Signal Recovery from Scattering Convolutional Networks

Joan Bruna
Dept. of Statistics, UC Berkeley

collaborators: Stephane Mallat (ENS), Yann LeCun (NYU), Pablo Sprechmann (NYU)
Complex data + Complex tasks

- compression, denoising
- source separation
- super resolution
- segmentation
- pattern recognition
- localization
- synthesis
- general object recognition
- synthesis
Complex data + Complex tasks

- Compression
- Denoising
- Source separation
- Super resolution
- Segmentation
- Pattern recognition
- Localization
- Synthesis
- General object recognition
- Pattern recognition
- Synthesis
Complex data + Complex tasks

- Compression
- Denoising
- Source separation
- Super resolution
- Segmentation
- Pattern recognition
- Localization
- General object recognition
- Synthesis

(SVHN)

(Mnist)

(from Aren Jensen)
Complex data + Complex tasks

- compression, denoising
- source separation
- super resolution
- segmentation
- pattern recognition
- localization
- general object recognition
- synthesis

(from Imagenet dataset)
Complex data + Complex tasks

- compression, denoising
- source separation
- super resolution
- segmentation
- pattern recognition
- localization
- general object recognition
- synthesis
- image captioning

Automatically captioned: “Two pizzas sitting on top of a stove top oven”

From Vinyals et al, CVPR’15
Complex Data + Complex tasks

• Spectrum of tasks with varying metric structure.
  – Metric invariances encoded into a non-linear signal representation
    \[ d(x, x') = \| \Phi(x) - \Phi(x') \| \]

• As we move towards the right, how much information do we lose? How to quantify what we keep/lose?
• Can we identify a “perceptual” metric?
Generative Models of Complex data

- $\Phi$ trained to reduce intra-class variability while preserving discriminability (e.g., a Deep Neural Network)
Generative Models of Complex data

• Sampling or Regressing in transformed space is easy

- Sampling
  e.g. $\sim \mathcal{N}(\mu_i, \Sigma_i)$

- high-dimensional space

class 1
class 2
class 3
Generative Models of Complex data

- How to perform high-dimensional density estimation via invariant representations?
- Applications to synthesis, inverse problems, unsupervised learning.
Plan

• Review of Scattering Convolutional Networks.

• Signal and Texture Recovery.

• Applications to high-dimensional Inverse Problems:
  – Synthesis,
  – Super-Resolution,
  – Audio Source Separation.
Geometric Variability Prior

\[ x(u),\ u: \text{pixels, time samples, etc.} \quad \tau(u),\ : \text{deformation field} \]

\[ L_\tau(x)(u) = x(u - \tau(u)) : \text{warping} \]

- Deformation “cost”: \( \|\tau\| = \lambda \sup_u |\tau(u)| + \sup_u |\nabla \tau(u)| \).
  - Model change in point of view in images
  - Model frequency transpositions in sounds
  - Consistent with local translation invariance
Geometric Variability Prior

- **Blur operator:** $Ax = x \ast \phi$, $\phi$: local average
- The only linear operator $A$ stable to deformations

$$\|AL_\tau x - Ax\| \leq \|\tau\|\|x\|.$$

[Bruna’12]
**Geometric Variability Prior**

- **Blur operator:** $Ax = x \ast \phi$, $\phi$: local average
  - The only linear operator $A$ stable to deformations:
    \[
    \|AL_\tau x - Ax\| \leq \|	au\|\|x\|
    \]  
    \[\text{[Bruna'12]}\]

- **Wavelet filter bank:** $Wx = \{x \ast \psi_k\}$, $\psi_k(u) = 2^{-j}\psi(2^{-j}R_\theta u)$
  \[\psi: \text{spatially localized band-pass filter.}\]
  \[W \text{ recovers information lost by } A.\]
**Geometric Variability Prior**

- **Blur operator**: $Ax = x \ast \phi$, $\phi$: local average
  - The only **linear** operator $A$ stable to deformations:
    $$\| AL_{\tau} x - Ax \| \leq \| \tau \| \| x \| .$$

- **Wavelet filter bank**: $W x = \{ x \ast \psi_k \}$, $\psi_k(u) = 2^{-j} \psi(2^{-j} R_{\theta} u)$
  - $\psi$: spatially localized band-pass filter.
  - $W$ recovers information lost by $A$.

- **Point-wise non-linearity** $\rho(x) = |x|$:
  - Commutes with deformations: $\rho L_{\tau} x = L_{\tau} \rho x$
  - **Demodulates** wavelet coefficients, preserves energy.

[Bruna’12]
Image and Audio descriptors

- **MFCC (audio)**  
  [Mermelstein, 76]

- **SIFT, Daisy**  
  [Lowe, 04, Fua et al’10]

- **ConvNets**  
  [LeCun et al, 98]
Image and Audio descriptors

- **MFCC (audio)**
  [Mermelstein, 76]

- **SIFT, Daisy**
  [Lowe, 04, Fua et al’10]

- **ConvNets**
  [LeCun et al, 98]

(figures from Jothilakshmi et al, Lowe, LeCun98)

[learnt 1st layer filters]
[Krizhevsky et al, 12]
Scattering Convolutional Network

Cascade of contractive operators.
Image Examples

Images

\[ f \]

\[ \hat{f} \]

Fourier

Wavelet Scattering

\[ |f \ast \psi_{\lambda_1} | \ast \phi \]

\[ ||f \ast \psi_{\lambda_1} \ast \psi_{\lambda_2} | \ast \phi \]

SIFT

window size = image size

[Bruna, Mallat, '11, '12]
Theorem: [Mallat ’10] With appropriate wavelets, $S_J$ is stable to additive noise,

$$\|S_J(x + n) - S_Jx\| \leq \|n\|,$$

unitary, $\|S_Jx\| = \|x\|$, and stable to deformations

$$\|S_Jx_\tau - S_Jx\| \leq C\|x\|\|\nabla \tau\|.$$

$x_\tau$ $|\hat{x}_\tau|$ $S_Jx_\tau$
$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)
Representation of Stationary Processes

\[ x(u): \text{realizations of a stationary process } X(u) \text{ (not Gaussian)} \]

\[ \Phi(X) = \{ E(f_i(X)) \}_i \]

Estimation from samples \( x(n) \): \( \hat{\Phi}(X) = \left\{ \frac{1}{N} \sum_n f_i(x)(n) \right\}_i \)

Discriminability: need to capture high-order moments

Stability: \( E(\| \hat{\Phi}(X) - \Phi(X) \|^2) \) small
Scattering Moments

$X$
Scattering Moments

$|W_J| \rightarrow X \rightarrow E(X)$

$|X \ast \psi_{j_1,\gamma_1}|$
Scattering Moments

\[ |W_J| \]

\[ |X \ast \psi_{j_1, \gamma_1}| \]

\[ |X \ast \psi_{j_1, \gamma_1} \ast \psi_{j_2, \gamma_2}| \]

\[ E(|X \ast \psi_{j_1, \gamma_1}|), \forall j_1, \gamma_1 \]

\[ E(X) \]
Scattering Moments

\[ |W_J| \]

\[ X \]

\[ E(X) \]

\[ |W_J| \rightarrow |X \ast \psi_{j_1, \gamma_1}| \rightarrow E(|X \ast \psi_{j_1, \gamma_1}|), \forall j_1, \gamma_1 \]

\[ |W_J| \rightarrow |X \ast \psi_{j_1, \gamma_1}| \ast \psi_{j_2, \gamma_2} \rightarrow E(|X \ast \psi_{j_1, \gamma_1}| \ast \psi_{j_2, \gamma_2}), \forall j_i, \gamma_i \]

\[ |W_J| \rightarrow \ldots \rightarrow |X \ast \psi_{j_1, \gamma_1}| \ast \ldots \ast \psi_{j_m, \gamma_m} \rightarrow E(|X \ast \psi_{j_1, \gamma_1}| \ast \ldots \ast \psi_{j_m, \gamma_m}), \forall j_i, \gamma_i \]

\[ |W_J| \rightarrow \ldots \rightarrow |X \ast \psi_{j_1, \gamma_1}| \ast \ldots \ast \psi_{j_{m+1}, \gamma_{m+1}} \]
Properties of Scattering Moments

- Captures high order moments:

\[ S_J[p] \hat{X} \]

Power Spectrum

\[ m = 1 \]

\[ m = 2 \]

[Bruna, Mallat, ’11,’12]
Properties of Scattering Moments

- Captures high order moments:
  \( m = 1 \)
  \( S_J[p]X \)
  \( m = 2 \)

  [Bruna, Mallat, ’11,’12]

- Cascading non-linearities is **necessary** to reveal higher-order moments.
Theorem: [B’15] If $\psi$ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - S X\|^2) = 0.$$
Consistency of Scattering Moments

**Theorem: [B’15]** If $\psi$ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - S X\|^2) = 0 .$$

**Corollary:** If moreover $X(t)$ is bounded, then

$$E(\|\hat{S}_N X - S X\|^2) \leq C \frac{|X|_\infty^2}{\sqrt{N}} .$$

- Although we extract a growing number of features, their global variance goes to 0.
- No variance blow-up due to high order moments.
- Adding layers is critical (here depth is $\log(N)$).
Classification with Scattering

• **State-of-the-art on pattern and texture recognition:**
  - MNIST [Pami’13]
  - Texture (CUREt, UIUC) [Pami’13]

• **Object Recognition:**
  - ~17% error on Cifar-10 [Oyallon&Mallat, CVPR’15]
  - General Object Recognition requires adapting the wavelets to the signal classes. Learning is necessary.
Signal and Texture Recovery Challenge

\[ S_J x = \{ x \ast \phi_J, |x \ast \psi_{j_1}| \ast \phi_J, ||x \ast \psi_{j_1}| \ast \psi_{j_2}| \ast \phi_J, \cdots \}_{j_i \leq J} \]

• [Q1] Given \( S_J x \) computed with \( m \) layers, under what conditions can we recover \( x \) (up to global symmetry)? Using what algorithm? As a function of the localization scale \( J \) ?
Signal and Texture Recovery Challenge

\[ S_J x = \{ x * \phi_J, |x * \psi_{j_1}| * \phi_J, ||x * \psi_{j_1}| * \psi_{j_2}| * \phi_J, \ldots \}_{j_i \leq J} \]

• [Q1] Given \( S_J x \) computed with \( m \) layers, under what conditions can we recover \( x \) (up to global symmetry)? Using what algorithm? As a function of the localization scale \( J \)?

\[ S x = \{ E(X), E(|X * \psi_{j_1}|), E(||X * \psi_{j_1}| * \psi_{j_2}|), \ldots \} \]

• [Q2] Given \( S x \), how can we characterize interesting processes? How to sample from such distributions?
Related Work

• [Q1] As $J \rightarrow \infty$, with depth fixed to $m$, we have measurements

  -- **Non-linear, invariant** compressed sensing.
  -- Eldar et al ['12]: Sparse Recovery from Fourier Magnitude
  -- Plan and Vershynin ['14]: Generalized Linear Model, 1-bit compressed sensing.

\[ O(|\log N|^m) \ll N \]
Related Work

• [Q1] As $J \to \infty$, with depth fixed to $m$, we have measurements
  
  \[ O(\left| \log N \right|^m) \ll N \]

  – **Non-linear, invariant** compressed sensing.
  – Eldar et al [’12]: Sparse Recovery from Fourier Magnitude
  – Plan and Vershynin [’14]: Generalized Linear Model, 1-bit compressed sensing.

• [Q1] For fixed $J$, it is a generalized phase-recovery problem
  – Balan et al [’06], Candes et al. [’11], Waldspurger et al [’12]: Phasecut
  – Bruna et al [’14]: Signal Recovery from lp pooling.
Related Work

• [Q1] As $J \to \infty$, with depth fixed to $m$, we have measurements

  - **Non-linear, invariant** compressed sensing.
  - Eldar et al ['12]: Sparse Recovery from Fourier Magnitude
  - Plan and Vershynin ['14]: Generalized Linear Model, 1-bit compressed sensing.

• [Q1] For fixed $J$, it is a generalized phase-recovery problem

  - Balan et al ['06], Candes et al. ['11], Waldspurger et al ['12]: Phasecut
  - Bruna et al ['14]: Signal Recovery from $l_p$ pooling.

• [Q2] Texture synthesis

  - Simoncelli & Portilla ['00], Simoncelli & McDermott ['11], Mumford et al ['98]: define statistical models using generalized wavelet moments.
  - Peyre et al ['14]: models on learnt dictionaries, Effros&Freeman ['01] Quilting

\[ O(|\log N|^m) \ll N \]
Problem Set-Up

• Given $y = S_J x_0$, (fixed $J$, fixed depth) consider

$$\min_x \|S_J x - y\|^2.$$ 

• When $J = \log N$, intersection of mixed $\ell_{1,2}$ balls:

$$\|x\|_1$$

$$\forall j_1, \|x \ast \psi_{j_1}\|_1$$

$$\forall j_1, j_2, \|\|x \ast \psi_{j_1} \ast \psi_{j_2}\|_1$$

• Non-convex optimization problem.
Theorem [B,M’14]: Suppose \( x_0(t) = \sum_n a_n \delta(t - b_n) \) with \(|b_n - b_{n+1}| \geq \Delta\), and \( S_Jx_0 = S_Jx \) with \( m = 1 \) and \( J = \infty \). If \( \psi \) has compact support, then

\[
x(t) = \sum_n c_n \delta(t - e_n), \quad \text{with } |e_n - e_{n+1}| \gtrsim \Delta.
\]
**Theorem [B,M’15]:** Suppose \( x_0(t) = \sum_n a_n \delta(t-b_n) \) with \(|b_n - b_{n+1}| \geq \Delta\), and \( \|x\|_1 = \|x_0\|_1 \), \( \|x * \psi_j\|_1 = \|x_0 * \psi_j\|_1 \) for all \( j \). If \( \psi \) has compact support, then

\[
x(t) = \sum_n c_n \delta(t-e_n), \quad \text{with} \quad |e_n - e_{n+1}| \gtrsim \Delta.
\]

- \( Sx \) essentially identifies sparse measures, up to log spacing factors.
- Here, sparsity is encoded in the measurements themselves.
- In 2D, singular measures (ie curves) require \( m = 2 \) to be well characterized.
Theorem [B,M’14]: Suppose $\hat{x}_0(\xi) = \sum_n a_n \delta(\xi - b_n)$ with $|\log b_n - \log b_{n+1}| \geq \Delta$, and $S_J x = S_J x_0$ with $m = 2$ and $J = \log N$. If $\hat{\psi}$ has compact support $K \leq \Delta$, then

$$\hat{x}(\xi) = \sum_n c_n \delta(\xi - e_n), \text{ with } |\log e_n - \log e_{n+1}| \geq \Delta.$$ 

- Oscillatory, lacunary signals are also well captured with the same measurements.
- It is the opposite set of extremal points from previous result.
Scattering Reconstruction Algorithm

\[ x_0 \sim \mathcal{N}(0, I) \]

\[ S = \{x \text{ s.t. } \hat{S}x = \hat{S}_0 \} \]

\[ \min_x \| \hat{S}x - \hat{S}_0 \|^2 \]

- Non-linear Least Squares.
- Levenberg-Marquardt gradient descent:

\[ x_{n+1} = x_n - \gamma (D\hat{S}x_n)^\dagger (\hat{S}x_n - \hat{S}_0) \]
Scattering Reconstruction Algorithm

$x_0 \sim \mathcal{N}(0, \mathbf{I})$

\[ S = \{x \text{ s.t. } \hat{S}x = \hat{S}_0 \} \]

\[
\min_x \|\hat{S}x - \hat{S}_0\|^2
\]

• Non-linear Least Squares.
  
  • Levenberg-Marquardt gradient descent:

  \[
x_{n+1} = x_n - \gamma (D\hat{S}x_n)^\dagger (\hat{S}x_n - \hat{S}_0)
\]

• Global convergence guarantees using complex wavelets:

  \( D\hat{S}x \) is full rank for \( m = 2 \) if \( x \) compact support.
Sparse Shape Reconstructions

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.
Multiscale Scattering Reconstruction

• For finite J and finite m, recovery depends on redundancy factor:
  \[ \dim(S_J x) = O(N2^{-2J} J^m) \]

• As J increases, redundancy decreases.

• No universal recovery guarantees.

• We use the same gradient descent algorithm.
Multiscale Scattering Reconstruction

Original Images
$N^2$ pixels

Scattering Reconstruction

$2^J = 16$
$1.4 N^2$ coeff.

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

$2^J = 64$

$2^J = 128 = N$
Related Work on CNN inversion

- Recently, interest in inverting Deep Convolutional Networks
  - The Learnt Representations are highly contractive: recovery is more “impressionistic”:

Reconstructions from a 5-layer CNN
(from Mahendran&Vedaldi, ’15)
Texture Synthesis

- Maximum Entropy Distribution from Scattering Moments: by Boltzmann Theorem, we have

\[ p(x) = \frac{1}{Z} e^{\sum_{|p| \leq m} \lambda_p (U[p]x \ast \phi_J)(0)} \]

- \( \lambda_p \) are Lagrange multipliers that guarantee that \( E_p(U[p]x) = \hat{S}X(p) \).
Texture Synthesis

• Maximum Entropy Distribution from Scattering Moments: by Boltzmann Theorem, we have

\[ p(x) = \frac{1}{Z} e^{\sum_{|\mathbf{p}| \leq m} \lambda_{\mathbf{p}}(\mathbf{U}[\mathbf{p}]x \ast \phi_J)(0)} \]

• \( \lambda_{\mathbf{p}} \) are Lagrange multipliers that guarantee that \( E_{\mathbf{p}}(\mathbf{U}[\mathbf{p}]x) = \hat{S}X(\mathbf{p}) \).

• When \( X(t) \) is ergodic, this distribution converges to the uniform measure on the set (the Julesz ensemble):

\[ \Omega(SX) = \{ x \text{ s.t. } \mathbf{U}[\mathbf{p}]x = SX(\mathbf{p}) \forall \mathbf{p} \} . \]

• Convergence in distribution is a hard problem (cf Chatterjee)

• We can sample approximately using previous algorithm.
Ergodic Texture Reconstruction

Original Textures

Gaussian process model with same second order moments

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

- Scattering Moments of 2nd order thus capture essential geometric structures with only $O((\log N)^2)$ coefficients.
- However, not all texture geometry is captured.
- Results using a deep VGG network from [Gathys et al, NIPS’15]
Ergodic Texture Reconstruction

• Scattering Moments of 2nd order thus capture essential geometric structures with only $O((\log N)^2)$ coefficients.
• However, not all texture geometry is captured.
• Results using a deep VGG network from [Gathys et al, NIPS’15]
Application: Super-Resolution

• Best Linear Method: Least Squares estimate (linear interpolation): $\hat{y} = (\hat{\Sigma}_x^\dagger \hat{\Sigma}_{xy})x$
Application: Super-Resolution

• Best Linear Method: Least Squares estimate (linear interpolation): \( \hat{y} = (\hat{\Sigma}_x^\dagger \hat{\Sigma}_{xy}) x \)

• State-of-the-art Methods:
  – Dictionary-learning Super-Resolution
  – CNN-based: Just train a CNN to regress from low-res to high-res.
  – They optimize cleverly a fundamentally unstable metric criterion:

\[
\Theta^* = \arg \min_{\Theta} \sum_i \| F(x_i, \Theta) - y_i \|^2, \quad \hat{y} = F(x, \Theta^*)
\]
Scattering Approach

- Relax the metric:

\[ S \xrightarrow{F} y \]

\[ S^{-1} \]

\[ x \]
Scattering Approach

- Relax the metric:
  - Start with simple linear estimation on scattering domain.
  - Deformation stability gives more approximation power in the transformed domain via locally linear methods.
  - The method is not necessarily better in terms of PSNR!
Some Numerical Results

Original  |  Best Linear Estimate  |  Scattering Estimate  |  state-of-the-art
Conclusions

• Geometric encoding with deformation stability
  – Convolutional Networks are good representations

• Inverse Scattering is a generalized phase recovery
  – Efficiently solved using back propagation

• Maximum Entropy Scattering Distributions
  – Capture non-gaussian properties

• Learning a metric contraction can break the curse of dimensionality.
Audio Source Separation
(joint work with P. Sprechmann and Y. LeCun, ICLR’15)

• Suppose we observe \( y(t) = x_1(t) + x_2(t) \).

• Goal: Estimate \( x_1(t), x_2(t) \).

• Ill-posed inverse problem. We need to impose structure in our estimates \( \hat{x}_1(t), \hat{x}_2(t) \).

• Different learning set-ups:
  – Blind/No learning: Construct priors via time-frequency local regularity ([Wolf et al,’14]).
  – Non-discriminative: We observe each source separately, learn a model of each source.
  – Discriminative: We train directly with input mixtures.
Audio Source Separation

- State-of-the-art methods:
  - $D$ is a synthesis operator, trained to estimate $\Phi x_i$ from $\Phi y$.
    - Non-negative Matrix Factorization
      \[
      \min_{z_i} \| \Phi y - \sum D_i z_i \|^2 + \lambda \left( \sum \| z_i \|_1 \right).
      \]
      - Can be trained either non-discriminative or discriminative.
    - DNN/ RNN / LSTM: $D$ is modeled as a Neural Net trained discriminatively.
  - $\Phi^{-1}$ is approximately linear if $\Delta$ small.
  - Long temporal structure is imposed on the $D$. 

\[
\begin{align*}
\Phi \quad \rightarrow \quad D \quad \rightarrow \quad \Phi^{-1} \quad \rightarrow \\
\hat{x}_1(t) \quad \quad \quad \quad \hat{x}_2(t)
\end{align*}
\]
Multi-Resolution Scattering Source Sep.

• Rather than adding structure to the unstable synthesis block, replace the analysis with a more invariant one.

• We use a multi-resolution pyramid CNN analysis $\Phi$
  – **Pros**: We relieve the synthesis from having to model uninformative variability.
  – **Pros**: The wavelets can be replaced by a learnt linear transformation that preserves informations.
  – **Cons**: Phase Recovery is more expensive, but approximate linear inverse still works well in practice.
Results on TIMIT

- 64 Speakers, gender-specific models.

<table>
<thead>
<tr>
<th></th>
<th>SDR</th>
<th>SIR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>scatt-NMF(1)</td>
<td>6.2 [2.8]</td>
<td>13.5 [3.5]</td>
<td>7.8 [2.2]</td>
</tr>
<tr>
<td>scatt-NMF(2)</td>
<td>6.9 [2.7]</td>
<td>16.0 [3.5]</td>
<td>7.9 [2.2]</td>
</tr>
<tr>
<td>CQT-DNN-1 frame</td>
<td>9.4 [3.0]</td>
<td>17.7 [4.2]</td>
<td>10.4 [2.6]</td>
</tr>
<tr>
<td>CQT-DNN-5 frame</td>
<td>9.2 [2.8]</td>
<td>17.4 [4.0]</td>
<td>10.3 [2.4]</td>
</tr>
<tr>
<td>CQT-CNN-scatt</td>
<td><strong>9.9 [3.1]</strong></td>
<td><strong>19.8 [4.2]</strong></td>
<td><strong>10.6 [2.8]</strong></td>
</tr>
</tbody>
</table>

- Learning long-range dependency with multi scale as an alternative to recurrent architectures.
Thank you!