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], [7]. This project investigates the efficacy of various image processing techniques at improving the robustness of image classifier models.

Requirements

1) Examine whether image preprocessing filters are effective at defending adversarial attacks
2) Compare the efficacy of different filters at different attack and filtering strengths
3) Investigate the transferability of image preprocessing defenses 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 many parameters would be untrackable by git

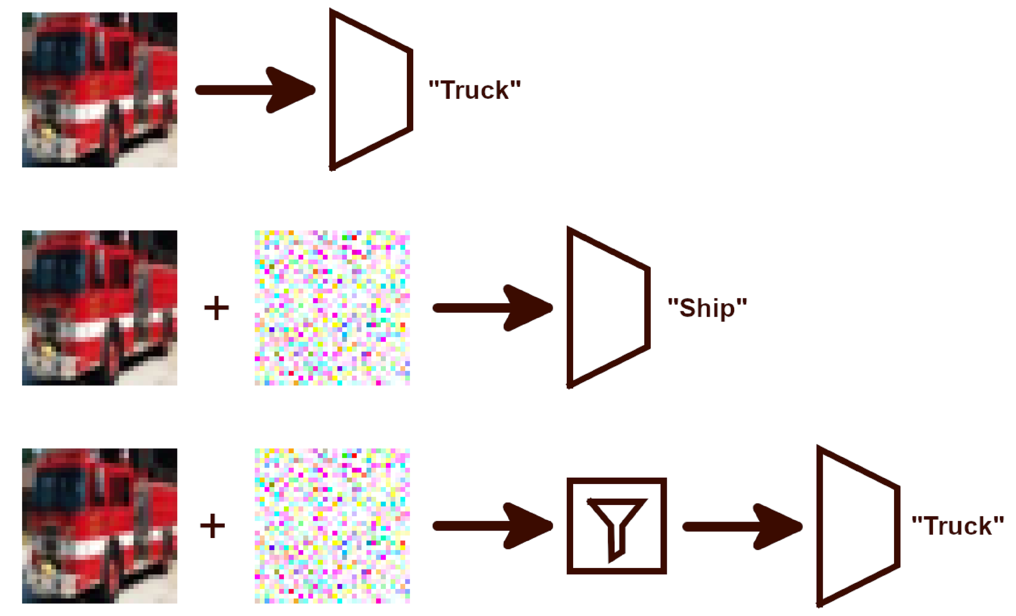
Engineering Standards

- ECMA 404 [2]
  - The JSON data interchange syntax
- IEEE 3129-2023 [4]
  - IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service

References

Source Code

Experimental Approach



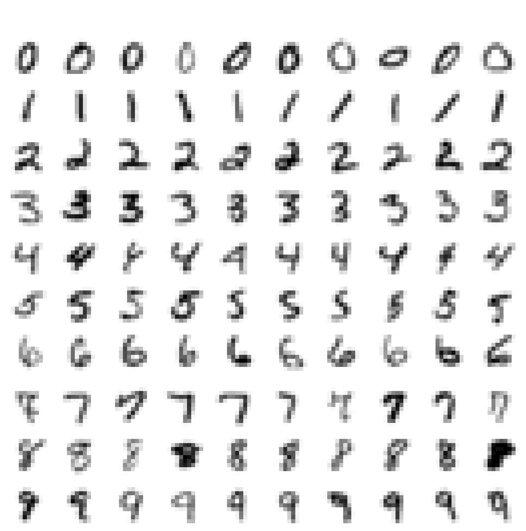
Block overview of adversarial attacks and filtering pipeline

1) Implement the FGSM attack [3]
2) Test FGSM attack on pre-trained MNIST classifier
3) Implement Gaussian Kuwahara filter as a defense
4) Create a standard "plug and play" interface to enable drop-in filters, model, and attacks
5) Evaluate each filter on different attack strengths with different values of 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 results in JSON format [2]
7) 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 [8]
   c) VGG16 trained on CIFAR-10 for 40 epochs scored over 80% accuracy on validation dataset [6]
8) Use the standard interface to test all filter alternatives on both MNIST and CIFAR-10

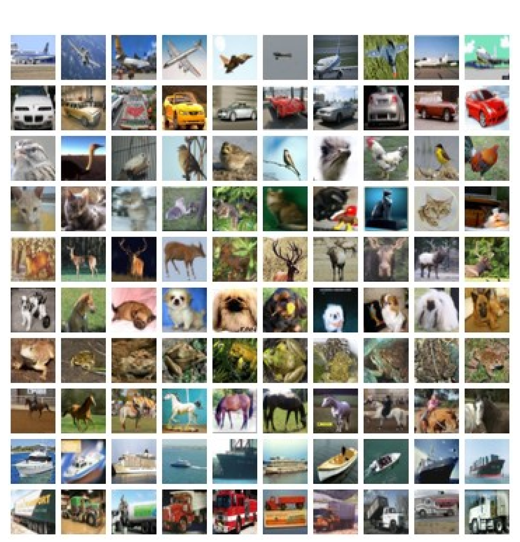
```
model = Net()
accuracies = {}

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)
```

Tested Datasets

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



MNIST Excerpt



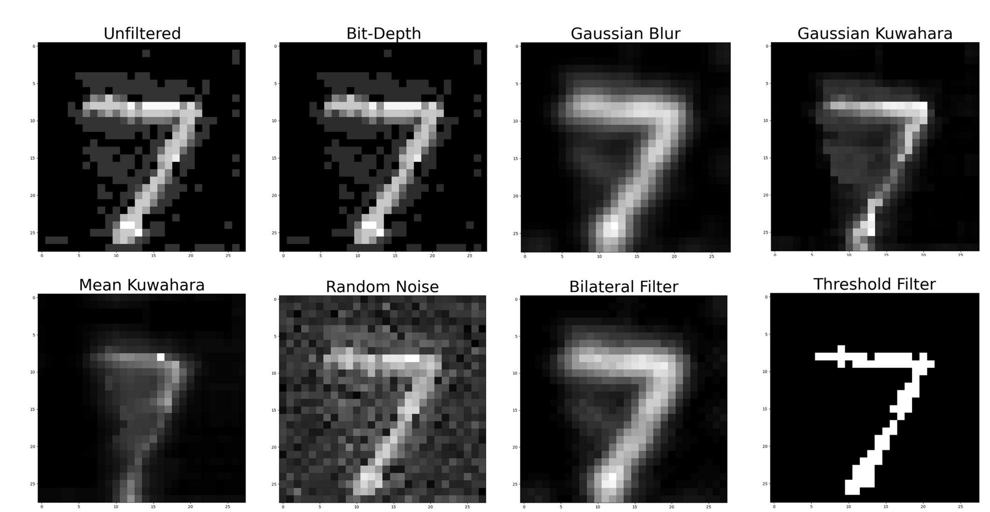
CIFAR-10 Excerpt

Tested Attacks

- Fast Gradient Sign Method (FGSM) [3]
- Carlini and Wagner (Planned) [1]

Alternative Filters

- Gaussian Blur
  - Blurs edges and smooth areas
  - Removes high frequency information (lowpass)
- Gaussian Kuwahara Filter
  - Blurs smooth area, but preserves edges
  - Has an oil painting-like effect
- Mean Kuwahara Filter
  - Similar effect as Gaussian Kuwahara
  - Slightly different way of calculating pixel values
- Bilateral Filter
  - Edge-preserving smoothing filter
- Random Noise
  - May outweigh effects of adversarial noise
- Threshold Filter
  - Removes all low-amplitude information
- Bit-Depth Reduction
  - Acts like multiple thresholds to multiple values



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

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

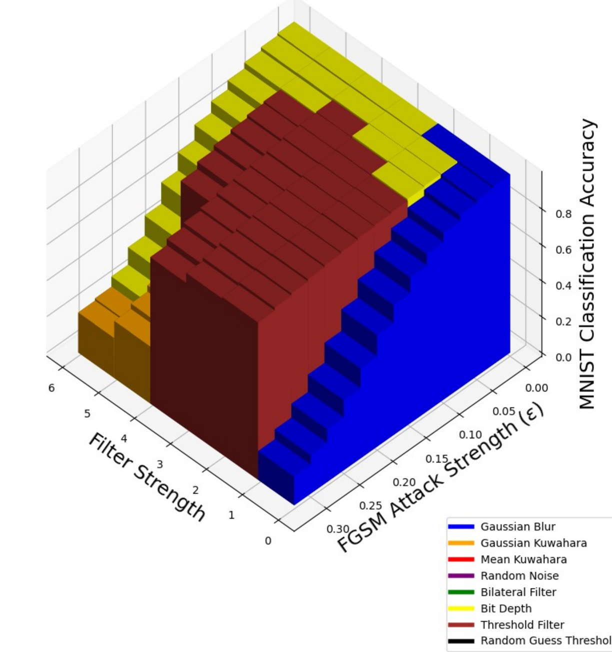
Environmental Considerations

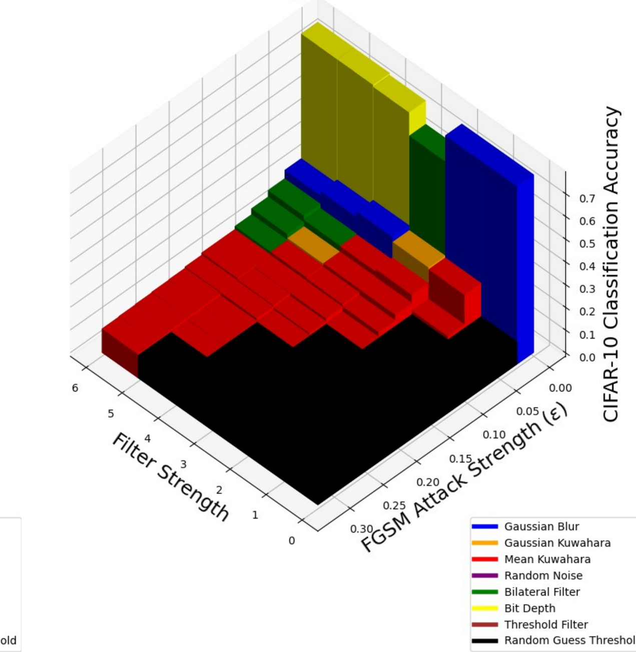
- Using image processing eliminates the computationally expenive training process found in ML-based defenses
- While untested, denoising filters may also be more energy-efficient than ML-based defenses during use

Economic Considerations

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

Experimental Results





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 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 [5]

Conclusions

- MNIST classifier does better than random guessing even without a defense (strength=0 case)
- CIFAR-10 is more strongly affected by FGSM (strength=0 case)
- MNIST filtering maintains accuracy at higher ε
- The threshold filter on MNIST is almost ideally effective for strengths 1, 2, and 3
- The most effective filters on CIFAR-10 are at best minimally effective regardless of strength

Future Work

- Implement and test Carlini and Wagner attack [1]
- 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
  - Lp norm-based definition

Rowan University