CS230

Facial Recognition for People Wearing Masks

Zhen Li
Department of Computer Science
Stanford University
zhnli@stanford.edu

Abstract

Traditional facial recognition (FR) algorithms have difficulties to handle masked faces. Masks obscure the important facial features that FR algorithms rely on to map to an embedding space. The aim of this project is investigate two new approaches to improve the performance of FR systems on masked facial images. In the 1st approach, we propose a residual model with a new quintuplets loss function which accounts for faces with and without masks. In the 2nd approach, we use generative model perform mask removal and image inpainting to restore certain facial features that can facial recognition. A synthetic dataset of facial images with masks is generated for training our models. Results show that both of our approaches can improve the performance of FR systems on masked facial images.

1 Introduction

Studies have shown that wearing masks helps prevent the spread of coronavirus during COVID-19 pandemic. However, the new situation has presented challenges to widely used Facial Recognition (FR) systems. NIST conducted facial recognition accuracy tests recently. Their report reveals that mainstream FR algorithms have increased error rates ranging from 5% to 50% on faces with masks.

Traditional FR systems have failed to handle this situation because they have been trained to rely on important facial features to map faces into an embedding space. And some of these features are obscured in cases that the lower half of a person's face is covered under a mask. As a result, people have to take off their masks to get identified successfully, which could lead to increased chances of virus infection. To contribute to the global battle against coronavirus, we evaluate the performance of FR systems on mask covered faces and investigate new methods to enhance the performance so that people do not have to take off their masks to open doors or unlock their devices.

We investigate two new approaches. One is to train a residual network with a quintuplets loss function, which takes into account anchor, positive unmasked, negative unmasked, positive masked and negative masked facial images. The idea of our approach is that by adding the new terms in the loss function, the model learns to create embeddings for masked covered facial images and these embeddings are closer in the latent space to the embeddings from the same identity than the ones from different identities.

The other approach is to unmask obscured facial images using General Adversarial Network (GAN) [1]. First, a binary segmentation is generate using an object detection and segmentation algorithm to mark the area of a mask in a facial image. Then this segmentation and the original facial image are both used in the input of a GAN model. A generator pre-trained with a discriminator unmasks the face in the image and performs an inpainting task. The unmasked and inpainted image then can used by a traditional facial recognition algorithm to match its identity.

CS230: Deep Learning, Fall 2020, Stanford University, CA.

2 Related Work

There are mainly two types approaches used by existing FR algorithms. One is to train a multi-class classifier like softmax which is able to distinguish face images of different identities [2, 3, 4]. Another is to learn embedding functions using CNN with Triplet Losses [5]. Both approaches can achieve impressive results in large scale facial recognition. There are also researches that proposed variant approaches [6, 7, 8]. For example, ArcFace [9] introduced angular margin loss to enhance discriminative power.

There are challenges in both of these approaches. In the softmax approach, the complexity of linear transformation increases linearly with the number of identities. The learned features can be used to separate the identities in a close-set problem but may not be discriminative enough for open-set applications. In the triplet loss approach, the techniques used for triplet mining in large scale dataset can be challenging and critical to the performance and convergence of the model.

In this project, we adopt the triplet loss method. A new quintuplets loss function is proposed to generate embeddings capable of discriminating examples of masked faces. We also explore different ways of triplet mining as choosing the appropriate triplets are important to prevent model collapse and achieve convergence, .

Studies have shown that GAN can be used to for object removal and image inpainting [10, 11]. In a two-stage process, the object to be removed is firstly marked with a segmentation mask. Then image completion is done on the region that is missing or obscured. The second step is performed by adversarial models that have been trained for image inpainting on a dataset that is suitable for the application.

3 Dataset

Large labelled datasets of people wearing face masks are difficult to come by. We decided to use synthesized data to train our models.

3.1 Facial Recognition

We generated our own synthesized dataset based on the widely used Vggface2 dataset [4]. A subset of 500 identities are chosen from the Vggface2 dataset due to limited resources. There are 169396 facial images in total. 460 of these identities are selected for training, 20 for validation and 20 for tests. We use a dlib based approach [12] to put masks on the pictures from the dataset. The algorithm can detect the face's position and angel and adjust the mask image so that a realistic result is achieved.

The Vggface2 dataset is loosely cropped which means that the images contains larger area than the face itself. These additional elements in the image can have negative impact on the training. To remove the distractions, We develop a script to crop the images. The script uses MTCNN [13] to locate a bounding box of the face. We expand the bounding box by a factor of 0.2 to include the entire face. The final image is cropped based on the bounding box. The size of the image after cropping is 224×224.

Figure 1 shows the dataset generation pipeline. From the left to right is the original face image, mask added image and cropped image.

Figure 1: Dataset Generation: Original => Masked => Cropped

3.2 Mask removal and image inpainting

For mask removal and image inpainting, we reuse the synthesized dataset that we created for Facial Recognition. While using dlib to generate the masks, we also record the locations of the masks and create binary segmentation maps. As shown in Figure 2, on the left is the synthesized facial image with mask, and on the right is the segmentation map in grayscale.

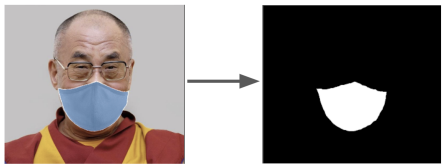

Figure 2: Mask Segmentation

4 Methods

4.1 Models

We train a Siamese Convolutional Neural Network on our synthesized dataset using a newly designed Loss Function. The resulting neural network is able to map a given image to a compact Euclidean space, where distances directly correspond to a measure of facial similarity.

We have run experiments using VGG and ResNet neural networks as the backbone of our model. VGG failed to converge on our dataset with the hyperparameters that we have tested. ResNet has worked better on this aspect. Hence, ResNet50 is chosen for as the backbone of our project. We extend ResNet50 with a Dropout layer and also use a new linear layer to create 512 bit embeddings. The final FC layer of ResNet is replaced by the layers in Table 1.

ID	Layers	In	Out	Kernel Size
0	Dropout			
1	AdaptiveAvgPool2d	2048	2048	1×1
2	Flatten			
3	Linear	2048	512	
4	BatchNorm1d			

Table 1: Extension layers

4.2 Loss Function

Traditional FR systems like FaceNet use a Triplet Loss function. We propose a new loss function called Quintuplet Loss to deal with masked and no-mask images. A Quintuplet includes 5 images which are the Anchor (A), Positive (P), Negative (N), Masked Positive (P_m) and Masked Negative (N_m). Similar to Triplet Loss, the Anchor is the image to be identified, Positive is another image of the same identity, Negative is an image of a different identity, P_m is a masked image of the same identity, N_m is a masked image of a different identity. α is the margin.

The Loss Function is:

$$\mathcal{L} = ||f(A) - f(P)||^2 + ||f(A) - f(P_m)||^2 - ||f(A) - f(N)||^2 - ||f(A) - f(N_m)||^2 + \alpha$$

4.3 Quintuplets Mining

Like training with Triplet Loss, to avoid model collapsing it is important not to select the Quintuplets that are too easy or too hard. If too easy, they do not help the neural network to train. If too hard, they may lead to local minima or model collapse [5]. In order to mitigate the problem, we choose quintuplets that:

$$0 < ||f(A) - f(N)||^2 - ||f(A) - f(P)||^2 < \alpha$$
$$0 < ||f(A) - f(N_m)||^2 - ||f(A) - f(P_m)||^2 < \alpha$$

4.4 Mask removal and image inpainting

As a different approach to improve the performance of facial recognition systems, we perform mask removal and image inpainting before feeding the images into a FR model. A pretrained generative model form Edgeconnect [10] is used to unmask our synthesized images of people wearing masks. In our experiments, the binary mask segmentation map is created together with the synthesized dataset. It also can be done using an object detection and localization algorithm. The process is illustrated in Figure 3, from left to right the images are: face without mask, face with synthesized mask and binary segmentation map and face with mask removed and inpainted. The resulting images are passed to FaceNet for predictions.

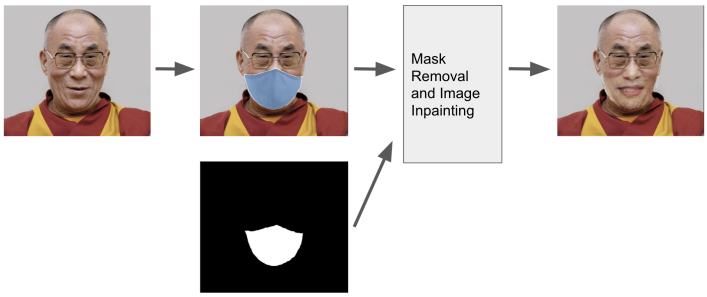

Figure 3: Mask removal and image inpainting

5 Experiments

For this project, we have written scripts to generate synthesized data. The model and training algorithm is implemented in PyTorch. We also developed a vectorized algorithm for the quintuplet loss function. It is known that models using triplet loss are difficult to train. The model tends to collapse if triplet mining is not done properly. We adopt a hybrid approach which combines hardest and semi-hardest quintuplet losses. We train our model using mini-batch gradient descent for 10 epoch. For the first 7 epochs, semi-hardest quintuplet loss are used and for the last 3 epochs, we use the hardest quintuplet loss. All of the loss are calculated via a vectorized algorithm based on the output embeddings of the current batch. Larger batch sizes is preferred when training with triplet loss function. We chose the batch size of 120 due to limited resources. Each batch contains 20 identities and 3 facial images with masks and 3 without masks are randomly chosen for each identity. We also practiced learning rate decay in the training. The learning rate starts with 0.05 and then divided by 10 every 2 epochs. We tried different dropout rates and settle with 0.6. We also experimented with early stopping and decided not to use it for better results.

6 Evaluation

We have tested our models using a test set from our synthesized dataset of people wear masks. FaceNet [5] is the baseline model which has been pretrained with dataset without masks. Our model (MaskFaceNet) is trained with our synthesized train set with quintuplet loss function. We also evaluate our approach of mask removal and image inpainting by using FaceNet to perform prediction on images that are unmasked. As we can see from the results shown in Table 2, FaceNet did not perform very well because it has not been trained with masked facial images. Our model (MaskFaceNet) performs better than FaceNet which proves that quintuplet loss function helps the model to recognize masked facial images. The overall accuracy of our model is not very high because we had to use a small batch size limited by GPU memory and also the training time is much less compared to main stream models like FaceNet. Given more compute resources, MaskFaceNet should have a better performance. It is interesting to see that our mask removal and image inpainting (MRII) approach significantly improved FaceNet's performance on masked facial dataset. It shows that some of the facial features restored by the image inpainting technique can help FR systems recognize masked facial images. It would be interesting to see if the accuracy can be further improved if the generative model is trained on our synthesized dataset.

Model	FaceNet	MaskFaceNet	MRII
Accuracy	67.28%	75.06%	82.39%

Table 2: Evaluation

7 Conclusion

In this project we investigated two approaches to improved the performance of FR systems on facial images of people wearing masks. In the first approach, we created our own synthesized masked facial images dataset and used it to train and evaluate our models. We proposed a modified model and a quintuplet loss function to generate embeddings for images with or without masks. In the second approach, we used generative model to perform mask removal and image inpainting. The evaluation results show that both approaches can improve the accuracy of FR systems on masked facial images. Further improvement can be achieved if more compute resources are available or retrain some of the models with our synthesized dataset.

8 Code

The code of this project can be found at: https://github.com/zhnli/MaskFaceNet

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[2] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[3] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[5] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[6] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[7] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 60–68, 2017.

[8] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[10] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[11] Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. A novel gan-based network for unmasking of masked face. *IEEE Access*, 8:44276–44287, 2020.

[12] Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.

[13] Jia Xiang and Gengming Zhu. Joint face detection and facial expression recognition with mtcnn. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, pages 424–427. IEEE, 2017.