High Dimensional Classification with Deep Networks

Joan Bruna, Stéphane Mallat, Edouard Oyallon, Laurent Sifre

École Normale Supérieure
www.di.ens.fr/data
• Estimate a label $y(x)$ of $x \in \mathbb{R}^d$ given examples $\{x_i, y_i\}_i$

• In high dimension $\|x - x'\|$ is not a good similarity measure

• Compute $\Phi x \in \mathbb{R}^D$ so that $\|\Phi x - \Phi x'\|$ measures similarity then a linear classifier applied to $\Phi x$ is highly effective.

\[ x \rightarrow \Phi \rightarrow \Phi x \rightarrow \text{Supervised Linear Classifier} \rightarrow \tilde{y}(x) \]

• How to define $\Phi$? Should we learn it?
Genericity: one network (Alex net) yields state of the art on very different image classification problems.
Overview of Outrageous Claims

- No need to learn deep net for structured signals (images) just wavelet filters derived from geometry.

- Deep wavelet networks are signal coders.

- One can learn physical interactions: quantum chemistry.
Image Metrics

- Low-dimensional "geometric shapes"

\[
x(u) \quad x'(u)
\]

\[
\begin{array}{cccc}
3 & 3 & 5 & 5
\end{array}
\]

Deformation metric: Grenander, Trouvé, Younes

Deformation: \( D_\tau x(u) = x(u - \tau(u)) \)

\[
\Delta(x, x') \sim \min_{\tau} \| D_\tau x - x' \| + \| \nabla \tau \|_\infty \| x \|
\]

Invariant to translations

diffeomorphism amplitude

High dimensional textures: ergodic stationary processes

What metric on stationary processes?

- Invariant to translations and stable to deformations
  \[ \Delta(x, x') \leq \min_{\tau} \| D_\tau x - x' \| + \| \nabla \tau \|_\infty \| x \| \]

Reverse inequality is wrong

- \( \Delta(x', x) = 0 \) for realisations of a "same stationary process"
Image Geometry and Metric

- High dimensional "structured" images

What metric on images?

- Invariant to translations and stable to deformations

- What else?
Embedding: find an equivalent Euclidean metric

\[ \| \Phi x - \Phi x' \| \sim \Delta(x, x') \]

with

\[ \Delta(x, x') \leq \min_{\tau} \| D_\tau x - x' \| + \| \nabla \tau \|_\infty \| x \| \]

Equivalent conditions on \( \Phi \):

- **Stable in \( L^2 \):** \( D_\tau = Id \) \( \Rightarrow \) \( \| \Phi x - \Phi x' \| \leq C \| x - x' \| \)

- **Lipschitz stable** to diffeomorphisms

\[ x' = D_\tau x \ \Rightarrow \ \| \Phi D_\tau x - \Phi x \| \leq C \| \nabla \tau \|_\infty \| x \| \]

\( \Rightarrow \) Invariance to translation

Failure of classical math invariants: Fourier, canonical...
Wavelet Transform of Images

- Complex wavelet: \( \psi(t) = g(t) \exp i \xi t \), \( t = (t_1, t_2) \)

- Rotated and dilated: \( \psi_\lambda(t) = 2^{-j} \psi(2^{-j} r_\theta t) \) with \( \lambda = (2^j, \theta) \)

- Wavelet transform: \( Wx = \begin{pmatrix} x \ast \phi_2^J(t) \\ x \ast \psi_\lambda(t) \end{pmatrix} \quad \lambda \leq 2^J \)

- Preserves norm: \( \|Wx\|^2 = \|x\|^2 \)
Fast Wavelet Transform

\[ |W_1| \]

\[ |x \ast \psi_{2^1, \theta}| \]

\[ 2^0 \]

\[ 2^1 \]

\[ 2^J \]

Scale

Figure 2.3: Three Morlet wavelet families with different sets of parameters. For each set of parameters, we show, from left to right, the gaussian window \( \phi_J \), the Morlet wavelets \( \psi_{\theta, j} \), and the associated Littlewood-Paley sum \( A(\omega) \). When the number of scales \( J \) increases, so does the width of the low pass wavelet \( \phi_J \). When the number of orientations \( C \) increases or when the number of scales per octave \( Q \) decreases, the Morlet wavelets become more elongated in the direction perpendicular to the orientation, and hence have an increased angular sensitivity.
Figure 2.3: Three Morlet wavelet families with different sets of parameters. For each set of parameters, we show, from left to right, the gaussian window $\phi_J$, the Morlet wavelets $\psi_{2^j, \theta}$, and the associated Little Wood-Paley sum $A(\omega)$. When the number of scales $J$ increases, so does the width of the low pass wavelet $\phi_J$. When the number of orientations $C$ increases or when the number of scales per octave $Q$ decreases, the Morlet wavelets become more elongated in the direction perpendicular to their orientation, and hence have an increased angular sensitivity.
First wavelet transform

\[ |W_1|_x = \left( \begin{array}{c} x * \phi_{2J} \\ x * \psi_{\lambda_1}^{1} \\ x * \psi_{\lambda_1}^{1} \end{array} \right) \lambda_1 \]

Modulus improves invariance:

\[ |x * \psi_{\lambda_1}^{1} (t)| * \phi_{2J} (t) \]

Second wavelet transform modulus

\[ |W_2|_x = \left( \begin{array}{c} |x * \psi_{\lambda_1}^{1} | * \phi_{2J} (t) \\ |x * \psi_{\lambda_1}^{1} | \psi_{\lambda_2} (t) \end{array} \right) \lambda_2 \]
Wavelet Scattering Network

\[ S_J x = |W_J| \cdots |W_4| |W_3| |W_2| |W_1| x \]
Scattering Neuronal Network

\[ |W_1| \]

\[ x \star \psi_{\lambda_1}(t) \quad x \star \psi_{\lambda_2}(t) \quad x \star \psi_{\lambda_3}(t) \quad x \star \psi_{\lambda_4}(t) \]
Scattering Neuronal Network

\[ |W_1| \]

\[ |x * \psi_{\lambda_1} * \phi_{2J}| \]

\[ |W_2| \]

\[ ||x * \psi_{\lambda_1} * \psi_{\lambda_2}(t)|| \]
Scattering Neuronal Network

\[ |W_1| \]

\[ |x \ast \psi_{\lambda_1} \ast \phi_{2^j}| \]

\[ |W_2| \]

\[ ||x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \phi_{2^j}|| \]

\[ |W_3| \]

\[ |||x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \psi_{\lambda_3}||| \]
Wavelet Scattering

Scattering operator:

\[ S_J x = \begin{pmatrix} x \ast \phi_{2^J} \\ |x \ast \psi_{\lambda_1}| \ast \phi_{2^J} \\ ||x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \phi_{2^J} \\ |||x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \psi_{\lambda_3} \ast \phi_{2^J} \\ \vdots \end{pmatrix} \xrightarrow{J \to \infty} \begin{pmatrix} \int x(u)du \\ ||x \ast \psi_{\lambda_1}| \ast \psi_{\lambda_2}| \ast \phi_{2^J} \\ |||x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \psi_{\lambda_3} \ast \phi_{2^J} \\ \vdots \end{pmatrix} \]

Theorem: The total energy of coefficients converge to 0 as the depth (number of modulus) increases.
Scattering Properties

\[ S_J x = \begin{pmatrix} x \ast \phi_{2J} \\ |x \ast \psi_{\lambda_1} \ast \phi_{2J}| \\ |x \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} \ast \phi_{2J}| \\ |x \ast \psi_{\lambda_2} \ast \psi_{\lambda_2} \ast \psi_{\lambda_3} \ast \phi_{2J}| \\ \vdots \end{pmatrix} = \ldots |W_3| |W_2| |W_1| \ x \lambda_1, \lambda_2, \lambda_3, \ldots \]

**Theorem:** For appropriate wavelets, a scattering is

1. **contractive** \( \| S_J x - S_J y \| \leq \| x - y \| \) (\( L^2 \) stability)
2. **preserves norms** \( \| S_J x \| = \| x \| \)
3. **stable to deformations** \( D_\tau x(u) = x(u - \tau(u)) \)

\[ \| S_J D_\tau x - S_J x \| \leq C \left( \| \nabla \tau \|_\infty \| x \| + 2^{-J} \| \tau \|_\infty \right) \]

\[ J \to \infty \| S x - S x' \| \leq C \left( \min_\tau \| x - D_\tau x' \| + \| \nabla \tau \|_\infty \| x \| \right) \]
Image Classification

\[ x \rightarrow S_J x \rightarrow \text{Linear Classif. SVM} \rightarrow y \]

MNIST: 6 \(10^4\) chiffres

0.4% errors
\[ 2^J = 2^3 \]

CUREt
61 classes

0.2% errors
\[ 2^J = \text{image size} \]
The scattering transform of a stationary process $X(t)$

$$S_J X = \begin{pmatrix}
X \ast \phi_{2J} \\
|X \ast \psi_{\lambda_1}| \ast \phi_{2J} \\
||X \ast \psi_{\lambda_1}| \ast \psi_{\lambda_2}| \ast \phi_{2J} \\
|||X \ast \psi_{\lambda_2}| \ast \psi_{\lambda_2}| \ast \psi_{\lambda_3}| \ast \phi_{2J} \\
... \\
\end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, ...}$$

is a low-variance estimator of the scattering moments of $X(t)$

$$\overline{S}X = \begin{pmatrix}
E(X) \\
E(|X \ast \psi_{\lambda_1}|) \\
E(||X \ast \psi_{\lambda_1}| \ast \psi_{\lambda_2}|) \\
E(|||X \ast \psi_{\lambda_2}| \ast \psi_{\lambda_2}| \ast \psi_{\lambda_3}|) \\
... \\
\end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, ...}$$

and $S_X \xrightarrow{J \to \infty} \overline{S}X$ if $X$ is ergodic.

- But does $\overline{S}X$ "characterize" $X$?
Adapt Convolutions to Invariants

Laurent Sifre

Translation

\[ x(t) \rightarrow |W_1| \rightarrow |x \ast \psi_{j,\theta}(t)| = x_1(j, \theta, t) \rightarrow |W_2| \rightarrow |x_1(j, \theta, .) \ast \psi_j(t)| \]

1st order translation

\[ W_2 \text{ computes wavelet convolutions along } (t_1, t_2) \]

4D space

\[ t = (t_1, t_2) \]

\[ \theta \]
Rotation-Translation Invariance

Laurent Sifre

Translation

1st order

Roto-translation

$W_2$ computes wavelet convolutions along $(t_1, t_2, \theta)$
Scalo-Roto-Translation Invariance

Laurent Sifre

\[ x(t) \xrightarrow{\text{translation}} |W_1| \xrightarrow{x \ast \psi_{j,\theta}(t)} |x \ast \psi_{j,\theta}(t)| = x_1(j, \theta, t) \xrightarrow{1\text{st order}} |W_2| \xrightarrow{|x_1 \ast \overline{\psi}_{j',\nu}(j, \theta, t)|} \]

\( W_2 \) computes wavelet convolutions along \((t_1, t_2, \theta, j)\)

4D space
UIUC database: 25 classes

Scattering classification errors

<table>
<thead>
<tr>
<th>Training</th>
<th>Translation</th>
<th>Transl + Rotation</th>
<th>+ Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20 %</td>
<td>2%</td>
<td>0.6%</td>
</tr>
</tbody>
</table>
Complex Image Classification

CalTech 101 data-basis:  

<table>
<thead>
<tr>
<th>Data Basis</th>
<th>2012</th>
<th>Deep-Net</th>
<th>Scat.-1</th>
<th>Scat.-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CalTech-101</td>
<td>80%</td>
<td>85%</td>
<td>50%</td>
<td>80%</td>
</tr>
<tr>
<td>CalTech-256</td>
<td>50%</td>
<td>70%</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>80%</td>
<td>90%</td>
<td>55%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Classification Accuracy  \[2^J = 2^5\]
Complex Image Classification

CalTech 101 data-basis:

\[
\begin{align*}
S_J x & \quad \text{Roto-Trans.} \\
\text{Linear Classif.} & \quad y
\end{align*}
\]

Classification Accuracy \(2^J = 2^5\)

<table>
<thead>
<tr>
<th>Data Basis</th>
<th>2012</th>
<th>Deep-Net</th>
<th>Scat.-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CalTech-101</td>
<td>80%</td>
<td>85%</td>
<td>80%</td>
</tr>
<tr>
<td>CalTech-256</td>
<td>50%</td>
<td>70%</td>
<td>50%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>80%</td>
<td>90%</td>
<td>80%</td>
</tr>
</tbody>
</table>
Given $S_J x$ we want to compute $\tilde{x}$ such that:

$$S_J \tilde{x} = \begin{pmatrix} \tilde{x} \ast \phi_{2^J} \\ |\tilde{x} \ast \psi_{\lambda_1} | \ast \phi_{2^J} \\ \vdots \\ || \tilde{x} \ast \psi_{\lambda_1} \ast \ldots \ast \psi_{\lambda_m} \ast \phi_{2^J} \end{pmatrix}_{\lambda_1, \ldots, \lambda_m} = \begin{pmatrix} x \ast \phi_{2^J} \\ |x \ast \psi_{\lambda_1} | \ast \phi_{2^J} \\ \vdots \\ || x \ast \psi_{\lambda_1} \ast \ldots \ast \psi_{\lambda_m} \ast \phi_{2^J} \end{pmatrix}_{\lambda_1, \ldots, \lambda_m} = S_J x$$

with $||\tilde{x}||$ minimum. Non convex optimisation problem.

For $m = 1$ and $2^J = \infty$, minimize $||\tilde{x}||$ subject to:

$$\int \tilde{x}(u) \, du = \int x(u) \, du$$

$$\forall \lambda_1 \ , \ || \tilde{x} \ast \psi_{\lambda_1} ||_1 = || x \ast \psi_{\lambda_1} ||_1$$

If $x(u)$ is a Dirac, or a straight edge or a sinusoid then $\tilde{x}$ is equal to $x$ up to a translation.
With a gradient descent algorithm:

Original images of $N^2$ pixels:

$m = 1$, $2^J = N$: reconstruction from $O(\log_2 N)$ scattering coeff.

$m = 2$, $2^J = N$: reconstruction from $O(\log_2^2 N)$ scattering coeff.
Gaussian process model with same second order moments

\( m = 2, 2^J = N: \) reconstruction from \( O(\log_2^2 N) \) scattering coeff.
Multiscale Scattering Reconstructions

Original Images
$N^2$ pixels

Scattering Reconstruction
$2^J = 16$
$1.4 \times N^2$ coeff.

$2^J = 32$
$0.5 \times N^2$ coeff.

$2^J = 64$

$2^J = 128 = N$
Scattering Reconstructions

Original Images

Scat-2. Reconstr.

$2^J = 32$
Energy of $d$ interacting bodies:

Can we learn the interaction energy $f(x)$ of a system with $x = \{\text{positions, values}\}$?
• Classic energy of $d$ interacting bodies:

If $x(u) = \sum_{k=1}^{d} q_k \delta(u - p_k)$ then $f(x) = \sum_{k=1}^{d} \sum_{k'=1}^{d} \frac{q_k q_{k'}}{|p_k - p_{k'}|^{\beta}}$

Each particle interacts with $O(\log d)$ groups

Theorem: For any $\epsilon > 0$ there exists wavelets with

$$f(x) = \sum_{m=0}^{M} \sum_{\lambda_1, \ldots, \lambda_m} \alpha(\lambda_1, \ldots, \lambda_m) S^2 x(\lambda_1, \ldots, \lambda_m)(1 + \epsilon)$$
Quantum Chemistry

- Complex orbital interactions: no analytical energy \( f(x) \).
  Invariant to translations, rotations, stable to deformations.
- Data basis \( \{ x_i, f(x_i) \}_i \) of 700 2D molecules (about 20 atoms).
- Best \( M \)-term scattering approximation \( f_M \) of \( f \):

\[
\log \| f - f_M \| \approx C M^{-1/2} \ll M^{-1/d}
\]

\[
\| f_M - f \| \approx C M^{-1/2} \ll M^{-1/d}
\]

where the \( \phi_n(x) \) is a 1st or 2nd order term.

\[
M = 80 \quad \| f - f_M \| = 9 \text{kcal/mole}
\]

\( \log M \)
Conclusion

• Do we need to learn deep net filters?

• Can we analyse geometry in Euclidean spaces?

• How much physics can we learn and why?

Looking for Post-Doc!

www.di.ens.fr/data/scattering