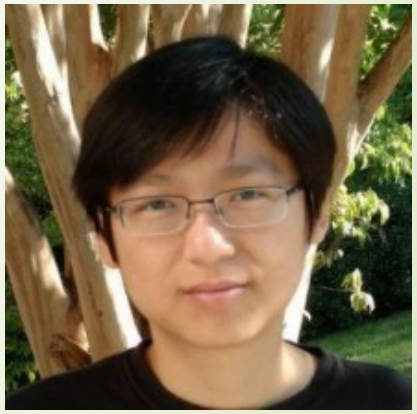
Harmonic Analysis of Deep Convolutional Networks

Yuan YAO

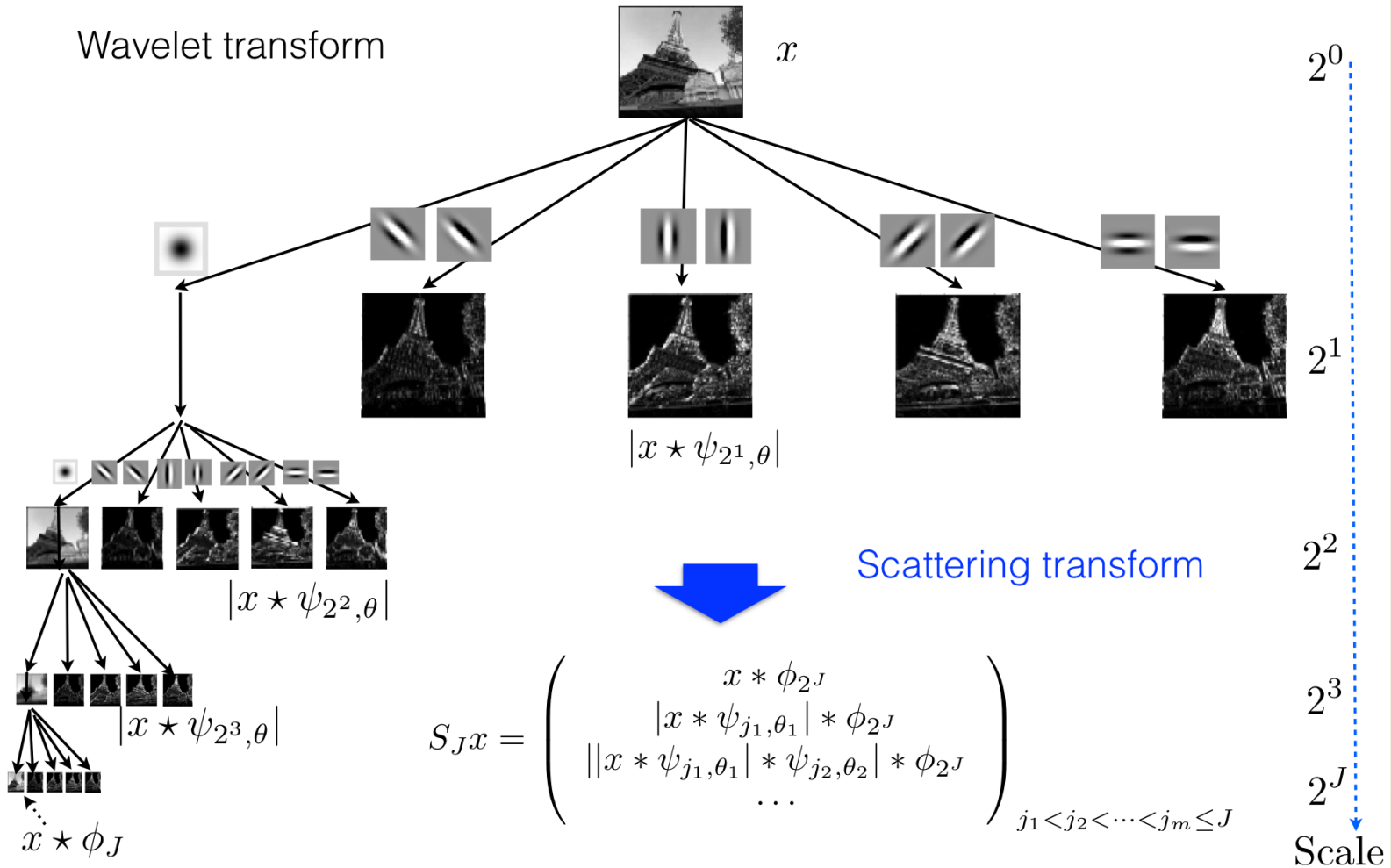
HKUST

Based on Mallat Bolcskei Cheng talks etc.



Scattering Networks

[Mallat '12]



Scattering Networks

[Mallat '12]

Stability of scattering representations

- Non-expansive mapping

$$\|S_J x - S_J y\| \leq \|x - y\|$$

- Deformation insensitivity

$$D_\tau x(u) = x(u - \tau(u)), \quad \|S_J D_\tau x - S_J x\| \leq C(\tau, J)\|x\|$$

No fitting,
Thus no overfitting!

Group Invariants/Stability

- Translation Invariance:
  - The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of ϕ.
  - Full translation invariance at the limit:

$$\lim_{\phi \to 1} |x \star \psi_{\lambda_1}| \star \phi(t) = \int |x \star \psi_{\lambda_1}(u)| \, du = \|x \star \psi_{\lambda_1}\|_1$$

- Stable Small Deformations:

$$stable\ to\ deformations\ \ x_\tau(t) = x(t - \tau(t))$$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)| \, \|x\|$$

Applications and extensions:

- Invertibility/completeness of representation [Waldspurger et al. '12]
- Extension to signals on graphs [Chen et al. '14] [Cheng et al. '16]
- With general family of filters [Bolcskei et al. '15] [Czaja et al. '15]

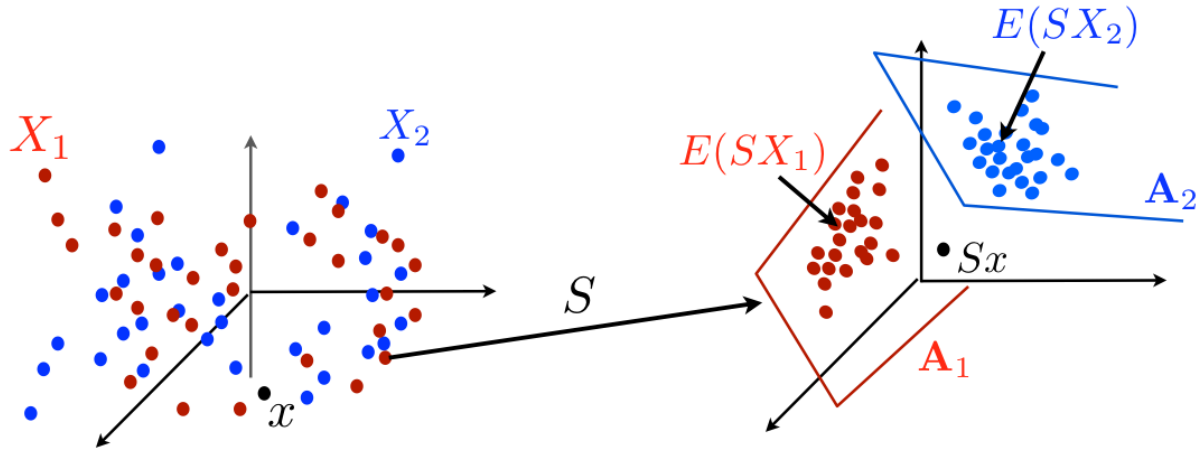
Feature Extraction

Linearized Classification

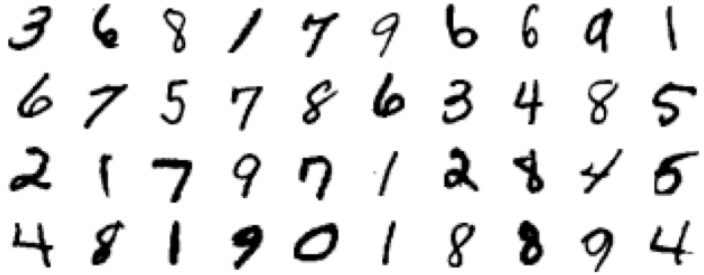
Joan Bruna

- Each class X_k is represented by a scattering centroid $E(SX_k)$

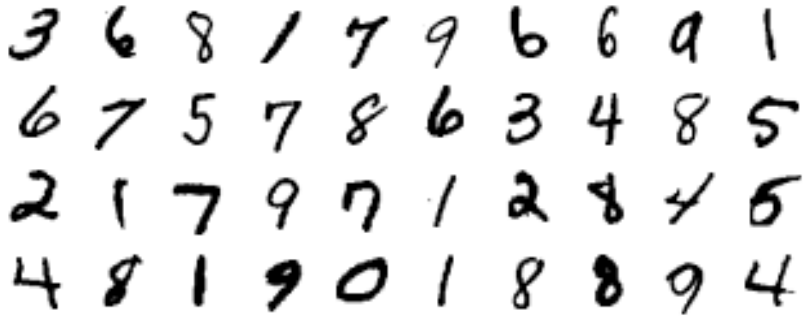
Affine space model $\mathbf{A}_k = E(SX_k) + \mathbf{V}_k$. computed with PCA.



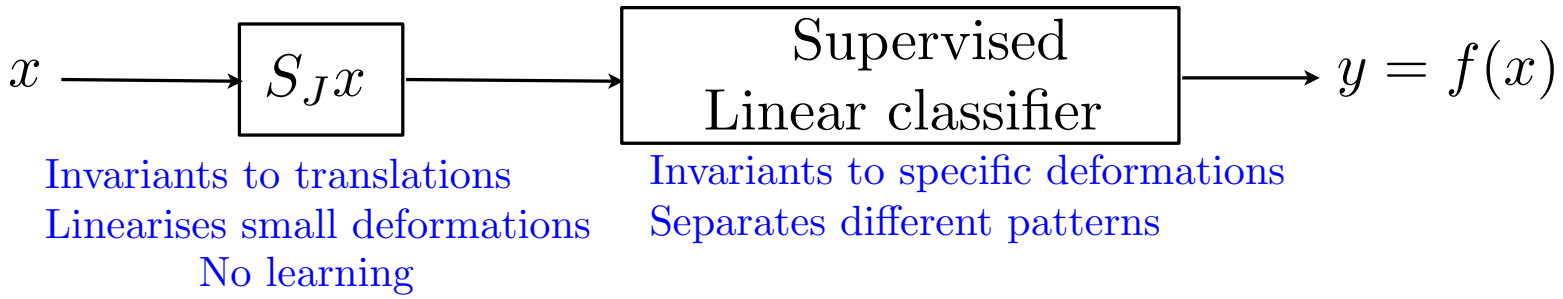
MNIST data basis:



Digit Classification: MNIST



Joan Bruna



Classification Errors

Training size	Conv. Net.	Scattering
50000	0.4%	0.4%

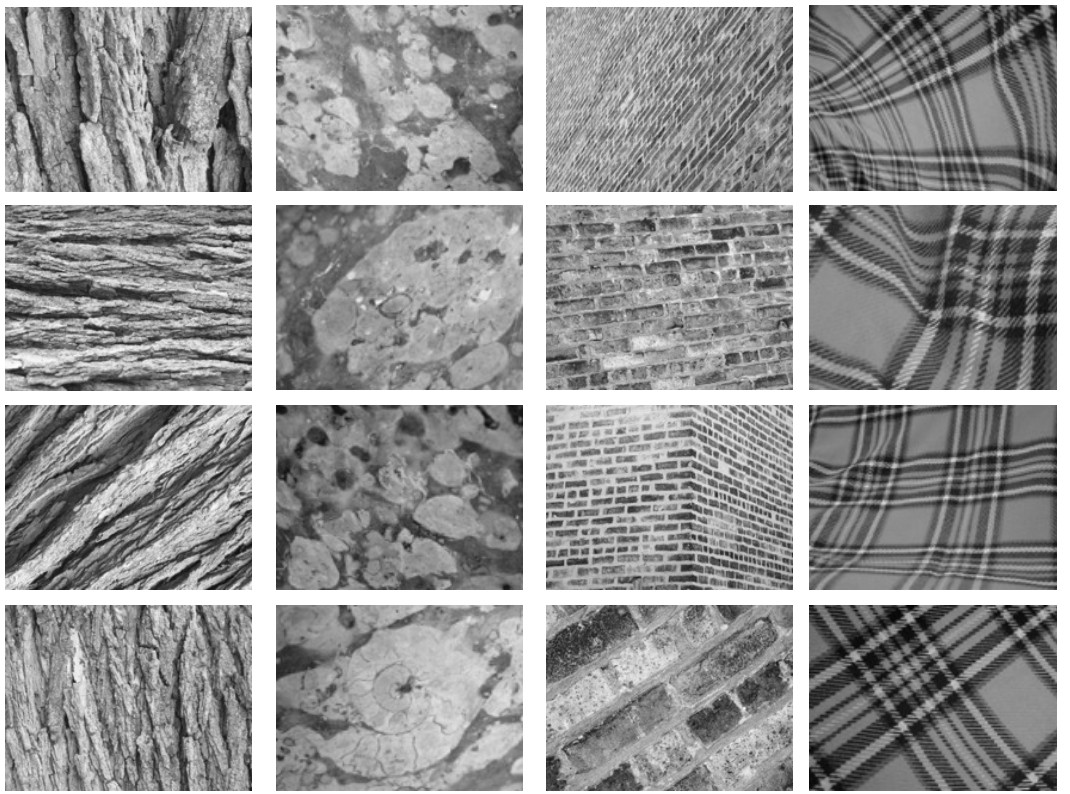
LeCun et. al.

Other Invariants?
Cross-channel pooling!

Rotation and Scaling Invariance

Laurent Sifre

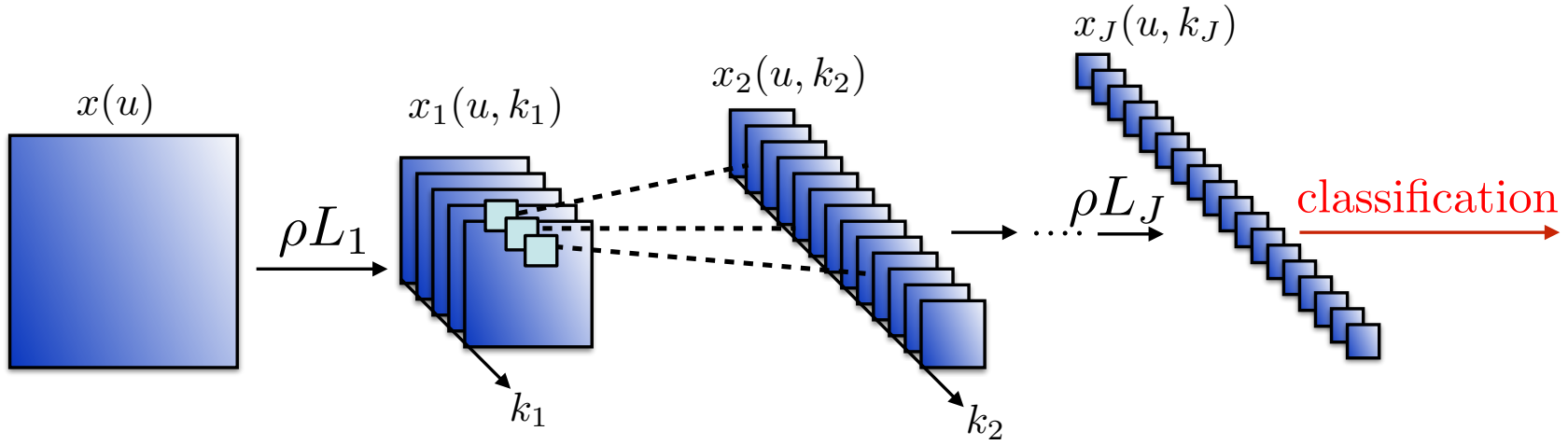
UIUC database:
25 classes



Scattering classification errors

Training	Scat. Translation
20	20 %

Deep Convolutional Trees



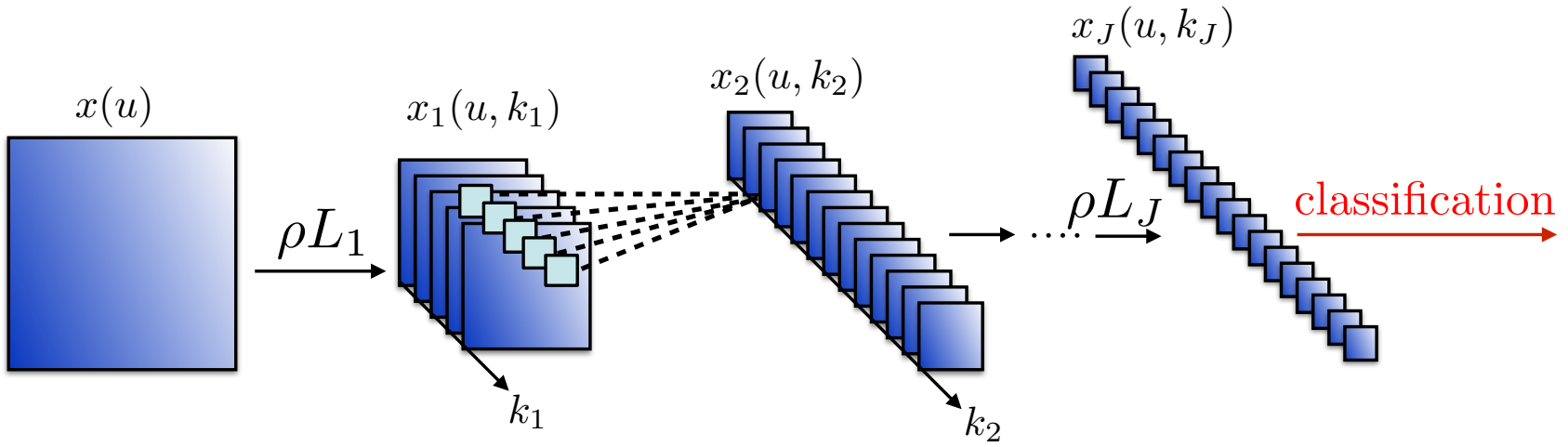
$$x_j = \rho\, L_j\, x_{j-1}$$

L_j is composed of convolutions and subs samplings:

$$x_j(u, k_j) = \rho\Big(x_{j-1}(\cdot, k) \star h_{k_j,k}(u)\Big)$$

No channel communication: what limitations ?

Deep Convolutional Networks



$$x_j = \rho\, L_j\, x_{j-1}$$

- L_j is a linear combination of convolutions and subsampling:

$$x_j(u, k_j) = \rho\Big(\sum_k x_{j-1}(\cdot, k) \star h_{k_j,k}(u) \Big)$$

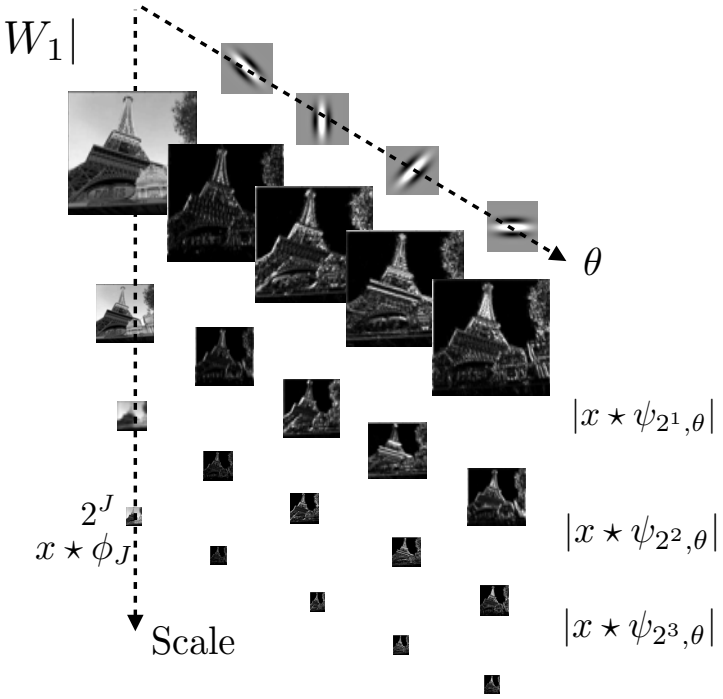
sum across channels

What is the role of channel connections ?

Linearize other symmetries beyond translations.

Rotation Invariance

- Channel connections linearize other symmetries.



- Invariance to rotations are computed by convolutions along the rotation variable θ with wavelet filters.

  $\Rightarrow$ invariance to rigid mouvements.

Wavelet Transform on a Group

Laurent Sifre

- Roto-translation group $G = \{g = (r, t) \in SO(2) \times \mathbb{R}^2\}$

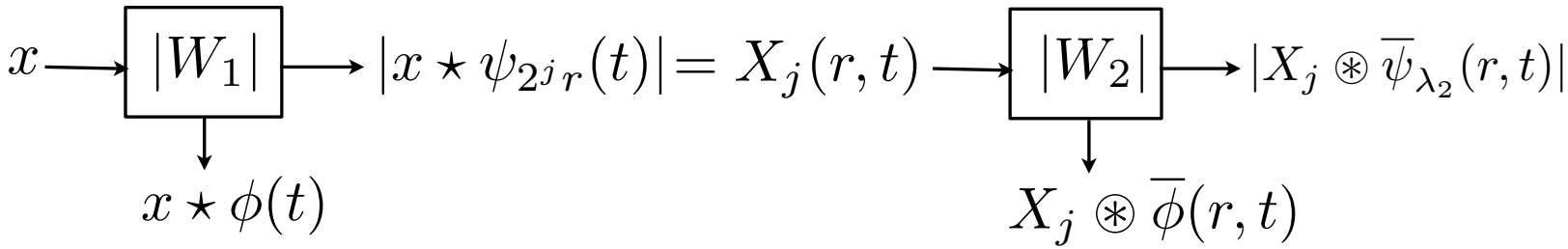
$$(r, t) \,.\, x(u) = x(r^{-1}(u - t))$$

- Averaging on G: $$X \circledast \overline{\phi}(g) = \int_G X(g')\, \overline{\phi}(g'^{-1} g)\, dg'$$

- Wavelet transform on G: $$W_2 X = \begin{pmatrix} X \circledast \overline{\phi}(g) \\ X \circledast \overline{\psi}_{\lambda_2}(g) \end{pmatrix}_{\lambda_2, g} .$$

translation

roto-translation



Wavelet Transform on a Group

Laurent Sifre

- Roto-translation group $G = \{g = (r, t) \in SO(2) \times \mathbb{R}^2\}$

$$(r, t) \,.\, x(u) = x(r^{-1}(u - t))$$

- Averaging on G: $\quad X \circledast \overline{\phi}(g) = \int_G X(g')\, \overline{\phi}(g'^{-1} g)\, dg'$

- Wavelet transform on G: $\quad W_2 X = \begin{pmatrix} X \circledast \overline{\phi}(g) \\ X \circledast \overline{\psi}_{\lambda_2}(g) \end{pmatrix}_{\lambda_2, g}$.

translation

scalo-roto-translation
+ renormalization

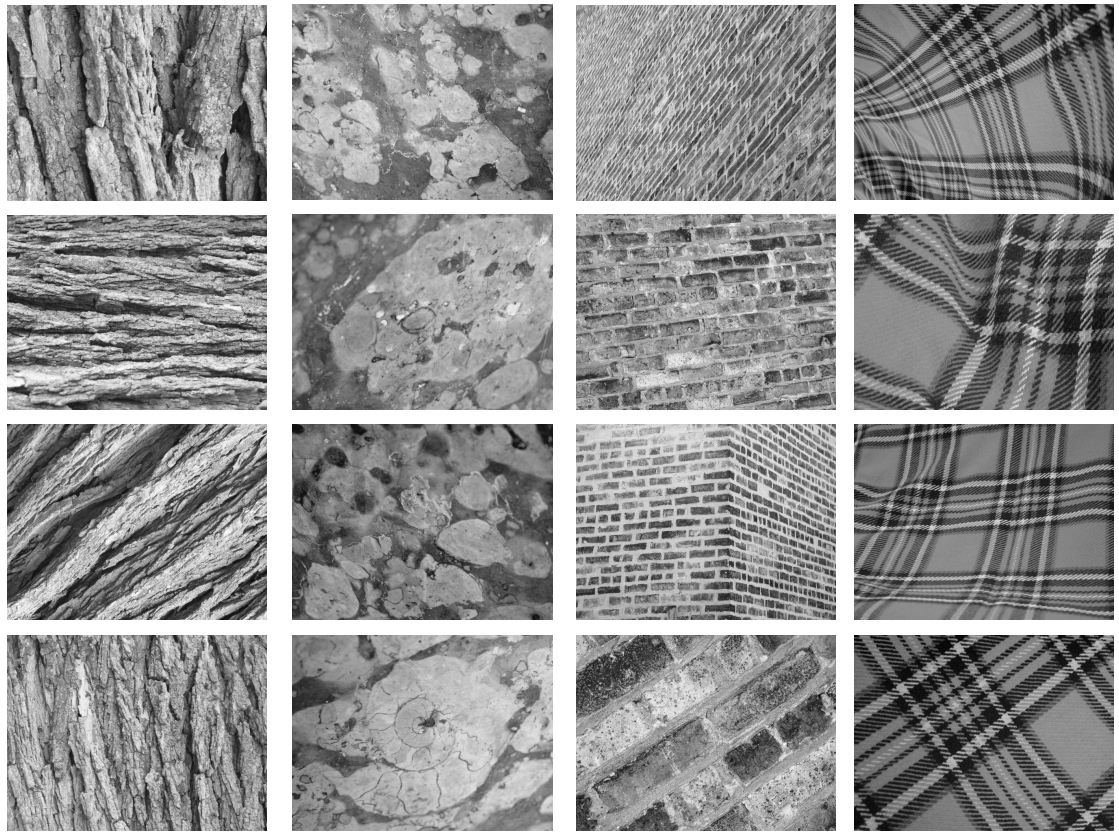
$$x \longrightarrow \boxed{|W_1|} \longrightarrow |x \star \psi_{2^j r}(t)| = X(2^j, r, t) \longrightarrow \boxed{|W_2|} \longrightarrow |X \circledast \overline{\psi}_{\lambda_2}(2^j, r, t)|$$

$$x \star \phi(t) \qquad\qquad X \circledast \overline{\phi}(2^j, r, t)$$

Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

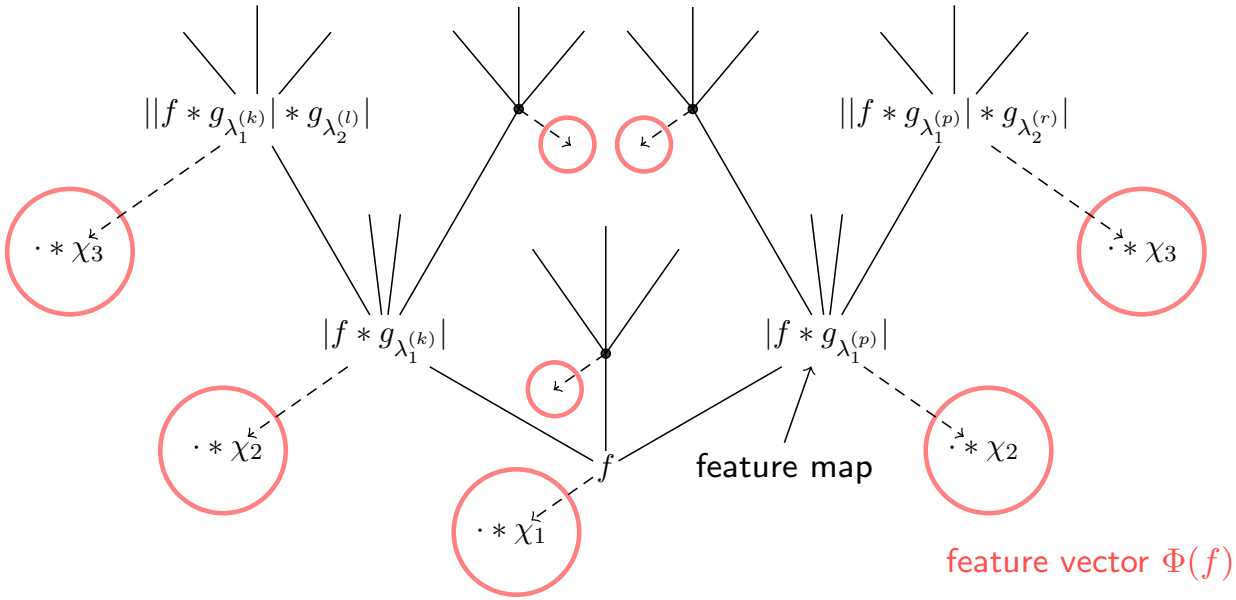
Training	Translation	Transl + Rotation	+ Scaling
20	20 %	2%	**0.6**%

Wiatowski-Bolcskei'15

- Scattering Net by Mallat et al. so far
  - Wavelet Linear filter
  - Nonlinear activation by modulus
  - Average pooling
- Generalization by Wiatowski-Bolcskei'15
  - Filters as frames
  - Lipschitz continuous Nonlinearities
  - General Pooling: Max/Average/Nonlinear, etc.

Generalization of Wiatowski-Bolcskei'15

Scattering networks ([*Mallat, 2012*], [*Wiatowski and HB, 2015*])



General scattering networks guarantee [*Wiatowski & HB, 2015*]

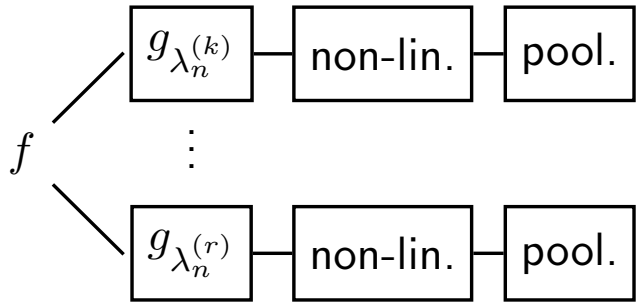
- (vertical) **translation invariance**
- **small deformation sensitivity**

essentially irrespective of filters, non-linearities, and poolings!

Wavelet basis -> filter frame

Building blocks

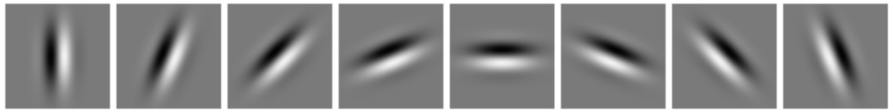
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

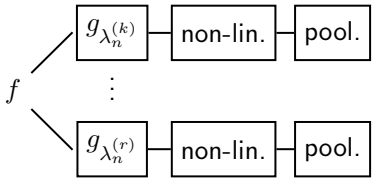
e.g.: Structured filters



Frames: random or learned filters

Building blocks

Basic operations in the n-th network layer



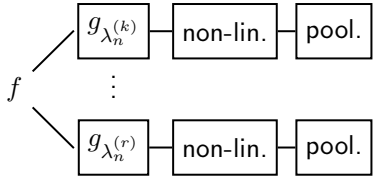
Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

e.g.: Unstructured filters

Building blocks

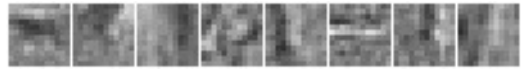
Basic operations in the n-th network layer



Filters: Semi-discrete frame $\Psi_n := \{\chi_n\} \cup \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$

$$A_n \|f\|_2^2 \leq \|f * \chi_n\|_2^2 + \sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

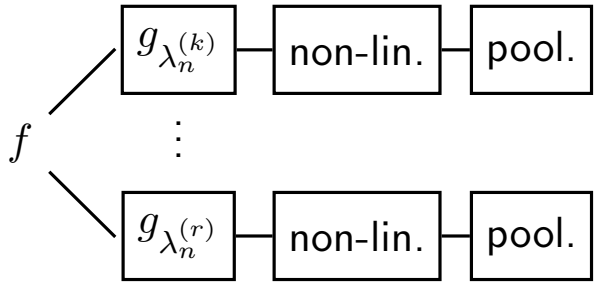
e.g.: Learned filters



Nonlinear activations

Building blocks

Basic operations in the n-th network layer



Non-linearities: Point-wise and Lipschitz-continuous

$$\|M_n(f) - M_n(h)\|_2 \leq L_n \|f - h\|_2, \quad \forall\, f, h \in L^2(\mathbb{R}^d)$$

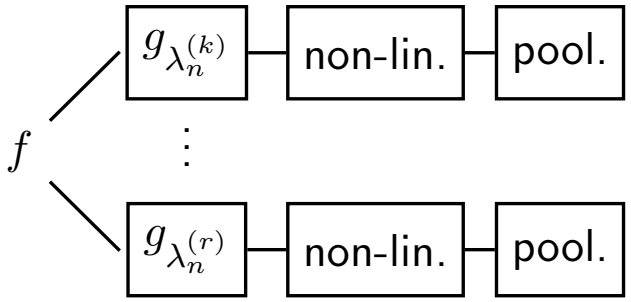
$\Rightarrow$ Satisfied by virtually **all** non-linearities used in the **deep learning literature**!

ReLU: $L_n = 1$; modulus: $L_n = 1$; logistic sigmoid: $L_n = \frac{1}{4}$; ...

Pooling

Building blocks

Basic operations in the n-th network layer



Pooling: In continuous-time according to

$$f \mapsto S_n^{d/2} P_n(f)(S_n \cdot),$$

where $S_n \geq 1$ is the **pooling factor** and $P_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is R_n-Lipschitz-continuous

$\Rightarrow$ **Emulates** most **poolings** used in the **deep learning literature**!

e.g.: Pooling by **sub-sampling** $P_n(f) = f$ with $R_n = 1$

e.g.: Pooling by **averaging** $P_n(f) = f * \phi_n$ with $R_n = \|\phi_n\|_1$

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. Then,

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

The condition

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N},$$

is **easily satisfied** by **normalizing** the filters $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$.

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. *Then,*

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

$\Rightarrow$ Features become **more invariant** with **increasing** network **depth**!

Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that the filters, non-linearities, and poolings satisfy

$$B_n \leq \min\{1, L_n^{-2} R_n^{-2}\}, \quad \forall\, n \in \mathbb{N}.$$

Let the pooling factors be $S_n \geq 1$, $n \in \mathbb{N}$. Then,

$$|||\Phi^n(T_t f) - \Phi^n(f)||| = \mathcal{O}\left(\frac{\|t\|}{S_1 \dots S_n}\right),$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Full translation invariance: If $\lim_{n\to\infty} S_1 \cdot S_2 \cdot \ldots \cdot S_n = \infty$, then

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0$$

Philosophy behind invariance results

Mallat's "horizontal" translation invariance [*Mallat, 2012*]:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$
- applies to wavelet transform and modulus non-linearity without pooling

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth
- applies to general filters, general non-linearities, and general poolings

Non-linear deformations

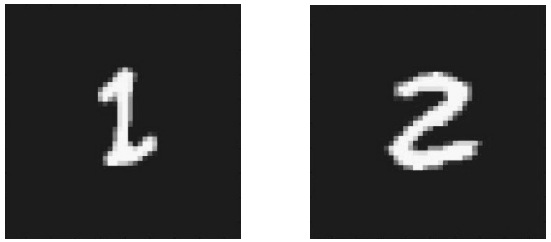
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

For "small" τ:

Non-linear deformations

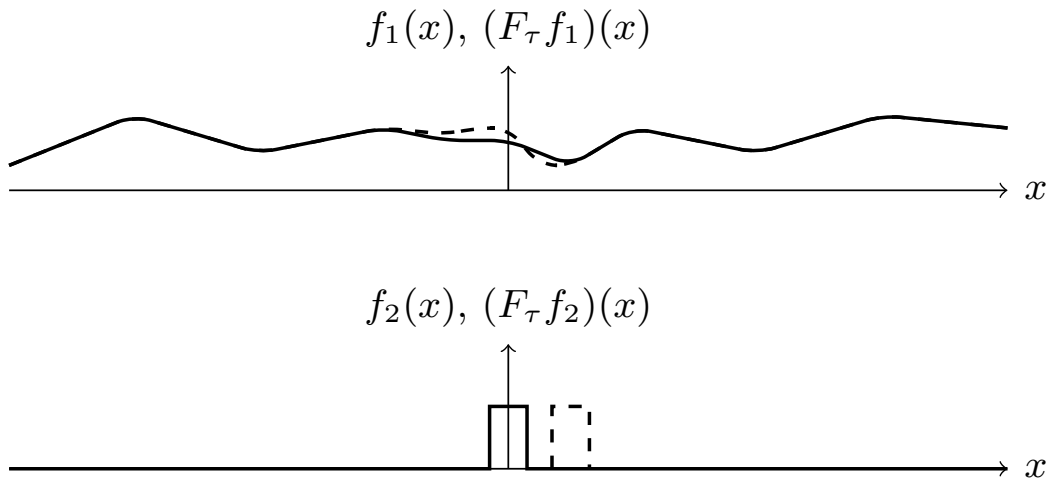
Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$, where $\tau : \mathbb{R}^d \to \mathbb{R}^d$

For "large" τ:



Deformation sensitivity for signal classes

Consider $(F_\tau f)(x) = f(x - \tau(x)) = f(x - e^{-x^2})$



For given τ the amount of deformation induced
can depend drastically on $f \in L^2(\mathbb{R}^d)$

Wiatowski-Bolcskei'15 Deformation Stability Bounds

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f)-\Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The signal class H_W and the corresponding norm $\|\cdot\|_W$ depend on the mother wavelet (and hence the network)

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \qquad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The signal class $\mathcal{C}$ (band-limited functions, cartoon functions, or Lipschitz functions) is independent of the network

Wiatowski-Bolcskei'15 Deformation Stability Bounds

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f)-\Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- Signal class description complexity implicit via norm $\|\cdot\|_W$

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \qquad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Signal class description complexity explicit via $C_{\mathcal{C}}$
  - L-band-limited functions: $C_{\mathcal{C}} = \mathcal{O}(L)$
  - cartoon functions of size K: $C_{\mathcal{C}} = \mathcal{O}(K^{3/2})$
  - M-Lipschitz functions $C_{\mathcal{C}} = \mathcal{O}(M)$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound depends explicitly on higher order derivatives of τ

Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \qquad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound implicitly depends on derivative of τ via the condition $\|D\tau\|_\infty \leq \frac{1}{2d}$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound [*Mallat, 2012*]:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound is *coupled* to horizontal translation invariance

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \qquad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

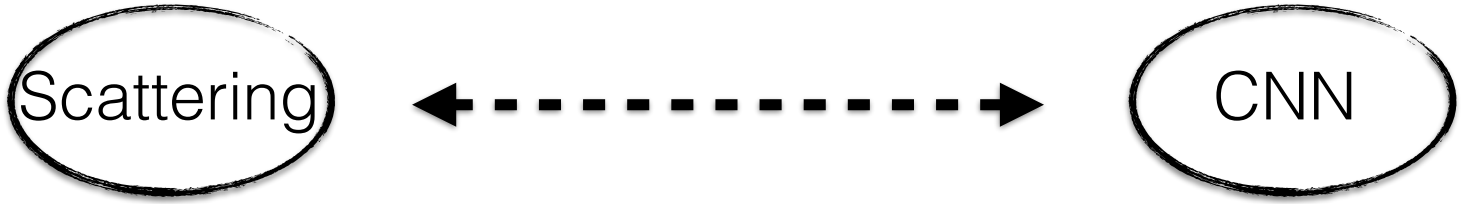
Our deformation sensitivity bound:

$$|||\Phi(F_\tau f) - \Phi(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \qquad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound is *decoupled* from vertical translation invariance

$$\lim_{n\to\infty} |||\Phi^n(T_t f) - \Phi^n(f)||| = 0, \qquad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

What is in between?



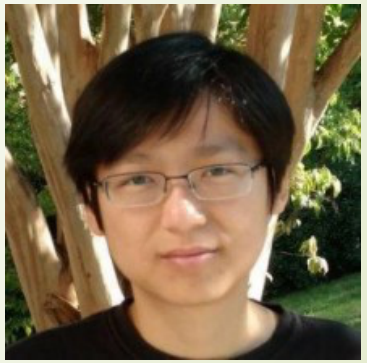
- No training until the classifier
- No parameters in the convolutional layers
- Most “control” of regularity and robustness
- Strong performance and explainable features

- Fully trained by large volume of data
- Lots of parameters (largest model capacity)
- Least “control” of regularity and robustness
- Best performance but not explainable

Decomposed Convolutional Filters (DCF)

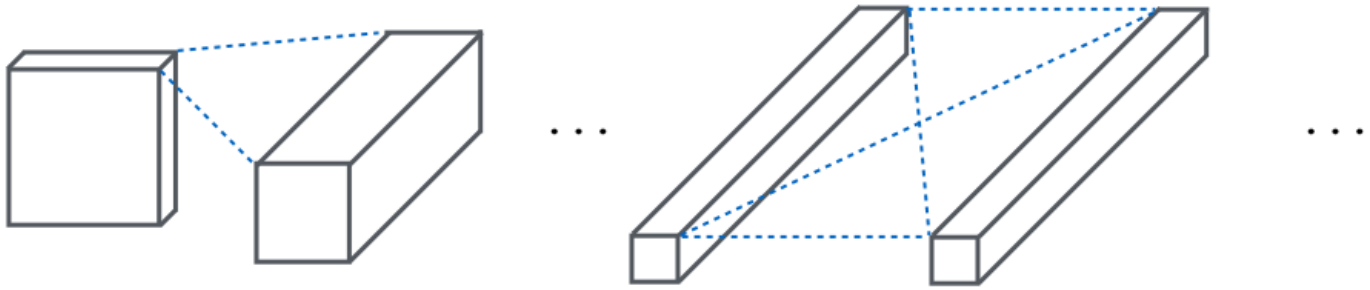
Xiuyuan Cheng et al.

https://arxiv.org/abs/1802.04145



Decomposition of Convolutional Filters

$$x^{(0)} \mapsto x^{(1)} \mapsto \cdots \mapsto x^{(l-1)} \mapsto x^{(l)} \mapsto \cdots$$



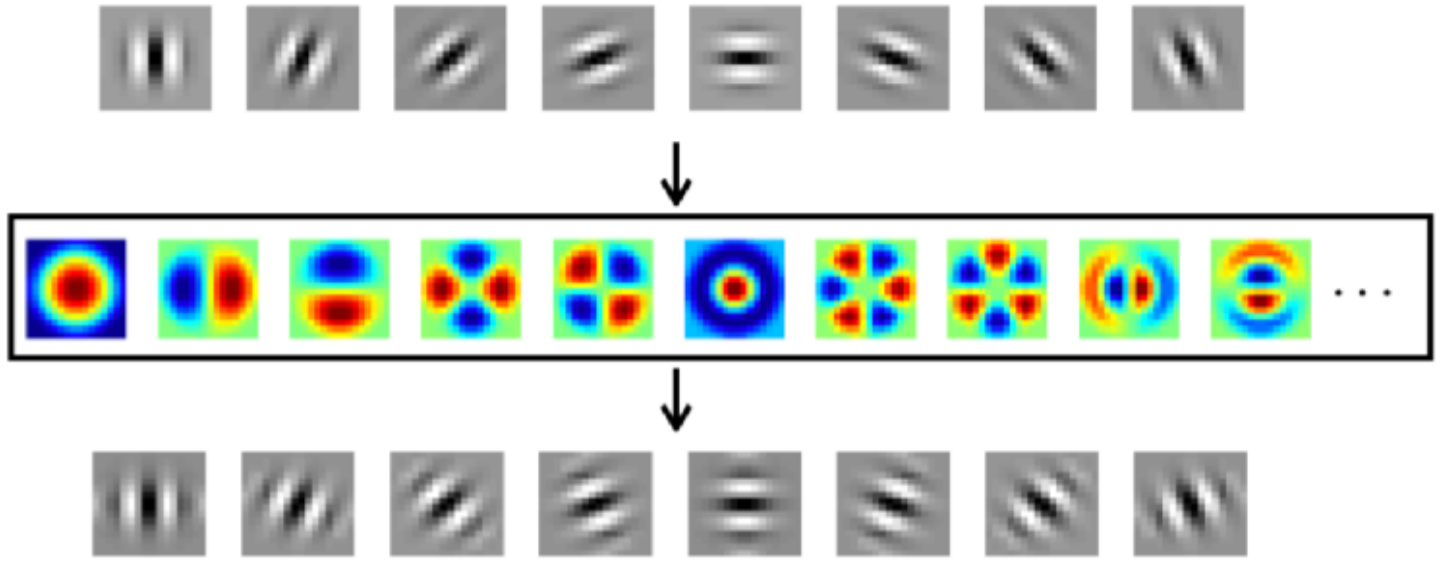
The mapping in a convolutional layer

$$x^{(l)}(u,\lambda) = \sigma\left(\sum_{\lambda'} \int W^{(l)}_{\lambda',\lambda}(v') x^{(l-1)}(u+v',\lambda') dv' + b^{(l)}(\lambda)\right)$$

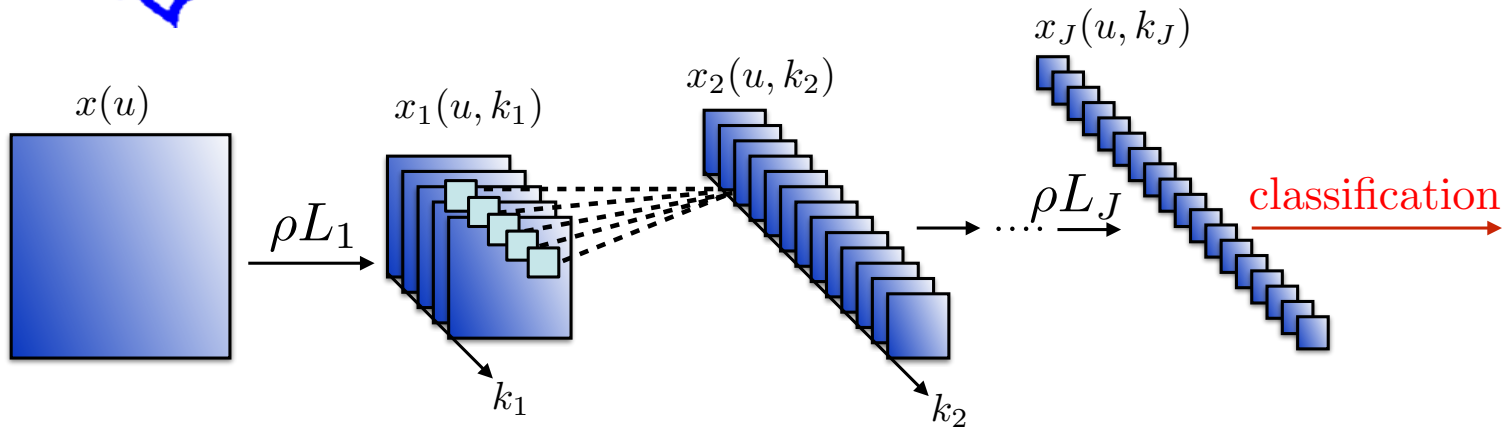
Decomposition of Convolutional Filters

Introducing bases ψ_k

$$W_{\lambda',\lambda}(u) = \sum_{k=1}^{K} (a_{\lambda',\lambda})_k \psi_k(u).$$



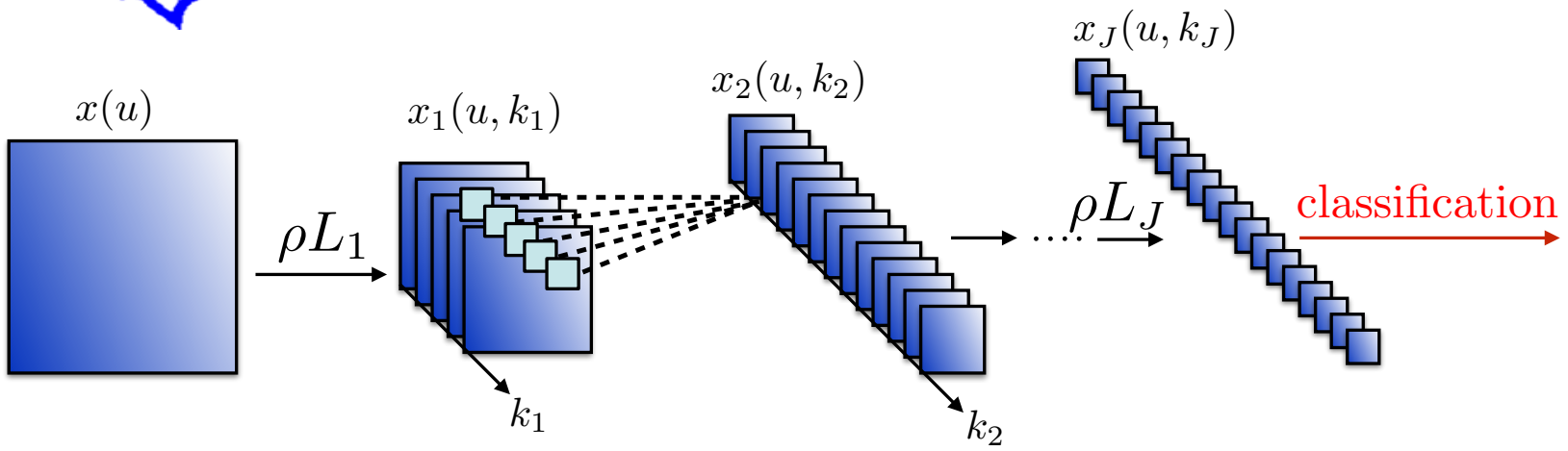
Deep Convolutional Networks



- The convolution network operators L_j have many roles:
  - Linearize non-linear transformations (symmetries)
  - Reduce dimension with projections
  - Memory storage of « characteristic » structures

- Difficult to separate these roles when analyzing learned networks

Open Problems



- Can we recover symmetry groups from the matrices *Lj* ?
- What kind of groups ?
- Can we characterise the regularity of *f(x)* from these groups ?
- Can we define classes of high-dimensional « regular » functions that are well approximated by deep neural networks ?
- Can we get approximation theorems giving errors depending on number of training exemples, with a fast decay ?

Thank you!