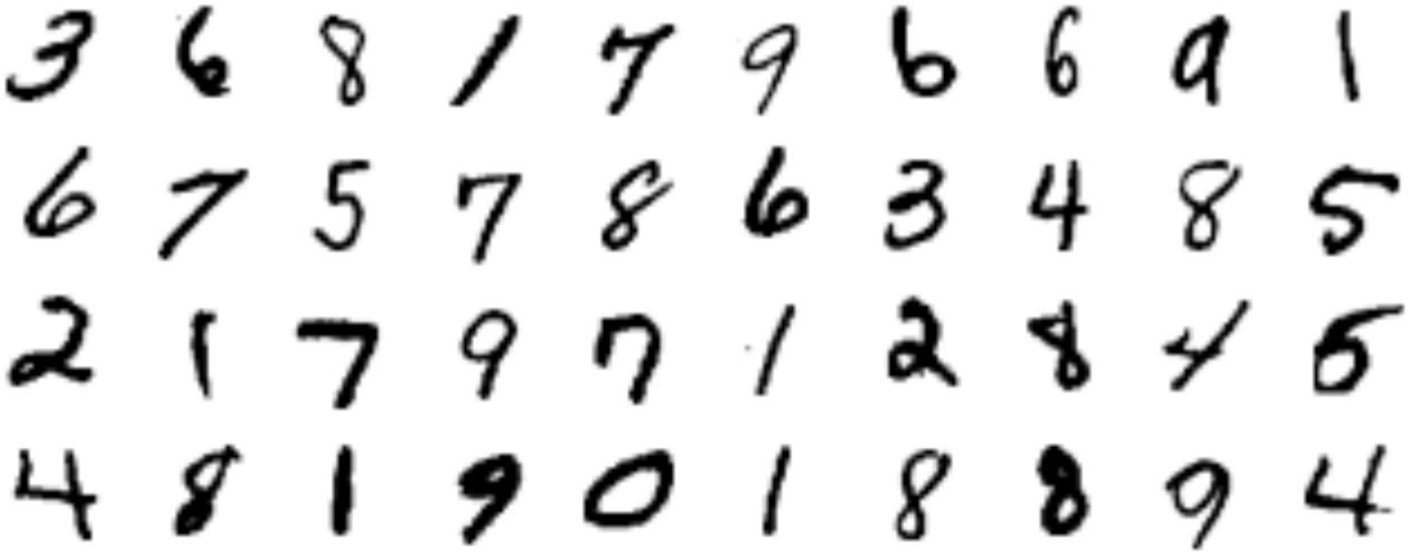
Image classification with Scattering Transform and Dictionary learning

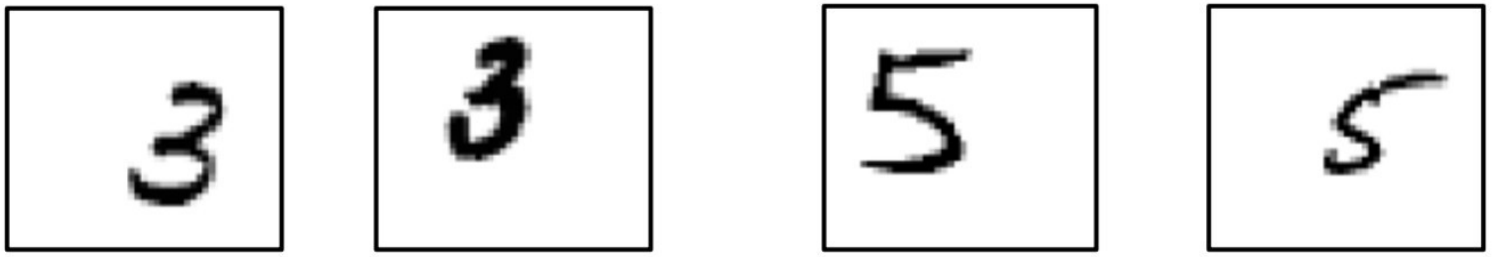
- https://openreview.net/forum?id=SJxWS64FwH
- J. Zarka, L. Thiry, T. Angles, S. Mallat
- Accepted at ICLR 2020
- Pytorch code soon published

Digits classification

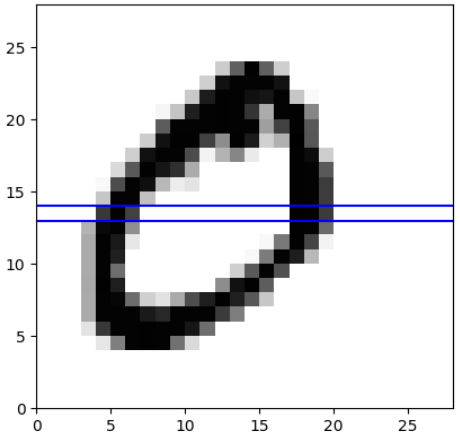
- MNIST database

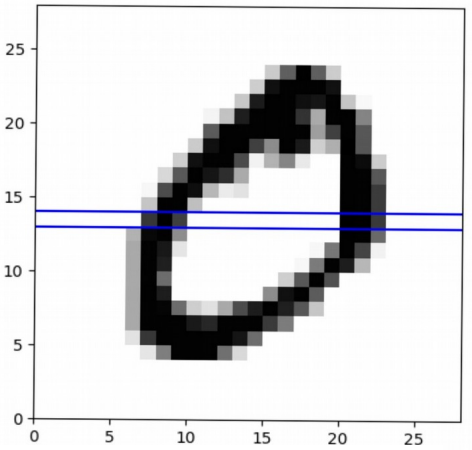


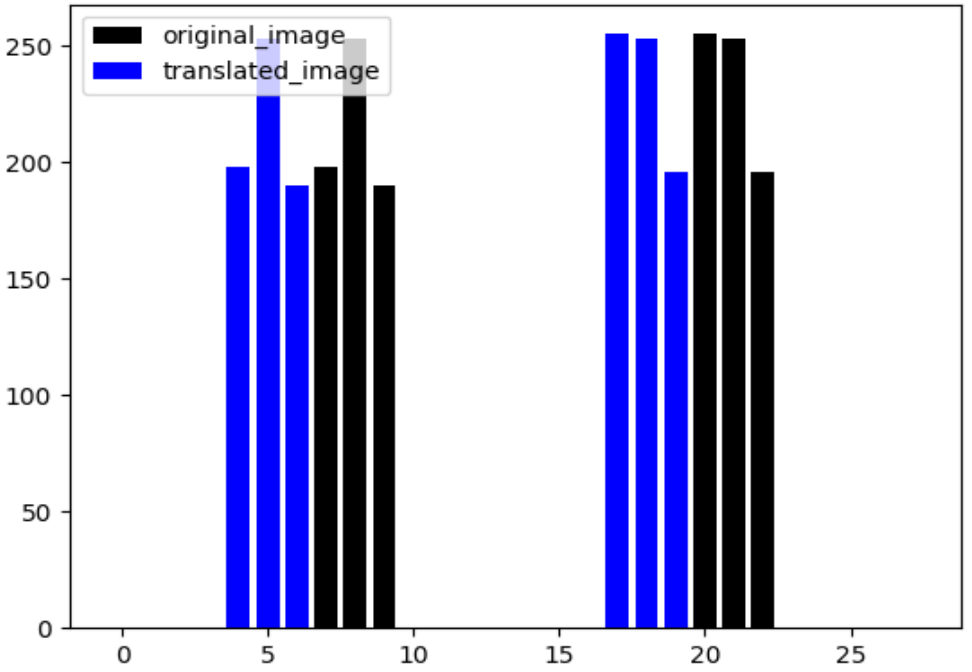
- Invariance to translations, stability to deformations



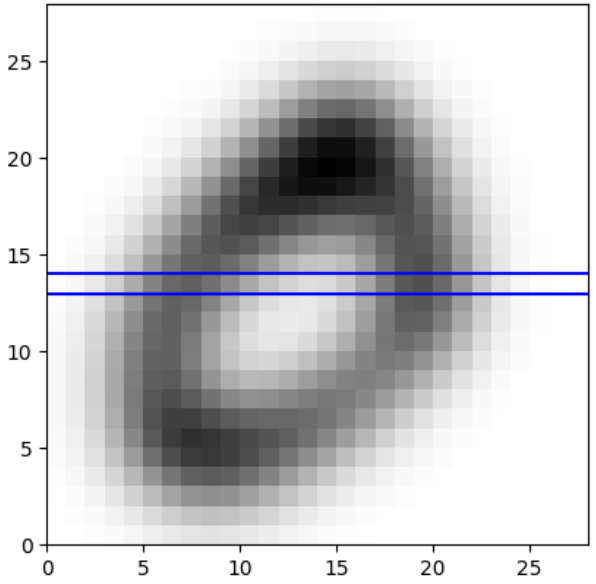
l_2 metric Instability to translations

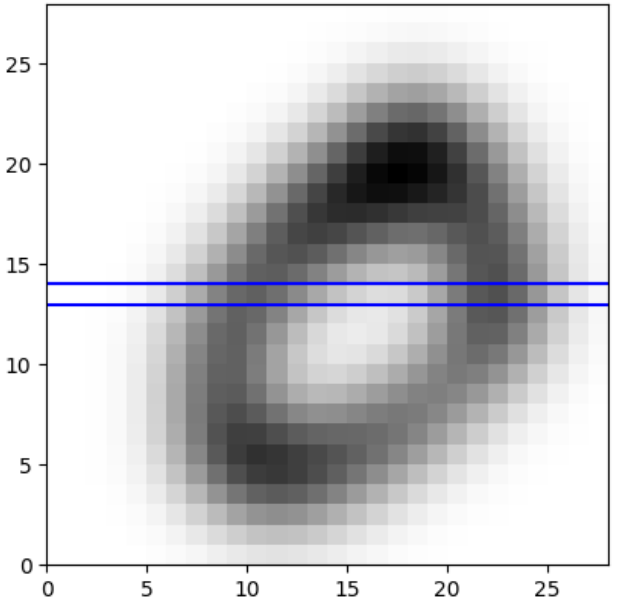


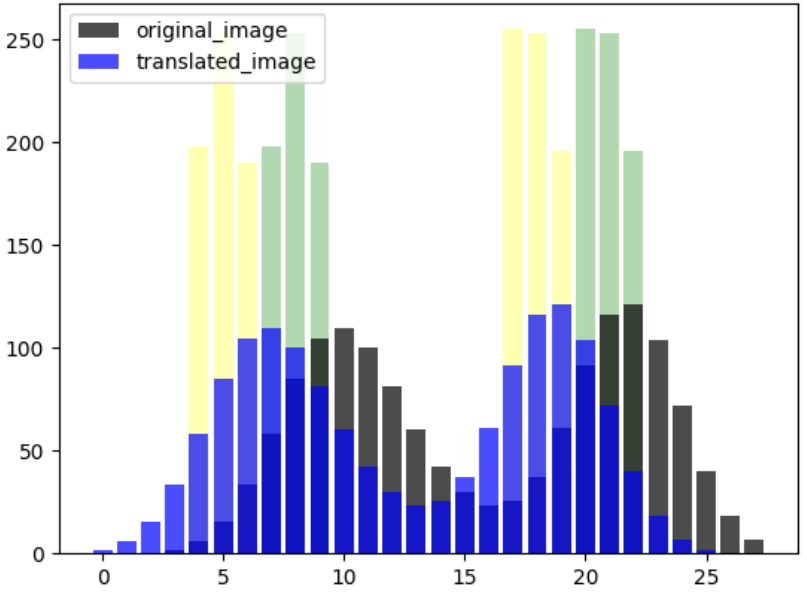




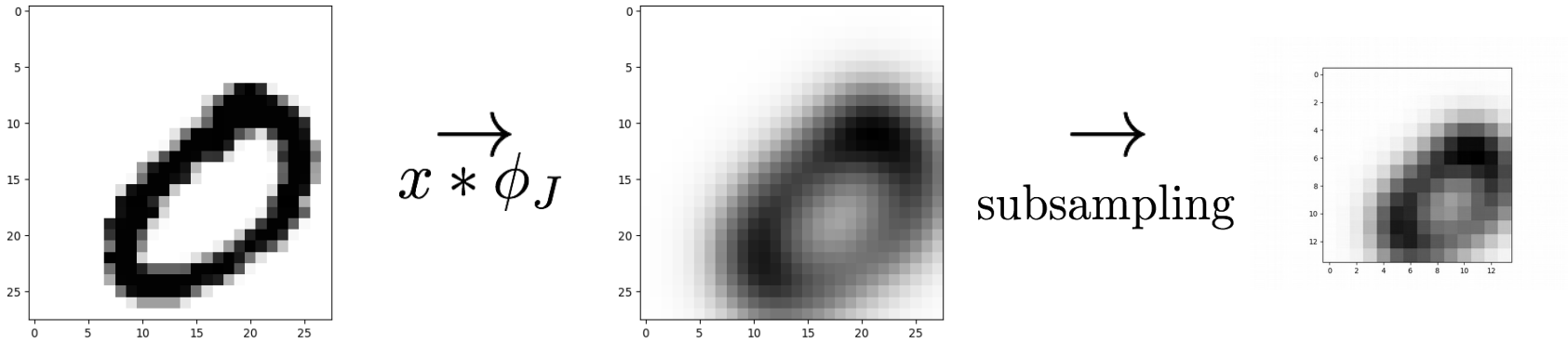
Local averaging







Stability to geometric transformations

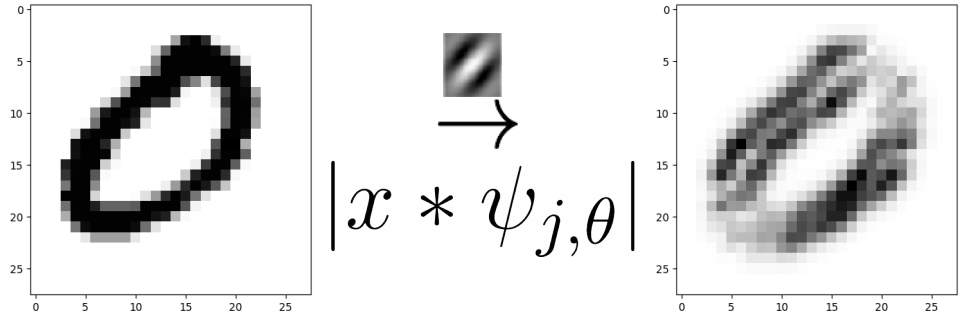


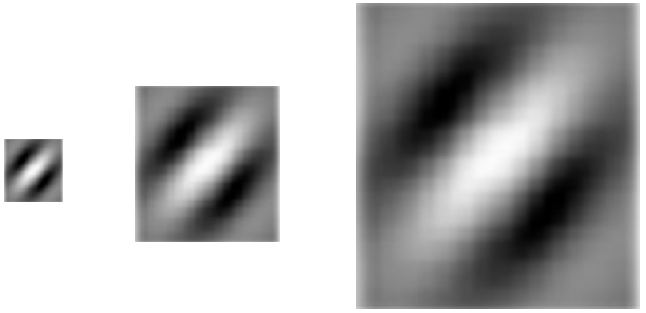
Convolution with Gaussian kernel ϕ_J :

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

Preserving signal information

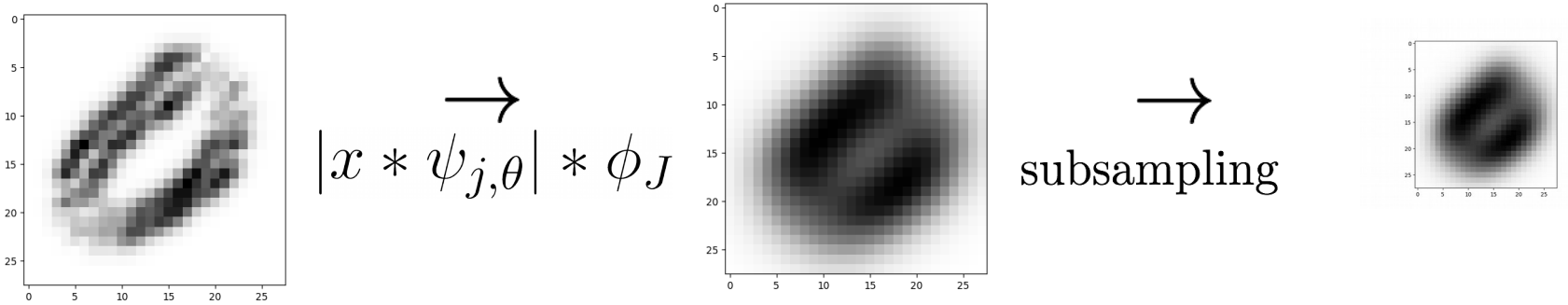
Recover information lost in averaging





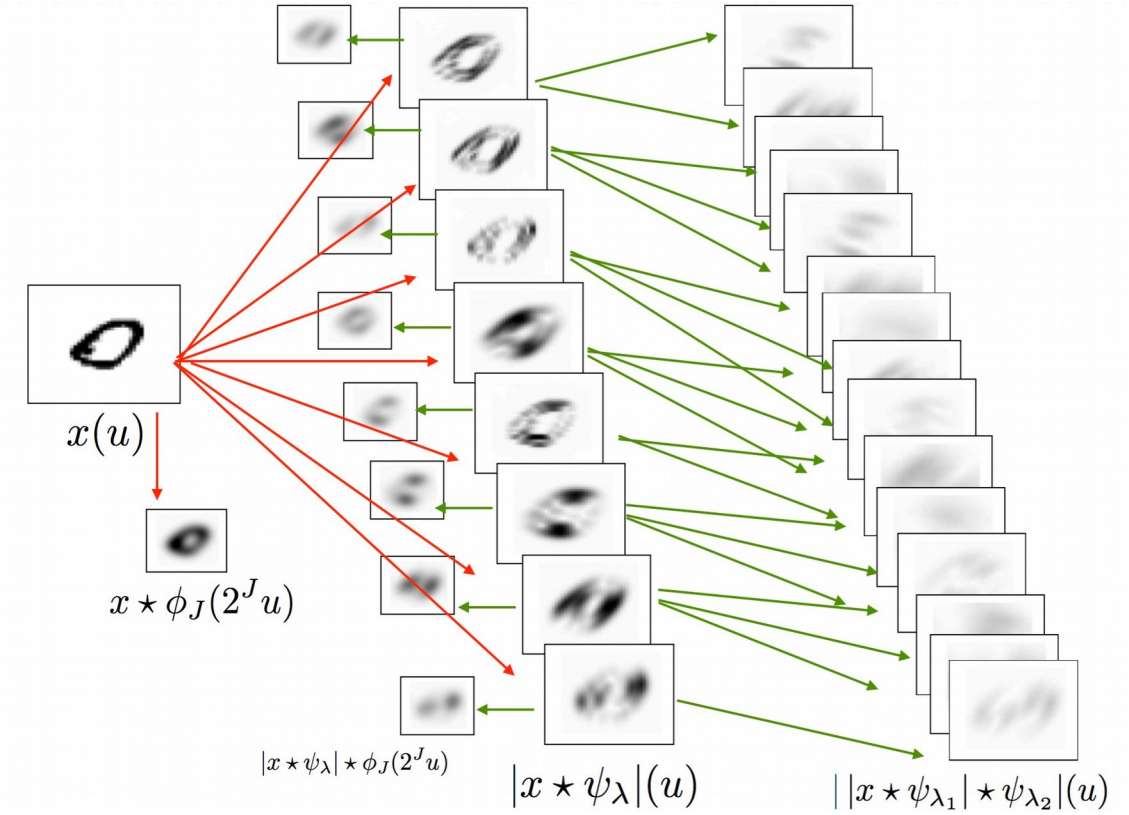
Gabor wavelets $\psi_{j,\theta}$

Stability to geometric transformations



Scattering transform

Mallat (2011), Mallat, Bruna (2012)

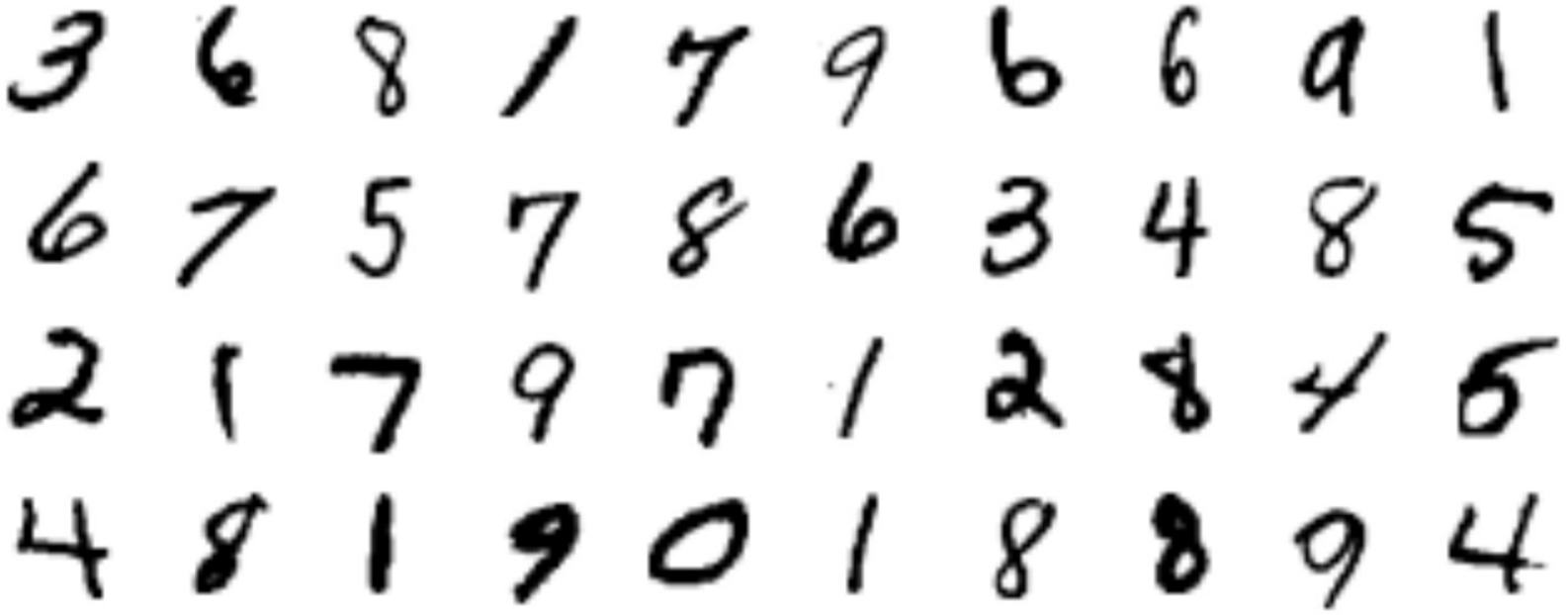


Theorem

$$\|Sx_\tau - Sx\| \leq K\|x\|\|\nabla\tau\|_\infty$$

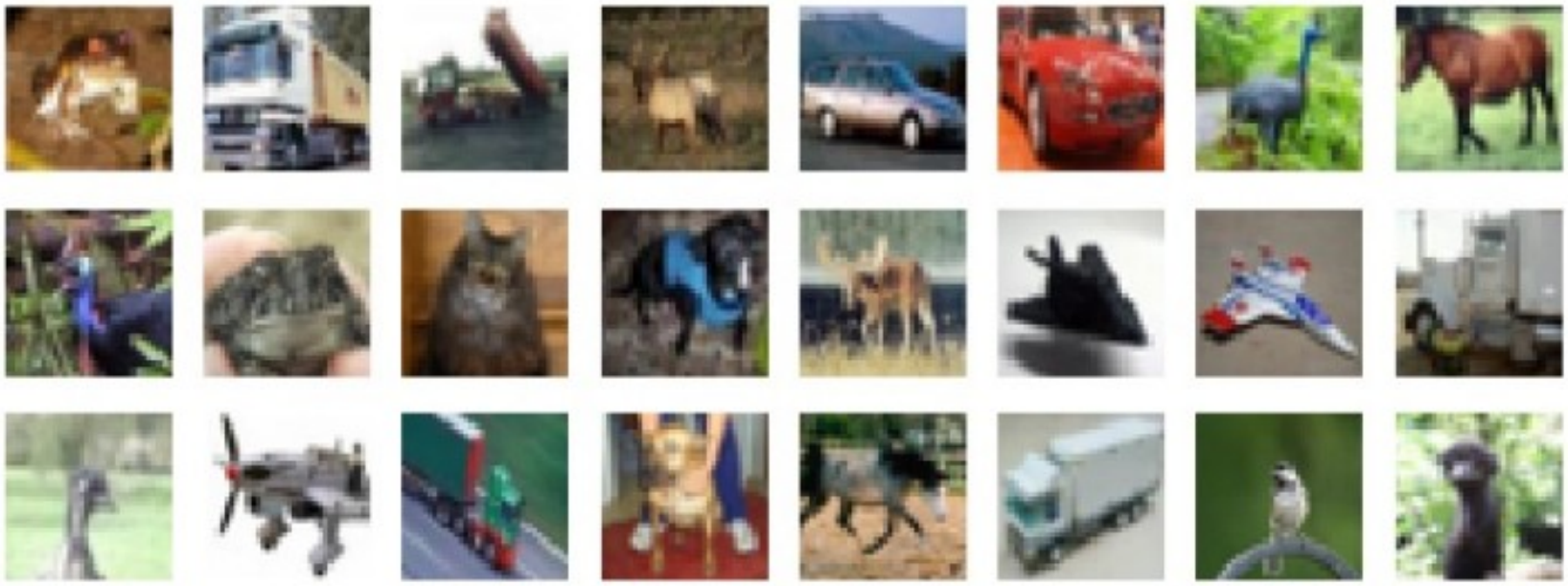
Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
MNIST 28^2 digit images 10 classes	>99 %	>99 %	>99 %



Scattering vs Deep ConvNets

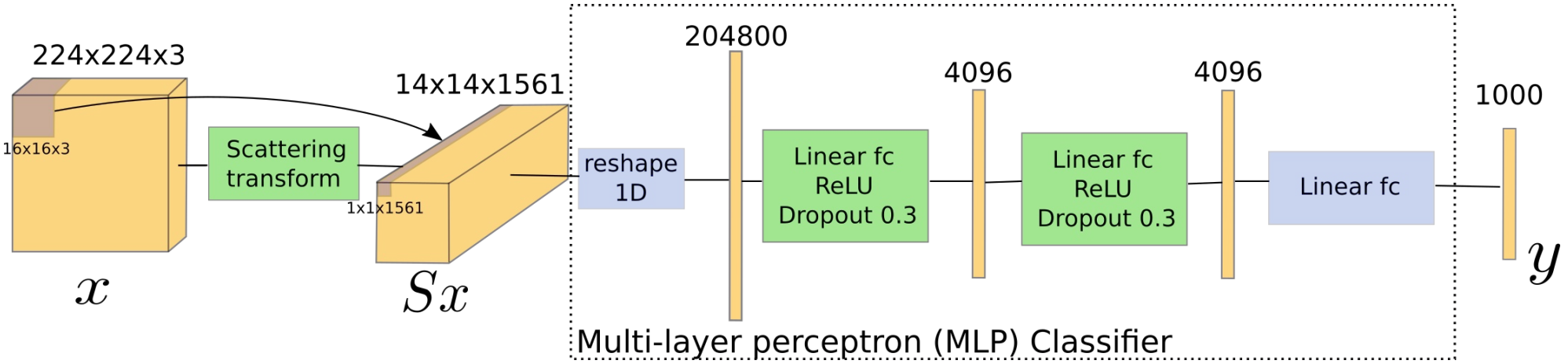
Dataset	Scattering Transform	AlexNet	ResNet
CIFAR-10 32^2 object images 10 classes	84.7 %	89.1 %	95.5 %



Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
ImageNet 224^2 object images 1000 classes	61.4 %	79.1 %	94.2 %

Scattering : ImageNet classification

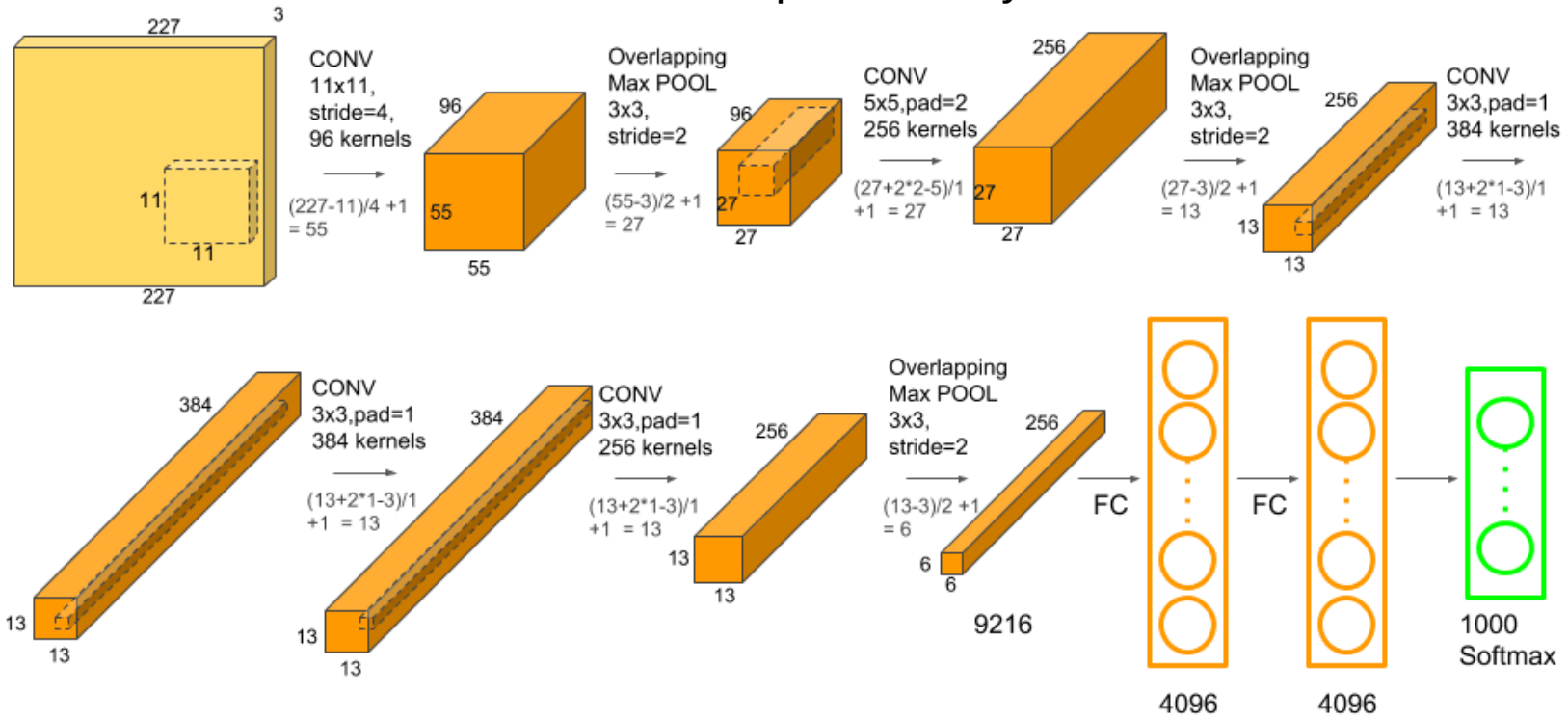


- RGB Images : Scattering Transform on each color channel
- Scale J=4, 8 angles, order 2
- $Sx[i, j]$ is a vector representing a patch of size 16x16x3
- 2 hidden layer classifier (MLP) as in AlexNet
- 38.1 % top1, 61.4 % top5 accuracy

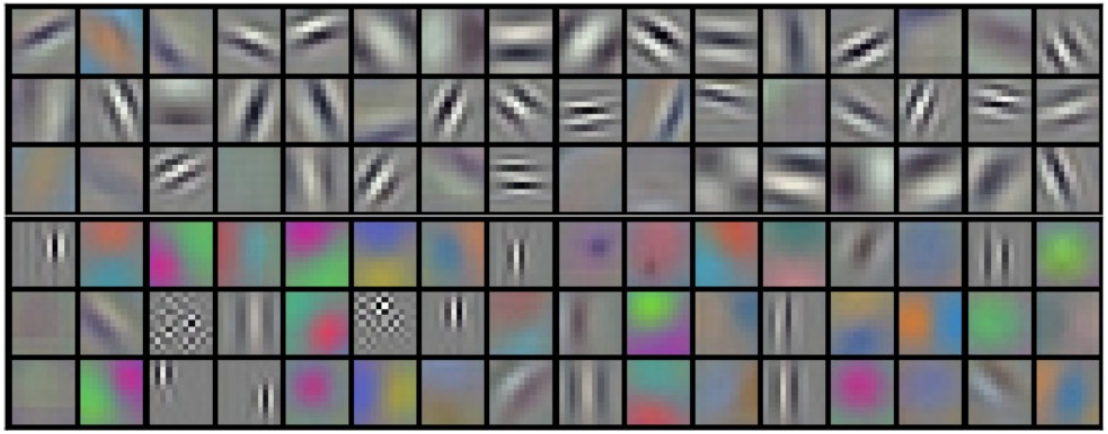
AlexNet

Krizhevsky et al. 2012

79.1 % top5 accuracy



First layer learned filters

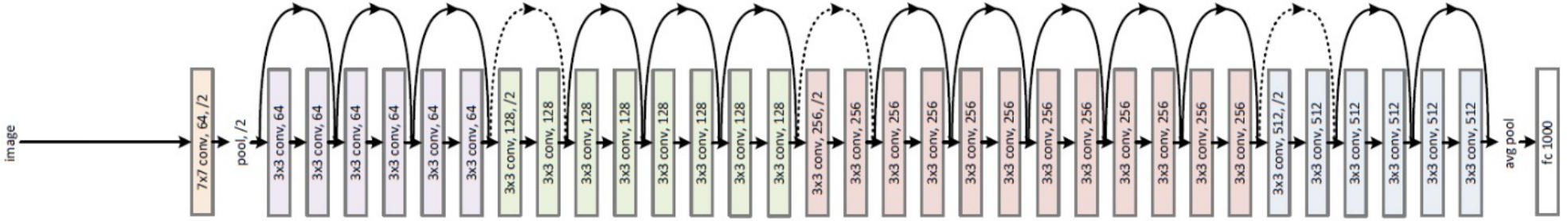


ResNet

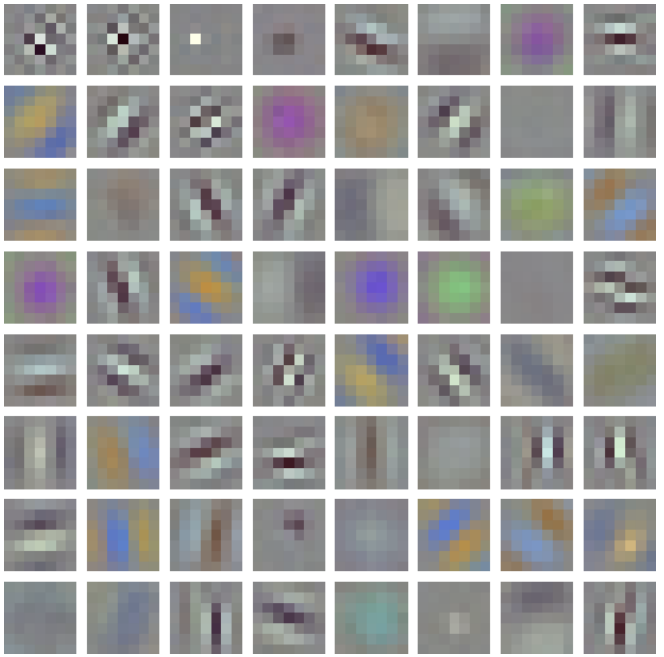
He et al. 2016

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

$$\alpha_*^0(D,\epsilon,x) = \underset{\alpha\geq 0,\ \|D\alpha - x\|<\epsilon}{\operatorname{argmin}} \|\alpha\|_0$$

Convex relaxation with l_1 norm (basis pursuit)

Chen, Donoho et al. 2001

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

Positive Iterated Soft Theshholding algorithm (ISTA)

Daubechies et al. 2003

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

Convolutional version with Scattering transform

$$\alpha_{n+1}[i,j] = \mathrm{ReLU}\left((Id - \frac{1}{L}D^TD)\alpha_n[i,j] + \frac{1}{L}D^TSx[i,j] - \frac{\lambda}{L}\right)$$

Supervised dictionary learning + LISTA

Mairal et al. (2008), Gregor and Lecun (2011)

Principle

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

example : $\mathcal{C}\left(y, f(\alpha, W)\right) = \|W^T \alpha - y\|_2^2$

Convolutional LISTA with N iterations

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

No guarantees that the output α_N is close to α_*

$$\alpha_*(D, \lambda, Sx) = \underset{\alpha \geq 0}{\mathrm{argmin}}\; \mathcal{L}(\alpha), \;\; \mathcal{L}(\alpha) = \|D\alpha - Sx\|_2^2 + \lambda \|\alpha\|_1$$

Task Driven dictionary learning + ISTC

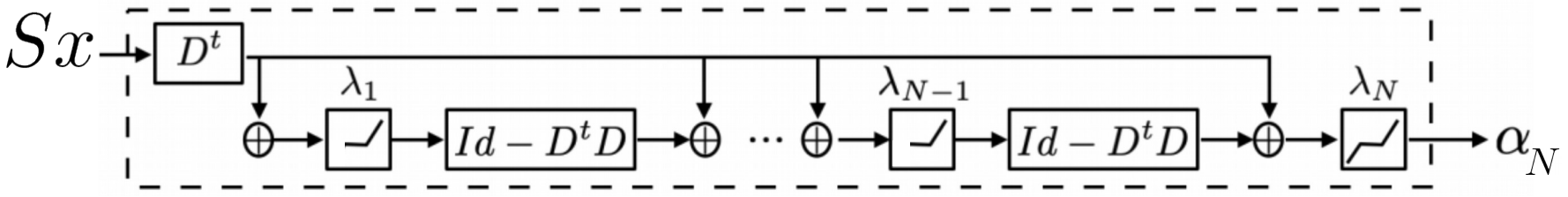
Mairal et al. (2011), ours

Principle

$$\min_{D,W,\alpha \geq 0} \mathcal{C}\left(y, f(\alpha_*(D,\lambda,x),W)\right)$$

Iterative Soft Trehsholding with continuation

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



Theorem

$\cdot\ s$ support size of $\alpha*$

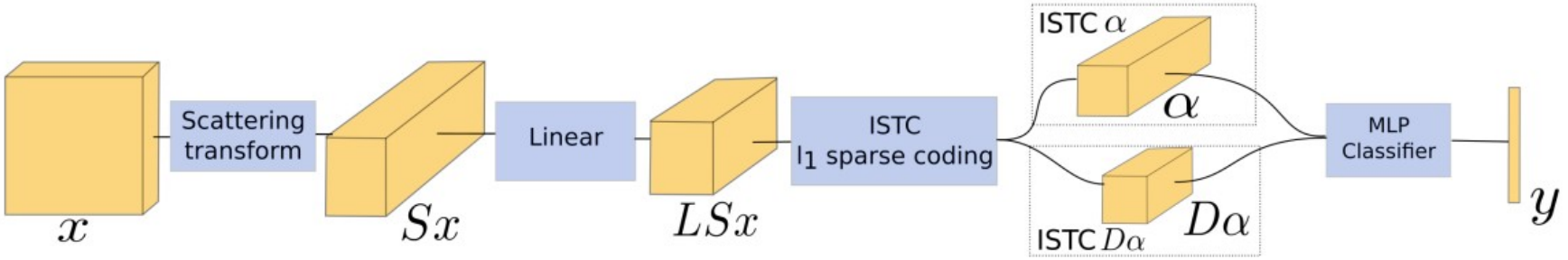
$\cdot\ \mu = \max_{m \neq m'} \langle D_m, D_{m'} \rangle$

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 convoltuional network



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

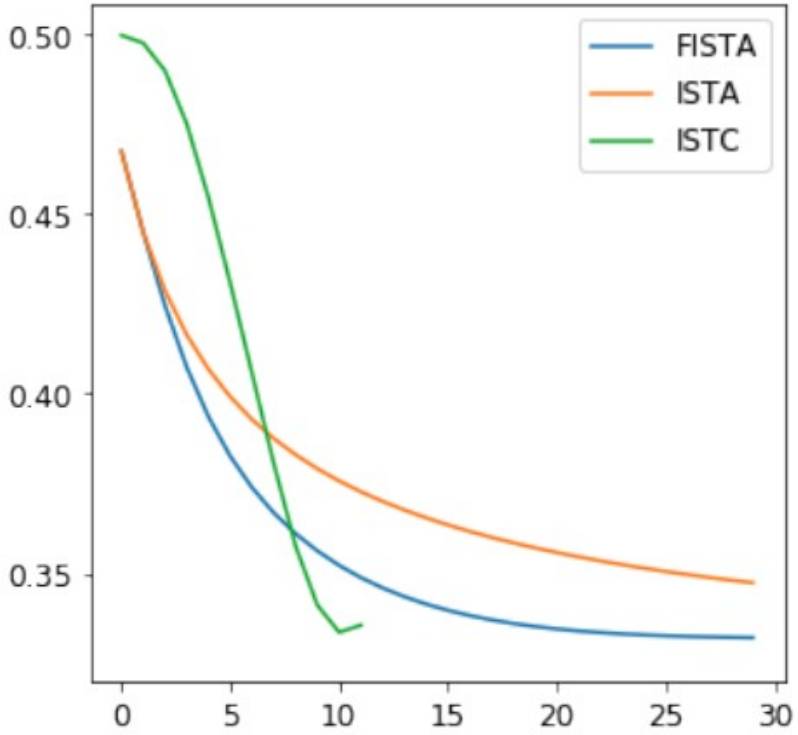
Results

l_1 algo	ISTC	ISTC	LISTA
Classifier input	α	$D\alpha$	α
Top 1	59.5	55.3	62.9
Top 5	81.3	78.3	83.9

Convergence analysis

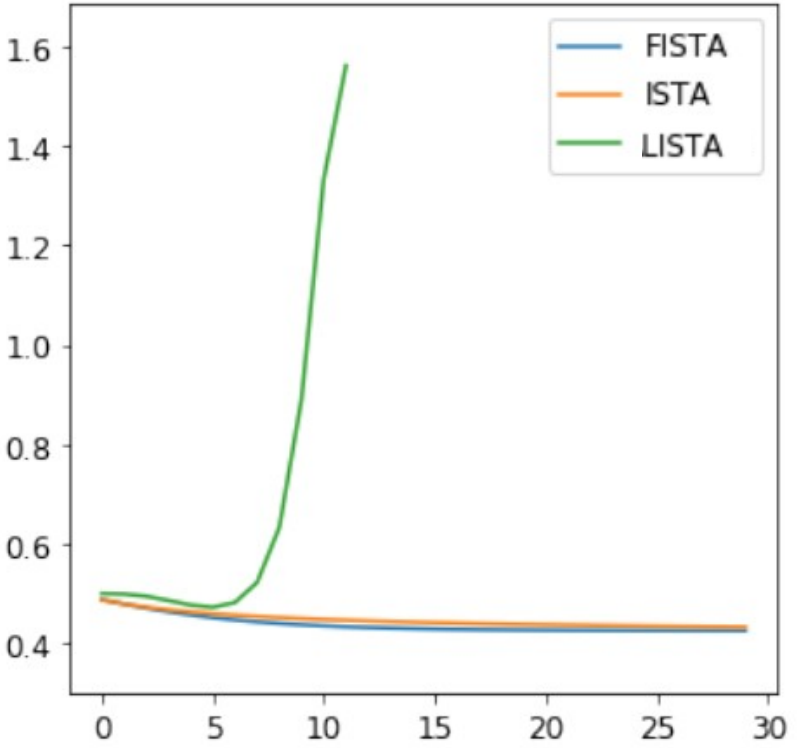
ISTC

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



LISTA

$$\frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 3.8$$



Comments

- Improvement of 20% over Scattering alone
- Large factor λ_*, reconstruction error $\|Sx - D\alpha\| / \|Sx\| = 0.5$
- Hard to reconstruct the original image from α
- 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 ?

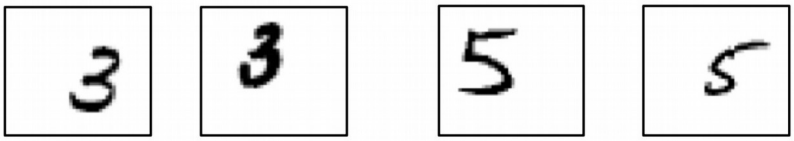
Image classification with Scattering Transform and Dictionary learning

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

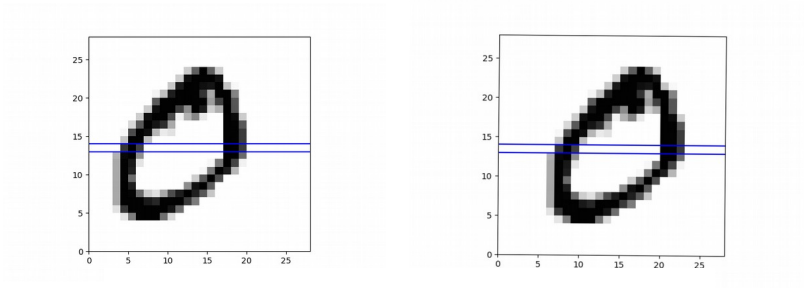
Digits classification

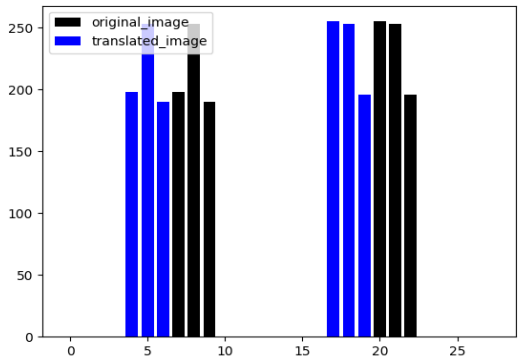
- MNIST database

- Invariance to translations, stability to deformations

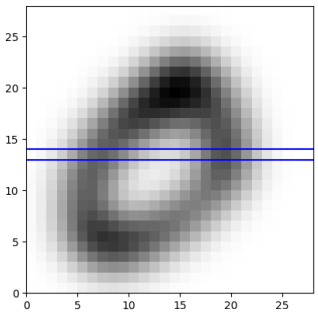


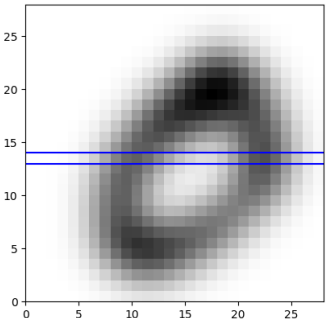
l_2 metric Instability to translations

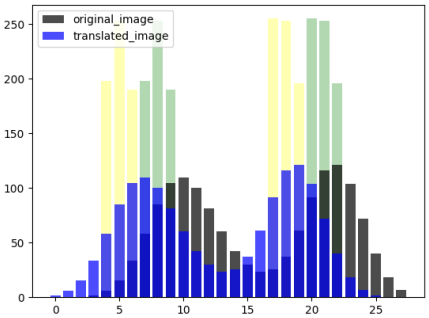




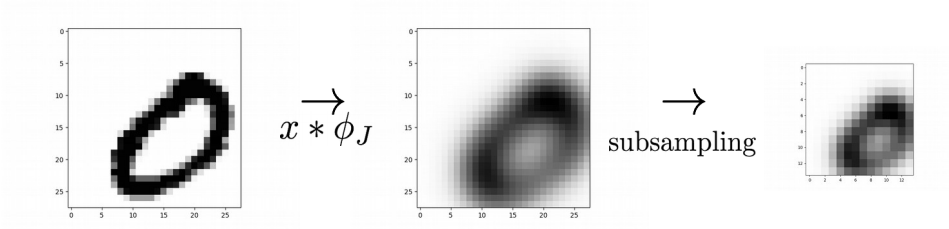
Local averaging







Stability to geometric transformations

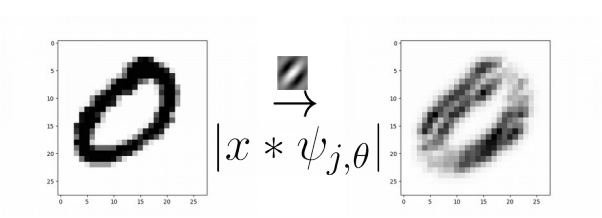


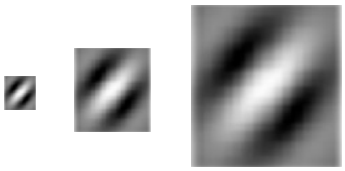
Convolution with Gaussian kernel ϕ_J :

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

Preserving signal information

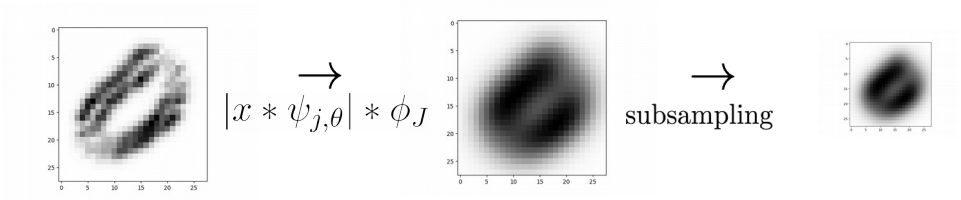
Recover information lost in averaging





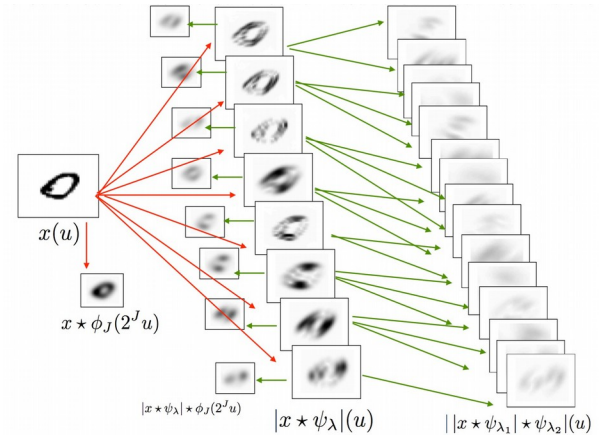
Gabor wavelets $\psi_{j,\theta}$

Stability to geometric transformations



Scattering transform

Mallat (2011), Mallat, Bruna (2012)



Theorem

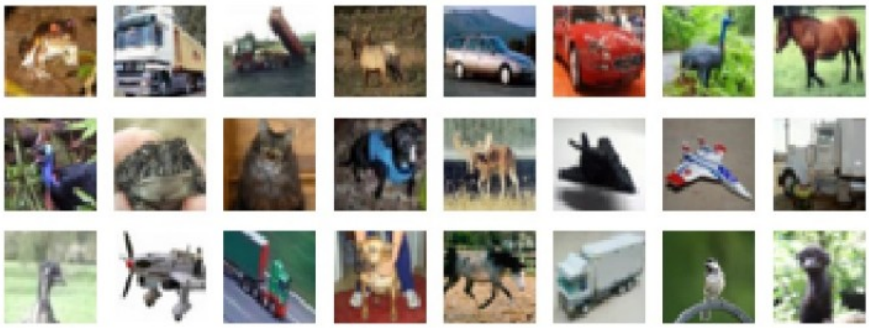
$$\|Sx_\tau - Sx\| \leq K\|x\|\|\nabla\tau\|_\infty$$

Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
MNIST 28^2 digit images 10 classes	>99 %	>99 %	>99 %

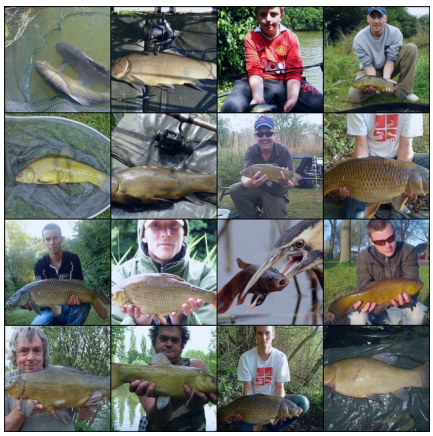
Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
CIFAR-10 32^2 object images 10 classes	84.7 %	89.1 %	95.5 %

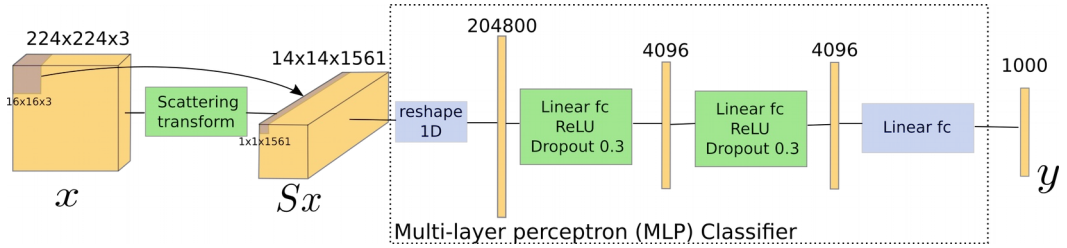


Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
ImageNet 224^2 object images 1000 classes	61.4 %	79.1 %	94.2 %



Scattering : ImageNet classification

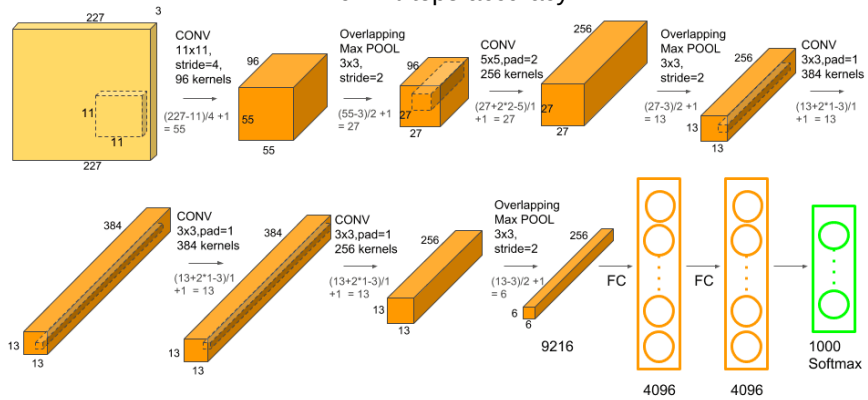


- RGB Images : Scattering Transform on each color channel
- Scale J=4, 8 angles, order 2
- $Sx[i,j]$ is a vector representing a patch of size 16x16x3
- 2 hidden layer classifier (MLP) as in AlexNet
- 38.1 % top1, 61.4 % top5 accuracy

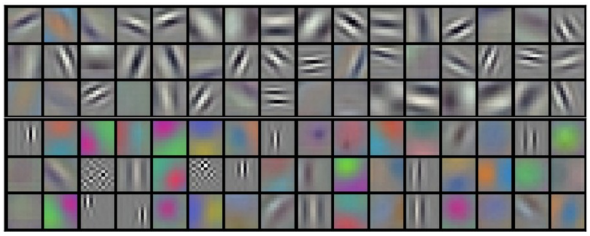
AlexNet

Krizhevsky et al. 2012

79.1 % top5 accuracy



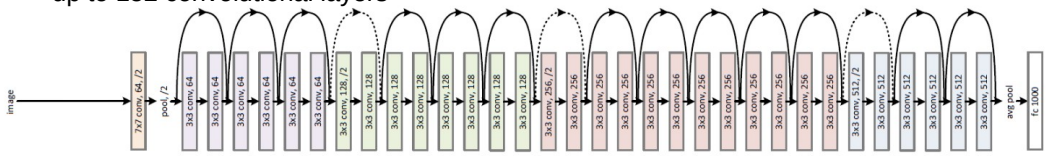
First layer learned filters



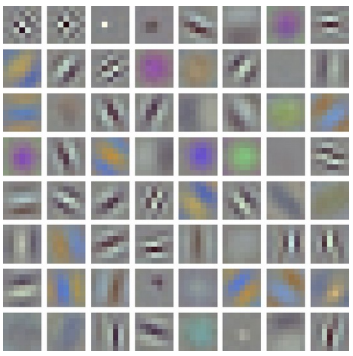
ResNet

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

$$\alpha_*^0(D, \epsilon, x) = \underset{\alpha \geq 0,\ \|D\alpha - x\| < \epsilon}{\operatorname{argmin}} \|\alpha\|_0$$

Convex relaxation with l_1 norm (basis pursuit)

Chen, Donoho et al. 2001

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

Positive Iterated Soft Theshholding algorithm (ISTA)

Daubechies et al. 2003

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

Convolutional version with Scattering transform

$$\alpha_{n+1}[i,j] = \mathrm{ReLU}\left((Id - \frac{1}{L} D^T D)\alpha_n[i,j] + \frac{1}{L} D^T S x[i,j] - \frac{\lambda}{L}\right)$$

Supervised dictionary learning + LISTA

Mairal et al. (2008), Gregor and Lecun (2011)

Principle

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

example : $\mathcal{C}\left(y, f(\alpha, W)\right) = \|W^T \alpha - y\|_2^2$

Convolutional LISTA with N iterations

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

No guarantees that the output α_N is close to α_*

$$\alpha_*(D, \lambda, Sx) = \underset{\alpha \geq 0}{\mathrm{argmin}}\ \mathcal{L}(\alpha),\ \mathcal{L}(\alpha) = \|D\alpha - Sx\|_2^2 + \lambda \|\alpha\|_1$$

Task Driven dictionary learning + ISTC

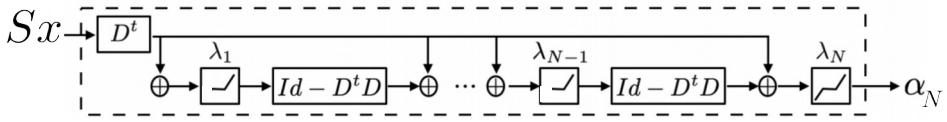
Mairal et al. (2011), ours

Principle

$$\min_{D,W,\alpha \geq 0} \mathcal{C}\left(y, f(\alpha_*(D,\lambda,x),W)\right)$$

Iterative Soft Trehsholding with continuation

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



Theorem

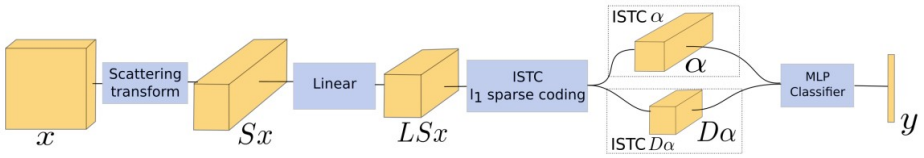
- s support size of $\alpha*$
- $\mu = \max_{m \neq m'} \langle D_m, D_{m'} \rangle$

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 convoltuional network



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

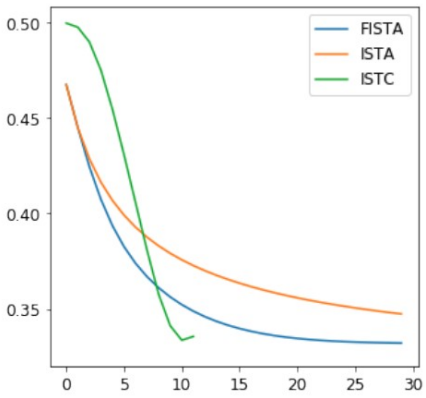
Results

l_1 algo	ISTC	ISTC	LISTA
Classifier input	α	$D\alpha$	α
Top 1	59.5	55.3	62.9
Top 5	81.3	78.3	83.9

Convergence analysis

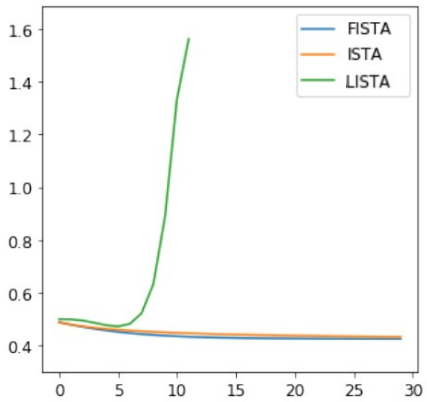
ISTC

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



LISTA

$$\frac{\mathcal{L}(\alpha_N)}{\mathcal{L}(\alpha_*)} = 3.8$$



Comments

- Improvement of 20% over Scattering alone
- Large factor λ_*, reconstruction error $\|Sx - D\alpha\| / \|Sx\| = 0.5$
- Hard to reconstruct the original image from α
- 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 ?