Real time face mask detection with SSD

Erick Sebastián Lozano Roa
Biomedical engineering department
Universidad de los Andes
Bogotá, Colombia
es.lozano@uniandes.edu.co

Juan Sebastián Urrea López
Biomedical engineering department
Universidad de los Andes
Bogotá, Colombia
js.urrea@uniandes.edu.co

Isai Daniel Chacón Silva
Biomedical engineering department
Universidad de los Andes
Bogotá, Colombia
id.chacon@uniandes.edu.co

***Abstract*—This paper proposes the Neural Network SSD architecture for the task of face mask detection. We trained this model to detect 3 different categories of people: those wearing a mask, those wearing it incorrectly and those not wearing it. For this proposal, a dataset with 853 images was used with annotations of these 3 classes. For the Neural Network, the weights were initialized from a pretrained model on the image dataset Pascal VOC and fine-tuned to finally achieve a mAP of 70.2% on the validation set. The mAP achieved on the test set was 66.7% which shows that despite the lack of a huge dataset, the architecture SSD is suitable for this task. Additionally, our experiments showed that the MultiBox Loss Function parameters can be modified to improve the performance on small and unbalanced datasets such as the one presented.**

Current work has focused on the task of classification or detection for a single class. Our work is a novel approach as it performs real-time detection and classification in 3 classes, making it suitable for natural scenarios such as public spaces.

***Index Terms*—Covid-19, Face Mask Detection, Single Shot Detector, MultiBox Loss.**

I. Introduction

Currently, the use of face masks has become part of everyday life. It is considered to be one of the main adaptive measures taken by the population throughout the world to prevent the transmission of SARS-CoV-2, since its use has shown a considerable decrease in the rate of transmission of the virus [1].

Around the world, social awareness regarding the use of face masks has not been fully developed, as it is common to find people in public places without or with face masks used incorrectly. For this reason, we created a Deep Learning model to solve the problem of detecting people's faces and classifying them into 3 main categories: with a face mask, without a face mask and with a face mask incorrectly worn.

The objective of the study is to automate the process of checking the proper use of the face mask to avoid possible contagions. This would have a direct application in the cameras of mass transit buses, hospitals, state buildings, shopping malls, gyms, universities, schools, companies, among many other public spaces. The motivation to do this is that people regain the confidence to be in public spaces.

The use of face masks to prevent transmission of COVID-19 was first adopted in response to the abrupt epidemiological outbreak that occurred all around the world in a few months, but its use was later shown to be fully justified. A study by The Royal Society demonstrated through the implementation of two mathematical models that the use of face masks by the population is a major contribution to reducing the impact of the pandemic. The study concluded that if the majority of the population wears a face mask with high effectiveness and all the time, very low values of effective reproductive rates are achieved. [2]

A usual approach to the problem is done via Deep Neural Networks of classification, such as the one presented by [3], [4] which are the state of the art since they presented an accuracy of 100% in test and 96% in the validation set, respectively. However, our approach is a novel way to see the problem because uses detection as the main problem to be solved and optimized. In this sense, our dataset is different because more than a person can appear in one image and it is essential to find them all and assign them 1 of the 3 possible categories. This will allow to make a more realistic model that can be used in public places as a way to avoid contagion.

The current state of the art for the database chosen in this project is the model proposed in [5]. In this article the experiments were performed with a database created from the combination of two public databases of face mask use: Medical Masks Dataset (682 images) and Face Mask Dataset (853 images). The latter is the one used for the development of this project. In the aforementioned article, a detection model based on the YOLO-v2 detector (Machine Learning) with ResNet-50 (Deep Learning) was proposed for detection and extraction in the training, validation and testing phases. They used two optimizers: with the Adam Optimizer obtained a mAP of 81% and with the SGD obtained a mAP of 61% [5].

This paper presents a novel approach to the task of detecting and classifying face masks, since 3 classes are examined (with a face mask, without a face mask and with a face mask incorrectly worn), moreover this process is so efficient that it allows performing these tasks in real time, as the detection process and the frame construction take less than a second, which allows implementing it in realistic environments where there are many people performing daily activities.

II. Methodology

A. Dataset

We used the Face Mask Detection dataset available at Kaggle [6]. This dataset has 853 images with bounding boxes of three categories of people: those wearing mask, those wearing it incorrectly and those not wearing it. Table I shows that these categories are not balanced: 80% of the bounding boxes are classified as *With mask*, 17% as *Without mask* and 3% as *Mask worn incorrectly*. Additionally, we split the data in subsets for training (70% of the images), validation (15%) and testing (15%).

Categories unbalance can be identified as one of the main problems, as it has been shown that unbalance can have a major impact on the value and significance of accuracy and some other well-known performance metrics [7]. This is one of the main problems, as models will tend to have low returns at the classes with less annotations. Another complication is that in some images there are a high number of annotations and each has a different size, in which many annotated faces are seen, but there are some so far apart that they are contained in a few pixels and even overlap with other annotations. Finally, most images in the dataset are from asian people, which could harm, as a hypothesis, the performance of the model for different parts of the world.

TABLE I
STATISTICS ABOUT DATASET IMAGES AND ANNOTATIONS

Annotation class	# Images present	Annotations
With mask	768	3232 (79,37%)
Worn incorrectly	97	123 (3,02%)
Without mask	286	717 (17,61%)
Total	**853**	**4072**

The figure 1 shows some images of the database and their corresponding annotations. Since each image can have several annotations of different classes, a color code was used to differentiate them.

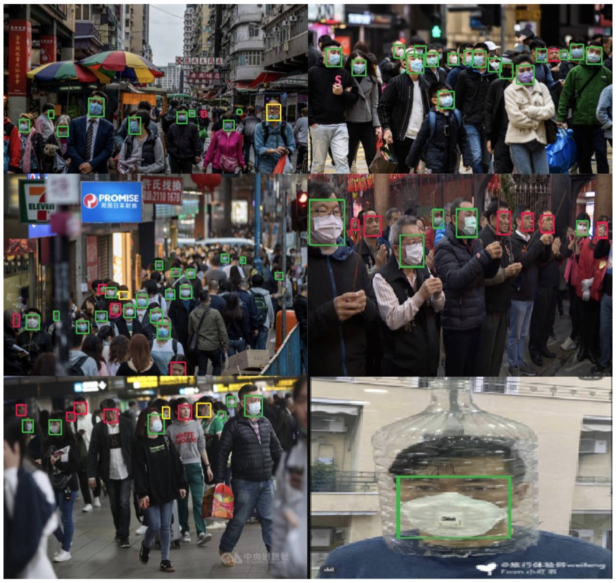

Fig. 1. Sample of images from [6] with corresponding annotations. Color code: *with mask* (green), *with mask used incorrectly* (yellow), *without mask* (red)

B. Metrics

The detection task for multiple categories is evaluated with the mean Average Precision (mAP). This is the mean of the Average Precision (AP) for each non-background class. The AP is a measure of the precision of the model along different recall levels; it is presented in percentage values ranging from 0 to 100%.

C. Architecture

The Single Shot Multibox Detector (SSD) architecture used for the task is a simple method for problems that require object detection, since it does not generate proposals, nor re-sampling of pixels or features, and also manages to compact the entire process in a single deep neural network that can be trained End-to-End. This network discretizes an image by a set of predetermined frames (also called *priors*) of different aspect ratios and scales [8]. For each frame, multiple scores are generated for each category and bounding boxes are adjusted with respect to the shape of the object of interest.

The architecture of SSD showed that by utilizing feature maps from different layers of the network can mimic the effect of processing the image at different scales [8], which is essential for our task, since all the images do not present the same scale and it is important to mimic that effect for a better performance in valid and test sets. Therefore, the Multibox problem can be presented as a regression problem, where a detected object bounding box is regressed to its ground truth's coordinates [9].

The Multibox loss can be formulated as follows:

$$multibox\ loss = confidence\ loss + \alpha * location\ loss \quad (1)$$

in which Alpha (α) can be tuned in order to give more importance to the task of localization over the task of classification.

These features allow SSD to be an easy-to-train architecture for detection tasks, such as the one proposed for face masks.

An important factor is that this architecture outperforms other highly regarded implementations such as Faster RCNN, YOLO and Fast-YOLO in accuracy on Pascal VOC [8] , as well as being faster than the aforementioned models, which is the reason why it works for real time detection and was the architecture proposed for this work.

D. Experimentation

Initially, we trained a benchmark model with the same learning parameters and data augmentation techniques as [8]; we only changed the number of epochs to 300 with no learning rate decay schedule. To initialize the weights, we used the pre-trained model on Pascal VOC provided by [10] that achieved 77.2 mAP on that dataset. It is important to emphasize that the experimentation is focused on the change of hyperparameters, not on the modification of the architecture of SSD network.

Beginning with this baseline model, we experimented with 5 different parameters from the learning process and the Multibox Loss. For each parameter, we trained several new models with different values in order to analyze their influence over the detector. In each epoch, the models were evaluated using the Multibox Loss in the validation set and the weights that yielded the lower loss were chosen as the best model for that value. Following this, these best models were evaluated in the validation set using the mAP metric and the value with the best performance was chosen.

The values used for the parameters are described below. First, we started with the parameters for the Multibox Loss. The default Negative/Positive ratio is 3 [8], but since the weights come from a pretrained model that already knows how to detect objects (and not background), we chose to test the values 1 and 2. Additionally, since the classification loss is calculated through Cross Entropy, we tested the effect of giving higher weights to the unbalanced classes to improve the AP on those classes. Finally, we tested alpha values different than 1.

On the other hand, we also experimented with the learning parameters. For instance, we increased the learning rate and also tried the Adam optimizer as opposed to the default Stochastic Gradient Descent optimizer.

Finally, we also tested the final model for the task of real time detection. For this, we sent the camera image from several devices to a central machine through TCP and ran the model on those images. We also measured the processing speed in frames per second (FPS).

E. Implementation

We modified the SSD implementation from [10] to perform detection in our 3 target classes with custom parameters as described before. We also implemented a simple client-server architecture to perform real time detection on a central machine (server) from multiple cameras (clients). For training and inference, we used a single Nvidia Titan X Pascal GPU on Ubuntu 20.04.2 LTS with Python 3.9.5, Pytorch 1.8.1 and CUDA 11.1

III. Results

TABLE II
DETECTION METRICS AFTER DIFFERENT EXPERIMENTS.
WOM: *Without mask*, MWI: *Mask worn incorrectly*, WM: *With mask*

Parameter	**mAP**	**WOM**	**MWI**	**WM**
Baseline	55.6	54.4	42.6	69.8
Negative/Positive Ratio	57.5	54.3	47.3	70.9
Learning Rate	63.9	57.1	56.0	78.4
Cross Entropy Weights	68.0	63.1	64.3	76.5
Alpha	**70.2**	62.4	71.4	76.8
Optimizer	61.1	54.2	53.6	75.4

The final model which showed the best performance, as shown in Table II, used a *Negative/Positive* Ratio of 2, which means that for every positive example the algorithm creates 2 hard negative examples to improve the overall performance; a *learning rate* of $1 \cdot 10^{-2}$ with no decay; and the following weights for the *Cross Entropy* Loss function: 1 for background, 2 for *With mask*, 4 for *Mask worn incorrectly* and 3 for *Without mask*, so that the larger the number, the larger the prioritization in the loss metric. Finally, the *Optimizer* used was SGD, and an *Alpha* (α) (eq 1) of 2. The only experiment which showed no improvement over the previous parameters was the change of optimizer from SGD to Adam.

We evaluated the final model from the validation experiments in the test set. We achieved an overall mAP of 66.7 with AP values of 69.2 for people wearing mask, 67.2 for people not wearing a mask and 63.6 for people wearing a mask incorrectly. This shows that the algorithm is learning to identify the masks and the people wearing them, as it is not overfitting, due to the fact that we chose the best model in the validation, also because SSD is a single neural network, which decreases the overfitting problem. Also we can see that the metrics in the test set are similar to the ones reported in validation set (Table II).

Therefore, in a qualitative way, we wanted to see what the algorithm was really detecting and found that the nose and its orifices were the main feature that changed the classification from a class to another. It was even seen that covering the nose (not necessarily with a mask) detected the class *With mask* as shown in Figure 2. It was also observed that the algorithm had a high performance with faces that are close to the camera, while with distant ones it tended to lower the performance, especially for the minority classes. Finally, we could observe some transition stages during the detection of the algorithm when using it in real time. When the mask was worn incorrectly under the nose, sometimes it predicted the classes with mask and worn incorrectly at the same time; but when we used the mask under the chin it predicted the classes mask worn incorrectly and without mask. Figure 3 shows the visual performance of the model on the dataset, while Figure 2 shows the performance on the client web camera.

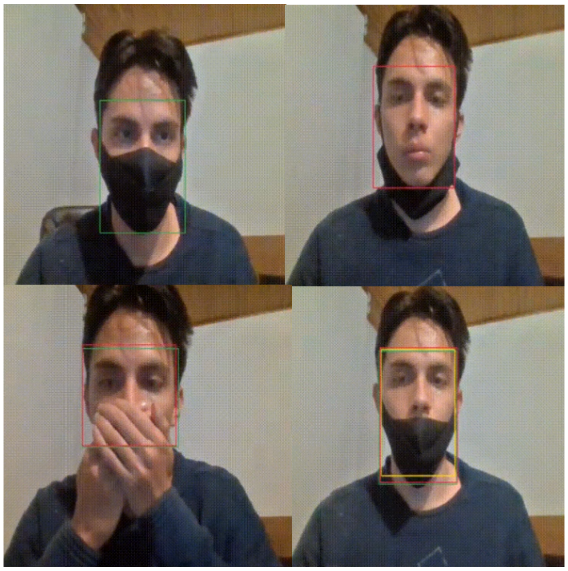

Fig. 2. Visual performance of the final model on client web camera. Color code: *with mask* (green), *with mask used incorrectly* (yellow), *without mask* (red)

The Figure 3 shows in a qualitative way the performance of the SSD Neural Network used in the test set. In this image it is possible to observe the 3 different classes being classified.

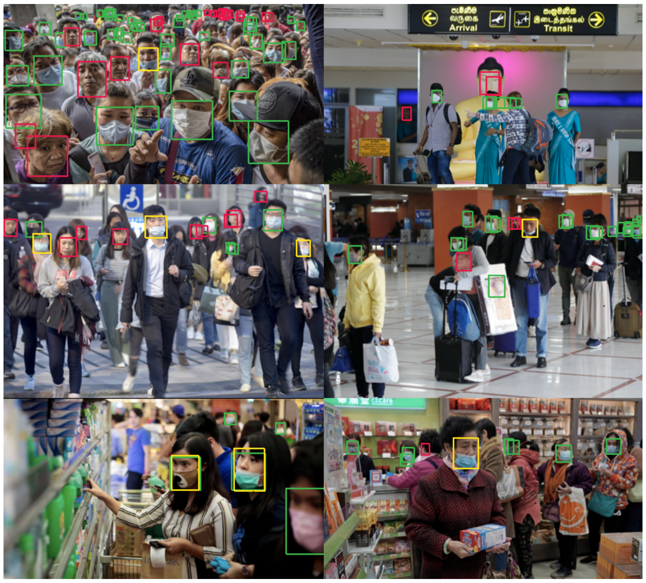

Fig. 3. Visual performance of the final model on dataset. Color code: *with mask* (green), *with mask used incorrectly* (yellow), *without mask* (red)

Also, it is possible to observe the transition classifications problem and that some drawings (of humans) were detected by the model; which are issues that can be tried to be solved in future experimentation.

When comparing the obtained result of 66.7% in mAP with the state of the art (61% in citeb5) with the SGD optimizer, an improvement of 11% over Yolo-v2 with Resnet50 is evident. It is important to note that this research used 1415 images, while this study used only 853 and obtained better performance with a smaller amount of data and more categories. Although the dataset is different and this is an obstacle for an objective comparison between works, it is important to note that with less data better results are obtained for the main class (with mask).

Finally, we measured the FPS in a real-time detection setup. Our model takes about 0.05 seconds on inference, which makes it suitable to an FPS rate of around 20. Yet, due to the time taken to transfer the camera images to the central machine, we measured between 10 and 15 FPS (depending on the internet connection) with our client-server architecture. Nonetheless, this makes our model suitable for real time applications.

IV. Conclusions

In this paper, we proposed SSD as a viable architecture to perform real-time face mask detection. Since it has a fast detection and frame creation (less than one second) and has an acceptable mAP (66.7%, 15% better than the baseline) and good qualitative results, this model can be implemented in different public spaces to identify people who do not wear the mask or not in the correct way. This is a novel approach to the mask detection problem, since it takes into account multiple classes and not only classifies, but also detects. Therefore, despite not having perfect metrics in AP classes, this algorithm may in fact have real utility in public spaces right now as the detection time in FPS is really fast and in real-time video it will be possible to determine whether a person is wearing a mask or not from different frames in a time sequence. The efficiency and accuracy of this algorithm would go a long way in helping people to wear masks correctly and thus decrease the contagion rate.

For future work, data augmentation is the main improvement to implement to deal with the imbalance problem of the dataset. It is essential to perform this augmentation in order to be able to use deeper Neural Networks that perform better without doing overfitting. While image transformations help, it is necessary to collect more annotated images, specially for the classes with less presence in the dataset. This could also improve the performance of the model outside the Asian people.

Another possible improvement when using the Neural Network live could be to use non-maximum suppression based on the probability of the classes, taking the maximum one. This could solve the transition stages problems observed during qualitative testing. We leave this for future work.

V. Acknowledgments

We thank Professor Pablo Arbeláez for allowing us to use the necessary hardware for the research.

References

[1] J. H. *et al.*, “Face masks against covid-19: An evidence review,” 4 2020. [Online]. Available: https://www.preprints.org/manuscript/202004.0203/v1

[2] R. O. Stutt, R. Retkute, M. Bradley, C. A. Gilligan, and J. Colvin, “A modelling framework to assess the likely effectiveness of facemasks in combination with ‘lock-down’in managing the covid-19 pandemic,” *Proceedings of the Royal Society A*, vol. 476, no. 2238, p. 20200376, 2020.

[3] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, “A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic,” *Measurement*, vol. 167, p. 108288, 2021.

[4] S. V. Militante and N. V. Dionisio, “Real-time facemask recognition with alarm system using deep learning,” in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*. IEEE, 2020, pp. 106–110.

[5] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, “Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection,” *Sustainable Cities and Society*, vol. 65, p. 102600, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670720308179

[6] M. ML, “Mask dataset,” Available: https://makeml.app/datasets/mask, May 2020.

[7] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognition*, vol. 91, pp. 216–231, 2019.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “Ssd: Single shot multibox detector,” in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[9] F. Eddie. Understanding ssd multibox — real-time object detection in deep learning. [Online]. Available: https://towardsdatascience.com/understanding-ssd-multibox-real-time-object-detection-in-deep-learning-495ef744fab

[10] S. Vinodababu, “Ssd: Single shot multibox detector — a pytorch tutorial to object detection,” https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Object-Detection, 2019.