Neural based models are seldom used for image classification since we need to come up with our own features which can be hard for a task such as image classification. Nevertheless, we attempt to tackle this problem by generating the following features. After loading each image, we first flatten and compress the image. We flatten the image to satisfy the input requirements of data being passed into the SVM model. Image compression is a key part so that we are within the RAM constraints of Google Colab (used to train the SVM model). We perform the image compression using Cubic Interpolation and reduce our original image (1334 X 1334) into a much smaller image (227 X 227). Following this we handle any missing values in the flattened array by replacing those pixelâs values with the median of that column. We handle missing values since we perform dimensionality reduction using PCA and any missing value can bias the PCA algorithm. We perform PCA primarily since SVM does not work on the entire fattened array which has over 155,000 features. Since SVM does not scale with such a large number of features, we reduce the dimensionality to have just 150 features. PCA works very well since it captures the most variance of all the features which is indicative of that featureâs importance. We finally feed our reduced dataset into the SVM model with any RBF kernel and report performance metrics such as F1 score and accuracy.

2.3 Deep Learning Baseline: Vanilla CNN

For our deep learning baseline model we started with the base architecture of AlexNet and trained the network from scratch. This includes five convolutional layers with kernel sizes of 11, 5, 3, 3, 3 and strides of 4, 2, 2, 1, 1 each followed by a ReLU nonlinearity. The first two layers have a padding of 2 and the last three have a padding of 1. The first, second and last convolutional layers are followed by a max pooling layer, each of which has a kernel size of 3 and stride of 2 and which are also followed by a ReLU nonlinearity. After these convolutional blocks there is an average pooling layer and then three fully-connected layers followed by ReLU nonlinearities (except the last layer). To keep things as vanilla as possible for our baseline model we did not include any regularization, meaning this implementation differs from the AlexNet architecture in that there is no dropout. We trained for 10 epochs and reported test accuracy at the end of this training.

2.4 Advanced Deep Learning Models

2.4.1 AlexNet Transfer Learning

For our first advanced deep learning method we started with a version of AlexNet that was pretrained on the ImageNet dataset (Krizhevsky et al., 2012). This network has the architecture described in section 2.2 with the addition of dropout layers with a probability of 0.5 before each of the first two fully connected layers. In addition to the use of transfer learning we also used data augmentation. Training images were horizontally flipped and shifted, scaled and rotated relative to the test image to better assess the ability of the network to generalize. We trained the pretrained architecture for an additional 10 epochs on the ternary mask image classification task and reported test accuracy at the end of this training period.

2.4.2 Symbolic Reasoning

Our second advanced model revisits the classic paradigm of Symbolic AI (Haugeland, 1989). In Symbolic Reasoning, the model learns the relationship between discrete representations of objects (Ungar et al., 2021). We propose that the task of identifying whether a mask is correctly worn can be represented and reasoned in a symbolic manner. Guidelines on correct use of masks is simple and straightforward: the mask should cover the person's nose, mouth, and chin (Center for Disease Control and Prevention, 2021). Therefore, assuming that the image contains a face and that the relevant facial features can be successfully identified, distinguishing correct use of masks can be represented by the following logical decision process: Provided that a face mask is seen in the image, if both the nose and the mouth are unseen, the mask is worn correctly.

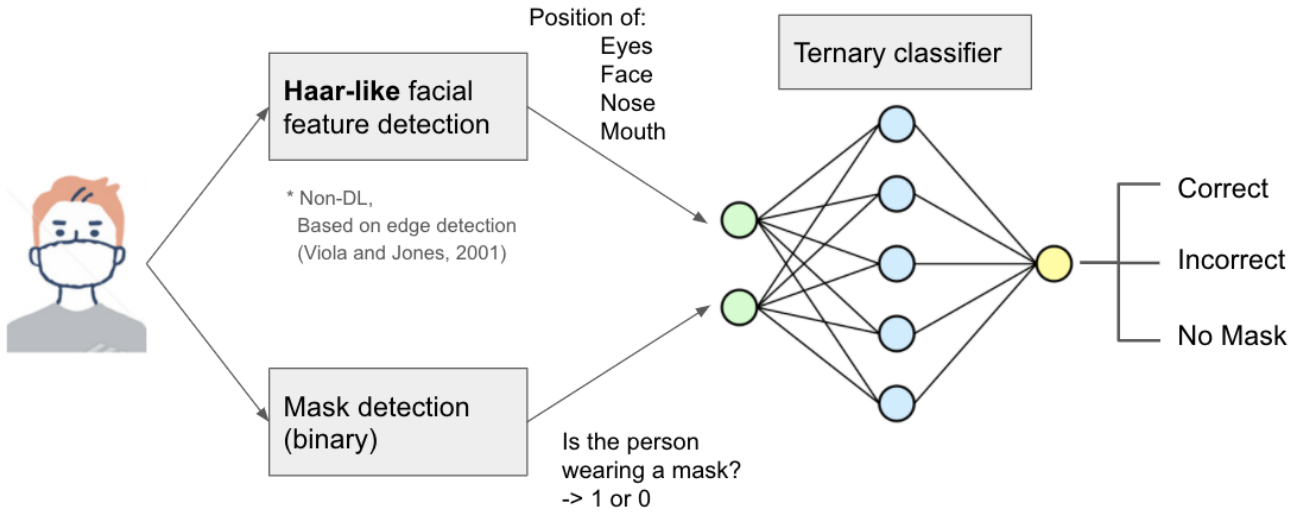

Figure 2: Left: Architecture of the advanced Symbolic Reasoning Model. Details of each component are given in text.

We designed a miniature ensemble model that takes in four main features as input and outputs ternary classification for mask usage. The first three inputs were related to the presence of a face, a nose, and a mouth in the image. For facial feature detection, we applied the Viola-Jones object detection framework, in which the edges in the images were scanned based on predefined Haar-like features (Jones Viola, 2003). The presence of the face was marked by a binary indicator variable, while the nose and the mouth were each represented by four coordinates that denote their respective positions in the image. We included an additional input from a state-of-the-art neural network model for binary mask classification (Nagrath et al., 2021) to detect the presence of a mask, in order to distinguish cases where the person is not wearing a mask but has their facial features obscured by other objects (hands, walls, etc.). These features are passed into a simple MLP with two hidden layers of size 36 and 12.

3 Results

Model	Test Accuracy	F1 Score
SVM	80.83%	0.77
Vanilla CNN	81.67%	0.82
AlexNet Transfer Learning	96.67%	0.97
Symbolic Reasoning	78.33%	0.78

Table 1: Test accuracy and F1 scores for the four models described in this report.

3.1 Non-Deep Learning Baseline: SVM

The non deep learning model achieved 80.83% total accuracy with an F1 score of 0.77 (Figure 3). The model was unsurprisingly making most of the errors in classifying incorrectly worn masks where it achieved an accuracy of only 43.75%. The majority of misclassifications were in labeling incorrectly worn mask images as correctly worn masks.

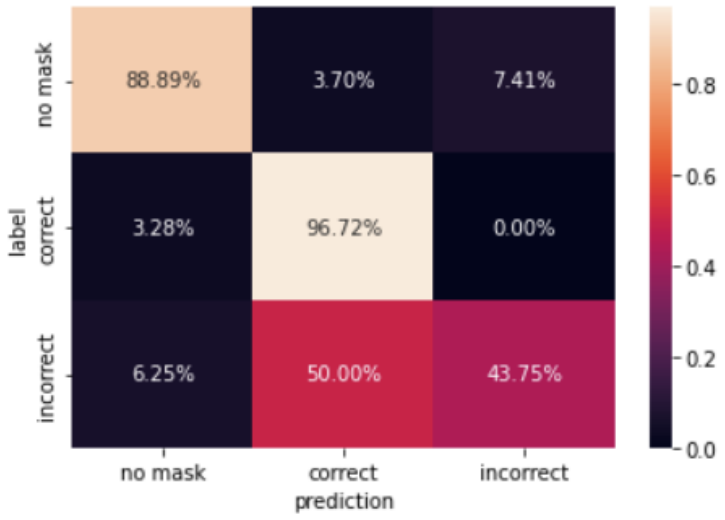

Figure 3: Confusion matrix for test of SVM.

3.2 Deep Learning Baseline: Vanilla CNN

The baseline Vanilla CNN achieved 81.67% total accuracy with an F1 score of 0.82 (Figure 4). The majority of errors made by the model were in the incorrectly masked category, with the model erroneously

labeling incorrectly masked images as no mask images 41% of the time.

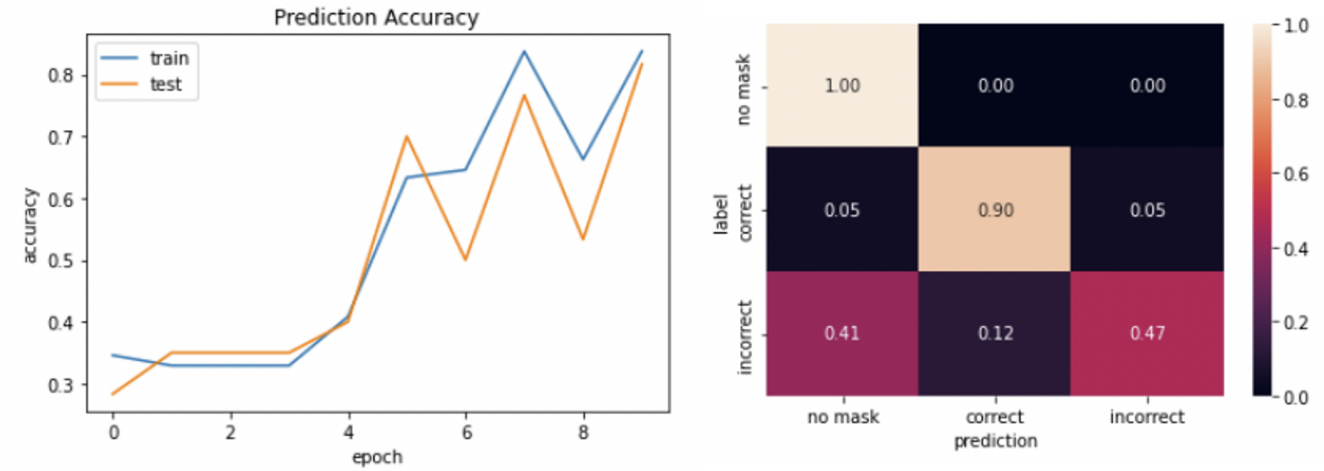

Figure 4: Left: Train and test accuracy for 10 epochs of training of the baseline deep learning model. Right: Confusion matrix for test of baseline deep learning model.

3.3 Advanced Deep Learning Models

3.3.1 AlexNet Transfer Learning

The AlexNet Transfer Learning model performed better than the baseline Vanilla CNN model with a total test accuracy of 96.67% and an F1 score of 0.97 (Figure 5). This network had perfect classification of no mask and correctly masked images, and 88% accuracy in classifying incorrectly masked images. Importantly, all of the errors made by this model were in erroneously classifying incorrectly masked images as depicting faces with no mask.

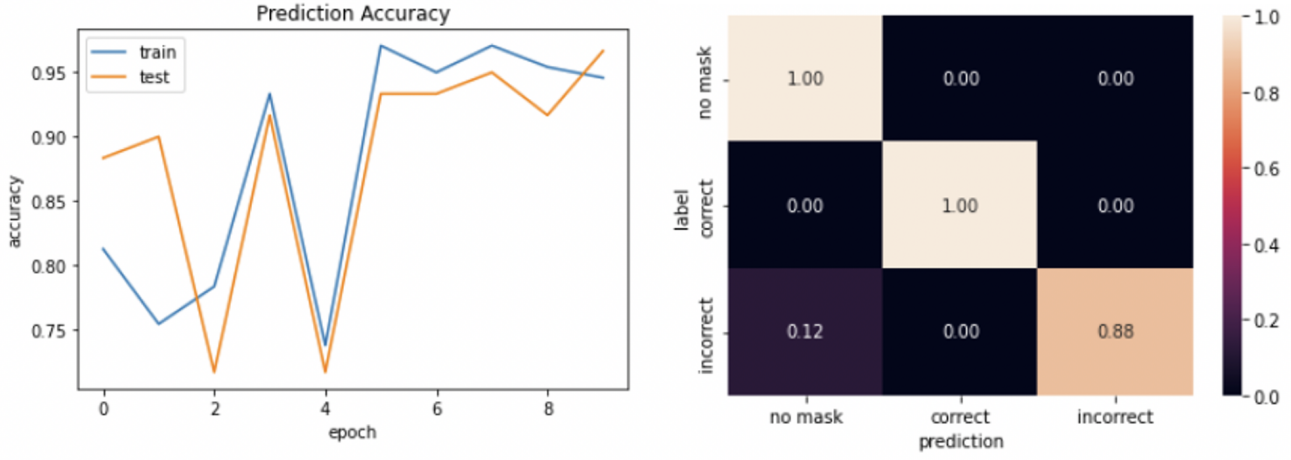

Figure 5: Left: Train and test accuracy for 10 epochs of training of the AlexNet Transfer Learning model. Right: Confusion matrix for test of AlexNet Transfer learning model.

3.3.2 Symbolic Reasoning Model

The Symbolic Reasoning Model did not beat the baseline Vanilla CNN model. It achieved a test accuracy of 78.33% with an F1 score of 0.78 (Figure 6). The network maintained near perfect classification for distinguishing between mask and no mask images, but struggled with the incorrectly masked images. Unlike the Vanilla CNN and Transfer Learning models, the majority of misclassifications were in labeling incorrectly worn masks as correctly worn.

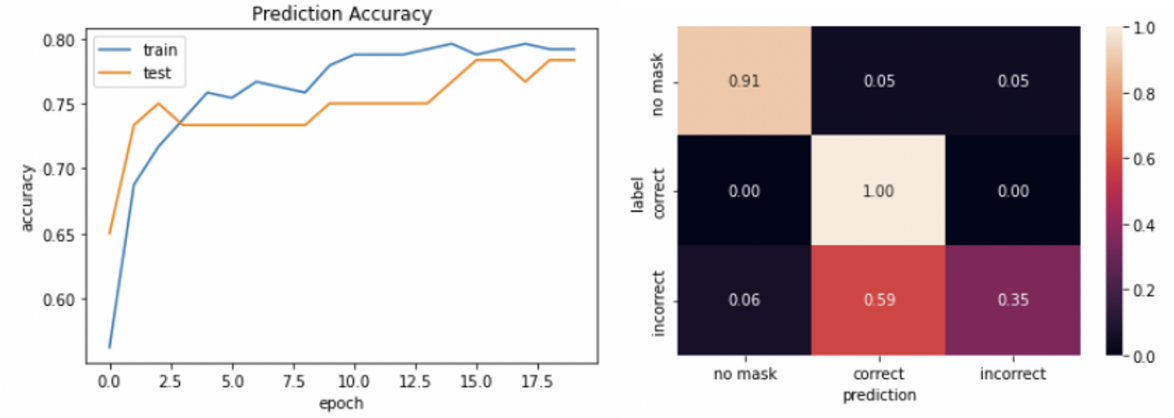

Figure 6: Left: Train and test accuracy for 10 epochs of training of the Ensemble Model. Right: Confusion matrix for test of Ensemble Model.

Curious as to what was causing the Symbolic Reasoning Model to struggle, we took a deeper look at performance. Images that were misclassified are shown in Figure 7. We noticed that images that were erroneously classified either didn't have face features detected or had a very high probability of a mask produced by the binary mask classifier. To determine whether the model was disproportionately overweighting the output of the binary mask classifier we computed the sum of the weights from each of the inputs to all 36 nodes in the first layer. We found that the weight was not much stronger for the output of the binary mask classifier than any of the other features (last value in the input). The sum of the weights was: [0.040, 1.208, 0.847, 0.797, -0.228, 1.208, 1.564, -0.311, 0.692, 0.140]. In fact, no feature seemed to be disproportionately more important than the others.

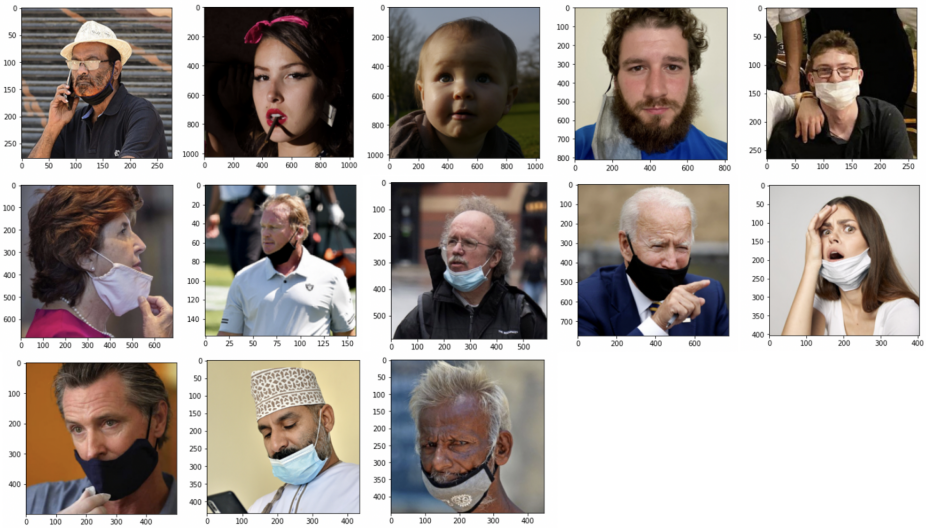

Figure 7: Left: All of the test images that were misclassified by the Ensemble Model.

4 Discussion

Overall, our results matched our theoretical expectations. The SVM model had the simplest architecture and had the lowest F1 score among the four models. The pretrained AlexNet was the clear winner and demonstrated the strengths of transfer learning on limited training data. Having a pretrained generalized model meant that the model started out with reasonable weights and finetuning did not require a large dataset.

The baseline models and the pretrained AlexNet both showed similar patterns in their errors in that they were more likely to classify incorrectly worn masks as no mask, rather than correctly worn. For the purposes of policy implementation, this parsimonious trait may be considered more desirable. Meanwhile, the Symbolic Reasoning Model had a tendency to classify incorrectly worn masks as correctly worn. We attribute the bias to the unreliability of the facial feature recognition stage. Figure 7 shows that the predefined Haar-like features fail to detect noses and mouths when the person is facing a different orientation or the face is partially obscured by shadows or a moustache.

While the performance of the Symbolic Reasoning Model leaves more to be desired, the systematic patterns in its misclassification show promise. Had the nose and the mouth been successfully identified, the final classifier could have correctly assigned the images as incorrectly worn. For future improvements, more advanced methods for individual facial feature detection would prove helpful. One approach would be to train CNN-based pretrained feature detectors for noses and mouths instead of the Haar cascades. While Haar-based features are still widely-used for facial features detection, we expect CNNs to outperform them for partially-obscured (masked) faces. Another extension would be to address face orientation using GAN models for image rotation (Wang et al., 2021). Not only would GAN-based rotation increase the detection rate of facial features, it could help further augment the dataset and improve generalizability. Lastly, the architecture of the miniature neural network can be improved from its current design of a Multi Layer Perceptron to one that may better capture the relationship between the conceptual representations of the inputs.

With the above issues ironed out, the Symbolic Reasoning Model has a potential to become a customizable and explainable classification model. For example, if we were to further classify incorrectly worn masks into subcategories (only the nose/mouth is exposed, both are exposed, etc.), the neural net model would require additional labels in the training data for the respective variations. The Symbolic Reasoning Model on the other hand may adjust to the new task with simple manual reweighting of the edges. Additionally, having fewer parameters in the model would facilitate explanation of its underlying logic and make the model robust to unrelated biases.

Our results show that simple models using symbolic representations have the potential to serve as explainable, data-efficient alternatives to popular deep learning models. The output of the Symbolic Reasoning Model also demonstrates a need for a greater variety of modular algorithms. In the duration of the project, the authors were unable to find reliable pretrained neural network classifiers for individual facial features, and had to resort to detection methods that could not take advantage of the latest developments in artificial intelligence. Expanding the library of high-performance feature generators would assist research efforts in finding novel and efficient solutions to new problems.

Acknowledgments

We would like to thank Anushree Hede for all of her guidance throughout the semester and especially for her advice as we worked through several stages of this project. We would also like to thank Professor Konrad Kording for his helpful insights during early stages of this work.

5 References

Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K., Taha, K. (2015). Efficient machine learning for big data: A review. Big Data Research, 2(3), 87-93.

Cabani, A., Hammoudi, K., Benhabiles, H., Melkemi, M. (2021). MaskedFace-NetâA dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health, 19, 100144.

Center for Disease Control and Prevention. (2021, January 30). How to Wear Masks. https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-to-wear-cloth-face-coverings.html

Crawford, K. Paglen T. (2019, September 19). Excavating AI: The Politics of Training Sets for Machine Learning. https://excavating.ai.

Fort, K., Adda, G., Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine?. Computational Linguistics, 37(2), 413-420.

Garvie, C. (2019, May 16). Garbage In, Garbage Out. Face Recognition on Flawed Data. Georgetown Law Center on Privacy Technology. https://www.flawedfacedata.com.

Haugeland, J. (1989). Artificial intelligence: The very idea. MIT press.

Howard, J., Huang, A., Li, Z., Tufekci, Z., Zdimal, V., van der Westhuizen, H. M., ... Rimoin, A. W. (2021). An evidence review of face masks against COVID-19. Proceedings of the National Academy of Sciences, 118(4).

Jones, M., Viola, P. (2003). Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96, 3(14), 2.

Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).

Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Larxel. (2020). Face Mask Detection. 853 images belonging to 3 classes., Version 1. Retrieved April 2021 from https://www.kaggle.com/andrewmvd/face-mask-detection.

Mercaldo, F., Santone, A. (2021). Transfer Learning for Mobile Real-Time Face Mask Detection and Localisation. Journal of the American Medical Informatics Association.

Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., Hemanth, J. (2021). SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. Sustainable cities and society, 66, 102692.

Smith, S. C. (2019, November 19). Dealing With Bias in Artificial Intelligence. The New York Times. https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html.

Ungar L., Kording K., Zhou M. (2021). CIS-522 Deep Learning Tutorial. Week 13 - The Future of Deep Learning. https://github.com/CIS-522/course-content/blob/main/tutorials/W13_FutureDL/W13_Tutorial.ipynb

World Health Organization. (2020). Advice on the use of masks in the context of COVID-19: interim guidance, 6 April 2020 (No. WHO/2019-nCov/IPC_Masks/2020.3). World Health Organization.