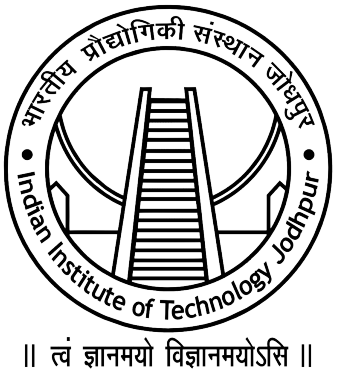
Indian Institute of Technology, Jodhpur, India

Department of Computer Science and Engineering

Advanced Bio-Metrics CSL7430

Course Project



Tejas Gaikwad

(MT19AI021)

Contents

1 Introduction

Deep learning-based systems have been shown to be vulnerable to adversarial attacks in both digital and physical domains. While feasible, digital attacks have limited applicability in attacking deployed systems, including face recognition systems, where an adversary typically has access to the input and not the transmission channel. In such a setting, physical attacks that directly provide a malicious input through the input channel pose a bigger threat. In this project, I have investigated methods to perform an attack which cannot be identified with naked eyes, which we can observe is a common trend in most of the previously published papers.

2 Literature Survey

There are a number of good articles which have attempted physical attacks using several approaches. Kurakin et al. [2] printed 2D adversarial patches containing objects overlaid by adversarial patterns to attack deep networks trained for the object recognition task. Several other researchers directly printed 2D adversarial patterns which are then manually attached to physical objects to attack object detection and classification algorithms [3, 5, 4, 6]. Similarly, Thys et al. [8] printed 2D adversarial patches to circumvent pedestrian detection classifiers. Athalye et al. [9] proposed a transformation-invariant adversarial pattern generation scheme called Expectation of Transformations (EOT) to fabricate 3D adversarial objects designed to fool object classifiers. More recently, Li et al. [10] printed adversarial dots on a 2D transparent paper to provide adversarial input via the camera to an object recognition system. Although the aforementioned methods succeed in achieving their stated objective, they usually require extensive calibration of each 2D or 3D-printed artifact before fabrication. In addition, they also require fabrication of physical artifacts. On the other hand, the camera projector setup used in the method presented here can be calibrated once based on the attack environment, and then subsequently used for conducting multiple real-time attacks targeting different enrolled users of a face recognition system. Similar to this work, Nichols and Jasper [11] used a camera-projector setup to generate 2D adversarial dot patterns that are then projected onto the physical scene to attack object recognition systems. However, they did not use the setup for conducting impersonation or obfuscation attacks on face recognition systems. Zhou et al. [12], on the other hand, fabricated a wearable cap with infrared LEDs to fool face recognition systems. Although this work is identical to the method presented here in terms of its objective, our method does not require creation of a wearable artifact and thus offers an easier alternative using off-the-shelf camera-projector setup for conducting physical attacks on facial recognition systems.

3 Methodology

The core approach used is inspired from Adversarial Generative Nets(AGN) and Generative Adversarial Networks(GANs). GAN and AGN have 2 main part during the training of the network
1) Generator
2) Discriminator
Steps/Metrics
1. Generator will try to generate 'lipstick' for the image using the latent space and will map to attackers image
2. If this successfully fools the targeted network then it returns the image otherwise it will return the loss. Loss function is given as
LossG = (z belongs to latent space)(log(1-D(G(z))
LossF(x+G(z)) = (Difference between attackers Image and Targeted Image
Fci(x+G(z)) - Fcx(x+G(z)) (For Dodging))
Fci(x+G(z)) - i!=tFcx(x+G(z)) (For Impersonating))
3. Discriminator gets updated by maximising the gain function
Gain = (x belongs to data) Maximise the score by the discriminator + (z belongs to latent space)Lower the score by Generator
The system developed is for Impersonating. The similar approach have been used to solved the given problem.

3.1 Requirements

1. Dataset Used
a. 5 Celebrity Dataset: The Model on which this dataset is trained is to be attacked by the impersonator.
b. Lips Identification: Halen Dataset with lips attribute (For ease, I have used dlib and imutils library to extract lips region from the image.)
2. Attacker Image I have used my image to attack the model.
3. Libraries used
a. Numpy
b. Matplotlib
c. Glob
d. Keras
e. tensorflow
f. PIL g. dlib
h. sklearn
i. sklearn.metrics
j. cv2
Note: all the versions used are latest updated by Nov,2021

3.2 Proposed System

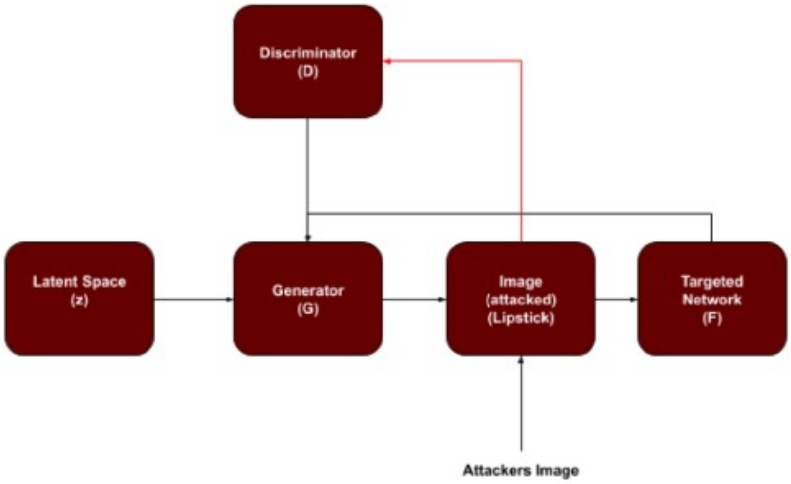


Figure 1: *Proposed System Block Diagram*

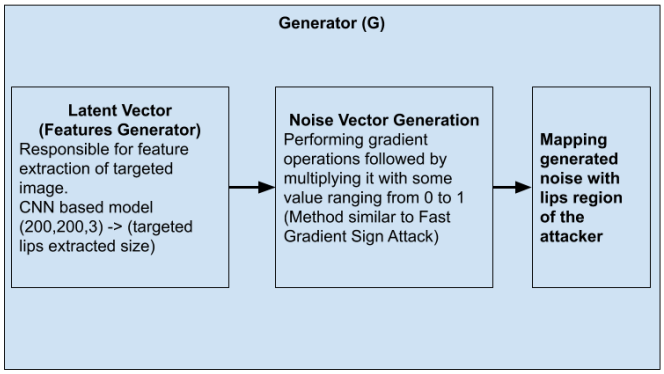


Figure 2: *Generator*

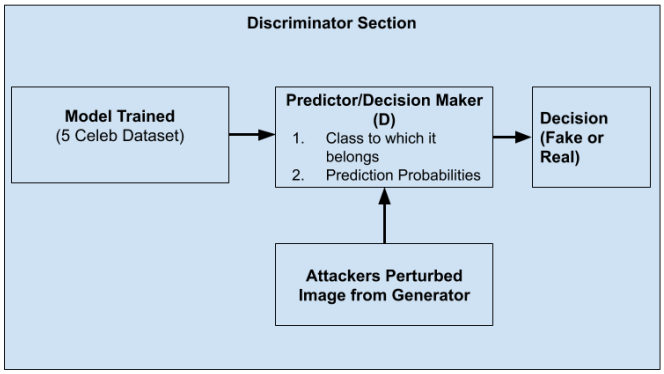


Figure 3: *Discriminator*

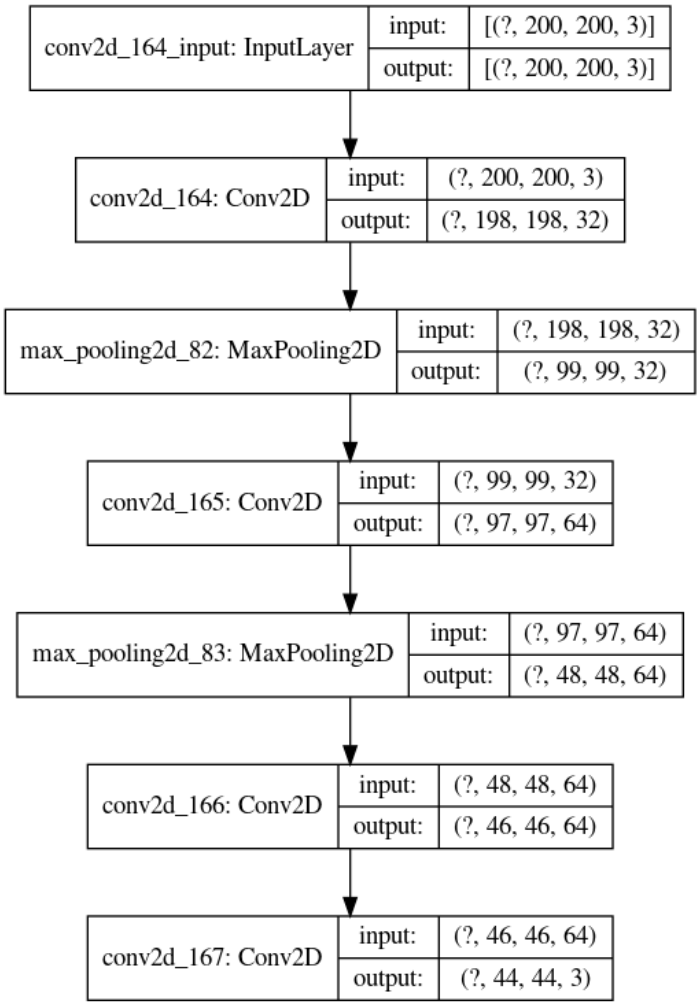


Figure 4: *Model Summary: Feature Extraction into lips size*

4 Results

4.1 Experiment 1 : Face Recognition

	precision	recall	f1-score	support
ben_afflek	0.80	0.80	0.80	5
elton_john	0.80	0.80	0.80	5
jerry_seinfeld	1.00	1.00	1.00	5
madonna	1.00	1.00	1.00	4
mindy_kaling	1.00	1.00	1.00	5
accuracy			0.92	24
macro avg	0.92	0.92	0.92	24
weighted avg	0.92	0.92	0.92	24

Figure 5: *Model Class-wise Prediction: Face Identification*

Figure 6: *Test Image Prediction: Face Identification*

4.2 Experiment 2 : Digital Attack Demonstration (FGSM Attack)

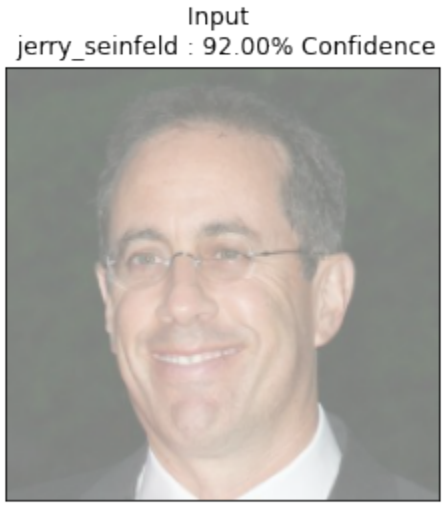


Figure 7: *Attacked Image Prediction [1]: Face Identification after adding Noise*

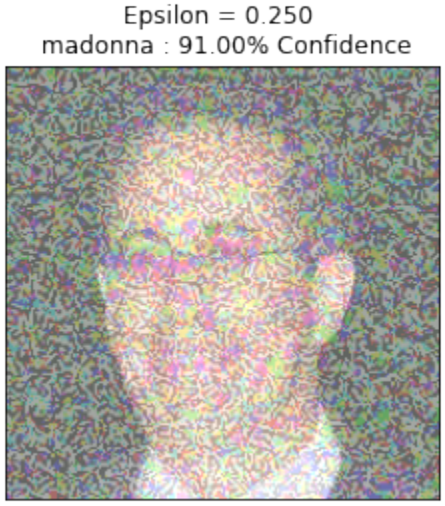


Figure 8: *Attacked Image Prediction [2]: Face Identification after adding Noise*

4.3 Experiment 3 : Lips Region Detection and Extraction of the Attacker

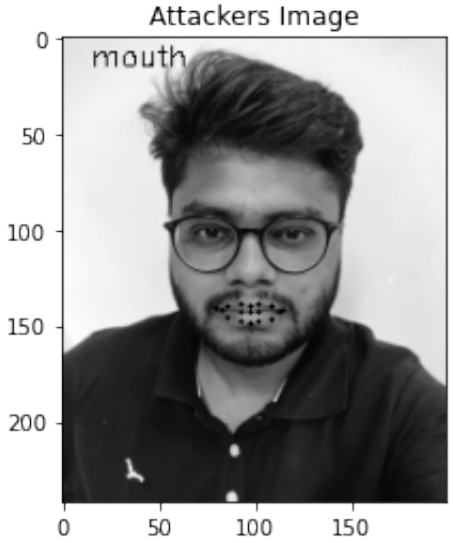


Figure 9: *Lips Region Detection on Attackers Image*

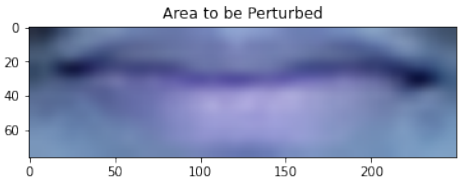


Figure 10: *Region of Index : Lips Region Extracted on Attackers Image*

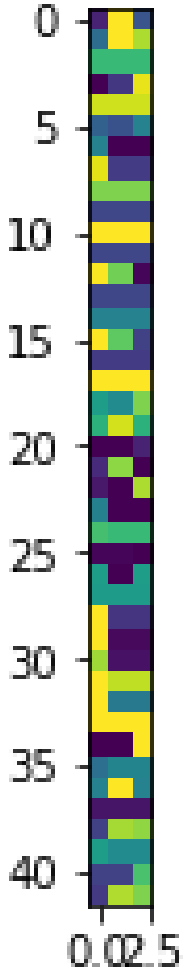


Figure 11: *Noise Generated for Lips Region*

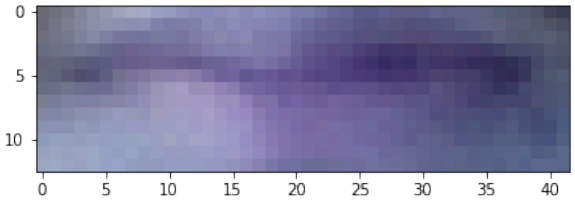


Figure 12: *Image of lips Region of the attacker with added noise*

4.4 Experiment 4 : Final Result - Attacker successfully Impersonate as other person

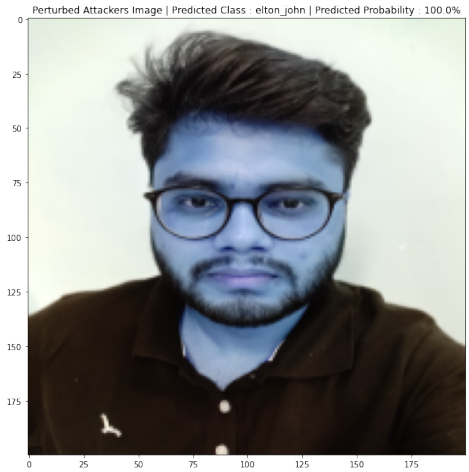


Figure 13: *Attackers Image : Successfully Impersonated as Elton John*

Figure 14: *Attacked Image : Elton John*

5 Discussions and Conclusions

In this paper I contributed a methodology that is similar to GAN's and AGN's to generate adversarial examples to fool DNN-based classifiers while meeting additional objectives. I focused on objectives imposed by the need to physically realize artifacts that, when captured in an image, result in miss-classification of the image. Future discussions could be using this method to physically attack the model. to fool face recognition as our driving example, we demonstrated the use of AGNs to improve robustness to changes in imaging conditions (lighting, angle, etc.) and even to specific defenses; inconspicuousness to human onlookers; and scalability in terms of the number of adversarial objects (eyeglasses) needed to fool DNNs in different contexts. AGNs generated adversarial examples that improved upon prior work in all of these dimensions, and did so using a general methodology. Our work highlights a number of features of AGNs. They are flexible in their ability to accommo- date a range of objectives, including ones that elude precise specification, such as inconspicuousness. In principle, given an objective that can be described through a set of examples, AGNs can be trained to emit adversarial examples that satisfy this objective. Additionally, AGNs are general in being applicable to various domains, which we demonstrated by training AGNs to fool classifiers for face and (handwritten) digit recognition. We expect that they would generalize to other applications, as well. For example, one may consider using AGNs to fool DNNs for street-sign recognition by training the generator to emit adversarial examples that resemble street-sign images collected from the internet. One advantage of AGNs over other attack methods is that they can generate multiple, diverse, adversarial examples for a given benign sample. A diverse set of adversarial examples can be useful for evaluating the robustness of models. Moreover, such adversarial examples may be used to defend against attacks .

6 References

[**1**] A General Framework for Adversarial Examples with Objectives, M. Sharif, S. Bhagavatula, L.Bauer, M. K. Reiter, arXiv:1607.02533, 2016

[**2**] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

[**3**] S. Chen, C. Cornelius, J. Martin, and D. Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In Joint European Conference on Machine Learn ing and Knowledge Discovery in Databases, pages 52–68.Springer, 2018.

[**4**] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C.Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical world attacks on deep learning visual classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition, pages 1625–1634, 2018.

[**5**] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno. Physical adversarial examples for object detectors. In 12th USENIX Workshop on Offensive Technologies (WOOT 18), 2018.

[**6**] Y. Zhao, H. Zhu, Q. Shen, R. Liang, K. Chen, and S. Zhang. Practical adversarial attack against object detector. arXiv preprint arXiv:1812.10217, 2018.

[**7**] S. Thys, W. Van Ranst, and T. Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition Workshops, pages 0–0, 2019.

[**8**] S. Thys, W. Van Ranst, and T. Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In Proceedings of the IEEE Conference on Com puter Vision and Pattern Recognition Workshops, pages 0–0, 2019.

[**9**] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017

[**10**] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016

[**11**] N. Nichols and R. Jasper. Projecting trouble: Light based adversarial attacks on deep learning classifiers. arXiv preprint arXiv:1810.10337, 2018

[**12**] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang. Invisible mask: Practical attacks on face recognition with infrared. arXiv preprint arXiv:1803.04683, 2018.