Attacks On CNN and Attention CNN - Brief Report

IIT JODHPUR

Problem Statement

The aim is to implement the FGSM, PGD and BIM attacks on the CNN model and the Attention CNN model we're using in our project.

README

The code is written in colab and thus can simple be run again by uploading the python files and the VisDrone dataset in colab and run normally everything will be automatically installed in colab and user only need to run all the cells to regenerate the result otherwise only to check the work if colab is uploaded results will be shown of last epochs they were trained on.

Attacks

3 attacks have been implemented. Here's an explanation of each attack:

FGSM (Fast Gradient Sign Method) Attack:

The FGSM attack is a simple and fast adversarial attack method. It perturbs the input data by adding a small perturbation in the direction of the gradient of the loss function with respect to the input. The attack is performed using a single step of perturbation, where the magnitude of the perturbation is determined by a hyperparameter called epsilon. The FGSM attack aims to maximize the loss to induce misclassification or misbehavior of the target model. However, FGSM might not be effective against models with strong defenses.

PGD (Projected Gradient Descent) Attack:

The PGD attack is an iterative version of the FGSM attack. It performs multiple iterations of perturbation, taking small steps in the direction that maximizes the loss. After each perturbation step, the perturbed data is projected back onto an epsilon-constraint region to ensure it stays within a permissible range. The PGD attack aims to find the maximum perturbation that can fool the model while satisfying the constraints. It is a stronger attack than FGSM and can overcome some defense mechanisms.

BIM (Basic Iterative Method) Attack:

The BIM attack is similar to the PGD attack but differs in the way perturbations are applied. It also performs multiple iterations of perturbation, but instead of taking a single step, it applies small perturbations along the gradient direction for each iteration. The step size for perturbation is determined by a hyperparameter called alpha. The BIM attack gradually accumulates the effect of perturbations over multiple iterations, allowing it to find more potent adversarial

examples. It is also more computationally expensive than FGSM but can be more effective against defenses.

Results

Models:

CNN - Baseline CNN Model
Attention CNN - CNN with Attention mechanism

Datasets:

CiFAR10 - CIFAR-10 dataset consists of 60,000 color images in 10 classes. Each image in CIFAR-10 is a low-resolution (32x32 pixels) RGB image representing one of the following classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The dataset is commonly used for image classification tasks and benchmarking deep learning models.

SVHN - The SVHN dataset is a real-world image dataset of house numbers captured from Google Street View. It contains over 600,000 digit images for training and 26,032 digit images for testing.

MNIST - MNIST is a widely known dataset that has been a standard benchmark for image classification tasks. It contains a collection of 60,000 handwritten digit images for training and 10,000 images for testing. The images in MNIST are grayscale and have a size of 28x28 pixels.

FashionMNIST - FashionMNIST is a dataset designed as a drop-in replacement for the traditional MNIST dataset. It consists of 70,000 grayscale images of 10 different fashion categories, including T-shirts, trousers, pullovers, dresses, coats, sandals, shirts, sneakers, bags, and ankle boots. Each image in FashionMNIST has a size of 28x28 pixels.

VisDrone19 - The VisDrone2019 dataset, which was put together by the AISKYEYE team at the Lab of Machine Learning and Data Mining, was used. This dataset includes: 6471 training samples, 548 validation samples, 1610 test samples. There are many things about the drone dataset. In the datasets we already have, objects are taken by people or by CCTV. A circle around a car is a path that can be taken based on the car. So, you can mostly see the front, back, and sides of the object, and you can also see a little bit of the top. Because the picture was taken up close, it is big and clear. On the other hand, drones can take pictures in a variety of car-shaped, semi-spherical shapes. It would have more information than traditional datasets, especially from a bird's-eye view and a plane. So, the object looks the same from the front, back, side, and top. But it looks different when taken at a 90-degree angle. Because they can only see the top, people look like dots, and street lamps look like straight lines. In this case, it's hard to put things into groups, which makes the classification task difficult.

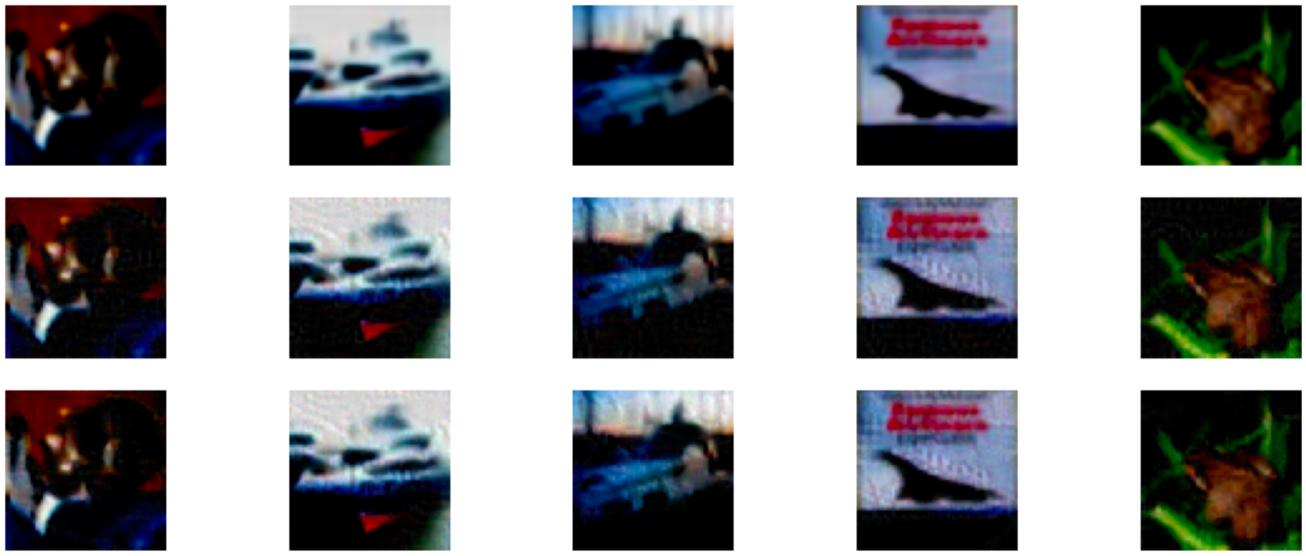
Table of test accuracies with and without attacks on the various datasets:

Dataset	Type of Model	Without Attack	FGSM Attack	BIM Attack	PGD Attack
CiFAR10	CNN	82.68%	14.88%	0.00%	0.06%
MNIST	CNN	99.04%	82.3%	6.76%	23.93%
Fashion MNIST	CNN	92.39%	15.03%	0.59%	0.62%
SVHN	CNN	94.0688384%	32.1834665%	0.2112784%	7.5599262%
VisDrone19	CNN	99.4409938%	99.4409938%	99.4409938%	99.4409938%
VisDrone19	Attention CNN	99.4409938%	99.4409938%	99.4409938%	99.4409938%

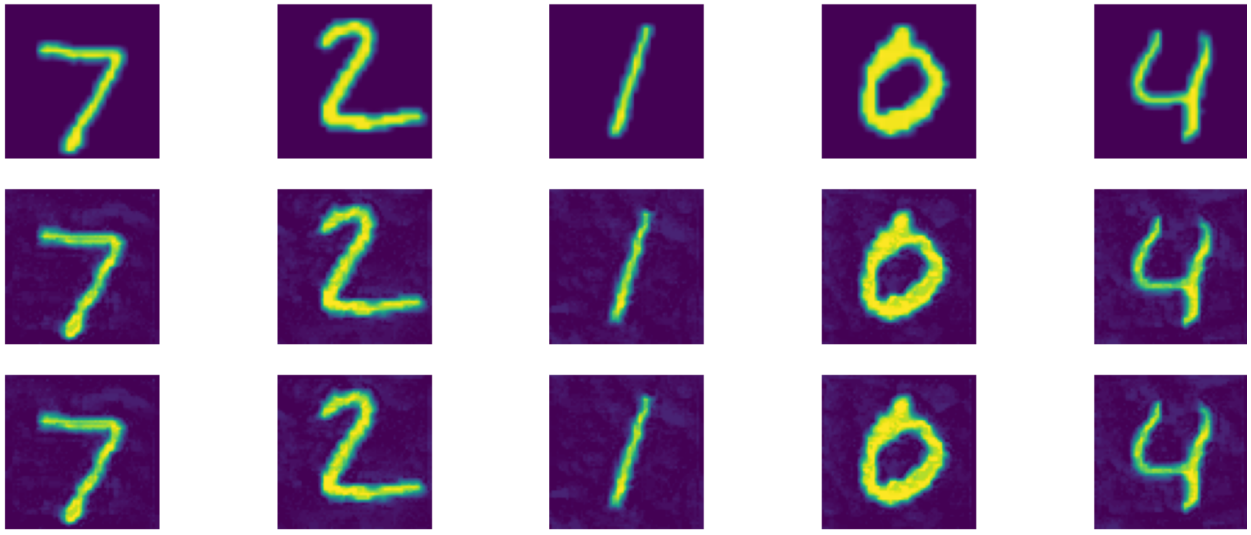
We can clearly see that the Attacks have no affect on the VisDrone19 dataset. VisDrone19 dataset may have been specifically designed to evaluate the robustness of models against adversarial attacks. It's possible that the models trained on this dataset have incorporated robust defense mechanisms that make them more resilient to attacks like FGSM. In such cases, traditional attack methods like FGSM may not be effective.

Image plots of each dataset's images before and after the attacks: (1st row - No Attack, 2nd row - BIM Attack, 3rd row - PGD Attack):

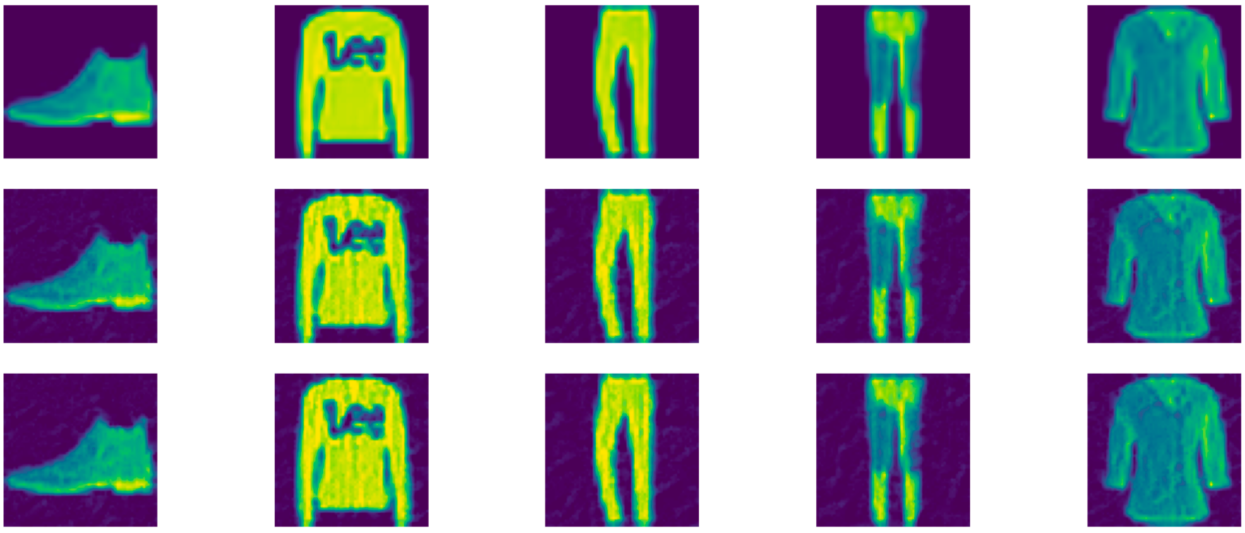
1. CiFAR10 Dataset:



2. MNIST Dataset:



3. FashionMNIST Dataset:



4. SVHN Dataset:

5. VisDrone19 Dataset: