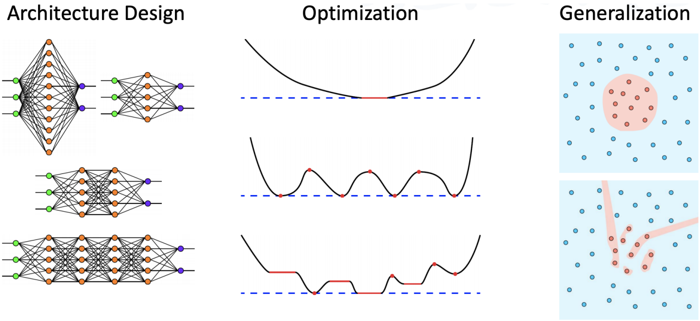
Wavelet-based Low Frequency Adversarial Attacks

Shpresim Sadiku

(Technische Universität Berlin & Zuse Institute Berlin)

BMS - BGSMath Junior Meeting · September 7, 2022

Three Problems in Deep Learning



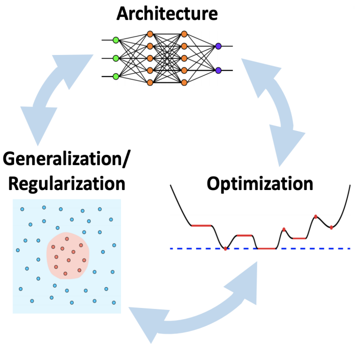
from: *Mathematics of Deep Learning*, René Vidal, DeepMath Plenary Lecture, 2020

The Three Problems are interrelated

↪ Easier to optimize some architectures than others (Haeffele et al., 2017)

↪ Generalization is strongly affected by architecture (Zhang et al., 2017)

↪ Optimization can impact generalization (Neyshabur et al., 2015, Zhou and Feng, 2017)



Error Decomposition

$$R(f) - R^* = \underbrace{(R(f) - R(\hat{f}))}_{\text{optimization error}} + \underbrace{(R(\hat{f}) - R_{\mathcal{F}})}_{\text{estimation error}} + \underbrace{(R_{\mathcal{F}} - R^*)}_{\text{approximation error}}$$

$R(f)$ - risk of a hypothesis f
$R^* = \inf_f R(f)$ - Bayes risk
$\hat{f}$ - minimizer of the empirical risk $\hat{R}(f)$

Interplay of

1. Learning
   ($\hookrightarrow$ Optimization, Optimal Control,...)
2. Generalization
   ($\hookrightarrow$ Statistics, Learning Theory, Stochastics,...)
3. Expressivity
   ($\hookrightarrow$ Approximation Theory, Applied Harmonic Analysis,...)

Generalization

Joint work

- Moritz Wagner (TU Berlin & ZIB)
- Sebastian Pokutta (TU Berlin & ZIB)

Image Representations

- Discrete Fourier Transform (DFT) basis
- Discrete Wavelet Transform (DWT) basis
  ↪ Captures frequency and location information
  ↪ Signals represented in the DWT basis have approximately sparse representations (Kutyniok and Lim, 2011)

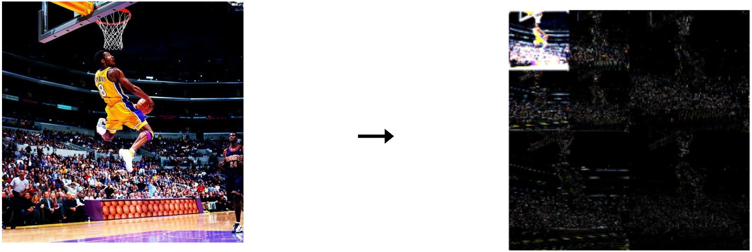


Figure 1: ImageNet image example and its 2D DWT representation.

Multiresolution Analysis (MRA)

Definition (Mallat, 1999)

An orthonormal Multiresolution Analysis (MRA) of $L^2(\mathbb{R})$ is an ordered chain of closed subspaces $\cdots \subseteq V_{-1} \subseteq V_0 \subseteq V_1 \subseteq \cdots$, satisfying

1. Completeness (C)
   $\hookrightarrow$ $\overline{\lim_{j\to\infty} V_j} = L^2(\mathbb{R})$ and $\lim_{j\to-\infty} V_j = \{0\}$
2. Dyadic Similarity (DS)
   $\hookrightarrow$ $u(x) \in V_j$ iff $u(2x) \in V_{j+1}$
3. Translation Seed (TS)
   $\hookrightarrow$ There exists $\varphi \in V_0$ such that $(\varphi(x-k))_{k\in\mathbb{Z}}$ is an orthonormal basis (ONB) of V_0

Father Wavelet

Definition (Mallat, 1999)

A function φ is defined as a father wavelet if φ generates an MRA.

Lemma

Let $\{V_i\}_{i\in\mathbb{Z}}$ denote an MRA of $L^2(\mathbb{R})$. Then for $\varphi_{j,k}(x) := 2^{\frac{j}{2}}\varphi(2^j x - k), j, k \in \mathbb{Z}$, the $\{\varphi_{j,k}\}_{k\in\mathbb{Z}}$ form an ONB of V_j.

$\hookrightarrow$ Scaled translates of φ are sufficient to represent all of L^2

- Signal $u \in L^2(\mathbb{R})$ can be approximated by its projection $u_j = P_j u = \sum_k \langle u, \varphi_{j,k}\rangle \varphi_{j,k}$ onto V_j
- E.g. $P_j : V_{j+1} \to V_j$. Details of $u_{j+1} \in V_{j+1}$: $u_{j+1} - P_j u_{j+1} = (I - P_j)u_{j+1}$
- Space of details $W_j := \{(I - P_j)u_{j+1} | u_{j+1} \in V_{j+1}\}$, i.e., $P_j W_j = \{0\}$, thus $V_{j+1} = V_j \oplus W_j$

Mother Wavelet

- (DS): $\eta(x) \in W_j \iff \eta(2x) \in W_{j+1}$
- (C): $L^2(\mathbb{R}) = \overline{V_0 \oplus \left(\bigoplus_{j=0}^{\infty} W_j\right)}$

$\hookrightarrow$ An element $u \in L^2(\mathbb{R})$ is the accumulated effect of its details

- A mother wavelet is a function $\psi \in W_0$ orthogonal to the father wavelet such that $\{\psi(x-k)\}_{k\in\mathbb{Z}}$ form an ONB of W_0
- (DS): $\{\psi_{j,k} = 2^{j/2}\psi(2^j x - k) | k \in \mathbb{Z}\}$ - ONB of W_j
  $\{\psi_{j,k} = 2^{j/2}\psi(2^j x - k) | j, k \in \mathbb{Z}\}$ - ONB of $L^2(\mathbb{R})$

$\hookrightarrow u(x) = \sum_k \underbrace{\langle u, \varphi_{0,k}\rangle}_{\text{approx coeffs}} \varphi_{0,k}(x) + \sum_{j=0}^{\infty} \sum_k \underbrace{\langle u, \psi_{j,k}\rangle}_{\text{detail coeffs}} \psi_{j,k}(x)$

2D Discrete Wavelet Transform (DWT)

- Generalization of the 1D MRA into $L^2(\mathbb{Z}^2)$
- Define $\psi^1 = \varphi\psi, \psi^2 = \psi\varphi$ and $\psi^3 = \psi\psi$ where

$$\psi^k_{j,(n_1,n_2)}(t_1,t_2) = 2^{j/2}\psi^k((2^j n_1 - t_1)/2^j, (2^j n_2 - t_2)/2^j), k \in \{1,2,3\}$$

$\hookrightarrow \{\psi^1_{j,n}, \psi^2_{j,n}, \psi^3_{j,n}\}_{j,n\in\mathbb{Z}^2}$ - ONB for $L^2(\mathbb{Z}^2)$ (Santamaria P. et al., 2021)

- Denote scaling function φ by H_0 and mother wavelet by H_1

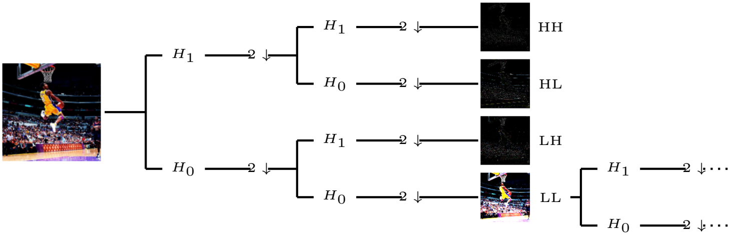


Figure 2: DWT decomposition tree for a basketball image from ImageNet dataset

Adversarial Attacks

- Input image $\mathrm{x} \in \mathcal{X} := [0,1]^{n\times c}$ of correct label t
- Neural network classifier $f_\theta : [0,1]^{n\times c} \to \mathbb{R}^k$
  \+ Softmax and classification loss $L(\theta, \mathrm{x}, t)$

- Adversarial attack problem (Szegedy et al., 2013)

$$\max_{\hat{\mathrm{x}} \in \mathcal{X} : \|\hat{\mathrm{x}} - \mathrm{x}\|_p \leq \varepsilon} L(\theta, \underbrace{\hat{\mathrm{x}}}_{\text{adv}}, t)$$

- Reformulate by defining $\mathrm{r} := \hat{\mathrm{x}} - \mathrm{x}$

$$\max_{\|\mathrm{r}\|_p \leq \varepsilon} L(\theta, \mathrm{x} + \mathrm{r}, t)$$

Adversarial Attack Methods

- Fast Gradient Sign Method (FGSM)

$$\hat{\mathrm{x}} = \mathrm{clip}_{\mathcal{X}}(\mathrm{x} + \varepsilon \ \mathrm{sign}(\nabla_{\mathrm{x}} L(\theta, \mathrm{x}, t)))$$

- Iterative Fast Gradient Sign Method (I-FGSM)

$$\hat{\mathrm{x}}^{(0)} = \mathrm{x}, \quad \hat{\mathrm{x}}^{(j)} = \mathrm{clip}_{\mathcal{X},\varepsilon}\left(\hat{\mathrm{x}}^{(j-1)} + \alpha \ \mathrm{sign}(\nabla_{\mathrm{x}} L(\theta, \hat{\mathrm{x}}^{(j-1)}, t))\right)$$

- Carlini-Wagner (C&W) - optimizes the Lagrangian formulation

$$\min_{\hat{\mathrm{x}}} \ [\|\mathrm{x} - \hat{\mathrm{x}}\|_2^2 + c \ \max(\max_{i \neq t}(f_\theta(\hat{\mathrm{x}})_i) - f_\theta(\hat{\mathrm{x}})_t, -\kappa)]$$

Adversarial Example

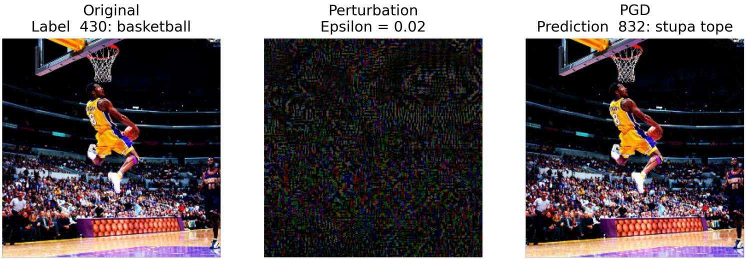


Figure 3: ImageNet example (left), the perturbation needed to change the image label (middle), and the perturbed image (right).

Defenses against Adversarial Attacks

- Adversarial Training Problem

$$\min_{\theta} \mathbb{E}_{(\mathrm{x},t)\sim D}\left[\max_{\hat{\mathrm{x}}\in\mathcal{X}:\|\hat{\mathrm{x}}-\mathrm{x}\|_p\leq\varepsilon} L(\theta,\hat{\mathrm{x}},t)\right]$$

$\hookrightarrow$ Trains classifiers to only defend against small norm ℓ_p attacks in the pixel domain

Idea:

Generate adversarial attacks in a different representation space

$\hookrightarrow$ Attacks generated in a different space circumvent adversarial training due to large ℓ_p norm in the pixel space

Circumventing Adversarial Training

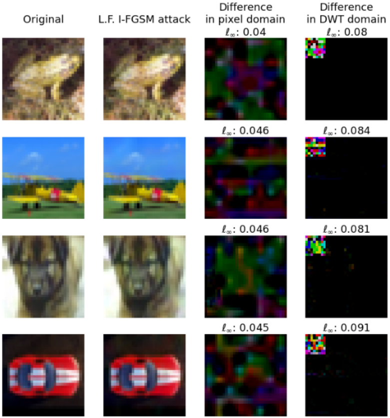


Figure 4: Images from the CIFAR-10 dataset with their corresponding adversarial examples generated by I-FGSM in the low frequency DWT domain (with a scale of 2), as well as their differences in the pixel and DWT domain.

Image Pre-processing Methods

Experiment with

- JPEG Compression
- PCA Denoising
- Soft-Thresholding
- Wavelet Denoising

↪ Do not modify the training procedure or the architecture
↪ Detect or remove adversarial attacks by smoothing the input data
↪ Rely on removing high frequency signal (Shaham et al., 2018a)

Can we also circumvent Compression Techniques?

- Adversarial attacks are made up of high frequency noise, regardless of the generation space
- Low frequencies are crucial for the SOA models to extract class-specific information from images

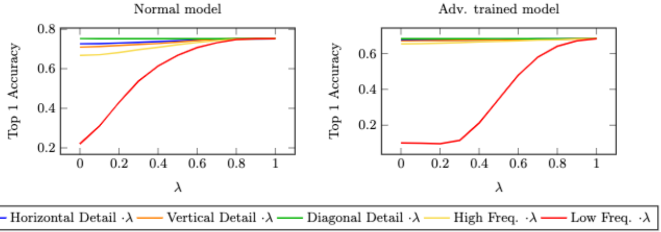


Figure 5: Accuracy of model trained on clean data and adversarially trained model. Some wavelet coefficients of the test images are multiplied by $0 \leq \lambda \leq 1$. Either the low frequency, HL, LH, HH, or all high frequency coefficients are multiplied by λ.

Circumventing Adversarial Training and Compression Techniques

Question:

$\hookrightarrow$ Can we generate adversarial attacks that circumvent both adversarial training and defense pre-processing methods?

Idea:

$\hookrightarrow$ Generate perturbations in the low frequency wavelet domain

Wavelet-based Low Frequency Adversarial Attacks

Wavelet-based Adversarial Attacks

- Representation space $\mathcal{R}$ - map given by the DWT basis
- $\mathrm{x} \in \mathbb{R}^{n \times c} \rightarrow \mathcal{R}(\mathrm{x})$

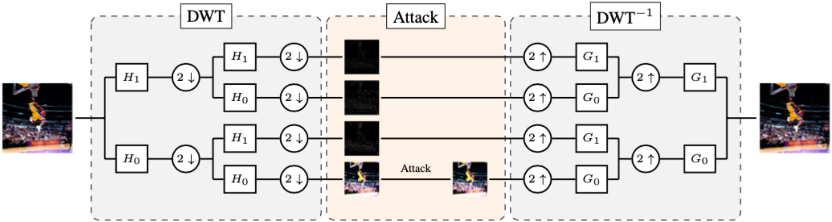


Figure 6: The low frequency I-FGSM attack with DWT scale 1 for a basketball image from ImageNet.

FGSM in the Wavelet Domain

1. FGSM problem in the wavelet domain $\mathcal{R}$

$$\arg\max_{\|\mathrm{r}\|_\infty \leq \varepsilon} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathrm{x}) + \mathrm{r}), t),$$

2. First order approximation

$$\arg\max_{\|\mathrm{r}\|_\infty \leq \varepsilon} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathrm{x})), t) + \mathrm{r}\nabla_{\mathcal{R}(\mathrm{x})} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathrm{x})), t)$$

3. Maximal perturbation

$$\mathrm{r} = \varepsilon \ \mathrm{sign}(\nabla_{\mathcal{R}(\mathrm{x})} L(\theta, \mathcal{R}^{-1}(\mathcal{R}(\mathrm{x})), t))$$

4. Linear $\mathcal{R}$

$$\mathrm{r} = \varepsilon \ \mathrm{sign}\left(\mathcal{R}\left(\frac{\partial L(\theta, \mathrm{x}, t)}{\partial \mathrm{x}}\right)\right)$$

Wavelet-based Low Frequency Adversarial Attacks

- Low Frequency FGSM

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

- Low Frequency I-FGSM

$$\hat{\mathrm{x}}^{(0)} = \mathrm{x}, \quad \hat{\mathrm{x}}^{(n)} = \mathrm{clip}_{\mathrm{x},\varepsilon}\left(\mathrm{clip}_{[0,1]}\left(\hat{\mathrm{x}}^{(n-1)} - \mathcal{R}^{-1}\left(\mathrm{r}^{(n)}\right)\right)\right)$$

with

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

Low frequency C&W ℓ_2

- $\tilde{\mathrm{x}} = \mathcal{R}(\tanh^{-1}(2\mathrm{x} - 1))$
- Define

$$\hat{\mathrm{w}} = \left[\begin{array}{c|c} \mathrm{w} & \tilde{\mathrm{x}}_{LH} \\ \hline \tilde{\mathrm{x}}_{HL} & \tilde{\mathrm{x}}_{HH} \end{array}\right]$$

- Choose

$$\delta = \mathcal{R}\left(\frac{1}{2}\left(\tanh\left(\mathcal{R}^{-1}\left(\hat{\mathrm{w}}\right)\right) + 1\right)\right) - \mathcal{R}(\mathrm{x}).$$

s.t. $\mathcal{R}^{-1}(\mathcal{R}(\mathrm{x}) + \mathrm{r}) \in [0, 1]^{n \times m}$

- Optimize over w

$$\min \|\mathcal{R}(\frac{1}{2}(\tanh(\mathcal{R}^{-1}(\hat{\mathrm{w}})) + 1)) - \mathcal{R}(\mathrm{x})\|_2^2 + cf(\frac{1}{2}(\tanh(R^{-1}(\hat{\mathrm{w}})) + 1)),$$

Experiments

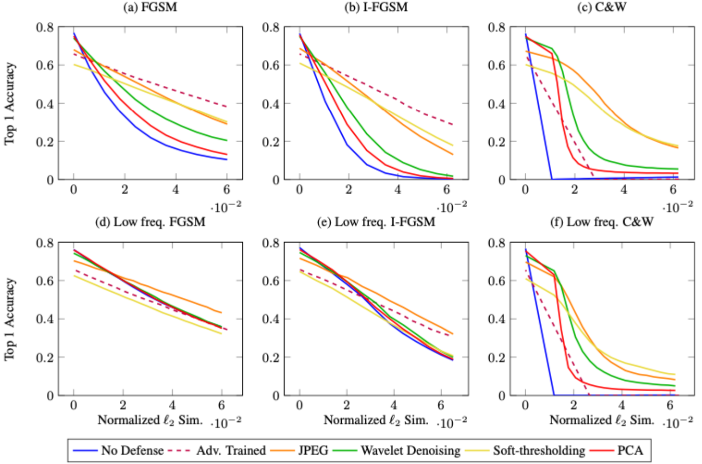


Figure 7: Accuracy of model with pre-processing defenses attacked by FGSM, I-FGSM and C&W ℓ_2 in pixel domain and low frequency DWT domain. Tested on 10,000 images from the CIFAR-10 dataset.

Future work

- Generate almost imperceptible low frequency adversarial attacks in a black box setting and for real-world scenarios
- Given this vulnerability of NNs, design SOA defense strategies
  ↪ Integrate low frequency adversarial attacks in the adversarial training procedure

Thank you!