# Masked Face Recognition with Barlow Twins

Omar Younis
University of Bologna
omargallal.younis@studio.unibo.it

Mattia Bertè
University of Bologna
mattia.berte@studio.unibo.it

## Abstract

*When Covid-19 appeared, the most immediate defensive tool for contrasting the spread was the use of face-mask. However this introduction made more evident one of the weaknesses of our facial recognition systems, they indeed were unable to correctly classify/recognize the identity of a person wearing a face mask. We tried to overcome this problem proposing a solution applicable to the standard systems made by a convolutional neural network as feature extractor plus a final classifier such as SVM or $k$-NN. Our method exploits the Barlow Twins technique for learning the invariance to the presence of the face mask. We used MLFW dataset, a new synthetic dataset created starting from CALF, so our network not only learned the invariance to the face mask but also to the age. Our results aren't comparable to state-of-the-art systems however seem promising for successive studies.*

## 1. Problem definition

When Covid-19 disease started to spread all over the world, face-mask become one of the first defensive strategy to counteract its diffusion. Face recognition is one of the most common biometric authentication methods, however masked face recognition is a highly challenging task since the mask occludes partially the face making impossible extracting some informative features. Previous systems indeed showed a drop in performances up to 20% when they had to deal with a masked face.

The most widespread solutions for facial recognition/identification were based on a features extraction phase, usually made by a convolutional neural network trained in a supervised way, plus a final classifier such as SVM or $k$-NN.

## 2. Proposed Solution

We tried to propose our solution exploiting a self-supervised learning technique called Barlow Twins [8]. The idea is pretty simple, they proposed to pass through a twins network architecture two distorted versions of the same image obtained by applying some randomly selected transformation from a set of predefined ones, and then with the help of a defined ad hoc loss, the system learns to produce the same embedding for the distorted images. It helps to produce very similar embedding for all the images of the same class since the network becomes invariant to all these transformations.

Starting from this idea we thought to consider wearing a face mask as a transformation and to train the network to learn the invariance to the presence of the mask.

At each step, the network receives two images of the same person, one with the mask and one without the mask. We found a tool for generating realistic masked-face starting from unmasked ones, however, to decrease the computational costs we used the pre-built dataset obtained using this tool.

## 3. Barlow Twins

This section contains a brief theoretical review of the Barlow Twins' loss.

This technique produces two distorted views for all images of a batch sampled from a dataset. The distorted views are obtained via a distribution of data augmentations. The two batches of distorted views $Y^A$ and $Y^B$ are then fed to a neural network which generates embedding $Z^A$ and $Z^B$ respectively.

They defined a new loss eq. 1.

$$\mathcal{L}_{BT} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{1}$$

where $\lambda$ is a positive constant trading off the importance of the first and second terms of the loss, and where $\mathcal{C}$ is the cross-correlation matrix computed between the outputs of the two identical networks along the batch dimension:

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,i}^B)^2}} \tag{2}$$

where $b$ indexes batch samples and $i, j$ index the vector dimension of the networks' outputs. $\mathcal{C}$ is a square matrix with

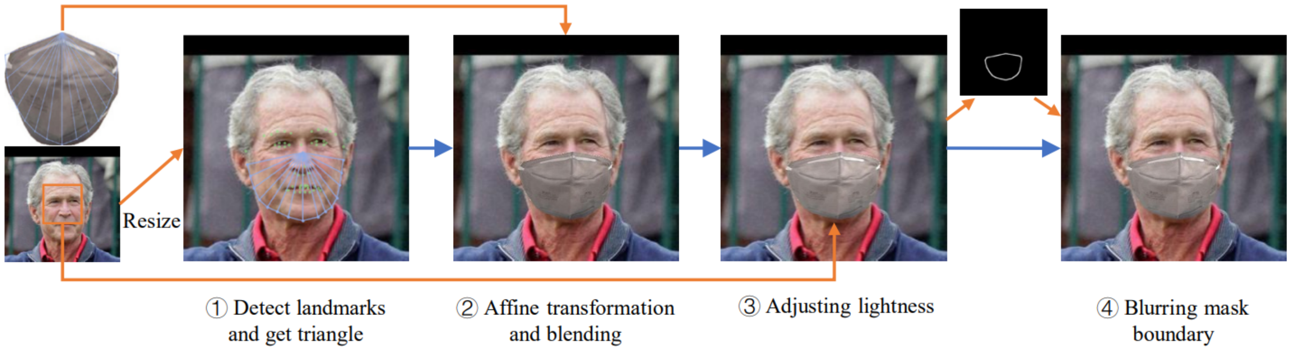

Resize
① Detect landmarks and get triangle
② Affine transformation and blending
③ Adjusting lightness
④ Blurring mask boundary

Figure 1. Face-mask application pipeline.

the size the dimensionality of the network's output, and with values comprised between $-1$ (i.e. perfect anti-correlation) and 1 (i.e. perfect correlation).

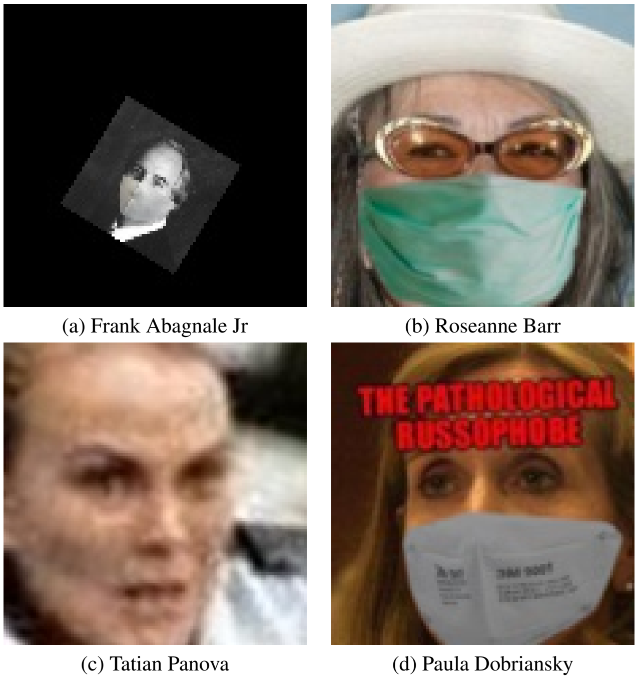
Intuitively, the invariance term of the objective, by trying to equate the diagonal elements of the cross-correlation matrix to 1, makes the embedding invariant to the distortions applied. The redundancy reduction term, by trying to equate the off-diagonal elements of the cross-correlation matrix to 0, decorrelates the different vector components of the embedding. This decorrelation reduces the redundancy between output units so that the output units contain non-redundant information about the sample.

## 4. Dataset

As explained in Section 2, the dataset we used is a synthetic one, called MLFW (Masked LFW) [6]. They started from CALFW (Cross-Aged LFW) [9] dataset which is a variation of the widespread dataset LFW (Labelled Face in the Wild) [4]. CALFW contains labeled faces of famous people in different scenarios, moreover, people are present at different ages making the classification harder. This fact made our task more complex since the network did not need just to learn the invariance to the face mask but also to the changes due to the age, making in a lot of cases the comparison between colored and gray-scale images. The dataset required also some cleaning due to duplicated images, similar masks applied to the same image, or bad-quality images (Fig. 2).

Starting from these images, they created a tool for generating realistic masked face through a complex algorithm that detects some key points and then applies the mask. To make more general as possible, researchers applied different mask templates to mimic the most common face masks.

Barlow Twins requires big batch sizes to perform well, however large batch sizes are known to lead to a drop in performance. To overcome this problem, they used LARS [7] as an optimizer which allows the use of large batch size thanks to different adaptive learning rates for each layer.



(a) Frank Abagnale Jr (b) Roseanne Barr (c) Tatian Panova (d) Paula Dobriansky

Figure 2. Example of low-quality images.

## 5. Distributed optimization

To speed up this computationally expensive training, we split the work on 8 Quadro RTX 6000.

Each GPU processes a different subset of the batch and computes the gradient; then all the gradients are averaged together to obtain the final update direction. In particular, we used a batch size of 512, thus every GPU process 64 samples. To speed up the communication of the gradient, we tried PowerSGD, a gradient compression algorithm [5]. PowerSGD compresses the gradient using a low-rank approximation. For speed purposes, the low-rank approximation is estimated using the power iteration algorithm.

## 6. Experiments and results

We finetuned a ResNet50 pre-trained on MS-Celeb-1M [3] and already finetuned on VGG2 [2] dataset [1].
We tested the network on identities not present in the train set. We split the test set into 2 parts, one for training the $k$-NN and one for testing it.
We generated the embedding for the $k$-NN using the backbone. We selected $k = 1$, which means that the predicted identity is one of the closest embedding in the projected space, it was also due to the fact that many identities present just a few images, so using $k = 1$ treat all the identities in the same way without favoring the ones with many images.
Since SGD presented the best results, we tried to exploit at most all the images at our disposal. Instead of sampling only 2 images for each celebrities, we sampled all $n/2$ couples for each identity where $n$ is the number of image for that id. Doing so the number of couples for each epochs almost doubled.
In Table 1 are presented the results of our experiments.

| **Technique** | **Accuracy** |
|---|---|
| SGD | 79.72 |
| Power SGD | 74.72 |
| SGD + new sampler | 76.67 |

Table 1. Accuracy for the three different techniques applied.

## 7. Ethical issues

Face identification presents always some issues related to the privacy of data and to the possibility of discrimination due to the different capabilities of the system to identify persons of different ethnicity. This is usually related to the unbalanced distribution of the training set.
We tried to derive the ethnicity of the people using an automatic method, such as skin tone estimation [1]. This algorithm [2] consists of two main parts :

- Foreground and background separation using Otsu's Binarization;
- Pixel-wise skin classifier based on HSV and YCrCb colorspaces.

Even if the algorithm performed very well, it was difficult to cluster the images, indeed the tone varies really gradually and it is impossible to estimate the ethnicity depending only on this feature. Moreover, the skin tone was too dependent on the light condition. A sample can be seen in Figure 3. We decided to analyze the performances of our system in each different group. We identified 5 main groups: White/Caucasian, African American, Asiatic, Arab, and Hispanic.

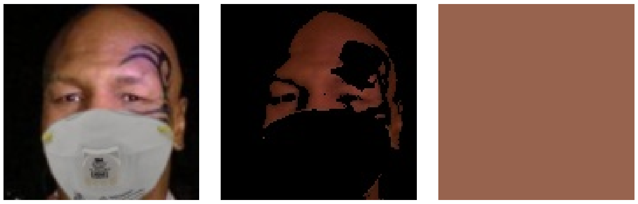

(a) Original image
(b) Segmented images with estimated tone
(c) Estimated tone

Figure 3. Example of skin tone estimation.

We labeled each person manually, however, this has not to be intended as a classification aim to create a new dataset but just to provide an idea of the dataset distribution. We are conscious of the possibility of making mistakes in this labeling process, our purpose was just for performance analysis and it wasn't intended to offend any of the people on the list.

| **Ethnicity** | **Train set percentage** | **Test set percentage** |
|---|---|---|
| African American | 8.04 | 5.22 |
| Arab | 4.06 | 6.71 |
| Asiatic | 5.49 | 2.24 |
| Hispanic | 5.11 | 6.72 |
| Caucasian | 77.30 | 79.11 |

Table 2. Ethnicity distribution of a dataset.

Once the dataset was labeled, we checked the performances of the single ethnicity to prevent biases. The performances are reported in Table 3.

| **Ethnicity** | **Accuracy** | | |
|---|---|---|---|
| | SGD | Power SGD | SGD + new sampler |
| African American | 94.44 | 94.44 | 88.89 |
| Arab | 75.00 | 85.72 | 78.57 |
| Asiatic | 83.33 | 83.33 | 83.33 |
| Hispanic | 80.00 | 70.00 | 70.00 |
| Caucasian | 79.02 | 72.72 | 75.87 |

Table 3. Accuracy results for each ethnicity.

Unexpectedly even if the Caucasian ethnicity is the most present, it isn't the one that presents the best accuracy. It means that this approach isn't particularly biased toward

[1] Link to the GitHub repository with the original weights: https://github.com/cydonia999/VGGFace2-pytorch
[2] Link to the GitHub repository with the algorithm: https://github.com/colin-yao/simple-skin-detection

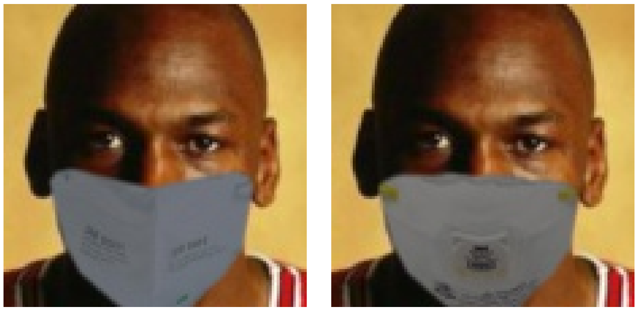
some specific ethnicity or it's unable to work with a specific minority.

## 8. Conclusions

We weren't interested in reaching state-of-the-art performances in masked face recognition, however, we proved the validity of our idea by exploiting a self-supervised technique combined with a metric learning approach. Our proposed solution first of all showed the effectiveness of the Barlow Twins technique moreover, we understood that also face masks and aging can be considered as simple transformations as it is for geometric or color transformations.

One of the biggest weaknesses of our solution is related to the dataset, it's not built so well for our purpose. It required some manual cleaning and a lot of images was really similar since produced from the same original unmasked face with the addition of the same face mask but with different colors (Fig. 4). It slows down the learning process and the generalization capability. This can be also the reason behind the decreasing in performances when we increased the dataset size. Sampling all the possible couple increased a lot the probability of sampling two really similar images which causes bad training.

Even if PowerSGD seemed a promising technique for speeding up the training, we noticed little to no speedup in using it. We believe this is because the GPUs were located on the same node, thus they had a very high bandwidth. In the end, a positive aspect we think is worth noticing is the behavior related to the different ethnicity. Of course, we can notice different performances for each subgroup, however, the majority of them show better performances than the Caucasian ethnicity. We cannot conclude for sure that this approach is free from biases however it seems a promising point for successive studies.



(a) Michael Jordan face mask template 1

(b) Michael Jordan face mask template 3

Figure 4. Example of different samples produced from the same original image but with different face masks.

## References

[1] Emir Buza, Amila Akagic, and Samir Omanovic. Skin detection based on image color segmentation with histogram and k-means clustering. In *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 1181–1186, 2017. 3

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. page 67–74. IEEE Press, 2018. 3

[3] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. volume 9907, pages 87–102, 10 2016. 3

[4] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 2

[5] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. In *Neural Information Processing Systems*, 2019. 2

[6] Chengrui Wang, Han Fang, Yaoyao Zhong, and Weihong Deng. Mlfw: A database for face recognition on masked faces. In *Chinese Conference on Biometric Recognition*, 2021. 2

[7] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv: Computer Vision and Pattern Recognition*, 2017. 2

[8] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021. 1

[9] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 2