Image classification with Scattering Transform and Dictionary learning

- https://openreview.net/forum?id=SJxWS64FwH
- J. Zarka, L. Thiry, T. Angles, S. Mallat
- Accepted at ICLR 2020
- Pytorch code soon published
Digits classification

- MNIST database

- Invariance to translations, stability to deformations
$l_2$ metric Instability to translations
Local averaging
Stability to geometric transformations

Convolution with Gaussian kernel $\phi_J$:

- stable to geometric deformations
- dimensionality reduction via subsampling
- lots of details are lost
Preserving signal information

Recover information lost in averaging

Gabor wavelets $\psi_{j,\theta}$

Stability to geometric transformations

$|x \ast \psi_{j,\theta}| \ast \phi_{J}$

subsampling
Scattering transform

Mallat (2011), Mallat, Bruna (2012)

Theorem

\[ \| Sx_\tau - Sx \| \leq K \| x \| \| \nabla \tau \|_\infty \]
### Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>&gt;99 %</td>
<td>&gt;99 %</td>
<td>&gt;99 %</td>
</tr>
<tr>
<td>28² digit images</td>
<td>10 classes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Example MNIST images:
Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>84.7 %</td>
<td>89.1 %</td>
<td>95.5 %</td>
</tr>
</tbody>
</table>

CIFAR-10: 32^2 object images, 10 classes
## Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>61.4 %</td>
<td>79.1 %</td>
<td>94.2 %</td>
</tr>
</tbody>
</table>

- **ImageNet**
  - 224² object images
  - 1000 classes

### Scattered vs Deep ConvNets

- **Scattering Transform**
- **AlexNet**
- **ResNet**

### AlexNet

### ResNet
Scattering : ImageNet classification

- RGB Images : Scattering Transform on each color channel
- Scale J=4, 8 angles, order 2
- $Sx[i, j]$ is a vector representing a patch of size 16x16x3
- 2 hidden layer classifier (MLP) as in AlexNet
- 38.1 % top1, 61.4 % top5 accuracy
AlexNet
Krizhevsky et al. 2012
79.1 % top5 accuracy

First layer learned filters
ResNet
He et al. 2016
94.2% top5 accuracy

- skip connections
- up to 152 convolutional layers

First layer learned filters
Research directions

What are in the convolutional layers of a Deep Networks?
→ Visualizing and Understanding Convolutional Networks, Zeiler, Fergus 2014

What’s needed to fill the gap between Scattering and DeepNets?
→ Sparse coding hypothesis
**$l_1$ sparse coding hypothesis**

Non negative sparse coding

$$
\alpha^0_*(D, \epsilon, x) = \arg\min_{\alpha \geq 0, \|D\alpha - x\| < \epsilon} \|\alpha\|_0
$$

Convex relaxation with $l_1$ norm (basis pursuit)

Chen, Donoho et al. 2001

$$
\alpha_*(D, \lambda, x) = \arg\min_{\alpha \geq 0} \mathcal{L}(\alpha), \quad \mathcal{L}(\alpha) = \|D\alpha - x\|_2^2 + \lambda \|\alpha\|_1
$$

Positive Iterated Soft Theshholding algorithm (ISTA)

Daubechies et al. 2003

$$
\alpha_0 = 0, \alpha_{n+1} = \text{ReLU} \left( (Id - \frac{1}{L} D^T D) \alpha_n + \frac{1}{L} D^T x - \frac{\lambda}{L} \right)
$$

Convolutional version with Scattering transform

$$
\alpha_{n+1}[i, j] = \text{ReLU} \left( (Id - \frac{1}{L} D^T D) \alpha_n[i, j] + \frac{1}{L} D^T Sx[i, j] - \frac{\lambda}{L} \right)
$$
Supervised dictionary learning + LISTA
Mairal et al. (2008), Gregor and Lecun (2011)

Principle

$$\min_{D,W,\alpha \geq 0} C(y, f(\alpha, W)) + \lambda_0 \|D\alpha - Sx\|_2^2 + \lambda_1 \|\alpha\|_1$$

example: $$C(y, f(\alpha, W)) = \|W^T \alpha - y\|_2^2$$

Convolutional LISTA with N iterations

$$\alpha_0[i, j] = 0, \quad \alpha_{n+1}[i, j] = \text{ReLU}(U\alpha_n[i, j] + VSx[i, j] - \lambda_{n+1})$$

No guarantees that the output $$\alpha_N$$ is close to $$\alpha_*$$

$$\alpha_*(D, \lambda, Sx) = \arg\min_{\alpha \geq 0} \mathcal{L}(\alpha), \quad \mathcal{L}(\alpha) = \|D\alpha - Sx\|_2^2 + \lambda \|\alpha\|_1$$
Task Driven dictionary learning + ISTC
Mairal et al. (2011), ours

Principle

\[
\min_{D, W, \alpha \geq 0} C(y, f(\alpha^*, (D, \lambda, x), W))
\]

Iterative Soft Treesholding with continuation

\[
\alpha_{n+1}[i, j] = \text{ReLU} \left( (I_d - \frac{1}{L} D^T D)\alpha_n[i, j] + \frac{1}{L} D^T Sx[i, j] - \lambda_\infty \gamma^n \right)
\]

Theorem

- \( s \) support size of \( \alpha^* \)
- \( \mu = \max_{m \neq m'} \langle D_m, D_{m'} \rangle \)

If \( s\mu \leq 1/2 \) and \( 2s\mu < \gamma < 1 \):

\[
\|\alpha_n - \alpha^*\|_\infty \leq K \gamma^n
\]
Scattering + ISTC classification

Implementation in a deep convolutional network

- $D$, $\lambda$, $W$ optimized by stochastic gradient descent to minimize the classification loss
- gradients computed by backpropagation

Results

<table>
<thead>
<tr>
<th>$l_1$ algo</th>
<th>ISTC $\alpha$</th>
<th>ISTC $D\alpha$</th>
<th>LISTA $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier input</td>
<td>59.5</td>
<td>55.3</td>
<td>62.9</td>
</tr>
<tr>
<td>Top 1</td>
<td>81.3</td>
<td>78.3</td>
<td>83.9</td>
</tr>
</tbody>
</table>
Convergence analysis

\[
\frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 1.01
\]

\[
\frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 3.8
\]
Comments

- Improvement of 20% over Scattering alone
- Large factor $\lambda_*$, reconstruction error $\|Sx - D\alpha\| / \|Sx\| = 0.5$
- Hard to reconstruct the original image from $\alpha$
- Classification works with the « denoised » $D\alpha$
  - 55% top1, 78% top5
- Atoms $D_m$ are in Scattering space, can not be visualised like usual dictionary atoms
- Sparse coding algorithm is not crucial (ISTC, FISTA, LARS)
- Still far from ResNet performance
Questions ?
Image classification with Scattering Transform and Dictionary learning

- https://openreview.net/forum?id=SJxWS64FwH
- J. Zarka, L. Thiry, T. Angles, S. Mallat
- Accepted at ICLR 2020
- Pytorch code soon published
Digits classification

• MNIST database

• Invariance to translations, stability to deformations
$L_2$ metric Instability to translations
Local averaging
Convolution with Gaussian kernel $\phi_J$:
- stable to geometric deformations
- dimensionality reduction via subsampling
- lots of details are lost
Preserving signal information

Recover information lost in averaging

Gabor wavelets $\psi_{j,\theta}$

Stability to geometric transformations

Subsampling
Scattering transform
Mallat (2011), Mallat, Bruna (2012)

Theorem

$$\| S_{\tau} x - S x \| \leq K \| x \| \| \nabla \tau \|_\infty$$
## Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>&gt;99 %</td>
<td>&gt;99 %</td>
<td>&gt;99 %</td>
</tr>
<tr>
<td>28² digit images</td>
<td>10 classes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32^2 object images 10 classes</td>
<td>84.7 %</td>
<td>89.1 %</td>
<td>95.5 %</td>
</tr>
</tbody>
</table>

The table above compares the performance of Scattering Transform, AlexNet, and ResNet on the CIFAR-10 dataset. Scattering Transform achieves 84.7%, AlexNet achieves 89.1%, and ResNet achieves 95.5% accuracy.
## Scattering vs Deep ConvNets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scattering Transform</th>
<th>AlexNet</th>
<th>ResNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>61.4 %</td>
<td>79.1 %</td>
<td>94.2 %</td>
</tr>
<tr>
<td>224x224 object images, 1000 classes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Images of fish and sharks are used to illustrate the concepts.*
- RGB Images: Scattering Transform on each color channel
- Scale J=4, 8 angles, order 2
- $Sx[i, j]$ is a vector representing a patch of size 16x16x3
- 2 hidden layer classifier (MLP) as in AlexNet
- 38.1% top1, 61.4% top5 accuracy
AlexNet
Krizhevsky et al. 2012
79.1% top5 accuracy

First layer learned filters
ResNet
He et al. 2016
94.2% top5 accuracy
- skip connections
- up to 152 convolutional layers

First layer learned filters
Research directions

What are in the convolutional layers of a Deep Networks?
→ Visualizing and Understanding Convolutional Networks, Zeiler, Fergus 2014

What’s needed to fill the gap between Scattering and DeepNets?
→ Sparse coding hypothesis
I\(_1\) sparse coding hypothesis

**Non negative sparse coding**

\[ \alpha^0_+ (D, \epsilon, x) = \arg\min_{\alpha \geq 0, \|D\alpha - x\|_2 < \epsilon} \|\alpha\|_0 \]

**Convex relaxation with I\(_1\) norm (basis pursuit)**

Chen, Donoho et al. 2001

\[ \alpha_* (D, \lambda, x) = \arg\min_{\alpha \geq 0} \mathcal{L}(\alpha), \quad \mathcal{L}(\alpha) = \|D\alpha - x\|_2^2 + \lambda \|\alpha\|_1 \]

**Positive Iterated Soft Theshholding algorithm (ISTA)**

Daubechies et al. 2003

\[ \alpha_0 = 0, \alpha_{n+1} = \text{ReLU} \left( (Id - \frac{1}{L} D^T D) \alpha_n + \frac{1}{L} D^T x - \frac{\lambda}{L} \right) \]

**Convolutional version with Scattering transform**

\[ \alpha_{n+1}[i, j] = \text{ReLU} \left( (Id - \frac{1}{L} D^T D) \alpha_n [i, j] + \frac{1}{L} D^T Sx[i, j] - \frac{\lambda}{L} \right) \]
Supervised dictionary learning + LISTA
Mairal et al. (2008), Gregor and Lecun (2011)

Principle

\[
\min_{D, W, \alpha \geq 0} C(y, f(\alpha, W)) + \lambda_0 \|D\alpha - Sx\|_2^2 + \lambda_1 \|\alpha\|_1
\]

example: \(C(y, f(\alpha, W)) = \|W^T\alpha - y\|_2^2\)

Convolutional LISTA with N iterations

\[
\alpha_0[i, j] = 0, \quad \alpha_{n+1}[i, j] = \text{ReLU} \left( U\alpha_n[i, j] + VSx[i, j] - \lambda_{n+1} \right)
\]

No guarantees that the output \(\alpha_N\) is close to \(\alpha_*\)

\[
\alpha_*(D, \lambda, Sx) = \arg\min_{\alpha \geq 0} \mathcal{L}(\alpha), \quad \mathcal{L}(\alpha) = \|D\alpha - Sx\|_2^2 + \lambda\|\alpha\|_1
\]
Task Driven dictionary learning + ISTC

Mairal et al. (2011), ours

Principle

\[ \min_{D, W, \alpha \geq 0} C(y, f(\alpha, D, \lambda, x), W) \]

Iterative Soft Thresholding with continuation

\[ \alpha_{n+1}[i, j] = \text{ReLU} \left( (I_d - \frac{1}{L}D^T D)\alpha_n[i, j] + \frac{1}{L}D^T Sx[i, j] - \lambda_\infty \gamma^n \right) \]

Theorem

\[ \cdot \text{s support size of } \alpha^* \]
\[ \cdot \mu = \max_{m \neq m'} \langle D_m, D_{m'} \rangle \]

If \( s \mu \leq 1/2 \) and \( 2s \mu < \gamma < 1 \):

\[ \| \alpha_n - \alpha^* \|_\infty \leq K \gamma^n \]
Scattering + ISTC classification

Implementation in a deep convolutional network

- $D$, $\lambda$, $W$ optimized by stochastic gradient descent to minimize the classification loss
- gradients computed by backpropagation

**Results**

<table>
<thead>
<tr>
<th>$l_1$ algo</th>
<th>ISTC</th>
<th>ISTC</th>
<th>LISTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier input</td>
<td>$\alpha$</td>
<td>$D\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Top 1</td>
<td>59.5</td>
<td>55.3</td>
<td>62.9</td>
</tr>
<tr>
<td>Top 5</td>
<td>81.3</td>
<td>78.3</td>
<td>83.9</td>
</tr>
</tbody>
</table>
Convergence analysis

**ISTC**

\[ \frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 1.01 \]

**LISTA**

\[ \frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 3.8 \]
Comments

- Improvement of 20% over Scattering alone
- Large factor $\lambda_x$, reconstruction error $\frac{\| Sx - D\alpha \|}{\| Sx \|} = 0.5$
- Hard to reconstruct the original image from $\alpha$
- Classification works with the « denoised » $D\alpha$
  - 55 % top1, 78 % top5
- Atoms $D_m$ are in Scattering space, can not be visualised like usual dictionary atoms
- Sparse coding algorithm is not crucial (ISTC, FISTA, LARS)
- Still far from ResNet performance
Questions ?