Objective

The state-of-art machine learning models, especially Neural Networks, are vulnerable to adversarial attacks: a small perturbation can make an image completely misclassified. In this project, we experimented on different inexpensive image processing defenses against different adversarial attacks. We found that different attacks generate different patterns of perturbation. The hyperparameters of image processing affect the defense robustness: aggressive processing makes models very astute to small perturbation, but perform poorly when perturbation is larger; conservative processing makes more robust to all perturbations.

Data and Model

Data: We used images with pixel 64x64 from ImageNet. Model: pretrained Inception V3 model to train adversarial attacks.

Attack

FGSM

\[ x' = x + \epsilon \cdot \text{sign} \left( \nabla_x \ell(x, h(x)) \right), \]

I-FGSM

\[ x^{(m)} = x^{(m-1)} + \epsilon \cdot \text{sign} \left( \nabla_{x^{(m-1)}} \ell(x^{(m-1)}, h(x)) \right), \]

Deepfool

\[ r_x(z_o) = \frac{f_i(z_o) - f_k(z_o)}{\|w_i(z_o) - w_k(z_o)\|_2} (w_i(z_o) - w_k(z_o)). \]

Image Processing Defense Methods

1. Depth-color-squeezing

Reduce number of bit for each pixel’s color.

\[ x(i, j) = x(i, j) \mod 2^x \]

where \( x \) is number of bit to reduce.

Hyperparameter: We tried on bit-depth 2, 4, and 6 bits.

2. Image compression with K-means

Substitute each pixel with nearest centroid value.

\[ x(i, j) = \text{centroid}_k \]

where \( k = \arg\min_k \|x(i, j) - \text{centroid}_k\| \)

Hyperparameter: \#centroid

3. Spatial smoothing

Substitute each pixel to the median in its sliding window.

\[ x(i, j) = \text{median}(\text{sliding window from } x(i-m, j-n)\rightarrow x(i+m, j+n)) \]

Hyperparameter: sliding window size \( m, n \)

4. Total variance minimization

Minimizes sum of local variance on adjacent pixels \( TV_p(z) \), plus to L1 regularization. Hyperparameter: \( \lambda_{TV}, \rho_{bernoulli} \)

\[ TV_p(z) = \sum_k \left[ \sum_j |w_j(z) - \text{centroid}| + \lambda_{TV} \cdot TV_p(z) \right] \]

Result

From the result graphs, we have the following conclusions:

- Deepfool poses stronger attack than FGSM wrt attack at same L2 norm.
- Depth-color-squeezing: The more compressed image, the more robust to attack.
- For spatial smoothing, we found using 3 pixels as the sliding window is better than 5 because it preserves more local features. Using square(3*3) is better using stride because we remove more attack features which are more likely be in square shapes.
- Total variance minimization outperforms K-means, due to it minimizes local variance, while K-means minimize global variance.

Processed Images

1. Processed image after we added noise

2. The left shows final image after we apply defense methods. The right shows its difference from the adversarial attack image.

Discussion

- Deepfool poses stronger attack than FGSM wrt attack at same L2 norm.
- Depth-color-squeezing: The more compressed image, the more robust to attack.
- For spatial smoothing, we found using 3 pixels as the sliding window is better than 5 because it preserves more local features. Using square(3*3) is better using stride because we remove more attack features which are more likely be in square shapes.
- Total variance minimization outperforms K-means, due to it minimizes local variance, while K-means minimize global variance.