Adversarial Patch Attacks on VGGFace2

Matthew Li GEN22, WH22

Advisor: Dr. Konrad Kording, PIK University Professor, Neuroscience, Perelman School of Medicine

Background

- Adversarial machine learning aims to exploit existing ML algorithms by taking advantage of inner workings of the model for malicious attacks. The first adversarial machine learning attacks were “evasion attacks” (Dalvi 2004) against spam filters and have grown considerably in sophistication
- Poisoning attacks entail contamination of training data
- Model stealing involved querying black box machine learning methods to reverse engineer the parameters and architecture of the model
- While most adversarial attacks are done maliciously, there is value to creating methods to avoid AI algorithms that are used to oppress and detain
- **Objective:** Conduct an adversarial patch attack on an established facial recognition algorithm to trick the network into confidently misclassifying the input image as the same class as desired class

Data

- VGGFace2 was chosen as the target database as it is one of the standards for facial recognition with minimal label noise (Cao 2017)
- With 9131 subjects and 3.31 million images, training a model on the entire dataset would have come at a high computational cost, so a pretrained model was chosen in order to direct computational power towards patch attacks

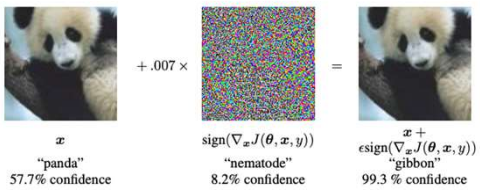
Fast Gradient Sign Method

- The fast gradient sign method is a white-box attack that assumes knowledge of the weights and biases of the target network (Goodfellow 2014)
- Computing the sign of the gradients of the loss of the input image, FGSM create a perturbated image that maximizes the loss by modifying every pixel in the input image
- Epsilon controls the intensity of the perturbations, creating an image with perturbations large enough to fool the neural network, but small enough that they are imperceptible to the human eye

$$x^{adv} = x - \varepsilon.sign(\nabla_x J(x, y_{target}))$$

where

x – Clean Input Image
x^{adv} – Adversarial Image
J – Loss Function
y_{true} – Target Label
ε – Tunable Parameter



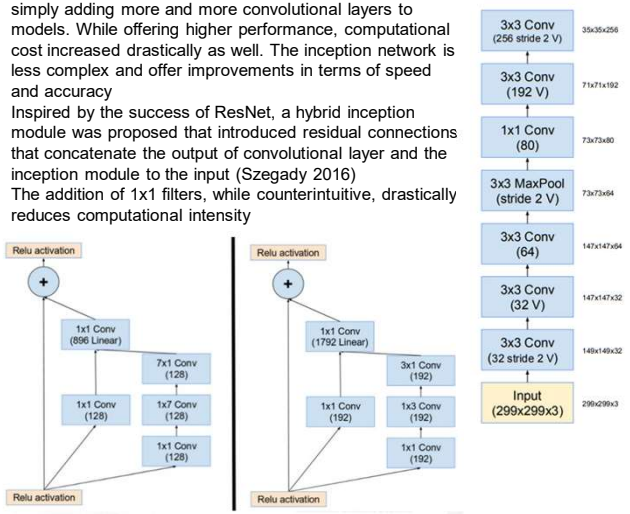
Patch Attacks

- Adversarial attacks can also be used to create image patches that are robust, targeted, and universal in the real world (Brown et. All 2017)
- A patch is created based on FGSM and inserted into the target image with random scaling, translation and rotation and optimized using gradient descent
- A variant of the Expectation over Transformation framework is used to option the patch as a solution to the unconstrained optimization problem
- As this patch is scene-independent, it allows attacks that is extremely salient to the target neural network even without knowledge of the angle or lighting conditions

$$\hat{z} = \arg\max_{z' \in \mathbb{R}^n} \mathbb{E}_{t \sim T} \left[\log \left(\mathbb{P} \left[\hat{y} \mid p_t(x, z') \right] \right) \right]$$

Inception-ResNet v1 Architecture

- Early Convolutional Neural Nets were improved upon by simply adding more and more convolutional layers to models. While offering higher performance, computational cost increased drastically as well. The inception network is less complex and offer improvements in terms of speed and accuracy
- Inspired by the success of ResNet, a hybrid inception module was proposed that introduced residual connections that concatenate the output of convolutional layer and the inception module to the input (Szegady 2016)
- The addition of 1x1 filters, while counterintuitive, drastically reduces computational intensity

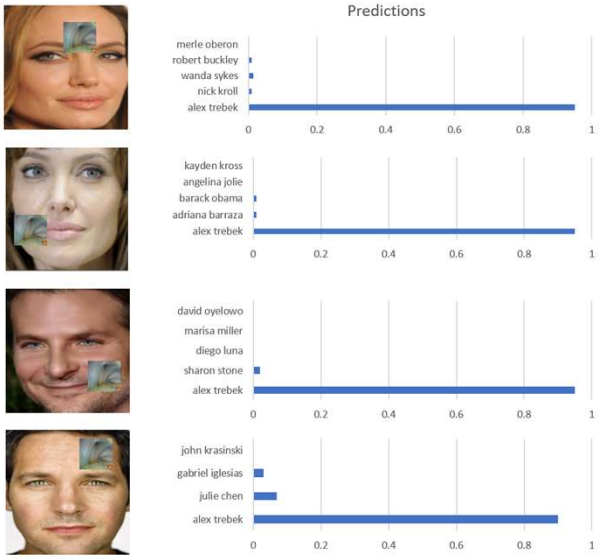


- The Top-5 and Top-1 error rate of the various inception models are shown on the ILSVRC validation sets

Network	Top-1 Error	Top-5 Error
BN-Inception [6]	25.2%	7.8%
Inception-v3 [15]	21.2%	5.6%
Inception-ResNet-v1	21.3%	5.5%
Inception-v4	20.0%	5.0%
Inception-ResNet-v2	19.9%	4.9%

Results

- An Inception-Resnet v1 model pretrained on VGGFace2 was chosen as the target model for its computational speed and accuracy in facial detection and recognition
- Running a targeted adversarial patch attack with Alex Trebek as the target class results in the following 64 x 64 patch
- Superimposing this patch to other classes within the VGGFace2 database and then classifying with Inception-Resnet v1 results in the following Top-1 and Top-5 results
- Our patch successfully fools Inception Resnet v1 96.7% of the time for Top-1 predictions and 99.2% for Top-5 predictions



Limitations and Further Research

- This approach would not generalize to the real world as the patches are not robust and limited to perfect squares. Further testing should be done to improve the robustness of the patch in different lighting conditions, as well with different size patches to limit the overall impact on the original image
- VGGFace2 is not representative of the global human population with racial composition being heavily skewed, and thus is not an equitable training set
- While the generalizability of this specific model is limited in scope, further, more robust versions could be used to protect user privacy in online databases