Detection of physical adversarial attack

Jiawei Zhang
jz3209@columbia.edu

Heetika Vipul Gada
heetika.gada@columbia.edu

May 12, 2020

Abstract

Recognizing a human face is a fairly easy task for humans. With the help of DNN, face recognition systems relying on deep learning can also fulfill this task readily. However, recent studies has shown that deep neural network is vulnerable to adversarial attack. Namely, an imperceptible perturbation on the input could drastically change the prediction of the model. Such weakness causes great concern in safety-critical areas. In this paper, we devise a detection technique that are able to distinguish adversarial examples in the context of face recognition. Our technique does not require altering the internal structure of the network. Thus it can be combined with any face recognition network. Our code is available at https://github.com/Jiawei955/detection-of-physical-adversarial-attack.
Keywords: *deep learning; adversarial attack; face recognition; physical attack; adversarial detection*

1 Introduction

Under almost all the natural conditions, we can easily recognize a person if we have seen him before. Deep neural network has been proved capable of addressing face recognition as well. However, Goodfellow et.al [1] proposed that deep neural network can be fooled by some perturbation imperceptible to the human eye. The vulnerability of network brings huge concern since deep learning is or will be widely used in many security-critical fields.

Recent studies has shown adversarial attack presents a threat to face recognition domain. There two possible scenarios: adversarial attack can either make a known identity not recognizable or make a unknown face recognized as someone else. [2] successfully confuses the state-of-art public face model Arcface by decreasing the similarity between the adversarial example and the clean example. Concretely, they generate an malicious stick using Iterative FGSM with momentum and then place the stick on the hat. Face recognition model assigns relatively low cosine similarity between the face with the hat and without the hat. It falls under the first situation that a authorized face being rejected. In this paper, We devise a novel detection mechanism which efficiently flags adversarial examples.

2 Methodology

Adversarial examples are usually close to the decision boundary, in other words, they are susceptible to random noise. While natural examples are relatively more robust to perturbations given that only intentionally crafted perturbations can fool the classifier. We exploit this inherent property of adversarial examples to design our detection methods. We make a reasonable assumption that physical attack also follows this characteristic.

Tao Yu et al. proposed a detection mechanism in 2019 that took advantage of the vulnerability of adversarial examples. They calculated the L_1 distance of the probability vector of original input and that of input corrupted by Gaussian noise. Our approach is similar in spirit. We compute the dissimilarity between embeddings of original face and corrupted face.

$$\epsilon \sim \mathcal{N}(0, \sigma) \; d = ||g(x) - g(x + \epsilon)||$$

where the metric of dissimilarity could be L_{inf}, L_1, or L_2 as long as it makes sense in the specific application. We use L_1 norm in the experiment but it does not mean it is better than other metrics. For a single input image, we make multiple samples and obtain the mean of the dissimilarity over the samples.

$$\frac{1}{n}\sum_{i=1}^{n}||g(x) - g(x+\epsilon_i)||_1 \tag{1}$$

If the distance surpasses the predefined threshold, our model would flag it as an adversarial example. It is worth noting that our method differs from Tao's in that we leverage the difference in the embedding space instead of the probability vector space. Moreover, we compute the distance multiple times rather than once as in Tao's algorithm.

3 Experiment

Modern face recognition system extracts the feature vector of the input face through a deep neural network, then determines the identity based on the distance to each face embedding in the database. Every single embedding corresponds to a known identity. The closet embedding has the greatest chance to share the same identity as the input and it can be obtained by, for example, Knn algorithm. The Arcface [] takes advantage of cosine similarity to find the closet embedding. There are two parts in our experiment. In the first part we construct the physical adversarial examples following the algorithm in [2]. [2] generates malicious samples which are far from its original embedding so it can regarded as an untargeted attack. We then introduce this adversarial example to our detection module to see if it can be detected as adversarial. In the second part, we apply this attack digitally without printing out the adversarial stick.

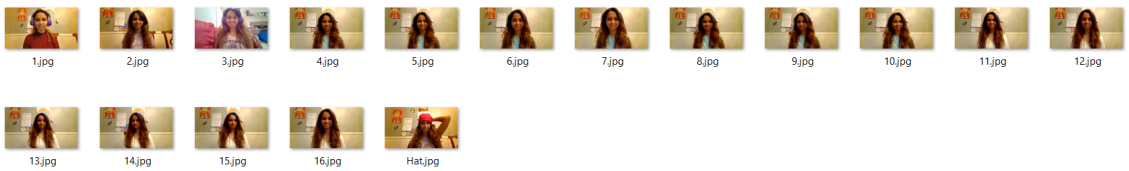

Figure 1: Dataset

As shown in the Figure 1, we initialize the dataset by taking front facing images only. It is a necessary condition that the images be front facing for the ArcFace model and Resnet model to work correctly.

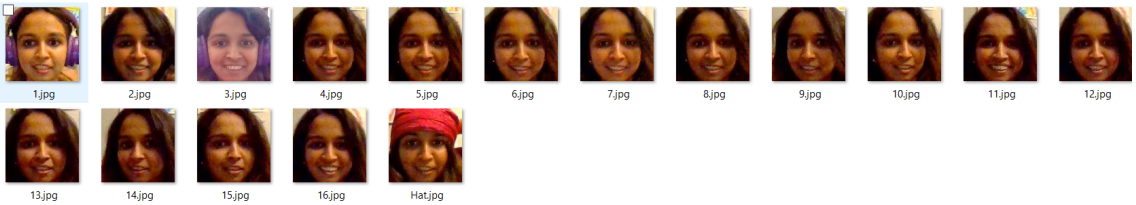

Figure 2: Dataset-Aligned

Next, we align the images to (600, 600). The downsizing is again done to set appropriate transformations for the ArcFace model.

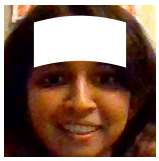

Figure 3: Sticker Position on Face

The position of the sticker is calculated by the detectface module. The detectface module in align, detects the face and locates the forehead. This alignment is then calculated and curved on the sides to accommodate the facial forehead feature.

Figure 4: Dataset-Aligned-Adversarial

Thus, we also align and size the adversarial images to (600, 600)

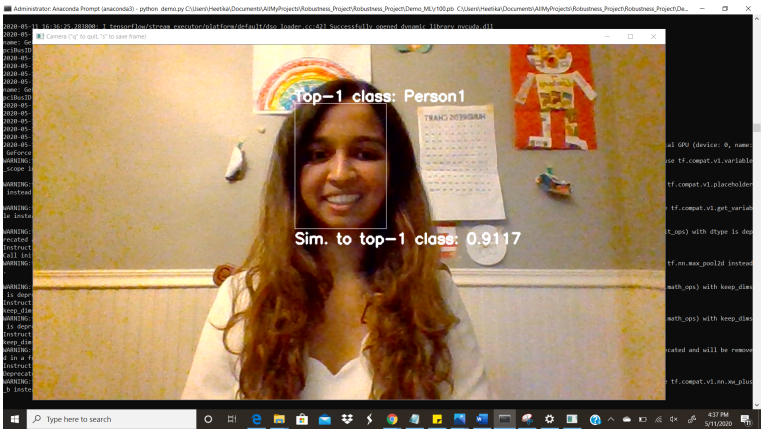

Figure 5: Similarity Detected: 91.17

The cos similarity between the person and the dataset in real time video was found to be ranging from 0.85 to 0.92.

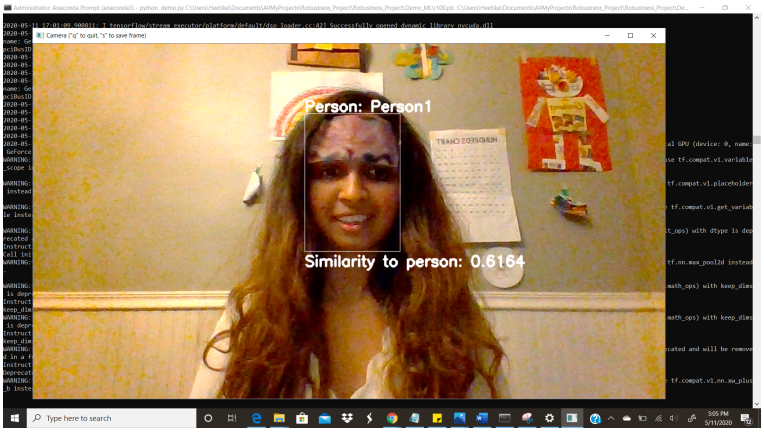

Figure 6: Similarity Detected: 61.64

The cos similarity between the person and the dataset in real time video was found to be ranging from 0.51 to 0.69 when it is an adversarial image.

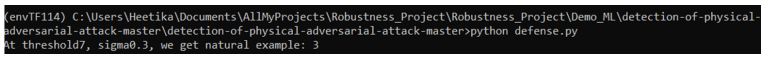

Figure 7: In a directory where 3 Natural examples are present, 3 is detected

The results of the detection of adversarial and natural images stated that the natural examples were 3. This was correctly calculated.

Person	Cos Similarity: Natural: Natural	Cos Similarity: Natural: Adversarial
1	0.76	0.69
2	0.79	0.61

Figure 8: Cos Similarity Results

We have taken the cosine similarity of two people. There was a difference of around 10 percent observed. To increase the amount of difference of cosine similarity, we tried changing the angle of our face and found that it does not make much of a differ

4 Discussion and Conclusions

We invent a novel detection technique by leveraging the inherent property of adversarial examples. In the digital context, there are already researches indicating that this can be easily circumvented by pushing the adversarial example deep inside the decision boundary. In other words, attackers could explicitly add the term to the loss function.

$$E_{\epsilon \sim \mathcal{N}(0,\sigma^2,I)}[||g(x) - g(x+\epsilon_i)||_1] \tag{2}$$

We observed that our results vary from advhat system vastly. Our cosine similarity between the adversarial image and person is quite higher compared to the paper. Although it is higher, it still recognizes the adversarial image correctly.

We also observed that the cosine similarity between real time results and static results vary vastly too.

With Sigma equal to 0.7, we also observe that the detection algorithm correctly counts the natural and adversarial images. The sigma value was varied from values ranging from 0.1 to 0.9. The Resnet model correctly identifies natural and adversarial at sigma equal to 0.7.

References

1 [1]C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. International Conference on Machine Learning (ICML), 2014.

[2] Stepan Komkov, Aleksandr Petiushko, "AdvHat: Real-world adversarial attack on ArcFace Face ID system," in IEEE Transactions on Automatic Control, arXiv:1908.08705 [cs.CV]

[3]J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition", arXiv preprint arXiv:1801.07698 (2018). [4] Fabio Valerio Massoli and Fabio Carrara and Giuseppe Amato and Fabrizio Falchi, "Detection of Face Recognition Adversarial Attacks,", arXiv:1912.02918 [cs.CV]