AI-Targeted Adversarial Attacks Research Project Overview

A Fast Gradient Sign Method (FGSM) attack is an adversarial technique where an input image is manipulated to fool an AI model to misclassify it. Researchers found that an FGSM attack that works effectively on one neural network model can successfully attack another model via transfer learning[1]. However, what isn't studied is how important the dataset (used to train the attack) is to transfer the attack onto another model.

National security implications/example use case

An adversary such as Russia may want to trick a US satellite to misclassify their submarine they deployed near the coast of Santa Monica as a fishing boat. Russia then decides to use a targeted FGSM attack- training it with datasets consisting of submarine and fishing boat images captured from Russian satellites. The question is- how similar do the images of submarines and fishing boats captured from **Russian satellites** have to be to the images of submarines and fishing boats captured from **US satellites** for Russia to successfully transfer an FGSM attack that tricks US satellite models?

What we already know/what is frequently studied:

- If Russia builds an attack against their own model, there is potential for it to be effective against US models

What we want to know:

- Does Russia have to use datasets very similar to America's to build a transfer attack against another model?

For this instance, find out:

  - Does an attack trained on Russia model work on US' model if training/Russian data is more blurry/grainy than US' data

Methodology:

Code a program that trains (via transfer learning) neural network models which classifies objects captured in satellite imagery. Find a labeled dataset of satellite imagery to train the model. Then, copy and manipulate the dataset to create multiple versions of it with varying blurs (7 versions: 1 clean and 6 blurred w/ varying intensities). Use the program to train a model for each variation of satellite imagery dataset (7 models- one for each version of dataset). Create a FGSM attack based on each model (7 attacks). For each FGSM attack, test the attack to fool all 7 models (including itself). Record its effectiveness to each model. In this next page is a table to demonstrate this process:

[1] Abdelkader, Ahmed, et al. "Headless Horseman: Adversarial Attacks on Transfer Learning Models." *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, https://doi.org/10.1109/icassp40776.2020.9053181.

Effectiveness of FGSM Attacks on Models Trained on Various Datasets

Datasets used to train the model being attacked

Datasets used to train the attack

Dataset	Clean	Blur 1	Blur 2	Blur 3	Blur 4	Blur 5	Blur 6
Clean	100%	?	?	?	?	?	?
Blur 1	?	100%	?	?	?	?	?
Blur 2	?	?	100%	?	?	?	?
Blur 3	?	?	?	100%	?	?	?
Blur 4	?	?	?	?	100%	?	?
Blur 5	?	?	?	?	?	100%	?
Blur 6	?	?	?	?	?	?	100%

Values inside the box represent the effectiveness of an FGSM attack based on one dataset against a model based on another dataset. The 100% values exist since an FGSM attack based on a dataset should theoretically be 100% effective against a model trained on the same dataset (although this is being tested). The goal is to figure out the effectiveness of the other attacks.

Completed project steps:

1. Create a program that trains a satellite imagery classifier AI model
2. Find an appropriate and labeled satellite imagery dataset of military and civilian vehicles
3. Copy and manipulate satellite imagery dataset to make varied versions of datasets (grainy, blur, white box, noise etc.)
    1. Create 7 versions (7 steps of intensity) for each type of manipulation
4. Train neural network models (via transfer learning on pre-trained models) for each of the varied datasets
5. Create an FGSM attack for each model

Steps currently in process:

6. Test each FGSM attack to evaluate how effectively it tricks each model into misclassifying an image
7. Record results and write a report intended for government audiences