Formalizing Convolutional Neural Networks:
Classification by alternating change of bases and simple nonlinearities

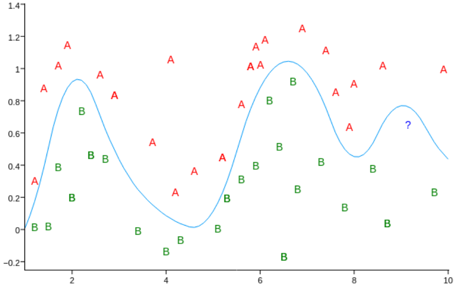
David Weber

Department of Mathematics
University of California, Davis

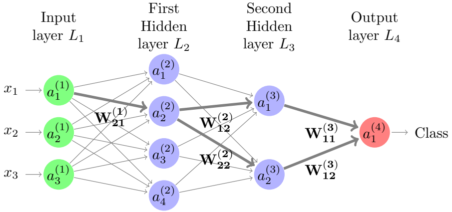
Davis Math Conference
UC Davis
January 12, 2016

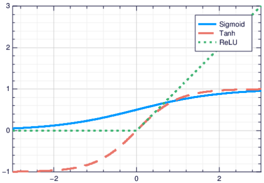
Classification

- $x \in X$ is the input
- $y \in Y$ is the output, usually one of a finite number of classes, e.g. A, B
- We have labelled training data $(x_i, y_i)_{i=1}^{N}$
- We are looking for a function $F : X \to Y$ which will classify new, unlabelled examples



Neural Networks





$$a_i^j = \sigma\left(\sum_{k=1}^{n_{j-1}} W_{ik}^{(j-1)} a_k^{(j-1)}\right) = \sigma\left(\vec{W}_i^{(j-1)} \cdot \vec{a}^{(j-1)}\right)$$

Convolutional Neural Networks

Instead of single values for each weight matrix we can output an entire vector by using convolution instead of a dot product:

$$a^j(k) = \sigma\left(\vec{W}^{(j-1)} \star \vec{a}^{(j-1)}(k)\right)$$

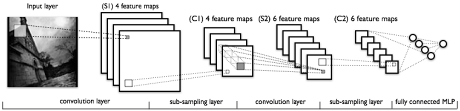


Figure: From http://deeplearning.net/tutorial/lenet.html

Visual system, CNNs, & wavelets

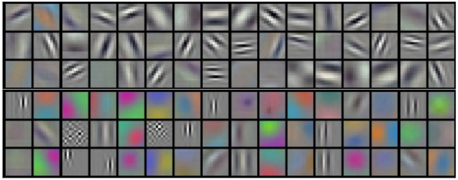


Figure: The filters from [Krizhevsky et al., 2012]

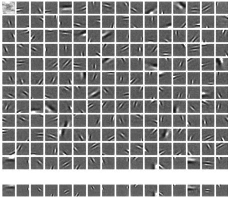


Figure: Sparsifying basis functions having similar structure to receptive fields, from [Bruno A Olshausen, 1996]

Wavelets

Definition

A *Wavelet Transform* uses wavelets which are translations and rescalings of a single mother wavelet ψ:

$$\psi_{n,j}(x) = a^{-n/2}\psi\big(a^{-n}(x - nb)\big)$$
$$W[n,j]f = f \star \overline{\psi}_{n,j} := \int f(x)a^{-n/2}\psi\big(a^{-n}(x - nb)\big)\,\mathrm{d}x$$

where the mother wavelet ψ satisfies $\|\psi\|_2 = 1$ and $\int \psi\,\mathrm{d}x = 0$.

The restrictions on the mother wavelet second part is our first example of an *admissibility condition*.

Morlet Wavelet

Example (Morlet Wavelet)

In the frequency domain, Morlet Wavelets are Gaussian modulated sinusoids shifted from the origin to make them almost analytic:

$$\psi(t) = c_\xi e^{-t^2/2}\left(e^{i\xi t} - \kappa_\xi\right) \quad \Leftrightarrow \quad \widehat{\psi}(\omega) = c_\xi\left(e^{-(\omega-\xi)^2/2} - \kappa_\xi e^{-\omega^2/2}\right) \tag{1}$$

κ_ξ *is used to make* ψ *admissible, while* c_ξ *is a normalization factor.*

Father and Mother wavelets

Paired with this mother wavelet is a “father wavelet”, or scaling function ϕ, which captures the remaining low frequency information.

Definition

The father wavelet ϕ (paired with mother wavelet ψ) is specified by its Fourier Transform

$$|\widehat{\phi}(\xi)|^2 = \int_{\xi}^{\infty} \frac{|\widehat{\psi}(\eta)|^2}{\eta}\,\mathrm{d}\eta$$

There is an admissibility condition on ϕ and ψ such that the set $\{\psi_{j,n}\}_{(j,n)\in\mathbb{N}^+\times\mathbb{Z}}$ forms an orthonormal basis of $L^2(\mathbb{R})$.

Signal invariants

The classes that are relevant in scattering problems have two easily identifiable invariants:

- Translation:
  - An operator Φ is translation invariant if $\Phi(T_c f)(t) = \Phi(f)(t)$ for $c \in \mathbb{R}$, where $T_c[f] = f(t-c)$
- Lipschitz continuity under small diffeomorphism
  - An operator Φ is Lipschitz-continuous relative to operators of the form $T_\tau[f](t) = f(t-\tau(t))$ if $\forall \Omega \in \mathbb{R}^d$, there is a universal bound C for $f \in L^2(\mathbb{R}^d)$

$$\|\Phi(f) - \Phi(T_\tau f)\|_{\mathscr{H}} \leq C\|f\|\left(\|\nabla\tau\|_\infty + \|H\tau\|_\infty\right) \tag{2}$$

Signal invariants

The classes that are relevant in scattering problems have two easily identifiable invariants:

- Translation:
  - An operator Φ is translation invariant if $\Phi(T_c f)(t) = \Phi(f)(t)$ for $c \in \mathbb{R}$, where $T_c[f] = f(t-c)$
- Lipschitz continuity under small diffeomorphism
  - An operator Φ is Lipschitz-continuous relative to operators of the form $T_\tau[f](t) = f(t-\tau(t))$ if $\forall \Omega \in \mathbb{R}^d$, there is a universal bound C for $f \in L^2(\mathbb{R}^d)$

$$\|\Phi(f) - \Phi(T_\tau f)\|_{\mathcal{H}} \leq C\|f\|\left(\|\nabla\tau\|_\infty + \|H\tau\|_\infty\right) \qquad (2)$$

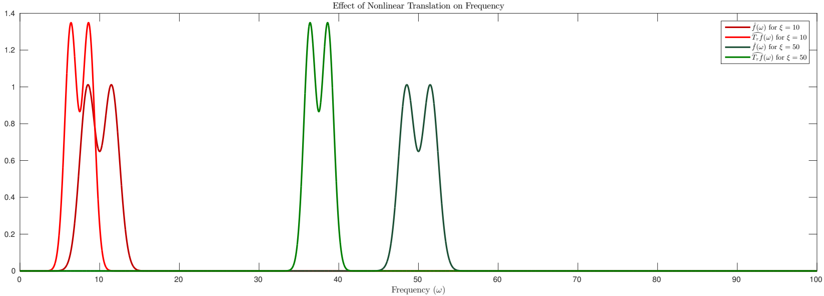
Why not just use the Fourier Transform?

The Fourier transform is translation invariant, but it is not Lipschitz continuous under diffeomorphisms:

Let $\tau(t) = st$, with $|s| < 1$, and $f(t) = e^{i\xi t}\theta(t)$, where θ is even and $O\left(e^{-x^2}\right)$ then $T_\tau[f](t) = f\big((1-s)t\big)$ translates the central frequency ξ to $(1-s)\xi$

$$\|\widehat{T_\tau f} - \widehat{f}\| \sim |s||\xi|\|\theta\| = |\xi|\|f\|\|\nabla\tau\|_\infty \tag{3}$$

No universal bound for arbitrary ξ!

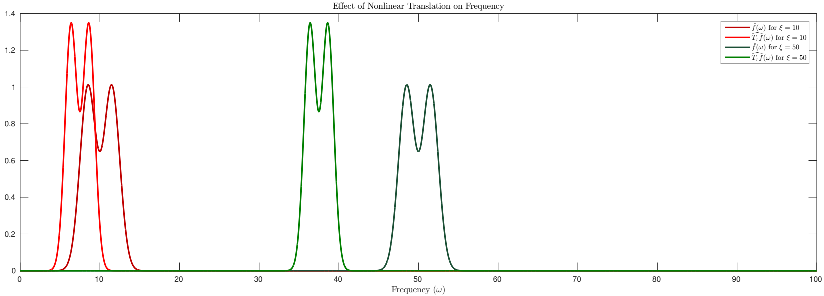


Why not just use the Fourier Transform?

The Fourier transform is translation invariant, but it is not Lipschitz continuous under diffeomorphisms:
Let $\tau(t) = st$, with $|s| < 1$, and $f(t) = e^{i\xi t}\theta(t)$, where θ is even and $O\left(e^{-x^2}\right)$
then $T_\tau[f](t) = f\big((1-s)t\big)$ translates the central frequency ξ to $(1-s)\xi$

$$\|\widehat{T_\tau f} - \widehat{f}\| \sim |s||\xi|\|\theta\| = |\xi|\|f\|\|\nabla\tau\|_\infty \tag{3}$$

No universal bound for arbitrary ξ!

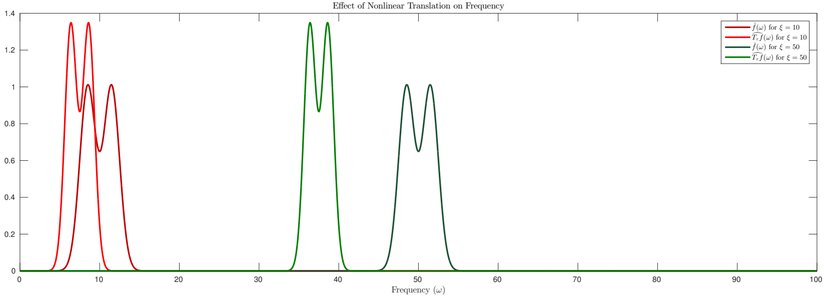


Why not just use the Fourier Transform?

The Fourier transform is translation invariant, but it is not Lipschitz continuous under diffeomorphisms:
Let $\tau(t) = st$, with $|s| < 1$, and $f(t) = \mathrm{e}^{\mathrm{i}\xi t}\theta(t)$, where θ is even and $O\big(\mathrm{e}^{-x^2}\big)$ then $T_\tau[f](t) = f\big((1-s)t\big)$ translates the central frequency ξ to $(1-s)\xi$

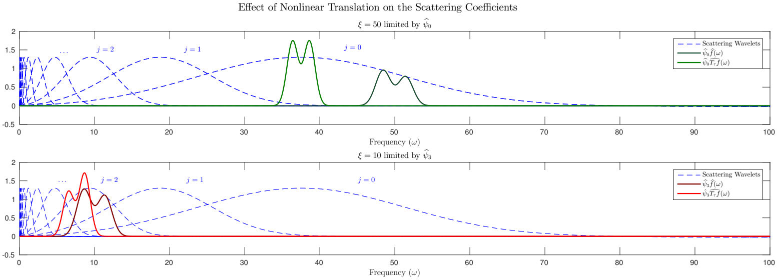
$$\|\widehat{T_\tau f} - \widehat{f}\| \sim |s||\xi|\|\theta\| = |\xi|\|f\|\|\nabla\tau\|_\infty \tag{3}$$

No universal bound for arbitrary ξ!



Wavelet Transform & T_τ

In the fourier domain, a wavelet transform $\psi_j \star f$ bandpasses the signal over windows whose width decreases exponentially with j, so that both f and $T_\tau f$ are captured within the same wavelet, regardless of ξ

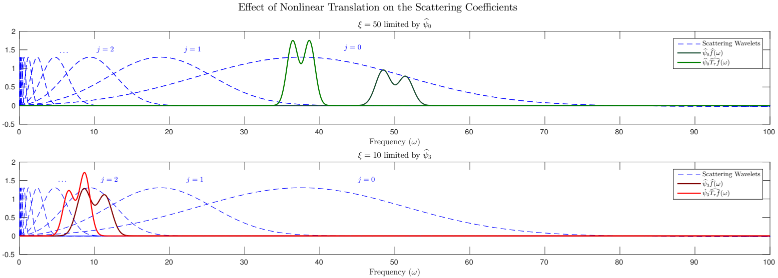


A Wavelet transform isn't translation invariant, but it does commute with the translation operator, i.e. if $W[j]f(n) = f \star \widehat{\psi}_{j,n}$, then

$$W[j]T_c f(n) = T_c W[j] f(n)$$

Wavelet Transform & T_τ

In the fourier domain, a wavelet transform $\psi_j \star f$ bandpasses the signal over windows whose width decreases exponentially with j, so that both f and $T_\tau f$ are captured within the same wavelet, regardless of ξ



A Wavelet transform isn't translation invariant, but it does commute with the translation operator, i.e. if $W[j]f(n) = f \star \widehat{\psi}_{j,n}$, then

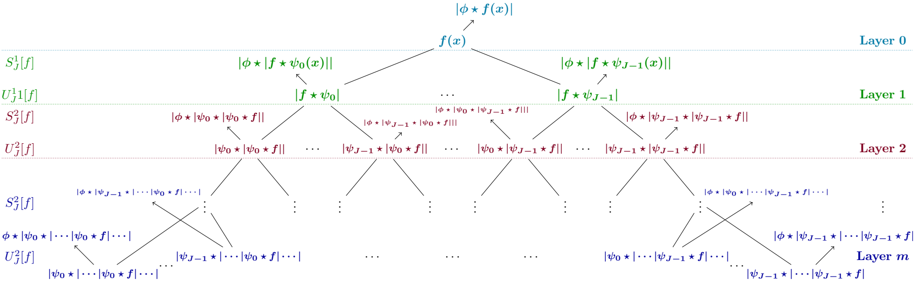
$$W[j]T_c f(n) = T_c W[j] f(n)$$

Scattering Transform

A single propagating layer $U_J^m[f]$ of the scattering transform is a vector consisting of alternating convolution with wavelets $\widehat{\psi}_j(\omega) = \hat{\psi}\left(2^{j/Q}\omega\right)$ with scales ranging from the finest 0 to the coarsest $J-1$ and a modulus $|\cdot|$:

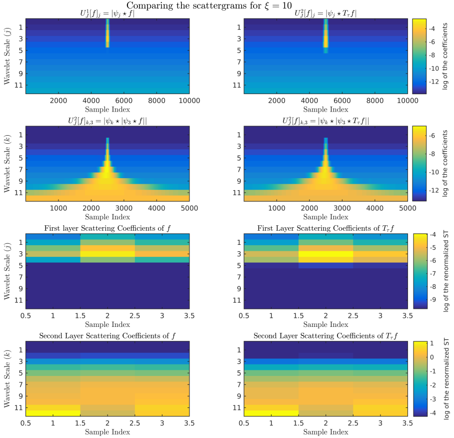
$$U_J^1[f] := \left(|\psi_0 \star f|, \ldots, |\psi_{J-1} \star f|\right)$$

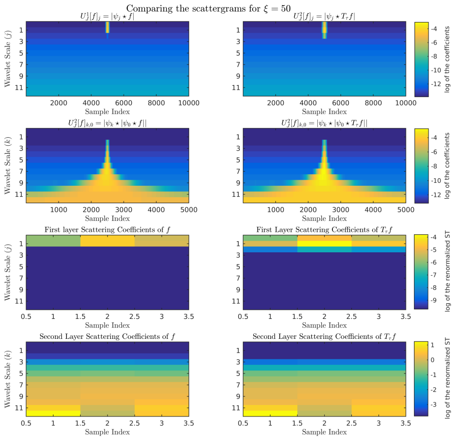
$$U_J^2[f] := \left(|\psi_0 \star |\psi_0 \star f||, |\psi_1 \star |\psi_0 \star f||, \ldots, |\psi_{J-1} \star |\psi_0 \star f||, \ldots\ldots, |\psi_{J-1} \star |\psi_{J-1} \star f||\right)$$



The output $S_J^m[f]$ is taken by averaging every term of $U_J^m[f]$ with the father wavelet ϕ corresponding to ψ, then subsampling.

Scattering Transform comparison of f and $T_\tau f$





Useful Properties

Theorem (Limit Translation Invariance from [Mallat, 2012])

For all $f \in L^2(Rf^d)$ *and* $c \in Rf^d$, *if* (ψ, ϕ) *are admissible, then*

$$\lim_{J \to -\infty} \|S_J[f] - S_J[T_c f]\|_2 = 0 \tag{4}$$

as the scale goes to infinite resolution, the scattering transform is translation invariant. In addition it preserves the total energy

Theorem (Energy conservation from [Mallat, 2012])

For all $f \in L^2(\mathbb{R}^d)$, *if* (ψ, ϕ) *are admissible, then*

$$\|f\|_2 = \|S_J[f]\|_2 \quad \textit{where} \quad S_J[f] := \left(S_J^0[f], S_J^1[f], \ldots, S_J^m[f], \ldots\right),$$

$$\|S_J[f]\|_2^2 := \sum_{m=0}^{\infty} \|S_J^m[f]\|_2^2$$

Useful Properties

Theorem (Limit Translation Invariance from [Mallat, 2012])

For all $f \in L^2(Rf^d)$ *and* $c \in Rf^d$, *if* (ψ, ϕ) *are admissible, then*

$$\lim_{J \to -\infty} \|S_J[f] - S_J[T_c f]\|_2 = 0 \tag{4}$$

as the scale goes to infinite resolution, the scattering transform is translation invariant. In addition it preserves the total energy

Theorem (Energy conservation from [Mallat, 2012])

For all $f \in L^2(\mathbb{R}^d)$, *if* (ψ, ϕ) *are admissible, then*

$$\|f\|_2 = \|S_J[f]\|_2 \quad \textit{where} \quad S_J[f] := \left(S_J^0[f], S_J^1[f], \ldots, S_J^m[f], \ldots\right),$$

$$\|S_J[f]\|_2^2 := \sum_{m=0}^{\infty} \|S_J^m[f]\|_2^2$$

Theorem (Lipschitz Continuity from [Mallat, 2012])

For all compactly supported $f \in L^2(\mathbb{R}^d)$ *satisfying* $\|\sum_m U_J^m f\|_1 < \infty$ *and* $\tau \in C^2(\mathbb{R}^d)$ *where* $\|\nabla\tau\|_\infty \leq \frac{1}{2}$ *and* $\|\tau\|_\infty / \|\nabla\tau\|_\infty \leq 2^J$, *there is a* C *such that:*

$$\left\| S_J[T_\tau f] - S_J[f] \right\|_2 \leq C \left\| \sum_m U_J^m f \right\|_1 \left(\|\nabla\tau\|_\infty + \|H\tau\|_\infty \right) \tag{5}$$

A more recent result is that for general frames, and not just admissible wavelets, that increasing the depth m increases translation invariance:

Theorem (Depth translation invariance, [Wiatowski and Bölcskei, 2015])

If R_n *is the subsampling rate layer* n, *as long as the wavelets have frame bounds* B_n *satisfying* $\max\{B_n, B_n R_n^d\} \leq 1$, *the features at depth* m *satisfy:*

$$S_m[T_c f] = T_{\frac{c}{R_1 \cdots R_{m-1}}} S_m[f] \tag{6}$$

Theorem (Lipschitz Continuity from [Mallat, 2012])

For all compactly supported $f \in L^2(\mathbb{R}^d)$ *satisfying* $\|\sum_m U_J^m f\|_1 < \infty$ *and* $\tau \in C^2(\mathbb{R}^d)$ *where* $\|\nabla\tau\|_\infty \le \frac{1}{2}$ *and* $\|\tau\|_\infty / \|\nabla\tau\|_\infty \le 2^J$, *there is a* C *such that:*

$$\left\| S_J[T_\tau f] - S_J[f] \right\|_2 \le C \left\| \sum_m U_J^m f \right\|_1 \left(\|\nabla\tau\|_\infty + \|H\tau\|_\infty \right) \tag{5}$$

A more recent result is that for general frames, and not just admissible wavelets, that increasing the depth m increases translation invariance:

Theorem (Depth translation invariance, [Wiatowski and Bölcskei, 2015])

If R_n *is the subsampling rate layer* n, *as long as the wavelets have frame bounds* B_n *satisfying* $\max\{B_n, B_n R_n^d\} \le 1$, *the features at depth* m *satisfy:*

$$S_m[T_c f] = T_{\frac{c}{R_1 \cdots R_{m-1}}} S_m[f] \tag{6}$$

References I

Bruno A Olshausen, D. J. F. (1996).
Emergence of simple-cell receptive field properties by learning a sparse code for natural images.
Nature, (381):607–609.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
ImageNet classification with deep convolutional neural networks.
In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Mallat, S. (2012).
Group invariant scattering.
Communications on Pure and Applied Mathematics, 65(10):1331–1398.

Wiatowski, T. and Bölcskei, H. (2015).
A mathematical theory of deep convolutional neural networks for feature extraction.

References II

Bruno A Olshausen, D. J. F. (1996).
Emergence of simple-cell receptive field properties by learning a sparse code for natural images.
Nature, (381):607–609.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
ImageNet classification with deep convolutional neural networks.
In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Mallat, S. (2012).
Group invariant scattering.
Communications on Pure and Applied Mathematics, 65(10):1331–1398.

Wiatowski, T. and Bölcskei, H. (2015).
A mathematical theory of deep convolutional neural networks for feature extraction.