Theory and Applications of Scattering Networks:

Classification by alternating change of bases and simple nonlinearities

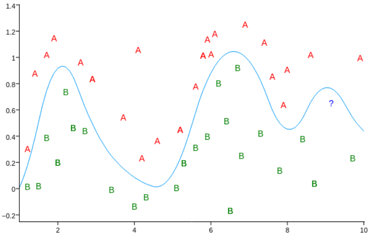
David Weber

Department of Mathematics
University of California, Davis

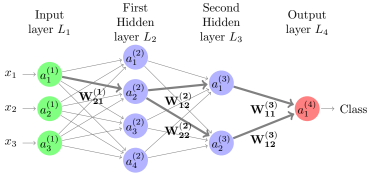
Davis Math Conference
UC Davis
January 12, 2016

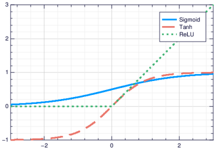
Classification

- $x \in X$ 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





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

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:

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

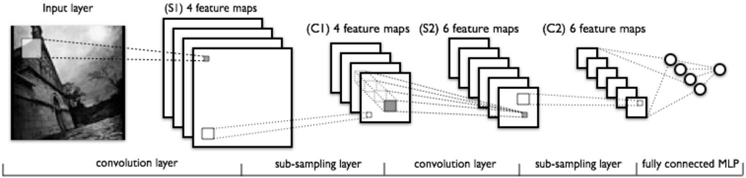
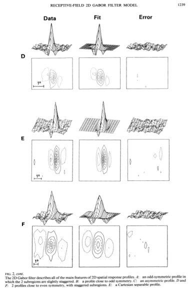
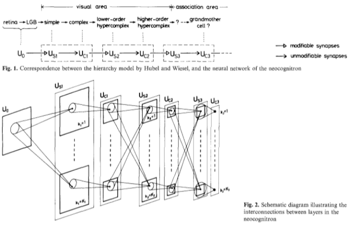


Figure: From http://deeplearning.net/tutorial/lenet.html

Visual system, CNNs, & wavelets
A little history

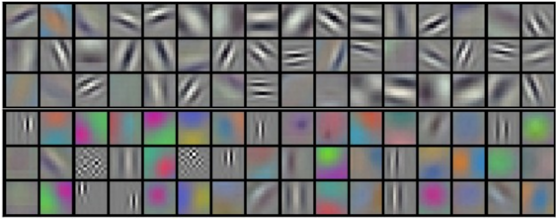


(a) Example of Gabor functions in modeling the simple cells of a cat from [Jones and Palmer, 1987]

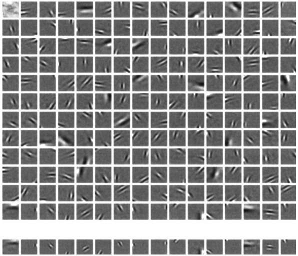


(b) The origin of CNN's, called the neocognitron[Fukushima, 1980]

Visual system, CNNs, & wavelets



(a) The filters from [Krizhevsky et al., 2012]



(b) Sparsified frames for real images have 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:

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

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 $T_c[f] = f(t-c)$

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 $f \in L^2(\mathbb{R}^d)$

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

Why not just use the Fourier Transform?

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

Why not just use the Fourier Transform?

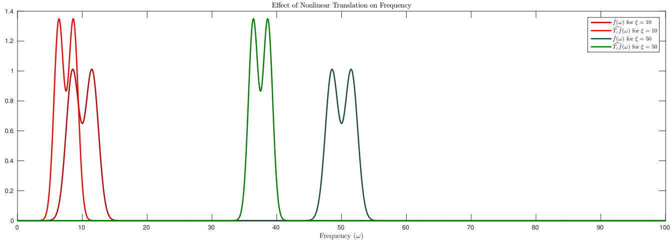
The Fourier transform is translation invariant, but it is not Lipschitz continuous under diffeomorphisms:
Let $\tau(t) = st$, with $|s| < 1$, and $f(t) = e^{i\xi t}\theta(t)$, where θ is even and $O\left(e^{-x^2}\right)$

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 $O\left(\mathrm{e}^{-x^2}\right)$
then $T_\tau[f](t) = f\big((1-s)t\big)$ translates the central frequency ξ to $(1-s)\xi$

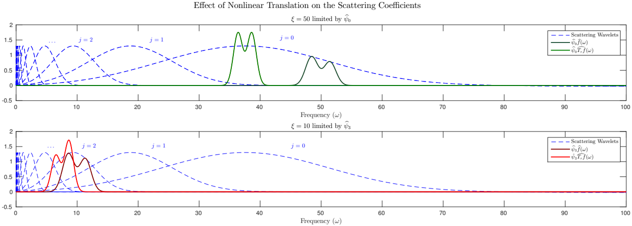
$$\|\widehat{T_\tau f} - \widehat{f}\| \sim |s||\xi|\|\theta\| = |\xi|\|f\|\|\nabla\tau\|_\infty$$

No universal bound for arbitrary ξ!



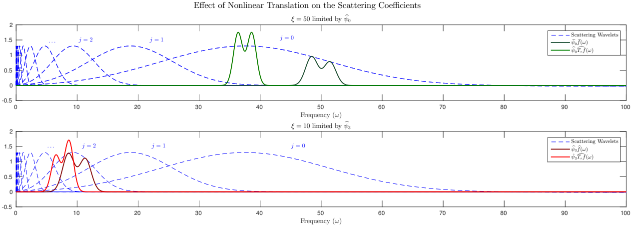
Wavelet Transform & T_τ

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



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



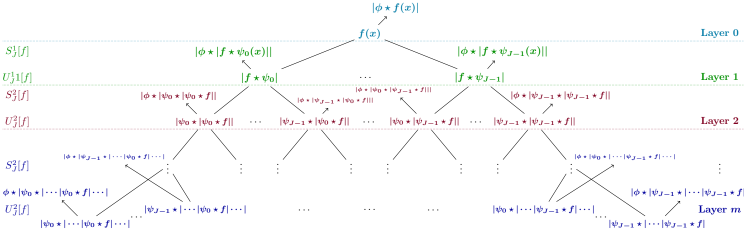
A Wavelet transform isn't translation invariant, but it does commute with the translation operator, i.e. if $W[j]f(n) = f \star \widehat{\psi}_{j,n}$, then

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

Scattering Transform

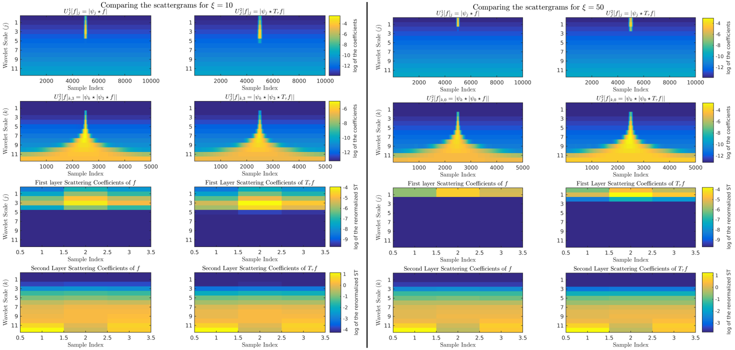
A single propagating layer $U_J^m[f]$ of the scattering transform is a vector consisting of alternating convolution with wavelets $\widehat{\psi}_j(\omega) = \hat{\psi}\left(2^{j/Q}\omega\right)$ with scales ranging from the finest 0 to the coarsest $J-1$ and a modulus $|\cdot|$: $U_J^1[f] := \left(|\psi_0 \star f|, \ldots, |\psi_{J-1} \star f|\right)$

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



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

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

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

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

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:

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

Sonar Scattering

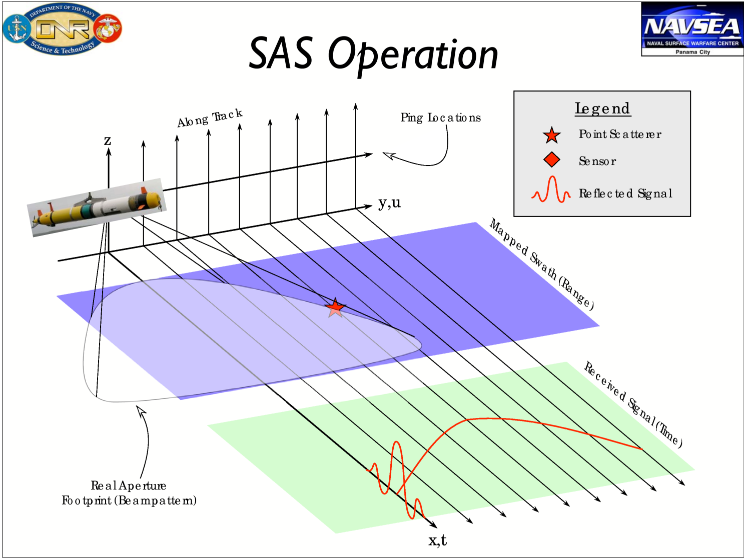
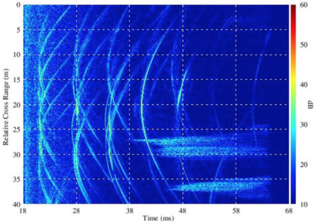
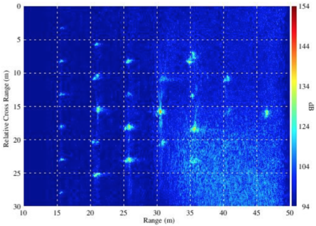


Figure: Various unexploded ordinance (UXO), replicas, and other sea debris

Sonar Scattering



Original signal



Reconstruction

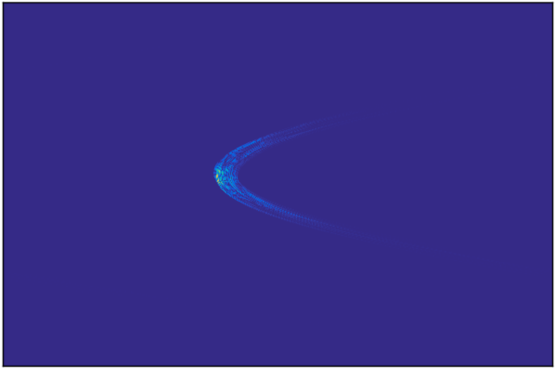


Figure: The scattering off of a 155mm Howitzer shell

Sonar Scattering

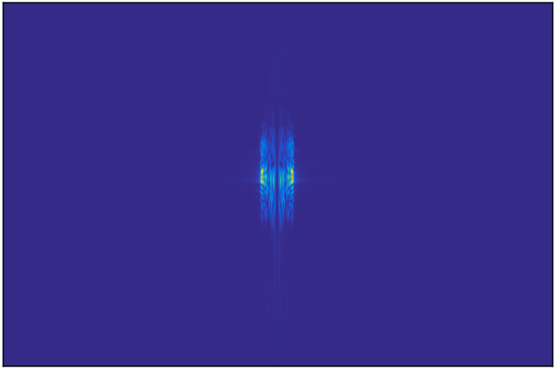


Figure: The 1D FFT of the shell

Sonar Scattering

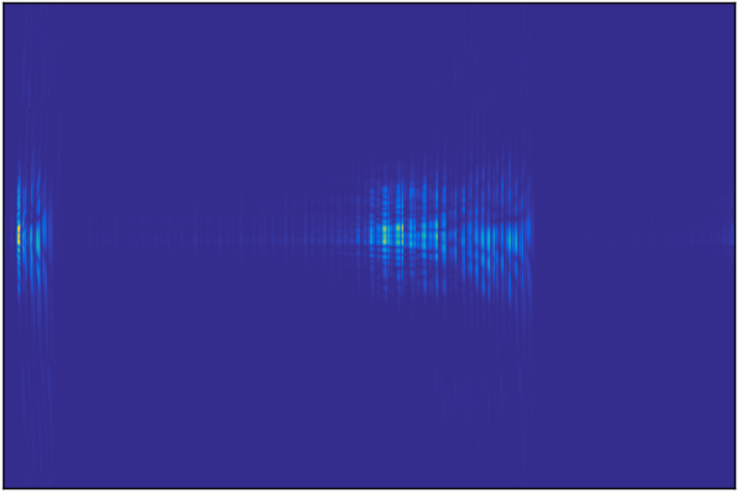


Figure: The Scattering transform of the shell

1D classifier

First plan: Consider each image as a set of 1D signals, perform a scattering transform using Morlet Wavelets, and use the resulting vectors as the input to a linear classifier, in our case Sparse logistic regression.

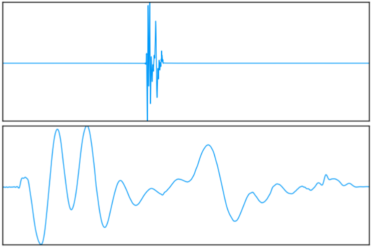


Figure: Central example from previous Data

Sparse Logistic Regression

Suppose we are trying to classify $\vec{x} \in \mathbb{R}^d$ into one of k classes. Then the sparse linear classifier is

$$\min_{\vec{\beta}_0 \in \mathbb{R}^k, \boldsymbol{\beta} \in \mathbb{R}^d \times \mathbb{R}^k} \frac{1}{N} \sum_{i=1}^{N} l(\vec{y}_i, \beta_0 + \boldsymbol{\beta}^T \vec{x}_i) + \lambda \big(||\boldsymbol{\beta}||_1 + \|\vec{\beta}_0\|_1 \big)$$

where l is the logit function:

$$l\big(\vec{y}_i, \vec{\beta}_0 + \boldsymbol{\beta}^T \vec{x}_i\big) = \sum_{k=1}^{K} y_{ik} \log \frac{e^{\beta_{0k} + \vec{\beta}_k^T \vec{x}_i}}{\sum_{\ell=1}^{K} e^{\beta_{0k} + \vec{\beta}_k^T \vec{x}_i}}$$

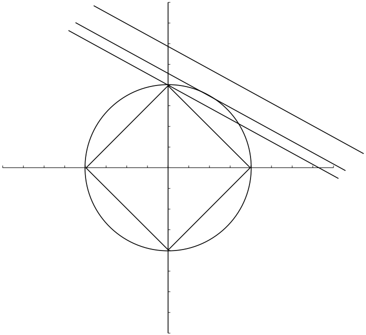
Which arises by maximizing

$$\max_{\vec{\beta}_0 \in \mathbb{R}^k, \boldsymbol{\beta} \in \mathbb{R}^d \times \mathbb{R}^k} -\log P\big(Y = \vec{y}_i = (0, \ldots, 1, \ldots, 0) | X = \vec{x}_i, \vec{\beta}_0, \boldsymbol{\beta}\big)$$

for the categorical distribution.

Sparse Logistic Regression

Why the $\|\cdot\|_1$? it induces sparsity:



The scattering transform is highly redundant, so we should only look for a subset of coefficients which are most important to classify.

14-way classification is difficult

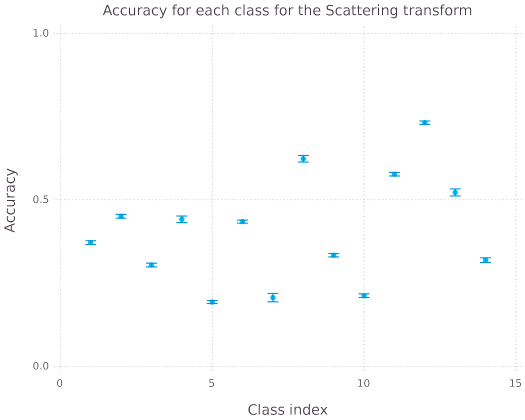


Figure: Averaging classification over 10 splits, standard error bars (Note that random guessing $1/14 \sim 7.1\%$)

Normalized data from BAYEX13, comparing 1 vs 1 classification

original \ classified as	DEU Trainer	Rock
DEU Trainer	$72.33 \pm .6\%$	27.67%
Rock	30.22%	$69.78 \pm .63\%$

Table: AVFT results

original \ classified as	DEU Trainer	Rock
DEU Trainer	$98.91 \pm .14\%$	1.09%
Rock	2.24%	$97.76 \pm .1\%$

Table: Scattering Transform, with $m = 2$ and quality factor $Q = 8$

Real Experiments

Figure: Various unexploded ordinance (UXO), replicas, and other sea debris

Normalized data from BAYEX13, grouped into two classes

original \ classified as	UXO-group	Others
UXO-group	$90.5 \pm .079\%$	9.49%
Others	50.53%	$49.47 \pm .26\%$

Table: AVFT results

original \ classified as	UXO-group	Others
UXO-group	$94.55 \pm .057\%$	5.45%
Others	24.28%	$75.71 \pm .19\%$

Table: Scattering Transform, with $m = 2$ and quality factor $Q = 8$

Synthetic Experiments
Helmholtz Equation Solver

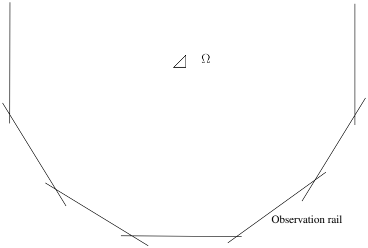


Figure: The triangle region and the observation paths

Mono frequency equation:

$$\begin{aligned}
\Delta u_\omega + k_1^2 u_\omega &= 0 && \text{in } \Omega \\
\Delta v_\omega + k_2^2 v_\omega &= 0 && \text{in } \Omega^c \\
u_\omega - v_\omega &= g && \text{on } \partial\Omega \\
\partial_\nu u_\omega - \partial_\nu v_\omega &= \partial_\nu g && \text{on } \partial\Omega \\
\sqrt{|x|}\left(\partial_{|x|} - \mathrm{i}k_2\right) v_\omega(x) &\to 0 \text{ as } |x| \to \infty
\end{aligned}$$

where $k_1 = \omega/c_{material}$ and $k_2 = \omega/c_{water}$. To approximate a more realistic signal $f(t)$ with finite support, use a discrete Fourier series $f(t) \approx \sum_{n=0}^{N-1} s_n \mathrm{e}^{\mathrm{i}2\pi}$ for $t \in [0, T]$

Synthetic Experiments
Shape Detection

Perturb each signal by Gaussian noise with $\mu = 0$ and $\sigma = 10^{-5}$ and try to discriminate the triangle from the Shark Fin

original \ classified as	Triangle	Sharkfin
Triangle	$69.59 \pm .2\%$	30.41%
Sharkfin	31.23%	$68.77 \pm .4\%$

Table: AVFT

original \ classified as	Triangle	Sharkfin
Triangle	$77.22 \pm .5\%$	22.78%
Sharkfin	19.50%	$80.50 \pm .2\%$

Table: Scattering Transform, with $m = 3$, and quality factor $Q = 1$

Fix the geometry of a triangle, and then vary c, which corresponds to different material properties.

original \ classified as	2000m/s	2500m/s
2000m/s	$95.81 \pm .2\%$	4.19%
2500m/s	5.16%	$94.84 \pm .2\%$

Table: AVFT results

original \ classified as	2000m/s	2500m/s
2000m/s	$96.48 \pm .3\%$	3.52%
2500m/s	4.10%	$95.9 \pm .2\%$

Table: Scattering Transform results

Frame Bounds

Definition

Frame A set of functions $\{\psi_k\}_{k=1}^{\infty}$ is a frame with frame bounds A and B if for all $f \in L^2(\mathbb{R})$

$$A\|f\|_2^2 \leq \sum_{k=1}^{\infty} |\langle f, \psi_k \rangle|^2 \leq B\|f\|_2^2$$

A frame is *tight* if $A = B$, and a **Parseval Frame** if $A = B = 1$

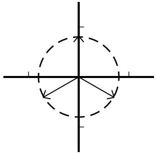


Figure: The Mercedes frame

Generalized Feature Extractor

A more recent result is that for "weakly admissible" 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 frames have frame bounds B_n satisfying $\max\{B_n, B_n R_n^d\} \le 1$, the features at depth m satisfy:

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

Generalized Feature Extractor

In addition to this quasi-translation invariance, this generalized feature extractor is stable under space and frequency modulations:

$$F_{\tau,\omega}[f](x) = \mathrm{e}^{2\pi \mathrm{i}\omega(x)} f\big(x - \tau(x)\big)$$

Theorem (Stability, [Wiatowski and Bolcskei, 2015])

if $f \in \{f \mid \text{supp}\,(\widehat{f}) \subseteq B_R(0)\}$ (f is a band limited function), ω and τ are continuous, τ is once differentiable and $\|\nabla\tau\|_\infty \leq 1/2d$, There is a C independent of S so that

$$\Big\| S[f] - S\big[F_{\tau,\omega}[f]\big] \Big\| \leq C \|f\|_2 \big(R\|\tau\|_\infty + \|\omega\|_\infty\big)$$

Alternative explanations

- Reproducing Kernel Hilbert Spaces (RKHS):
  [Daniely et al., 2016]
  Developed a framework where random initial weights are shown to be close with high probability to a kernel constructed based on the network's skeleton.
- Manifold approximation [Cloninger et al., 2016]
  Demonstrated that for all classification functions on some smooth manifold (a subspace of $\mathbb{R}^m$), they constructed a convolutional neural network that well approximates it.

Where to go from here?

- Fully implement and test a shearlet-based classifier on synthetic and real data
- create a synthetic database of small changes in material properties and geometry
- See if translation and deformation results can be found in the *object* domain in the specific case of the helmholtz equations.
- Implement a CNN with specific frame bounds in each layer
- Explore the connection between the RKHS theory and the fact that band-limited functions form a RKHS.

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.

Cloninger, A., Shaham, U., and Coifman, R. R. (2016).
Provable approximation properties for deep neural networks.
Applied and computational harmonic analysis.

Daniely, A., Frostig, R., and Singer, Y. (2016).
Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2253–2261. Curran Associates, Inc.

References II

Fukushima, K. (1980).
Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.
Biological cybernetics, 36(4):193–202.

Jones, J. P. and Palmer, L. A. (1987).
An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex.
Journal of neurophysiology, 58(6):1233–1258.

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.

References III

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

Wiatowski, T. and Bolcskei, H. (2015).
Deep convolutional neural networks based on semi-discrete frames.
In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1212–1216.

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