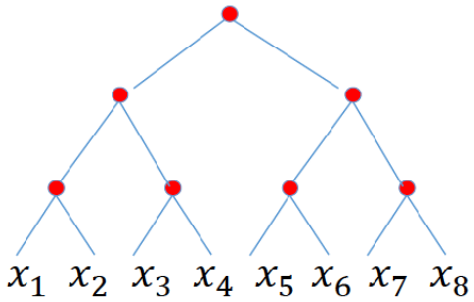
On Mathematical Theories of Deep Learning: II.

Yuan YAO

HKUST

Hierarchically local compositionality

$$f(x_1, x_2, ..., x_8) = g_3(g_{21}(g_{11}(x_1, x_2), g_{12}(x_3, x_4))g_{22}(g_{11}(x_5, x_6), g_{12}(x_7, x_8)))$$



Theorem (informal statement)

Suppose that a function of d variables is hierarchically, locally, compositional . Both shallow and deep network can approximate f equally well. The number of parameters of the shallow network depends exponentially on d as $O(\varepsilon^{-d})$ with the dimension whereas for the deep network dance is $O(d\varepsilon^{-2})$

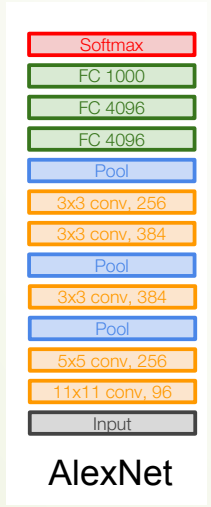
CENTER FOR Brains Minds + Machines

Mhaskar, Poggio, Liao, 2016

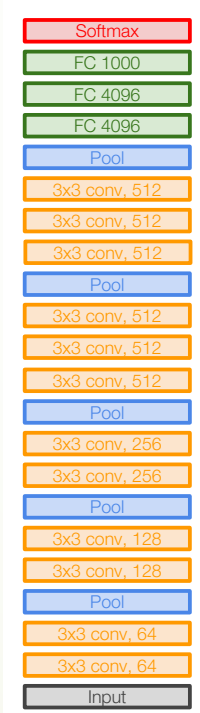
Convolutional Neural Networks (VGG, ResNet etc.) are of this type.

Local filters of small receptive fields (sparsity) are the key to avoid the curse-of-dimensionality

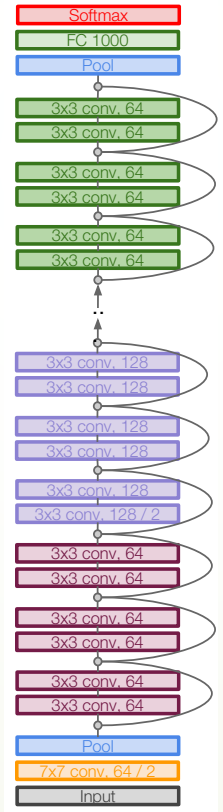
Stacking local filters -> large receptive fields



AlexNet

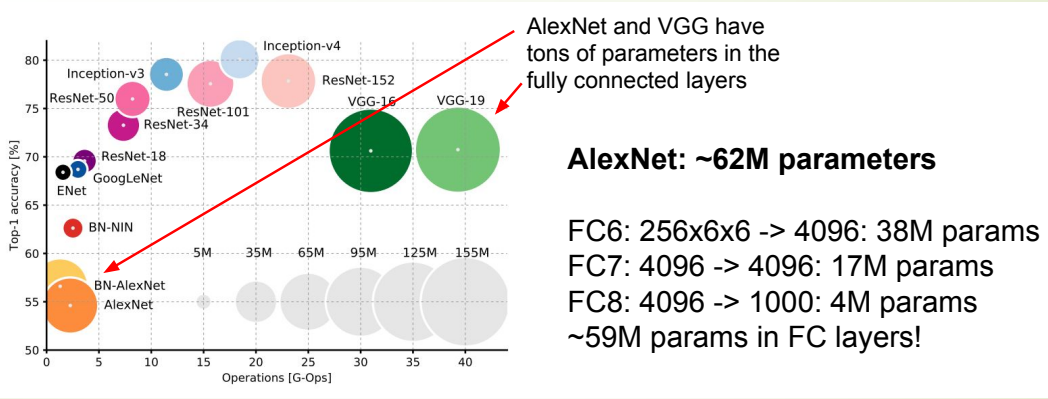


VGG16



ResNet

Fully connected layers -> explosion of parameters



AlexNet: ~62M parameters

FC6: 256x6x6 -> 4096: 38M params
FC7: 4096 -> 4096: 17M params
FC8: 4096 -> 1000: 4M params
~59M params in FC layers!

Important Special Cases in Statistics

Minimax rates of estimations (Stone 1982): if a regression function f is Lipschitz on $\mathbb{R}^d$ α with $0 < \alpha < 1$, then the optimal minimax rate of statistical regression estimators with N samples is $N^{-\frac{2\alpha}{2\alpha+d}}$.

Additive models (Stone 1985): $f(x_1, \ldots, x_d) = f_1(x_1) + \ldots + f_d(x_d)$ with minimax rate $N^{-\frac{2\alpha}{2\alpha+1}}$.
Raskutti-Wainwright-Yu (IEEE TIT, 2011), Yuan-Zhou (AoS, 2016), ...

Interaction models (Stone 1994): $f = \sum_{I \subseteq \{1,\ldots,d\}, |I|=d^*} f_I(x_I)$ with minimax rate $N^{-\frac{2\alpha}{2\alpha+d^*}}$. Here $d^* \in \{1, \ldots, d\}$ and for $I = \{i_1, \ldots, i_{d^*}\} \subseteq \{1, \ldots, d\}$ with $|I| = d^*$, $x_I = (x_{i_1}, \ldots, x_{i_{d^*}})$.

Single index models (Härdle and Stoker 1989): $f = g(a \cdot x)$ for some $a \in \mathbb{R}^d$ and $g : \mathbb{R} \to \mathbb{R}$

Projection pursuit (Friedman and Stuetzle 1981): $f(x_1, \ldots, x_d) = \sum_{k=1}^{K} g_k(a_k \cdot x)$ with $K \in \mathbb{N}$, $a_k \in \mathbb{R}^d$ and univariate functions g_k

Hierarchical interaction models (Kohler1 and Krzyzak 2016)

Simple case: $f = g(f_1(x_{I_1}), f_2(x_{I_2}), \ldots, f_{d^*}(x_{I_{d^*}}))$

Generalized hierarchical model: $f = g(a_1 \cdot x, \ldots, a_{d^*} \cdot x)$

Generalized hierarchical interaction model: $f = \sum_k g_k(f_{1,k}, \ldots, f_{d^*,k})$ with $f_{i,k}(x)$ generalized hierarchical model

All models are wrong, but some are useful ... blessing-of-dimensionality?

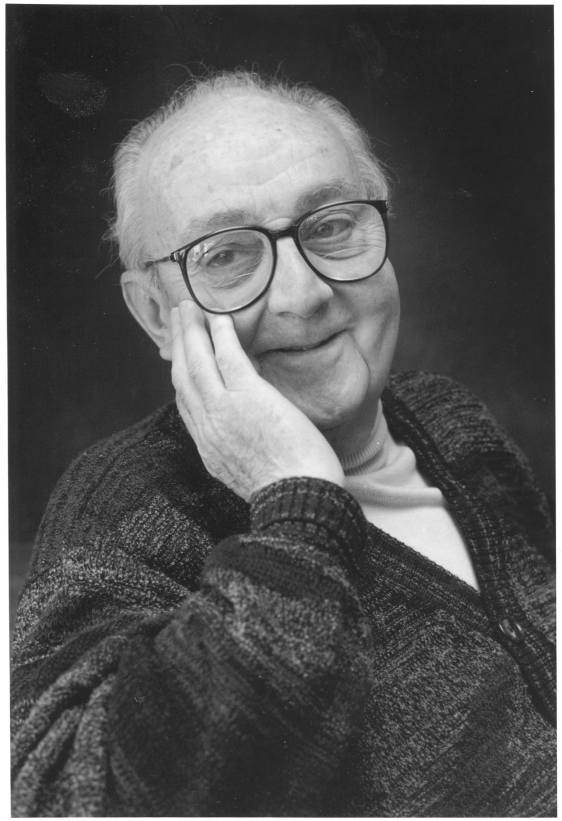


Figure 7: George Box: "Essentially, all models are wrong, but some are useful."

Some Historical Results

- A classical theorem **[Sipser, 1986; Hastad, 1987]** shows that deep circuits are more efficient in representing certain Boolean functions than shallow circuits. Hastad proved that highly-variable functions (in the sense of having high frequencies in their Fourier spectrum) in particular the parity function cannot even be decently approximated by small constant depth circuits
- **Chui-Li-Mhaskar (1994)** shows that multilayer networks can do localized approximation while single layer ones can not. Older examples exist: consider a function which is a linear combination of n tensor product Chui–Wang spline wavelets, where each wavelet is a tensor product cubic spline. It was shown by **Chui and Mhaskar** that is impossible to implement such a function using a shallow neural network with a sigmoidal activation function using O(n) neurons, but a deep network with the activation function $(x_+)^2$ do so. In this case, as we mentioned, there is a formal proof of a gap between deep and shallow networks.
- The main result of **[Telgarsky, 2016, Colt]** says that there are functions with many oscillations that cannot be represented by shallow networks with linear complexity but can be represented with low complexity by deep networks.
- **Eldan and Shamir (2016)** show an example of a function expressible by a 3-layer feedforward neural network cannot be approximated by any 2-layer neural network to certain accuracy unless the width is exponential in the dimension.
- **Shaham-Cloningen-Coifman (2018)**: functions on manifolds and order of approximation by fully connected deep neural networks

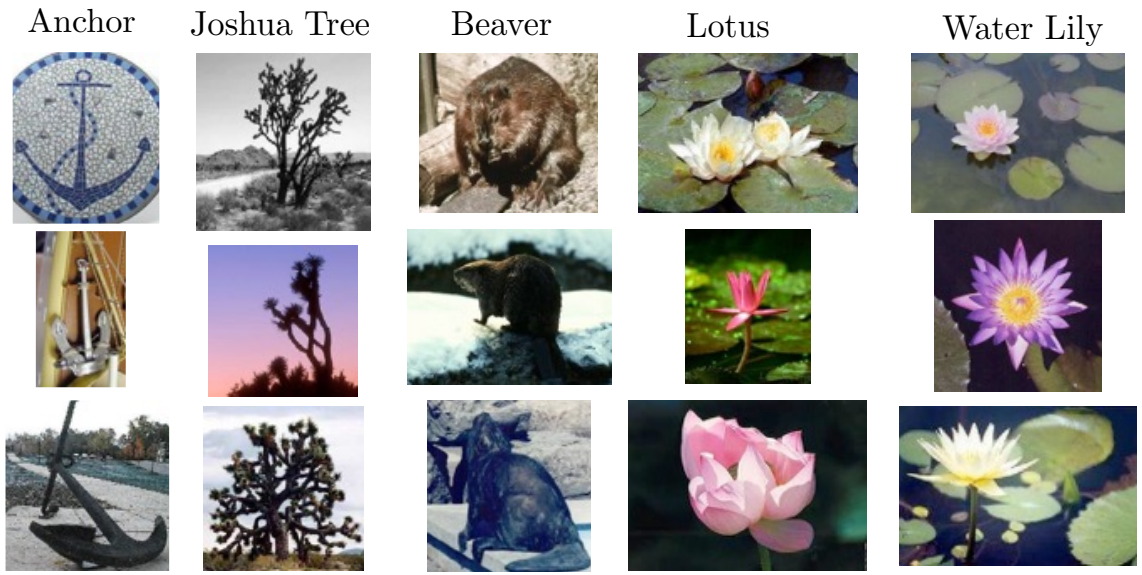
What are the geometric properties of deep networks?

Contraction within-class level sets toward invariants when depth grows, while keeping the separation between classes

High Dimensional Natural Image Classification

- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:
- **Classification:** estimate a class label $f(x)$ given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \leq n}$

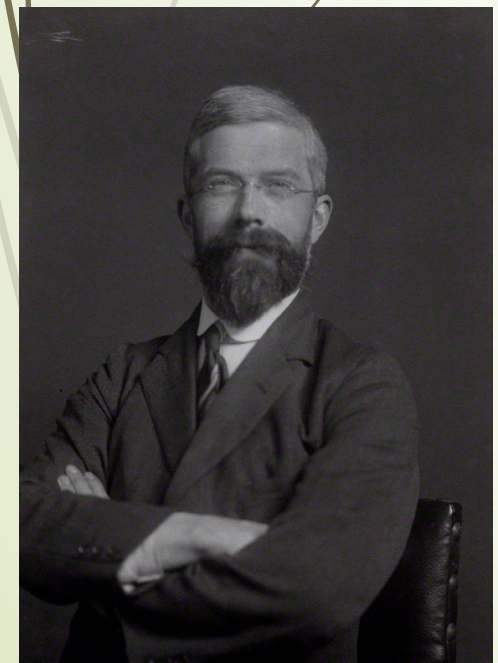
Image Classification $d = 10^6$



Huge variability inside classes

Find invariants

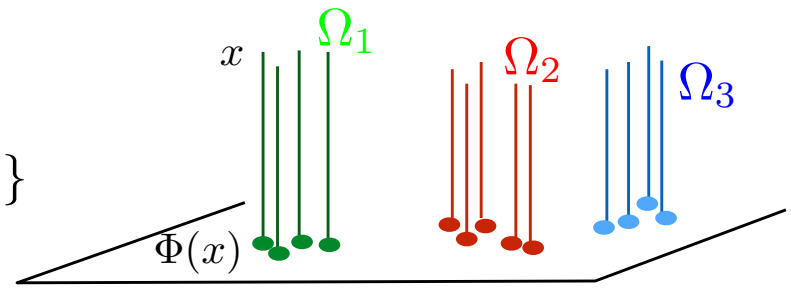
Fisher's Linear Discriminant (1936) (Linear Dimensionality Reduction)



Classes

Level sets of $f(x)$

$$\Omega_t = \{x \ : \ f(x) = t\}$$



If level sets (classes) are parallel to a linear space then variables are eliminated by linear projections: *invariants*.

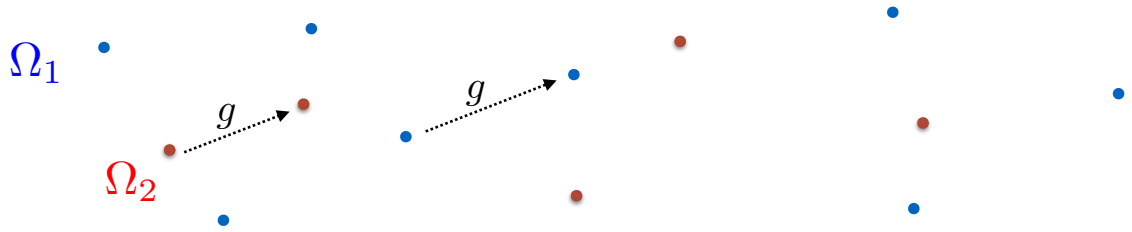
$$\Phi(x) = \alpha \hat{\Sigma}_W^{-1}(\hat{\mu}_1 - \hat{\mu}_0)$$

$$\hat{\mu}_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i \qquad \hat{\Sigma}_W = \sum_k \sum_{i \in C_k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Nonlinear Level Set Group Symmetries

Level Set Geometry: Symmetries

- Curse of dimensionality $\Rightarrow$ not local but global geometry
  Level sets: classes, characterised by their global symmetries.



- A symmetry is an operator g which preserves level sets:

$$\forall x \quad , \quad f(g.x) = f(x) \; : \; \text{global}$$

If g_1 and g_2 are symmetries then $g_1.g_2$ is also a symmetry

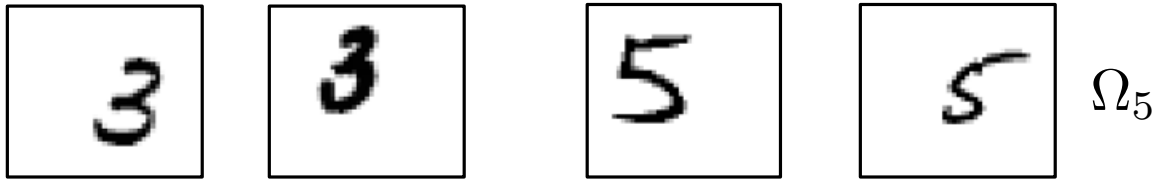
$$f(g_1.g_2.x) = f(g_2.x) = f(x)$$

Level set symmetries lead to groups…

Translation and Deformations

- Digit classification:

$x(u)$ $\quad x'(u) = x(u - \tau(u))$

Ω_3  Ω_5

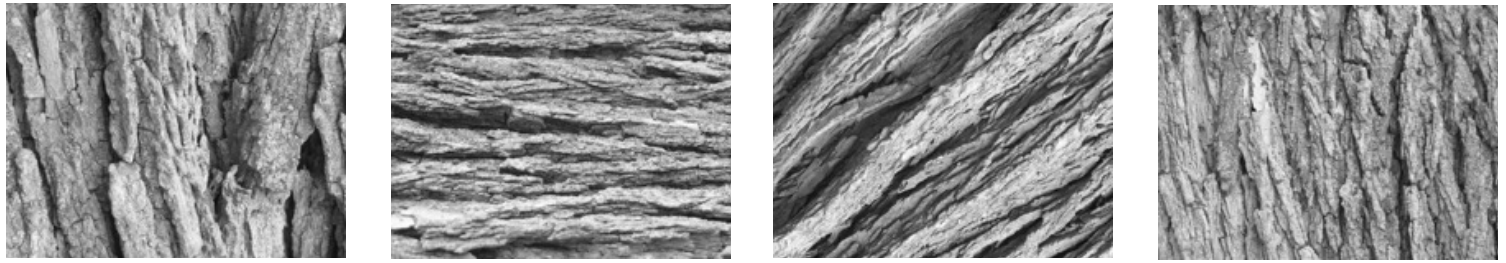
- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Video of Philipp Scott Johnson

https://www.youtube.com/watch?v=nUDIoN-_Hxs

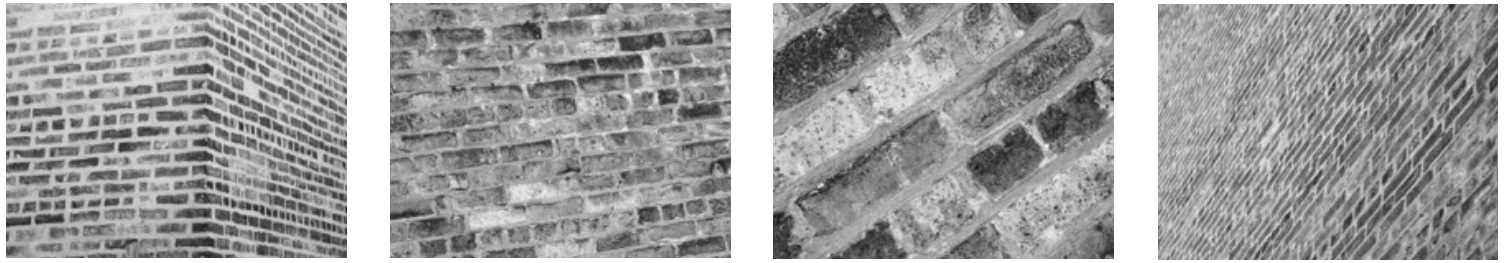
Rotation and Scaling Variability

- Rotation and deformations



Group: $SO(2) \times \mathrm{Diff}(SO(2))$

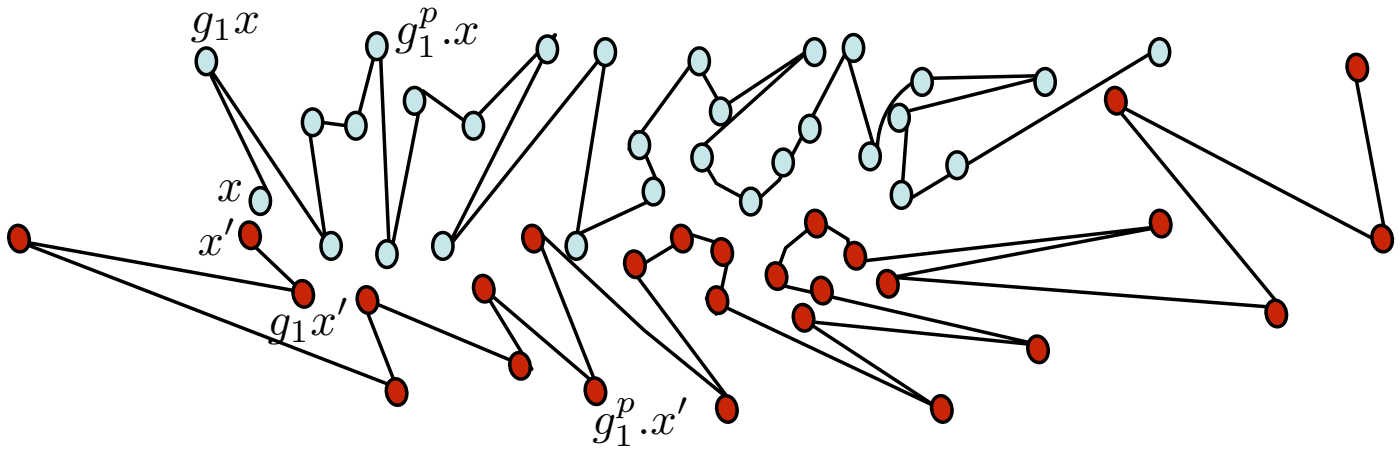
- Scaling and deformations



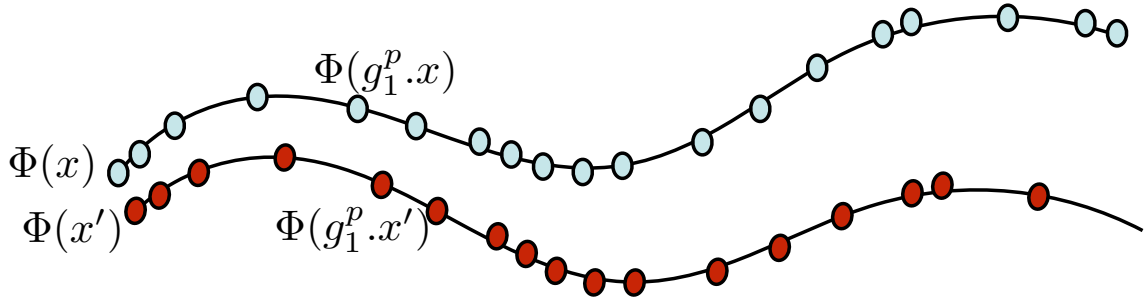
Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

Linearize Symmetries

- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g \in G}$



- Linearise symmetries with a change of variable $\Phi(x)$



- Lipschitz: $\forall x, g \ : \ \|\Phi(x) - \Phi(g.x)\| \leq C \|g\|$

Wavelet Scattering Net

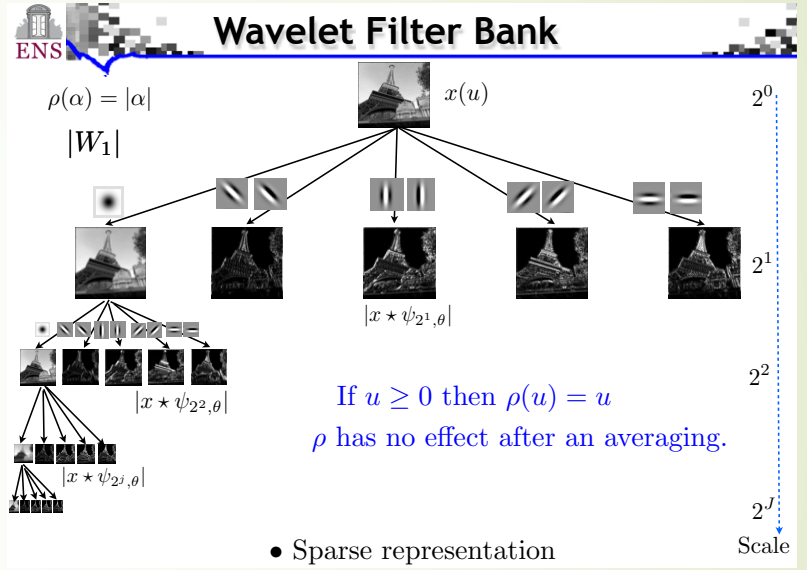
Stephane Mallat et al. 2012

- Architechture:
  - Convolutional filters: band-limited complex wavelets
  - Nonlinear activation: modulus (Lipschitz)
  - Pooling: averaging (L1)
- Properties:
  - A Multiscale Sparse Representation
  - Norm Preservation (Parseval's identity):

$$\|Sx\| = \|x\|$$

  - Contraction:

$$\|Sx - Sy\| \leq \|x - y\|$$



$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \cdots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Invariants/Stability of Scattering Net

- **Translation Invariance** (generalized to **rotation** and **scaling**):

  - 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:

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

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

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.
  - As depth grows, the multiplicative pooling factors leads to full invariances.

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

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

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

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

Summary

- All these works partially explains the success of CNNs
  - Contraction within level set symmetries toward invariance when depth grows (invariants)
  - Separation kept between different levels (discriminant)
- Other questions?
  - Can one adaptively learn some networks with the same invariant properties as the scattering net?
  - How deep networks generalize well without overfitting?
  - What's the landscape of empirical risks and how to efficiently optimize?