Alden Kane

Notre Dame Department of Electrical Engineering

**Proposal:** Examine adversarial attacks on facial recognition systems by making use of sub-visual patterns on images.

**Literature Review:**

**Paper:** “Can we still avoid automatic face detection?”

**Authors:** Wilber, MJ, Shmatikov, V, and Belongie, S

**Institution:** Cornell University

**Date:** 2015

**Objective:** Use a variety of techniques to evade facial recognition of Facebook. Methods explored include “synthetic image transformations using various image filters” and occlusion methods, i.e. sunglasses, scarves, hats, and so on.

**Findings:** Several filters substantially degraded photos quality. High levels of Gaussian noise proved to be highly effective in decreasing Facebook’s ability to recognize an image, while not heavily degrading a human’s ability to do so. On the highest setting for Gaussian noise, Facebook recognized just 22% of images. Drawing white lines through eyes proved to be a very effective method of preventing Facebook from recognizing images, with just a 2% success rate of face recognition. Other methods, such as warping, swirling, and putting leopard spots over the image proved to be very effective. Nevertheless, the highly degrade image quality for human detection. Ultimately, occlusions and noise are effective means of preventing recognition.

**Comments:** Good experiment design. A good number of image filters explored. Could be room to see how Facebook’s responsivity to Gaussian noise, etc., has changed. However, no advanced algorithmic perturbations were explored in this experiment.



**Paper:** “Adversarial Attacks on Face Detectors using Neural Net base Constrained Optimizaiton”

**Authors:** Bose, JA and Arabi, P

**Institution:** University of Toronto

**Date:** 31 May 2018

**Objective:** To “craft adversarial examples by solving a constrained optimization problem using an adversarial generator network.”

**Findings:** Ultimately, only 0.5% of originally detected faces were detected after this attack. This team attacked a Faster R-CNN based faced detector using “small perturbations that when added to an input face image causes the pretrained face detector to fail.” Their algorithm consisted of a generator *G* that produces a small perturbation that when added to an image *x*, produces an adversarial image *x’*. *G’s* loss function is dependent on how effectively it can trick the face detector into misclassifying *x’*. Their method of perturbations was also fairly robust to JPEG compression.

**Comments:** Good design in exploring defense against different methods of evading adversarial attacks. Robust methodology in using a neural network based approach to producing perturbations.



**Paper:** “Eluding Mass Surveillance: Adversarial Attacks on Facial Recognition Models”

**Authors:** Milich, A and Karr, M

**Institution:** Stanford University

**Date:**

**Objective:** Use random noise and random noise targeted around facial landmarks (thus obstruction of facial landmarks) to create adversarial black-box attacks against deep neural network facial recognition models.

**Findings:** The researchers began by training an existing facial recognition model, using Carnegie Mellon’s facial landmark predictor, Inception ResNet v1, and the Scikit-learn SVM classifier. For experimentation with noise, the team added RGB salt-and-pepper noise and also perturbed facial landmarks using a DNN. All in all, obscuring landmarks with targeted noise proved to be more effective than simply adding noise to images. Both noise and targeted noise decreased the accuracy at which the facial recognition classifier was able to recognize faces, and also the confidence that it did so with.

**Comments:** The researchers used different datasets to train their facial landmark recognizer, which they hypothesize decreased the efficacy of the targeted landmark approach. This was done as undergraduate research, and I believe that it is fairly high-quality.



**Paper:** “Explaining and Harnessing Adversarial Examples”

**Authors:** Goodfell, IJ, Shlens, J, and Szegedy, C

**Institution:** Google Inc.

**Date:** 20 Mar 2015

**Objective:** Argue that the primary cause of neural networks’ vulnerability to adversarial attacks is their linear nature. Also, set forth a method for generating adversarial examples.

**Findings:** They examine the fast gradient sign method (FGSM), which creates an optimal perturbation by adding the sign of the elements of the gradient of the cost function, w.r.t. the input. This provided high misclassification with little human-perceptible change to the image. Ultimately, they put forth that adversarial attacks result from linearity and high-dimensional dot products and the FGSM is a solid method of devising adversarial examples.

**Comments:** This proves to be an effective way of producing adversarial perturbations, and is relevant here because of the low distortion to the original image.



**Paper:** “Going deeper with convolutions”

**Authors:** Szegedy, C, Liu, W, and Jia, Y et. al

**Institution**: Google Inc.

**Date:**

**Objective:** Propose a new neural network architecture, named Inception, that improves computing resources.

**Findings:** With Inception, the research team used increased network depth. The team sought the optimal local sparse structure in a CNN, then repeat it spatially. The team applied dimension reductions to save computing power. The team stacks 1x1, 3x3, and 5x5 convolutions

**Comments:** Gives context to different methods of neural networks for computer vision, but not extremely pertinent for adversarial attacks. Nevertheless, the Inception network has popped up as the neural network used to train some facial detection algorithms.



**Paper:** “Intriguing properties of neural networks”

**Authors:** Szegedy, C, Wojeich, Zaremba, and Sutskever, I et. al

**Institution:** Google Inc. / New York University / University of Montreal / Facebook

**Date:** 2014

**Objective:** Examine (1) the lack of distinction between individual high level units and random linear combinations of high level units in neural networks and (2) neural networks’ tendencies to have discontinuous input-output mappings. These discontinuities allow for perturbations to misclassify inputs, regardless of network training (i.e. the same perturbation stumps the different networks).

**Findings:** I focused on the second property examined here. This project calculated a minimum distortion function *D* by minimizing the loss function. Here, the researchers varied models (hyper-parameters), and training sets to find that adversarial examples do not result from overfitting of a specific model or training set selection.

**Comments:** Their findings agree with [4] and give motivating information on why adversarial examples arise. This paper was one of the first to point out the weaknesses of deep networks to adversarial attacks.



**Paper:** “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey”

**Authors:** Akhtar, B and Mian, A

**Institution:** University of Western Australia

**Date:** 26 February 2018

**Objective:** Provide a comprehensive survey of adversarial attacks and review works to design adversarial attacks, as well as defenses against them. This is the first comprehensive survey of adversarial attacks in computer vision.

**Findings:** Despite high success, deep networks are vulnerable to adversarial attacks that present as imperceptible changes. This paper reviews different attack types, such as the Box-constrained L-BFGS, Fast Gradient Sign Method (FGSM), Basic & Least-Likely-Class Methods, Jacobian-based Saliency Map Attack (JSMA), One-Pixel Attacks, Carlini and Wagner Attacks, DeepFool, and others. They rated Carline and Wagner Attacks and Universal perturbations among the highest strength attacks. They also explore the debate around why adversarial examples arise –– Goodfellow et al. points to linearity as a reason that they appear, while others (Tanay and Griffin) point out that there are image classes that do not have adversarial examples.

**Comments:** This paper, currently, is the Holy Grail of adversarial attacks in Computer Vision. It examines leading models for generating adversarial examples, as well as leading defense mechanisms against adversarial attacks. A thorough study of this paper will put one in touch with the current progresses, and holes, in adversarial machine learning research.



**Paper:** “DeepFool: a simple and accurate method to fool deep neural networks”

**Authors:** Moosavi-Dezfooli, S, Fawzi, A, and Frossard, P

**Institution:** École Polytechnique Fédérale de Lasusanne

**Date:** 4 July 2016

**Objective:** Propose an algorithm, DeepFool, to compute adversarial perturbations that fool deep networks.

**Findings:** The team defines the minimal perturbation, *r,* that is sufficient to change the label in a classifier. It uses iterative linearization of the classifier to create false classifications.

**Comments:** They were robust in their methodology, in testing different three different data sets and eight different classifiers against adversarial attacks.



**Paper:** “The Limitations of Deep Learning in Adversarial Settings”

**Authors:**

**Institution:**

**Date:**

**Objective:**

**Findings:**

**Comments:**



**Paper:** “Houdini: Fooling Deep Structured Prediction Models”

**Authors:**

**Institution:**

**Date:**

**Objective:**

**Findings:**

**Comments:**

**Paper:** “One Pixel Attack for Fooling Deep Neural Networks”

**Authors:**

**Institution:**

**Date:**

**Objective:**

**Findings:**

**Comments:**

**Paper:** “Facial Attributes: Accuracy and Adversarial Robustness”

**Authors:**

**Institution:**

**Date:**

**Objective:**

**Findings:**

**Comments:**

**Paper:** “Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition”

**Authors:**

**Institution:**

**Date:**

**Objective:**

**Findings:**

**Comments:**

**My Experiment:**

* Will focus on adversarial attacks with respect to facial recognition. A majority of the work that I’ve encountered has been on adversarial attacks w/ object classification, so I believe that there is a hole in the research with respect to adversarial attacks and facial recognition. Not much work has been done with social media black-box models (e.g. Facebook sandbox mode), so incorporating this into my study could be worthwhile.
* I could do a comparative study: take 5-10 leading methods from the literature and test them against both a trained recognition algorithm or Facebook’s sandbox mode (possibly both). Not all methods would be extremely robust, and I could include some simpler methods (Gaussian Noise, etc.).
  + Is this robust enough to stand on its own? Extensive enough for a thesis?
* I could try to devise a new way of generating adversarial examples. My ideas here so far include:
  + Drastically scaling down pixel intensity from other faces and adding it to existing faces
  + Taking key features from other faces (noses, eyes, and mouths), scaling down the pixel intensity, and adding it around the facial landmarks of other faces.
* Could use a classifier such OpenFace, FaceNet w/ Keras. LFW data set
* Start small, break up into sub-experiments so that I can start making progress, then build on it over time 🡪 could allow for me to do a whole lot in 7 months.

**Questions:**

* Testing: Use someone else’s classifier or use Facebook Developer Mode? 🡪 I’ve only found the one paper using Facebook Developer Mode, worried that this could be a bit more time intensive, but would probably be a very robust form of testing.
* Any key works in the area? Any people to talk to?
* Any thesis writing advice? Advice for research projects? Should I look into adding some credit to my schedule for this with Dr. Stevenson?
* If I pursue a comparative study, should I start reaching out to researchers to ask for the codes that they used? Or, just ask for their feedback and start developing contacts in the area of adversarial learning?

**Timeline:**

* **09/27/19:** 
  + 12+ more papers reviewed
  + Facial recognition algorithm selected and implemented
  + Data set selected and downloaded
  + Gotten Access to BxGrid
* **10/18/19:** 
  + Filtering Algorithm for Gaussian Noise Written
  + 1 More Adversarial Attack Designed
* **April, 2020:** Defend Thesis