Enhancing Image Classifiers with Denoising Filters

AIDAN SHARPE
Advisor: Dr. Robi Polikar
Rowan University,
Henry M. Rowan College of Engineering,
Glassboro, NJ 08028

Problem Statement

Neural networks are vulnerable to adversarial attacks [1], [2]. This project investigates the efficacy of various image processing techniques at improving the robustness of image classifier models.

Requirements

1) Show that image preprocessing filters are an effective defense against adversarial attacks
2) Compare the efficacy of different filters at different strengths
3) Show the transferability of image preprocessing across different datasets and classifier architectures

Constraints

- Limited computing resources
  - Restricted the resolution of datasets used
  - Limited model complexity (parameters, epochs, etc.)
- Maximum file size of 100 MB
  - Models with too may parameters would be untrackable by git

Engineering Standards

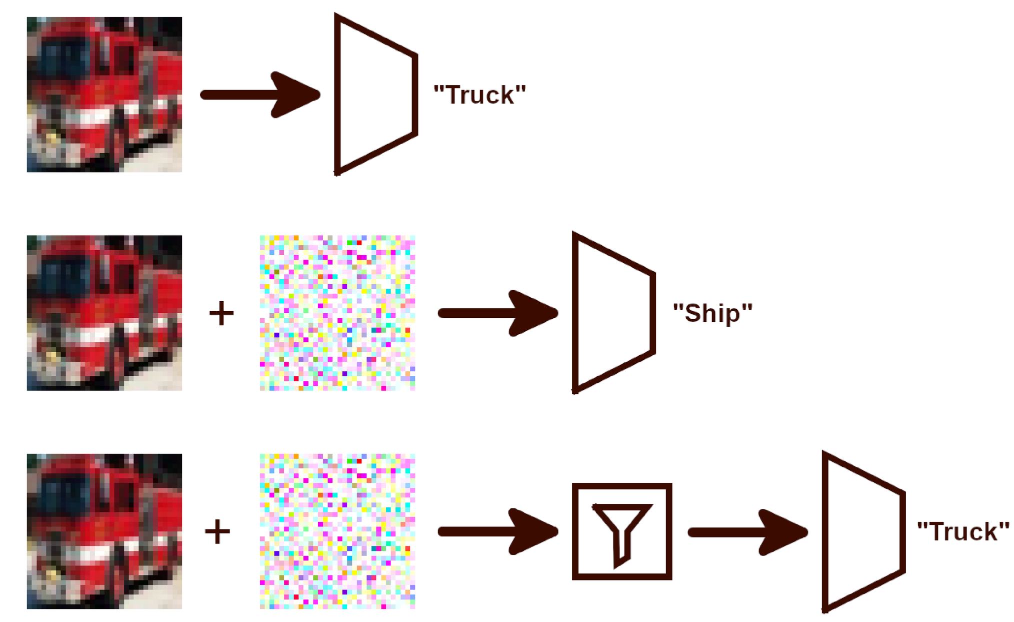
- ECMA 404 [3]
- IEEE 3129-2023 [4]

References

Source Code

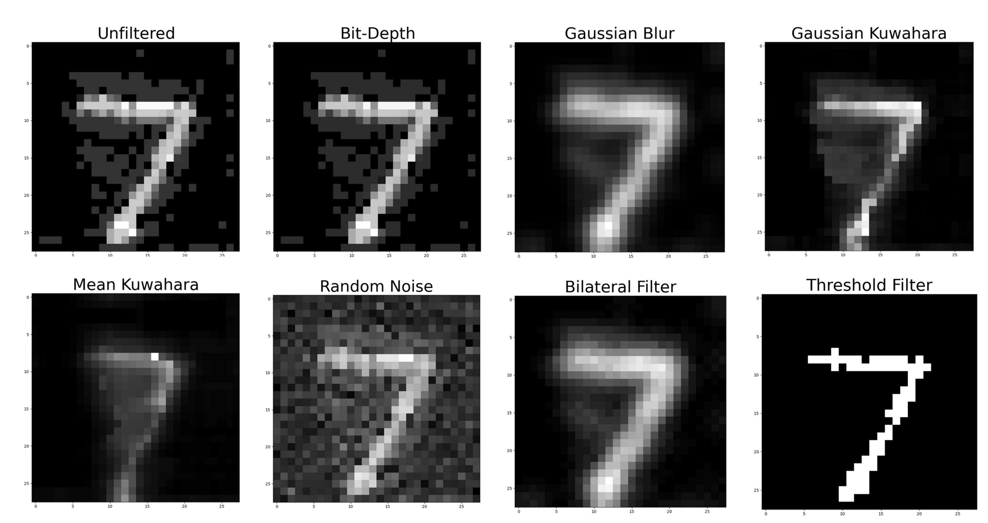
Experimental Approach

1) Implement the FGSM attack [5]
2) Test FGSM attack on pre-trained MNIST classifier
3) Implement Gaussian Kuwahara filter between attack generation and model input stages



Block overview of adversarial attacks and filtering pipeline

4) Create a standard "plug and play" interface to enable drop-in filters, model, and attacks
5) Evaluate each filter on different attack strengths and varying the filter's free parameter
   a) This free parameter is referred to generically as "strength", although some filters have a greater impact on images at lower "strength"
6) Enable saving output data in JSON format [3]
7) Use the previously designed standard interface to test all filter alternatives on MNIST classifier



Effect of filtering a sample from MNIST attacked with FGSM at ε=0.2

8) Train CIFAR-10 classifier
   a) Initial CNN could only achieve ~65%-70% accuracy on validation dataset
   b) DLA trained on CIFAR-10 was more promising [6]
   c) VGG16 trained on CIFAR-10 for 40 epochs scored over 80% accuracy on validation dataset [7]
9) Use the previously designed standard interface to test all filter alternatives on VGG16 classifier trained on CIFAR-10
10) Create a program that reads the saved JSON data and generates custom views of the results to compare the efficacy of filters across datasets

```python
model = Net()
accuracies = {}

for filter in filters:
    for epsilon in epsilons:
        for strength in range(5):
            correct = 0
            total = 0
            for data, target in dataset:
                atk_data = fgsm_attack(data, epsilon)
                filt_data = filtered(atk_data, filter, strength)
                prediction = model(filt_data)

                total += 1
                if prediction == target:
                    correct += 1

            accuracies[filter][epsilon][strength] = correct/total

save_json("results.json", accuracies)
```

Alternative Filters

- Gaussian Blur
- Gaussian Kuwahara Filter
- Mean Kuwahara Filter
- Bilateral Filter
- Random Noise
- Threshold Filter
- Bit-Depth Reduction

Alternative Datasets

MNIST – High contrast, greyscale, 28x28
CIFAR-10 – Med. contrast, RGB, 32x32

MNIST Excerpt   CIFAR-10 Excerpt

Alternative Attacks

- Fast Gradient Sign Method (FGSM) [5]
- Carlini and Wagner (Planned) [2]

Health & Safety Considerations

- Self-driving systems must respond rapidly and accurately to ensure passenger safety
- A lightweight filtering approach was chosen over an ML-based defense to reduce the time between perception and classification

Social Considerations

- All software & data is free and open source (FOSS)
- Ensures full and equal access to all who wish to recreate the results or defend their own models

Environmental Considerations

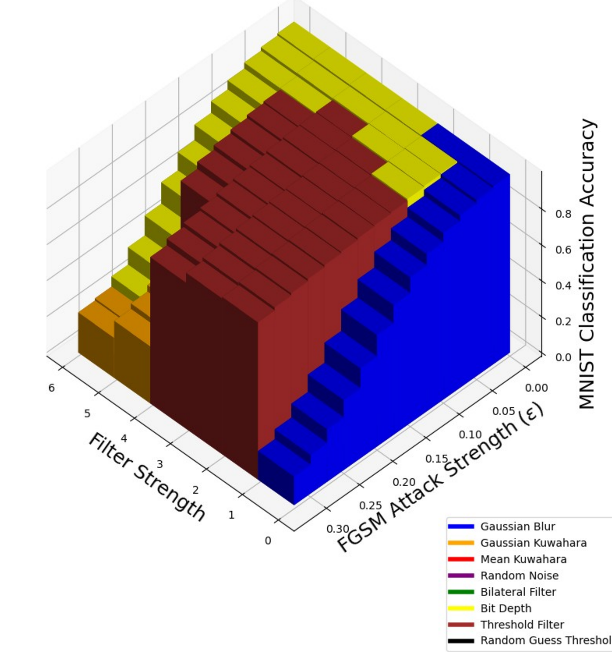
- Using image processing eliminates the computationally expensive training process found in ML-based defenses
- While untested, denoising filters may also be more energy-efficient than ML-based defenses during use

Economic Considerations

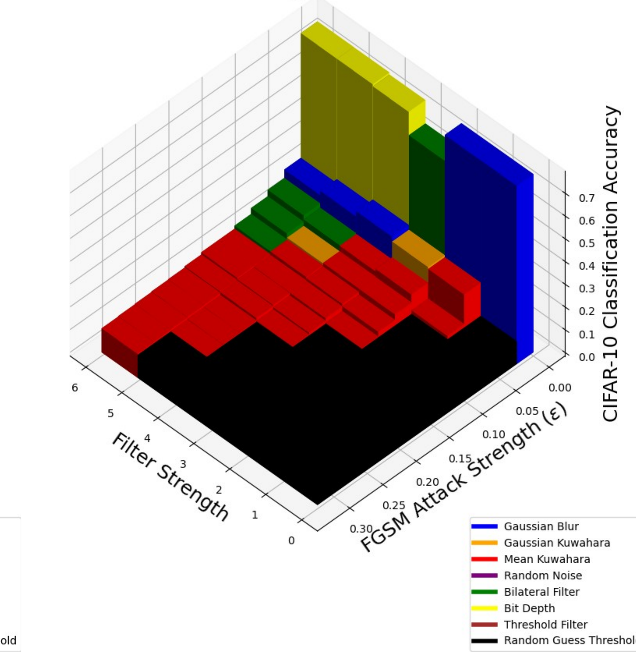
- Costs are minimized by prioritizing lightweight, power-saving algorithms
- Less computationally intense filters with similar results should rank higher

Experimental Results

Filter Efficacy for MNIST



Filter Efficacy for CIFAR-10



Evaluation Criteria

- The **accuracy** of a classifier model is given by:

$$\text{Accuracy} = \frac{\text{Correct Classifications}}{\text{Total Classifications}}$$

- The **random guessing threshold** is the expected accuracy if a class was guessed at random
- A **weak learning algorithm** produces a prediction rule that performs just slightly better than random guessing [8]
- A filter is deemed **ideally effective** if it prevents the accuracy of the classifier from changing with increasing attack strength
- A filter is deemed **minimally effective** if it keeps accuracy above the random guessing threshold
  - Being at least minimally effective means that a boosting technique can be used

Conclusions

- MNIST classifier does better than random guessing even without a defense
- MNIST filtering maintains accuracy at higher ε
- CIFAR-10 is more strongly affected by FGSM
- The threshold filter on MNIST is almost ideally effective
- The most effective filters CIFAR-10 are only minimally effective

Future Work

- Implement and test Carlini and Wagner attack [2]
- Implement and test ImageNet dataset
- Implement more filters
  - Median blur
  - JPEG compression
  - Anisotropic diffusion
- Test the power consumption of an image processing defense against an ML-based defense
- Standardize the meaning of strength
  - SNR-based definition [9]
  - Lp norm-based definition

Rowan University