Wavelet-based Low Frequency Adversarial Attacks

S. Sadiku, M. Wagner, S. Pokutta

ZUSE INSTITUTE BERLIN

Berlin Mathematics Research Center MATH+

Motivation

- Deep Neural Networks are vulnerable to adversarial attacks.
- Adversarial training augments the training data with adversarial examples and has shown effective against small ℓ_p norm attacks in the pixel domain.
- Image compression techniques have long utilized the fact that low frequency signal consists of the most crucial content-defining information in natural images, whereas high frequency spectrum often represents the noise. Such methods smooth the data and rely on removing high-frequency signal.
- The goal is to directly target the class-defining information by designing white-box attacks generated in the low frequency domain given by the DWT basis while preserving the high frequency coefficients of an image $\mathbf{x}$. Generated perturbations in the new basis are still imperceptible but do circumvent both training-based and basis-manipulation defense methods.

2D Discrete Wavelet Transform (DWT)

Among many possibilities to represent image data, a popular representation is the two-dimensional Discrete Wavelet Transform (DWT) basis (2), which captures both frequency and location information, unlike, for example, the Fourier Transform.
Let $\mathscr{R}$ denote a 2D DWT map. Given an image $\mathbf{x} \in [0,1]^{n\times c}$, its 2D DWT coefficients are given by

$$\mathscr{R}(\mathbf{x}) = \left[\begin{array}{c|c} \mathbf{x}_{LL} & \mathbf{x}_{LH} \\ \hline \mathbf{x}_{HL} & \mathbf{x}_{HH} \end{array}\right] \in \mathbb{R}^{n\times c}.$$

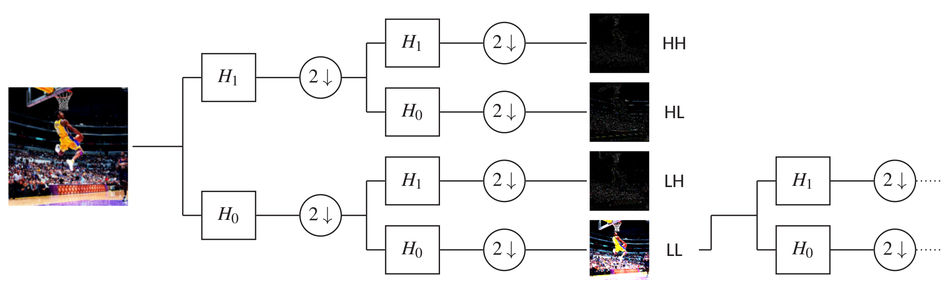

Figure 1: The DWT decomposition tree of scale 2 for a basketball image from ImageNet dataset.

Low Frequency Adversarial Attacks

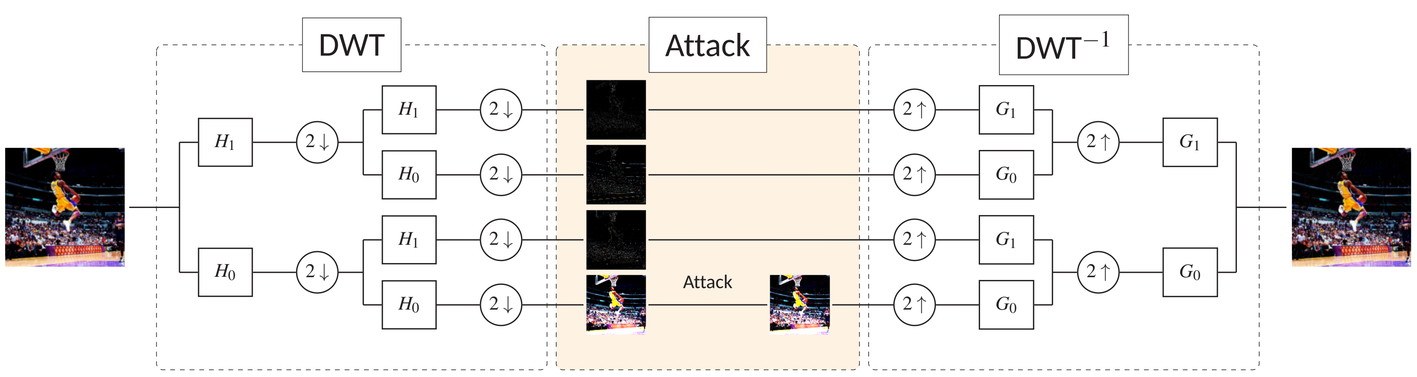

Figure 2: The low frequency FGSM attack with DWT of scale 1 for a basketball image from ImageNet.

Wavelet-based Adversarial Attacks

The ultimate goal of an adversary is to succeed under minimal distortion (1). The adversarial attack problem in the representation space whose corresponding map is given by $\mathscr{R}$, the 2D DWT basis of Daubechies mother wavelet, aims to solve

$$\mathbf{r}' = \underset{\|\mathbf{r}\|_\infty \leq \varepsilon}{\arg\max}\, L(\theta, \mathscr{R}^{-1}(\mathscr{R}(\mathbf{x}) + \mathbf{r}), t).$$

We design practical low frequency adversarial attacks in the wavelet domain from three popular white-box attacks, namely FGSM, I-FGSM, and C&W ℓ_2 (3).

Low Frequency FGSM, I-FGSM, C&W ℓ_2

- **Low Frequency FGSM**

$$\mathbf{r}' = \varepsilon \text{ sign}\left(\left[\begin{array}{c|c} \left[\mathscr{R}\left(\frac{\partial L(\theta,\mathbf{x},t)}{\partial \mathbf{x}}\right)\right]_{LL} & 0 \\ \hline 0 & 0 \end{array}\right]\right)$$

- **Low Frequency I-FGSM**

$$\mathbf{r}^{(n)} = \varepsilon \text{ sign}\left(\left[\begin{array}{c|c} \left[\mathscr{R}\left(\frac{\partial L(\theta,\hat{\mathbf{x}}^{(n-1)},t)}{\partial \hat{\mathbf{x}}^{(n-1)}}\right)\right]_{LL} & 0 \\ \hline 0 & 0 \end{array}\right]\right)$$

- **Low Frequency C&W ℓ_2** - Define $\tilde{\mathbf{x}} = \mathscr{R}(\tanh^{-1}(2\mathbf{x}-1))$ and $\hat{\mathbf{w}} = \left[\begin{array}{c|c} \mathbf{w} & \tilde{\mathbf{x}}_{LH} \\ \hline \tilde{\mathbf{x}}_{HL} & \tilde{\mathbf{x}}_{HH} \end{array}\right]$. Choose

$$\mathbf{r} = \mathscr{R}\left(\frac{1}{2}\left(\tanh\left(\mathscr{R}^{-1}\left(\hat{\mathbf{w}}\right)\right)+1\right)\right) - \mathscr{R}(\mathbf{x}).$$

The new objective function is given by

$$\begin{aligned}\min \;\; \Big\{ &\|\mathscr{R}(\frac{1}{2}(\tanh(\mathscr{R}^{-1}(\hat{\mathbf{w}}))+1)) - \mathscr{R}(\mathbf{x})\|_2^2 \\ &+ c\cdot f(\frac{1}{2}(\tanh(R^{-1}(\hat{\mathbf{w}}))+1))\Big\},\end{aligned}$$

which we optimize over $\mathbf{w}$ and set $\hat{\mathbf{x}} = \frac{1}{2}(\tanh(\mathscr{R}^{-1}(\hat{\mathbf{w}}))+1)$.

Defenses against Adversarial Attacks

We generate perturbations using FGSM, I-FGSM, and CW ℓ_2 in the pixel basis and in the low frequency DWT basis. Next, we apply traditional defense methods such as adversarial training (4) and image processing methods, such as JPEG compression, PCA/wavelet denoising, and soft-thresholding (5) to the adversarial examples, feed them back to the model, and measure the top 1 accuracy against the normalized ℓ_2 similarity between the adversarial and the original images.

Results

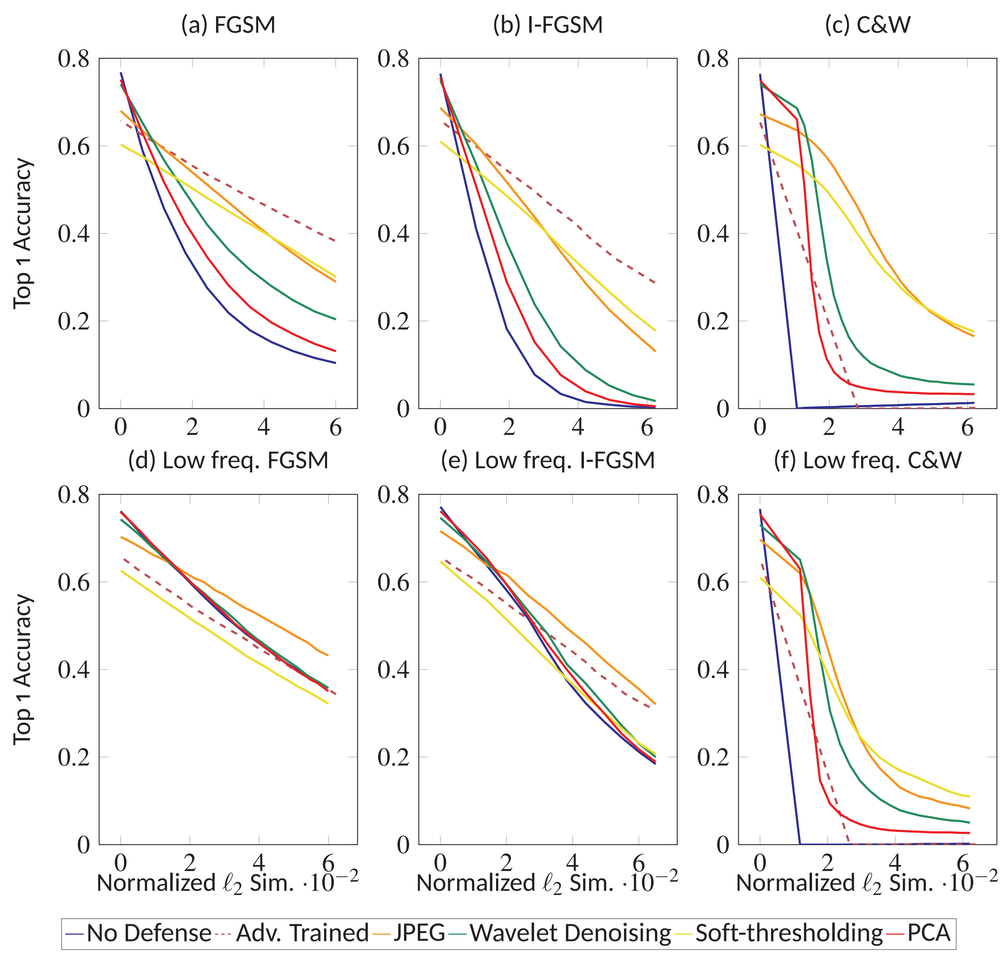

Figure 3: Model accuracy with pre-processing defenses attacked by FGSM, I-FGSM and C&W ℓ_2 in pixel domain (a), (b), (c), and low frequency DWT domain (d), (e), (f). Tested on 10,000 images from the CIFAR-10 dataset.

Literature

Bibliography

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.

[2] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41, 1988.

[3] C. Guo, M. Rana, M. Cisse, L. van der Maaten. Countering Adversarial Images using Input Transformations. *International Conference on Learning Representations*, 2018.

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

[5] U. Shaham, J. Garritano, Y. Yamada, E. Weinberger, A. Cloninger, X. Cheng, K. Stanton, Y. Kluger. Defending against Adversarial Images using Basis Functions Transformations. *arXiv:1803.10840*, 2018.