Stability to diffeomorphisms and translations.

Edouard Oyallon
edouard.oyallon@cnrs.fr
CNRS, ISIR

Machine Learning & Deep Learning for Information Access

Reminders and complements

Wavelet Transform

$$\psi, \phi \text{ regular with } \int_{\mathbb{R}^2} \psi(u)du = 0$$

A Wavelet Transform is given by:

$$Wx = \{x \star \psi_{j,\theta}, x \star \phi_J\}_{\theta, j \leq J}$$

where

$$\psi_{j,\theta} = \frac{1}{2^{2j}} \psi(\frac{-r_\theta(u)}{2^j})$$

$$\phi_J(u) = \frac{1}{2^{2j}} \phi(\frac{u}{2^J})$$

such that:

$$(1-\epsilon)\|x\|^2 \leq \|Wx\|^2 \leq \|x\|^2$$

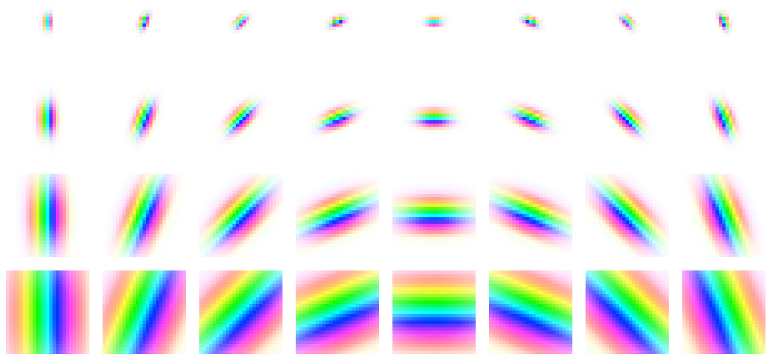
$$\Leftrightarrow$$

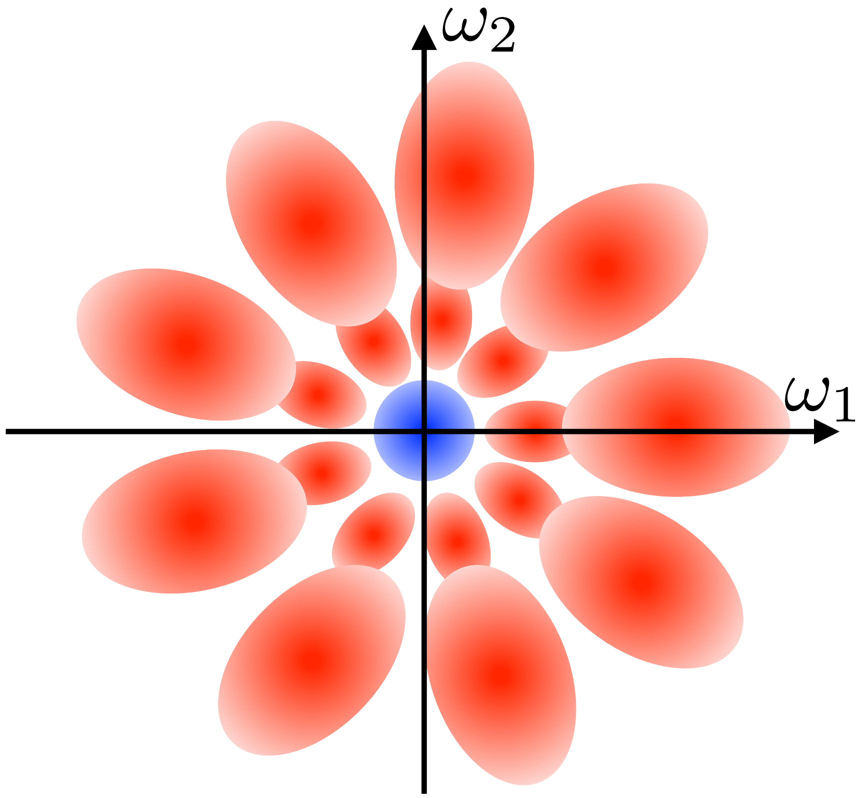
$$1-\epsilon \leq \sum_{j,\theta} |\hat{\psi}_{j,\theta}(\omega)|^2 + |\hat{\phi}_J(\omega)|^2 \leq 1$$

$$\Leftrightarrow$$

$$1-\epsilon \leq \sum_{j,\theta} |\hat{\psi}(2^j(r_{-\theta}\omega))|^2 + |\hat{\phi}(2^J\omega)|^2 \leq 1$$

Gabor wavelets





basis for sparsifying a signal
+observed in DNNs

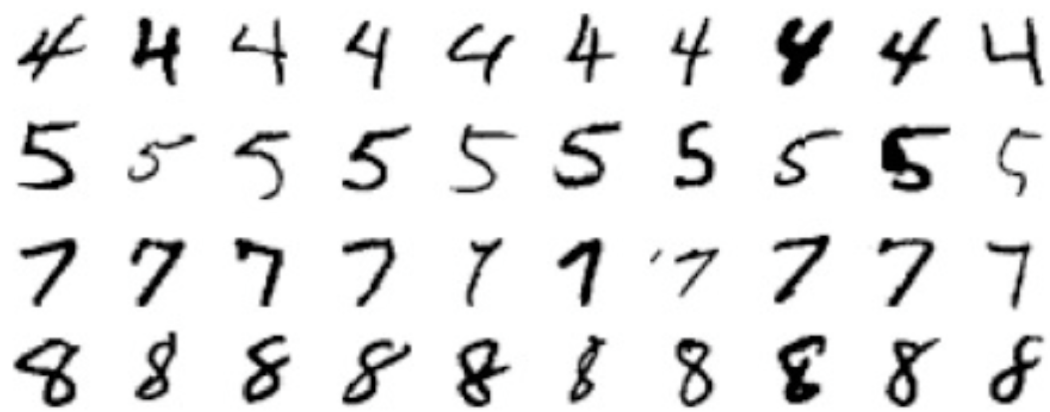
We will discuss widely the Scattering Transform.

Ref.: Invariant Convolutional Scattering Network, J. Bruna and S Mallat

- Successfully used in several applications:

All variabilities are known

  - Digits

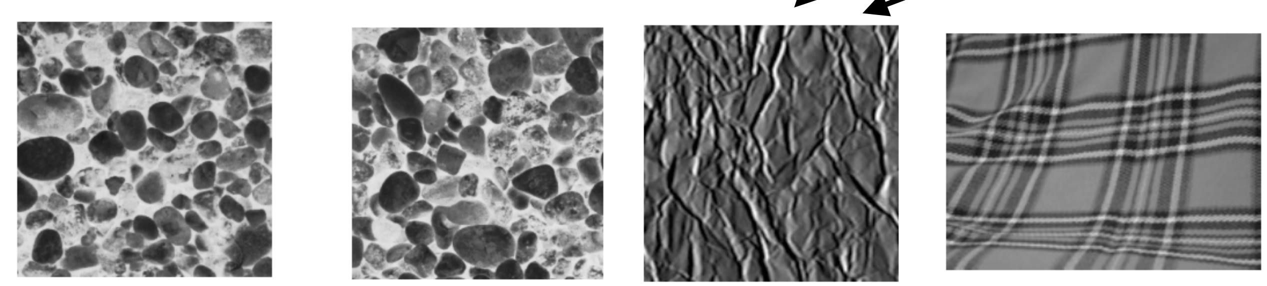


Small deformations +Translation

  - Textures

Ref.: Rotation, Scaling and Deformation Invariant Scattering for texture discrimination, Sifre L and Mallat S.

Rotation+Scale



- The design of the scattering transform is guided by the euclidean group.

- A scattering transform is a combination of complex-valued wavelets and modulus non-linearity.

Motivation to introduce the Scattering Transform

- Compared to a Deep Neural Network, the Scattering Transform is well-understood mathematically when all variabilities are deduced from symmetries.

- Involves no-learning: allows to obtain guarantees on the filters.

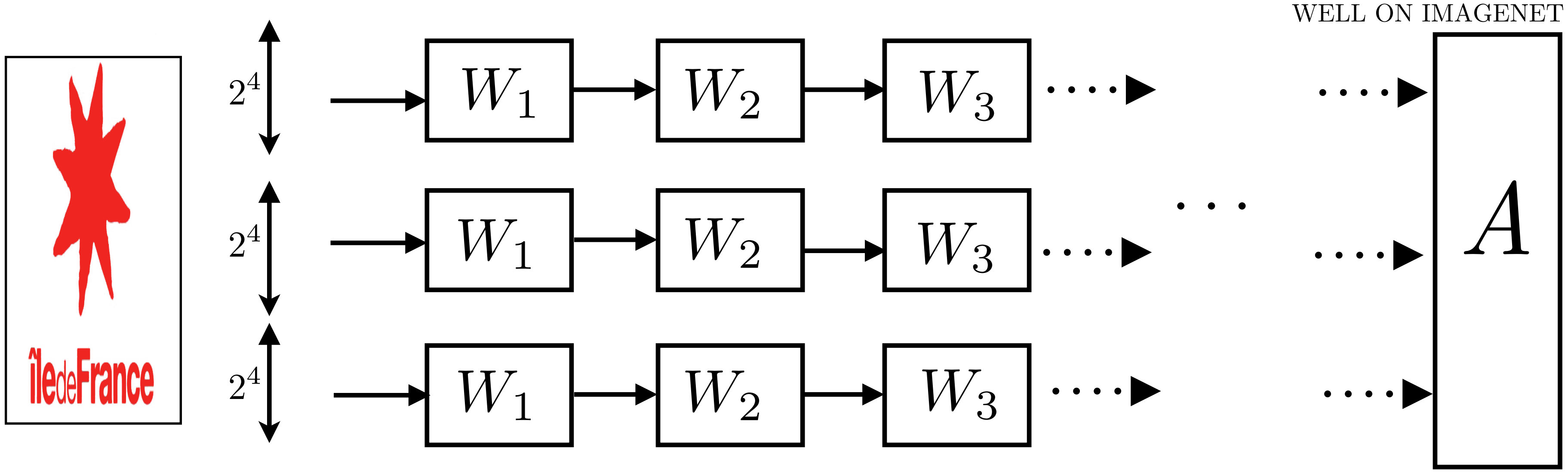
- State-of-the-art on simple benchmarks… what about complex ones?

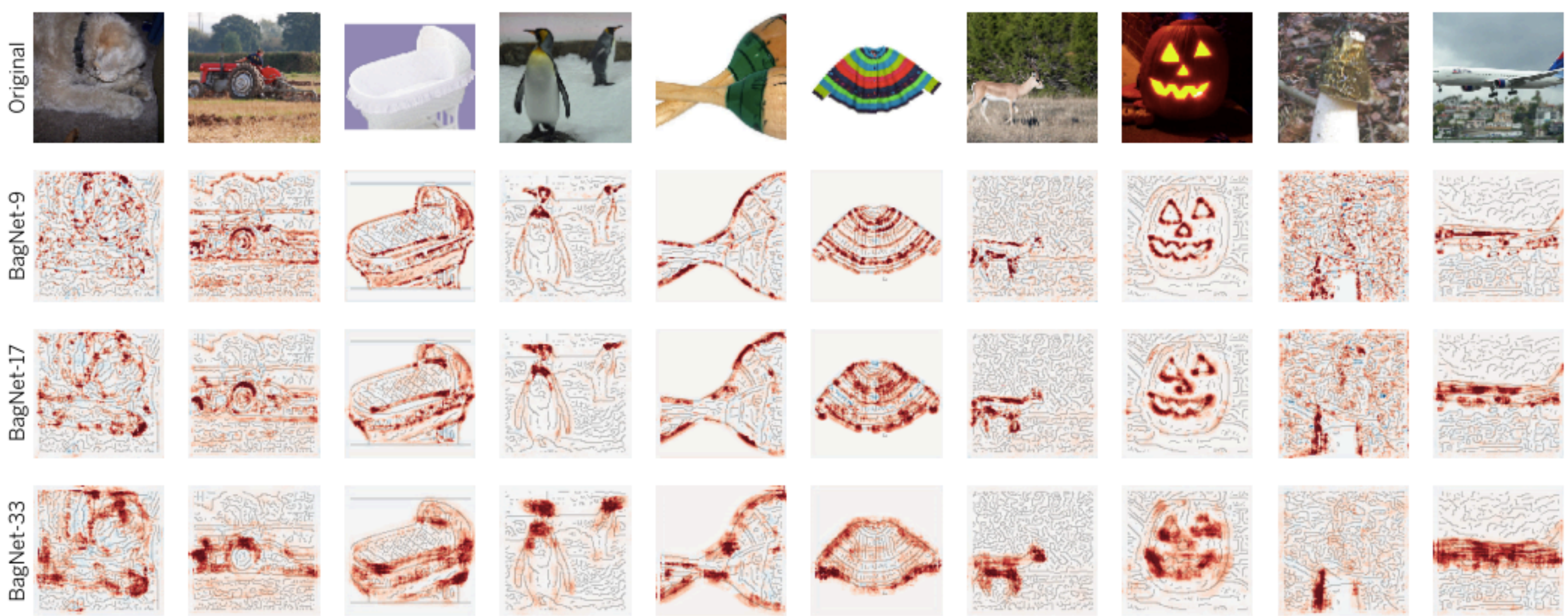
- Let us discuss some benchmarks.

Two surprising architectures (1)

BagNet (2019):

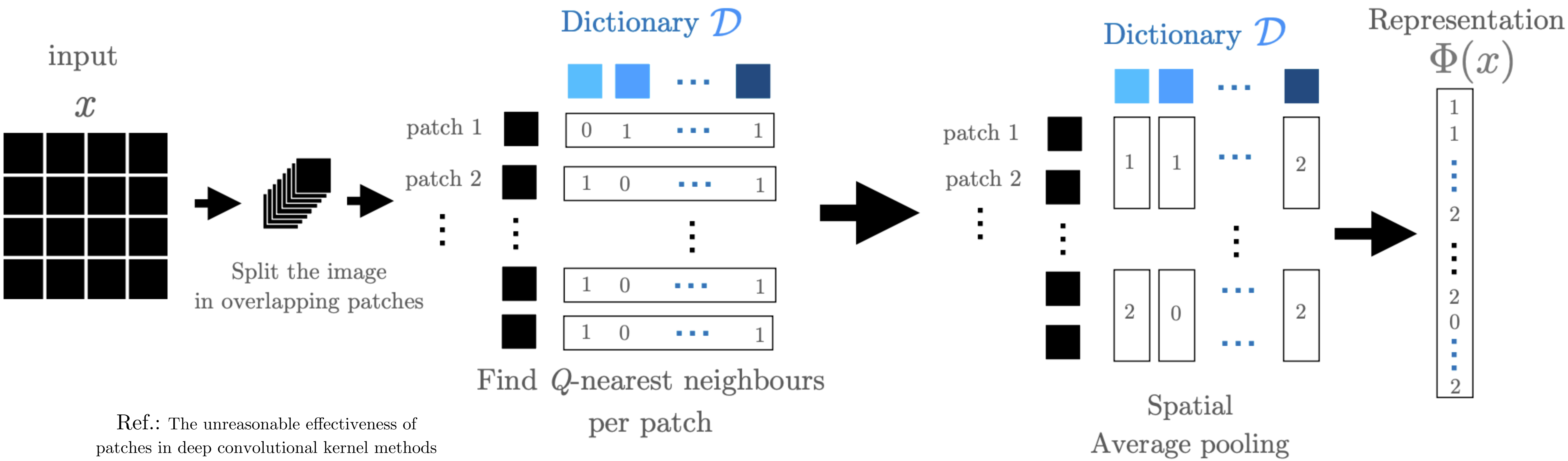
Ref.: APPROXIMATING CNNS WITH BAG-OF-LOCALFEATURES MODELS WORKS SURPRISINGLY WELL ON IMAGENET



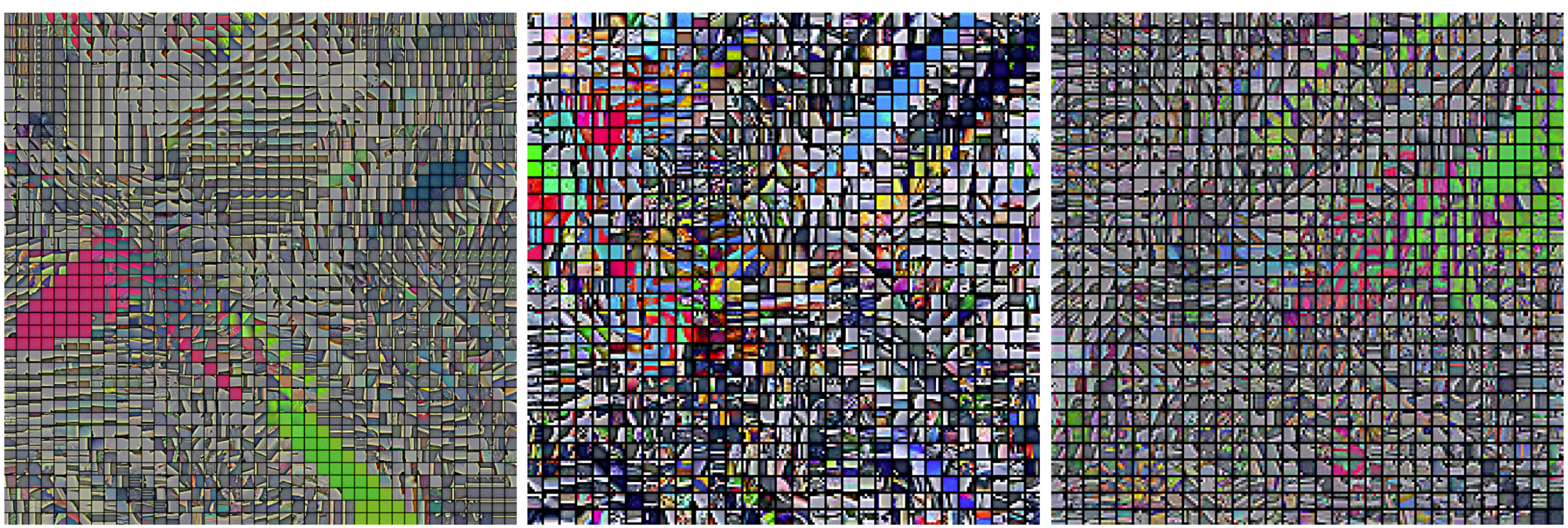


Two surprising architectures (2)

Patches (2021)

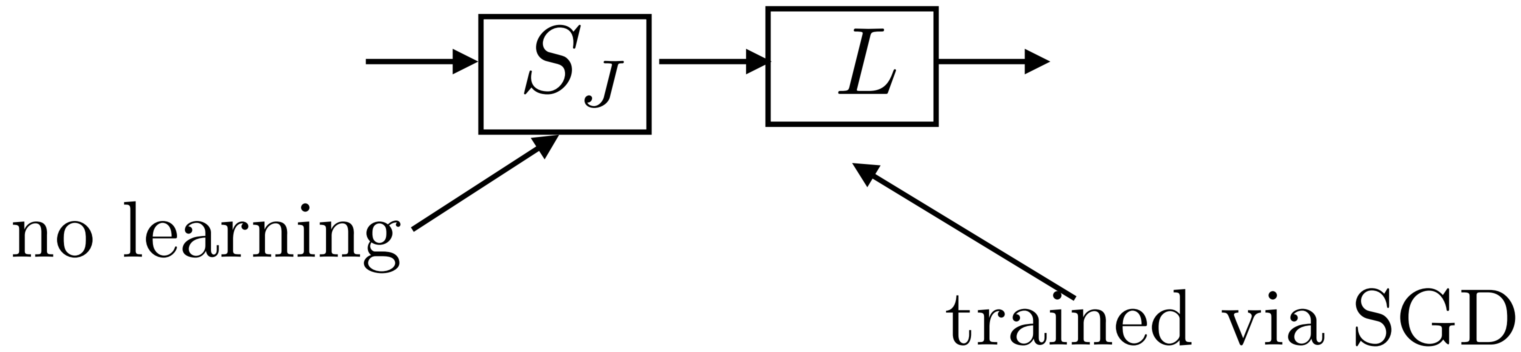


Ref.: The unreasonable effectiveness of patches in deep convolutional kernel methods

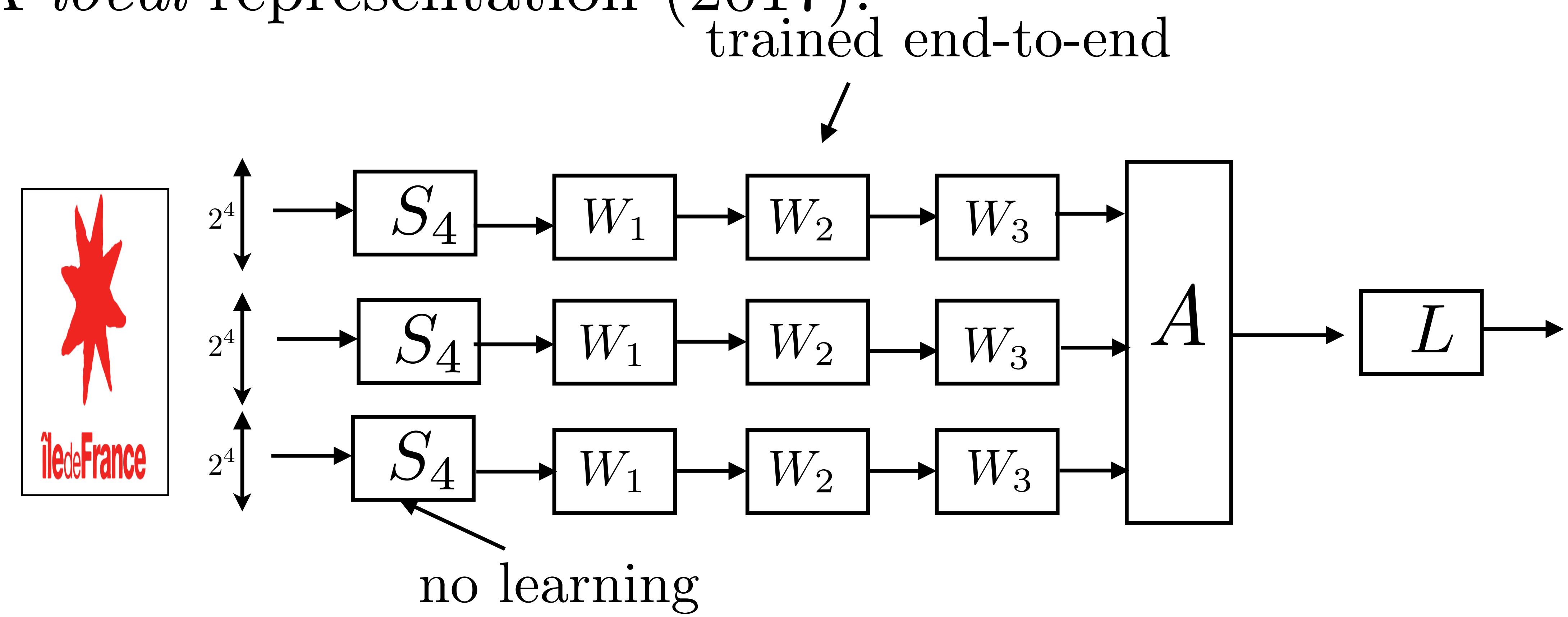


Scattering Classification Pipeline

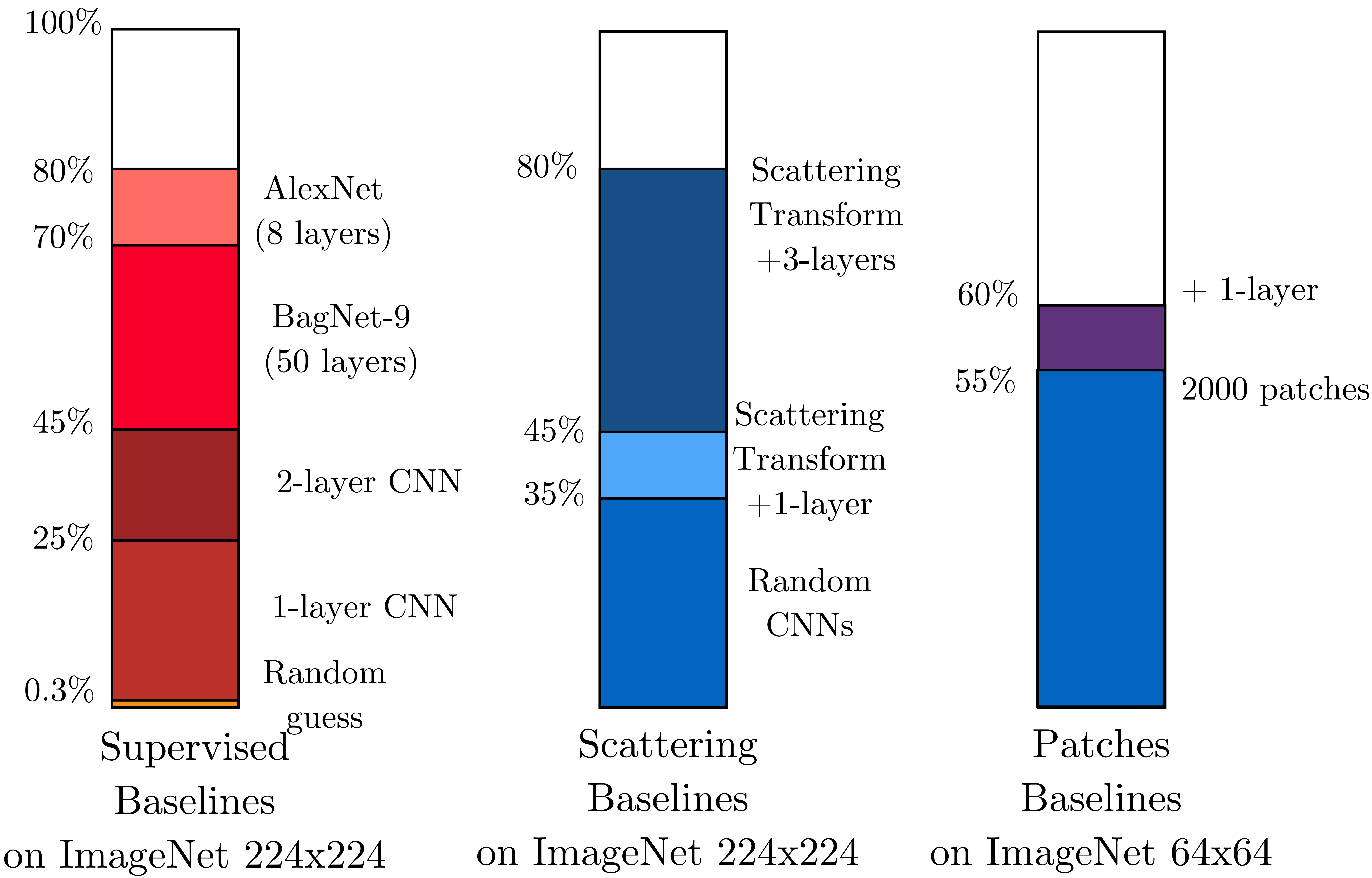
- A *global* representation (2013):



- A *local* representation (2017):



ImageNet Top5-Accuracy



Stability to groups of symmetries

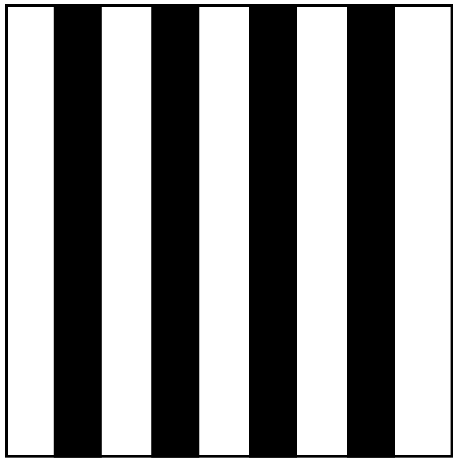
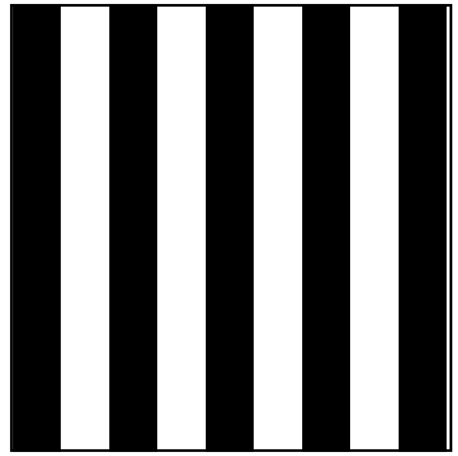
Group Invariance?

3 methods to get invariance w.r.t. translations

- Linear averaging
- Non-linear Fourier modulus
- Wavelets

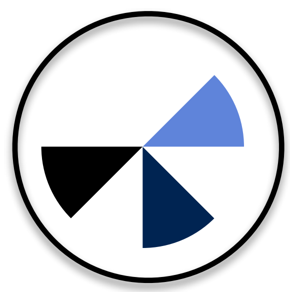
High dimensionality issues

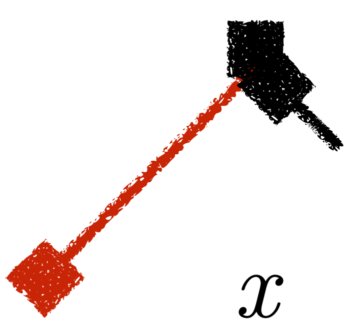
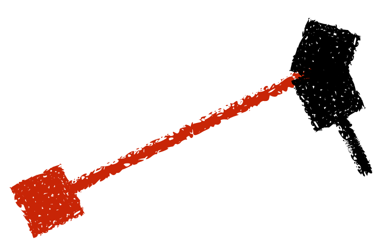
Translation

x y

Rotation

x y

$$\|x - y\|_2 = 2$$

Averaging is the key to get invariants

An example: translation

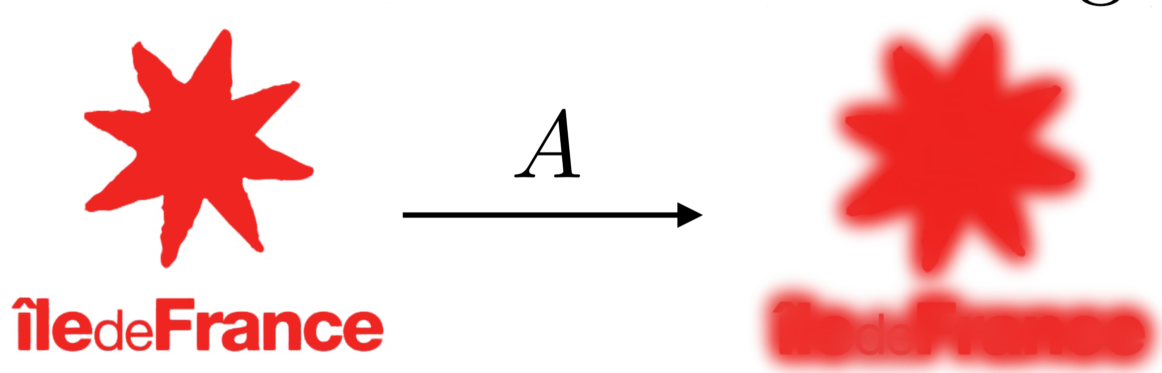
- Translation is a linear action:

$$\forall u \in \mathbb{R}^2, L_a x(u) = x(u - a)$$

- In many cases, one wish to be invariant globally to translation, a simple way is to perform an averaging:

$$Ax = \int L_a x da = \int x(u) du$$

- Even if it can be localized, the averaging keeps the low frequency structures: the invariance brings a loss of information!

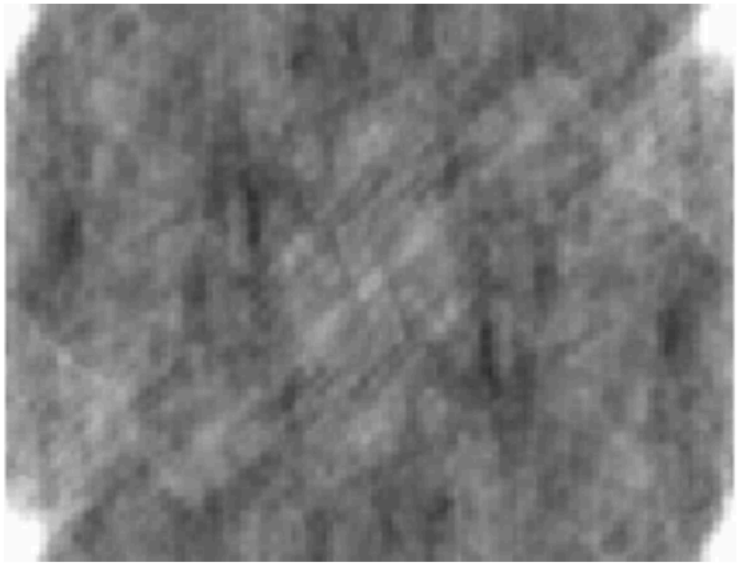


Another example
Fourier moduli

- Consider: $x \to |\mathcal{F}x|$

- This is clearly invariant to translations but…

x



$\mathcal{F}^{-1}|\mathcal{F}x|$

Modulus reconstruction

$\mathcal{F}^{-1}(\frac{\mathcal{F}x}{|\mathcal{F}x|})$

Phase reconstruction

An example of non trivial-invariants on non-trivial groups: the roto-translation

Ref.: PhD of L. Sifre

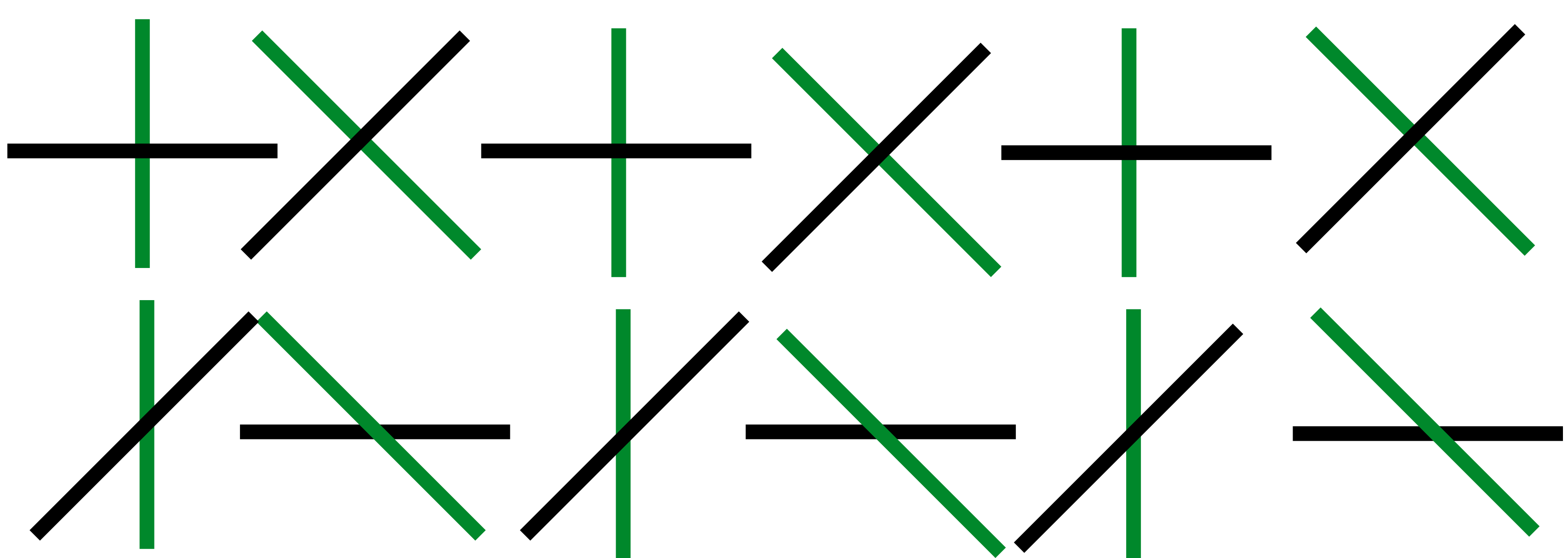
- Roto-translation $SL(E) = \mathbb{R}^2 \ltimes SO_2(\mathbb{R})$ is a non commutative group:

$$(u, \theta).(v, \varphi) = (u + r_\theta v, \theta + \varphi)$$

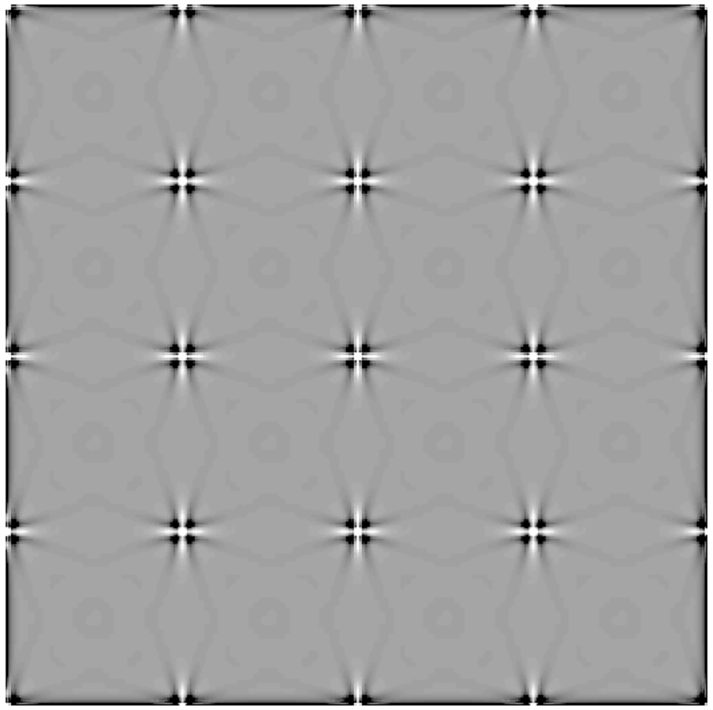
$$g.g' = \mathcal{L}_u r_\theta \mathcal{L}_v r_{-\theta} r_\theta r_\varphi = (u + r_\theta v, \theta + \varphi)$$

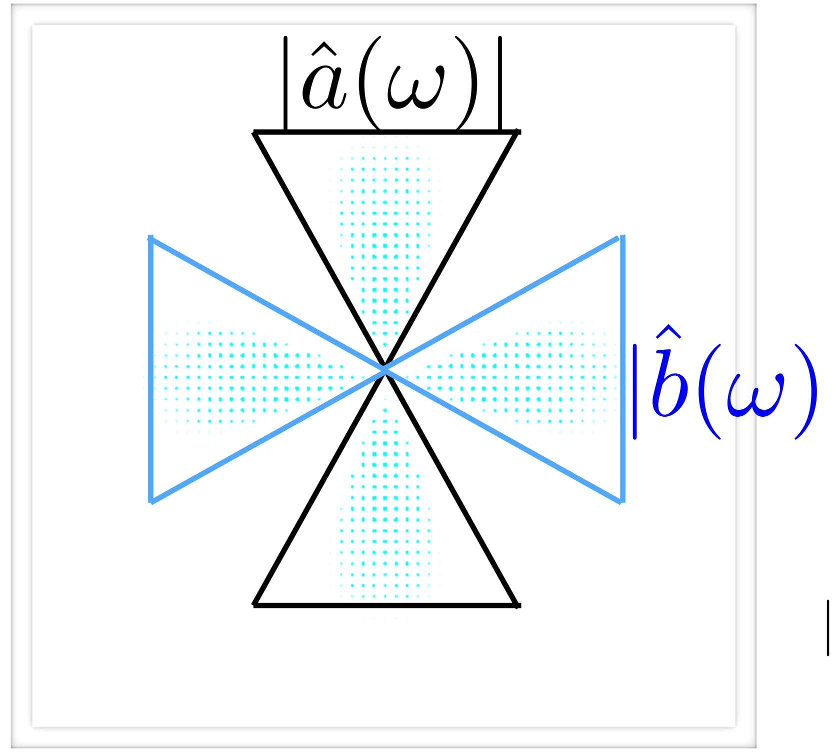
- We can define convolutions, Fourier, along this group!

Why?
Difficult to distinguish:



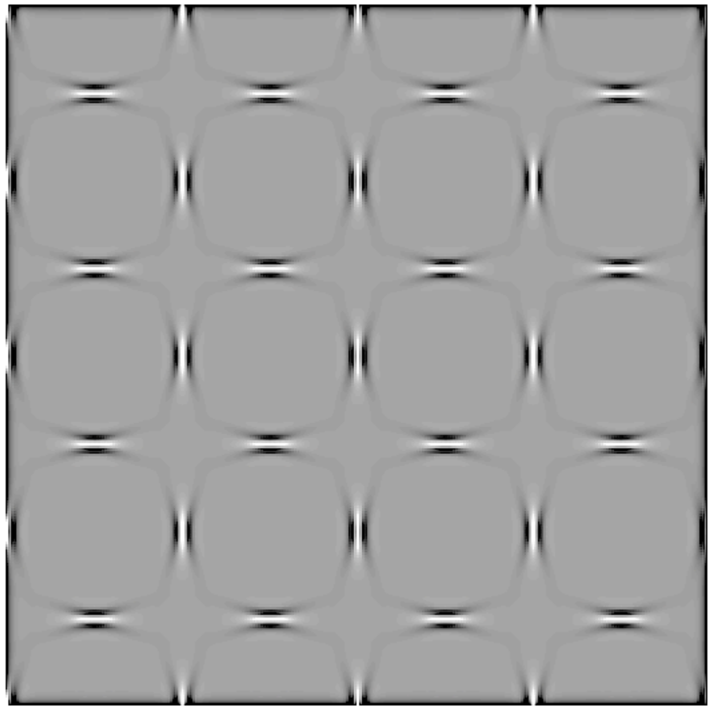
Fourier modulus of the same images. . .

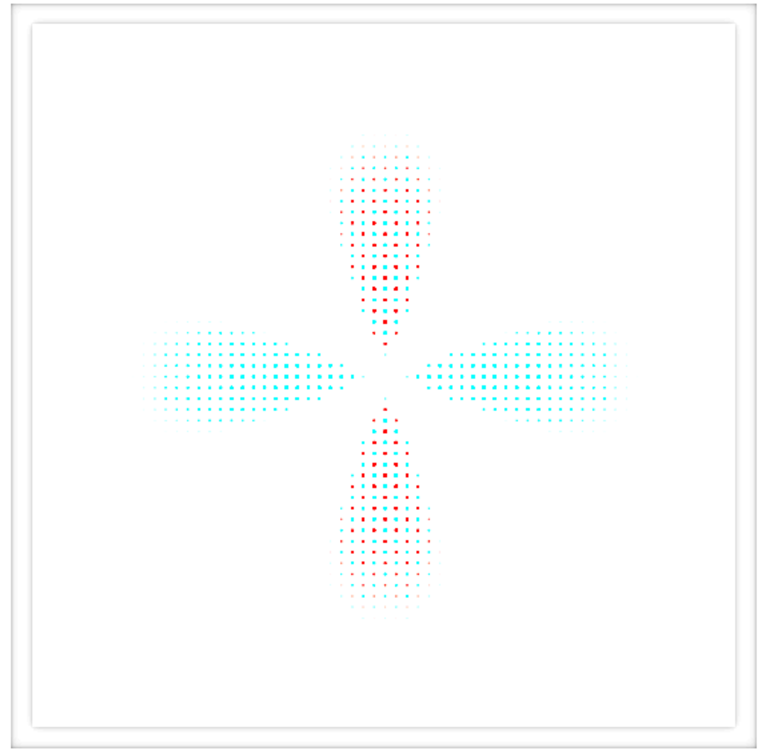




here:

$$|\hat{a}(\omega) + \hat{b}(\omega)| = |\hat{a}(\omega)| + |\hat{b}(\omega)|$$





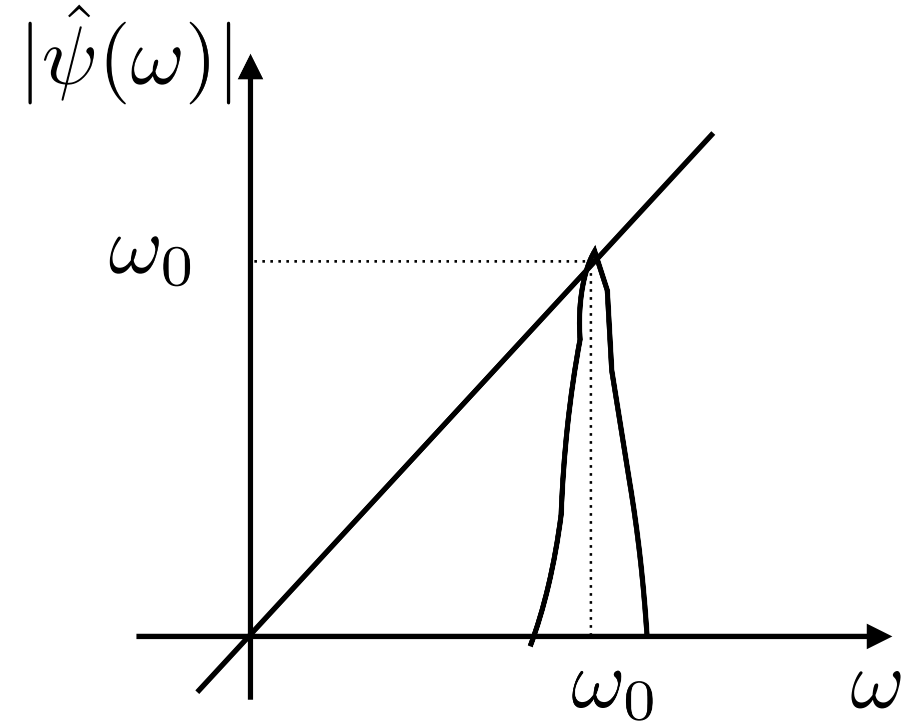
$$\tilde{a} = L_u(a) \Rightarrow |\hat{\tilde{a}}| = |\hat{a}|$$

Invariances via analytic wavelets

- Analytic wavelets permit to build stable invariants to small translations by a:

Ref.: Group Invariant Scattering, Mallat S

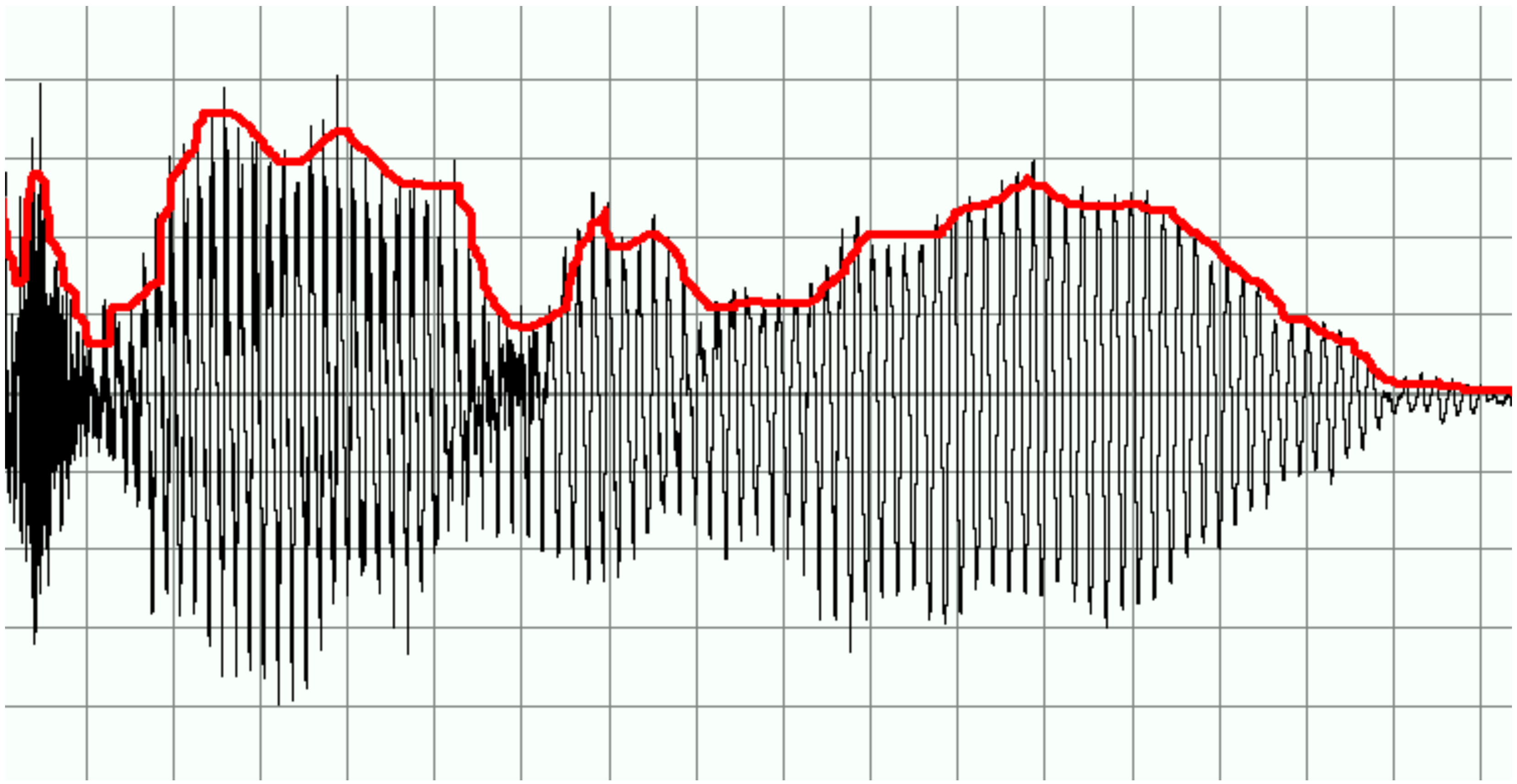
$$L_a\widehat{(x \star \psi)} =$$



$$\omega^T a\hat{\psi}(\omega) \approx \omega_0^T a\hat{\psi}(\omega)$$

Link with the AlexNet?

Demodulation:



For a Gabor wavelet:

average pooling

$$\int_0^T |x \star \psi_\lambda(u)|\, du \approx \int_0^T \mathrm{ReLU}(x \star \mathcal{R}(\psi_\lambda)(u))\, du$$

up to constants

Stability to deformations?

Diffeomorphism

- Let $E, F \subset \mathbb{R}^d$, $\phi : E \to F$,E, F open. ϕ is a diffeomorphism if:
  -ϕ is bijective
  - both ϕ, ϕ^{-1} are differentiable.
  It is said $\mathcal{C}^k$ if ϕ, ϕ^{-1} are $\mathcal{C}^k$. Smooth if ϕ, ϕ^{-1} are $\mathcal{C}^k$ for any k.

- <u>Theorem (Local inversion)</u>: Let $\phi : \Omega \to \mathbb{R}^d$ a $\mathcal{C}^k$ function with $\Omega \subset \mathbb{R}^d$ open .
  If $\det(\partial\phi(x)) \neq 0$ then there is $\mathcal{U}, \mathcal{V}$ open sets with $x \in \mathcal{U}$ such that $\phi : \mathcal{U} \to \mathcal{V}$ is a $\mathcal{C}^k$ -diffeomorphism.

- <u>Theorem (Global inversion):</u> Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ a $\mathcal{C}^k$ function, then ϕ is a $\mathcal{C}^k$-diffeomorphism with $\phi(\mathbb{R}^d) = \mathbb{R}^d$ if and only if for any x: $\det(\partial\phi(x)) \neq 0$

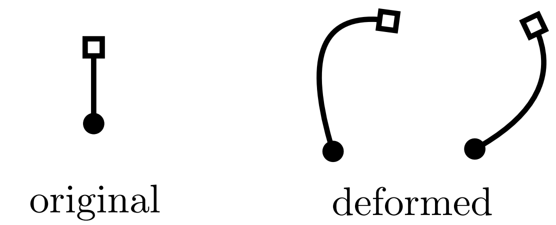
$$\text{and} \quad \lim_{\|x\| \to \infty} \|\phi(x)\| = \infty$$

Deformations:

- Consider: $\tau \in C^\infty$ and define:

$$\|\tau\|_\infty = \sup_u \|\tau\| \quad \text{and} \quad \|\nabla\tau\|_\infty = \sup_u \|\nabla\tau\|$$

If $\|\nabla\tau\|_\infty < 1$ then: $\mathbf{I} - \tau \in \mathrm{Diff}^\infty$

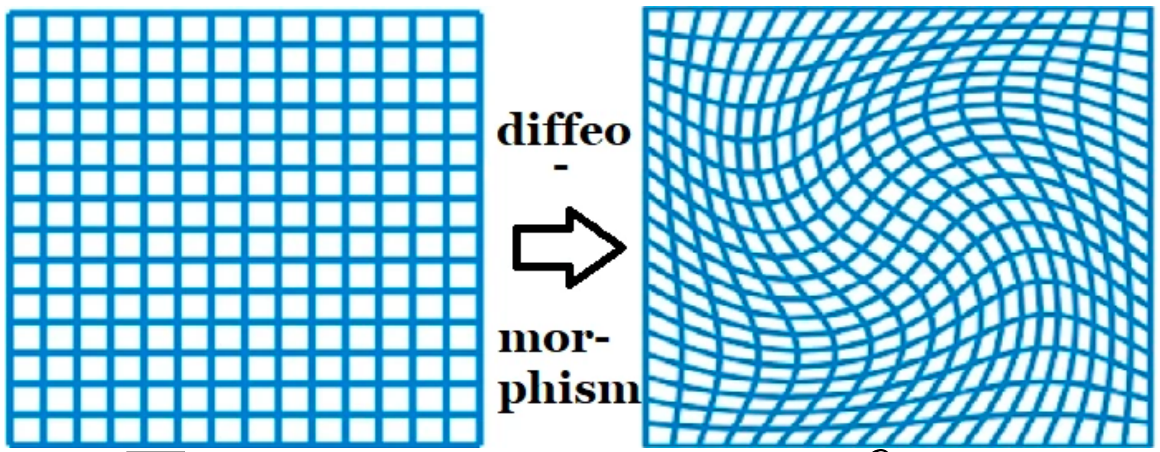


In this case, we can introduce:

$$L_\tau x(u) \triangleq x(u - \tau(u))$$

Small deformations are locally group which is quite high-dimensional.
And, you don't want to be fully invariant to diffeomorphisms!

Stability to group of symmetries



$$d(\mathbf{I}-\tau,\mathbf{I}) \triangleq \boxed{\sup_{x\in\mathbb{R}^d}\|\nabla\tau(x)\|} + \sup_{x\in\mathbb{R}^d}\|\tau(x)\| + \sup_{x,y\in\mathbb{R}}\|\tau(x)-\tau(y)\|$$

(introduce: $\|\Delta\tau\|_\infty = \sup_{x,y}\|\tau(x)-\tau(y)\|$)

- Two groups are of interest: **diffeomorphisms** and **translations.**

- Stability means thus here:

$$\|\Phi(x)-\Phi(L_\tau x)\| \leq C\|x\|d(\mathbf{I}-\tau,\mathbf{I})$$

- **Lemma:** If $\|\nabla\tau\| < \frac{1}{2}$ then $\mathbf{I}-\tau$ is a diffeomorphism and:

$$\|L_\tau\| \leq 2^d$$

$$|1-\det(\mathbf{I}-\nabla\tau(u))| \leq d\|\nabla\tau\|_\infty$$

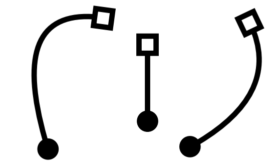
$$2^{-d} \leq |\det(\mathbf{I}-\nabla\tau)(u)| \leq 2^d$$

A motivating example

- Translation invariance and stability to deformations? Why not:

Deformations

$$L_\tau x(u) = x(u - \tau(u))$$



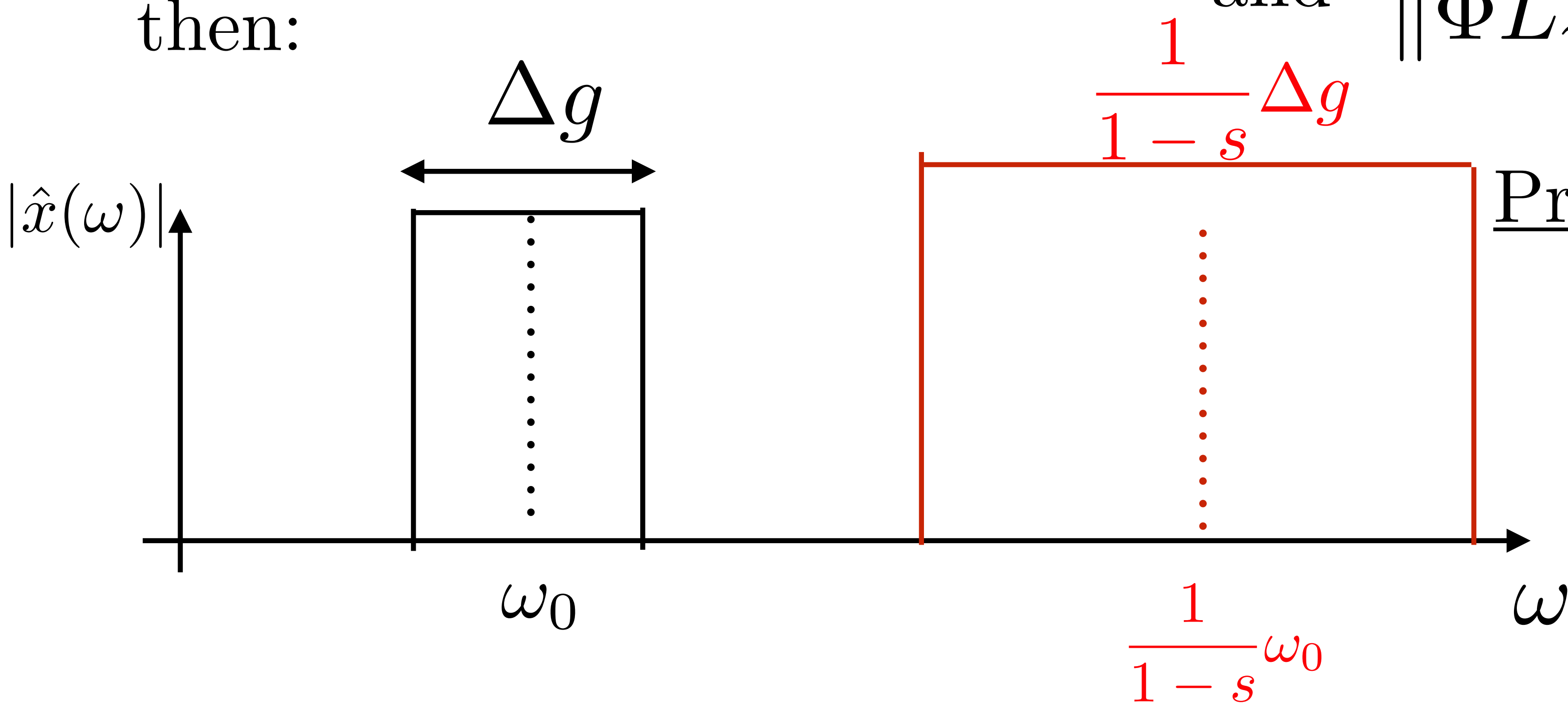
$$\Phi x(\omega) = |\hat{x}(\omega)|$$

Again, doesn't work!

Consider: $\tau(u) = su, 1 > s > 0$

Let $x(u) = e^{i\omega_0 u}\hat{g}(u)$ thus: $\Phi L_\tau x(\omega) = \frac{1}{1-s} g(\frac{\omega - \omega_0}{1-s})$

then: and $\|\Phi L_\tau x\| = \frac{1}{\sqrt{1-s}} \|g\|$



Proof: Construct (x_n, τ_n)

s.t. $\|x_n\| = 1$

$\|\nabla \tau_n\| \to 0$

but:

$$\|\Phi L_{\tau_n} x_n - \Phi x_n\| \geq 1$$

Remark on the non-linearity

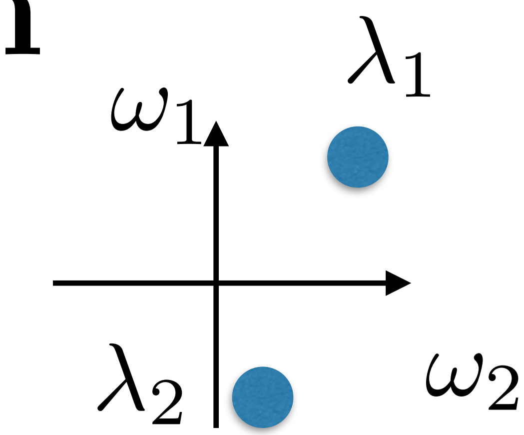
- Theorem (Bruna): Let $M : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$s.t. M is a non expansive operator,$\|M(x) - M(y)\| \leq \|x - y\|$ and assume that $L_\tau M = ML_\tau$, meaning that M commutes with the action of diffeormophisms. Then:

Ref.: PhD Joan Bruna + Your homework!!!

$$\exists \rho : \mathbb{R} \to \mathbb{C} : \forall u \in \mathbb{R}^d, Mx(u) = \rho(x(u))$$

The Scattering Transform

Definition of the Scattering Transform



Define a path of length n as $(\lambda_1, \ldots, \lambda_n)$ where $\lambda = (\theta, 2^{-j}), |\lambda| = 2^{-j}$

Let us fix mother wavelet ψ and low-pass filter ϕ, smooth, with fast decay.

Definition: The Scattering path of $S_J[\lambda_1, \ldots, \lambda_n]x$ is given by:

$$S_J[\lambda_1, \ldots, \lambda_n]x \triangleq ||\ldots|x \star \psi_{\lambda_1}| \star \ldots| \star \psi_{\lambda_n}| \star \phi_J$$

Definition: Scattering Transform of order n:

$$S_J^n x \triangleq \{S_J[\lambda_1, \ldots, \lambda_k]x\}_{\lambda_1, \ldots, \lambda_k, k \leq n}$$

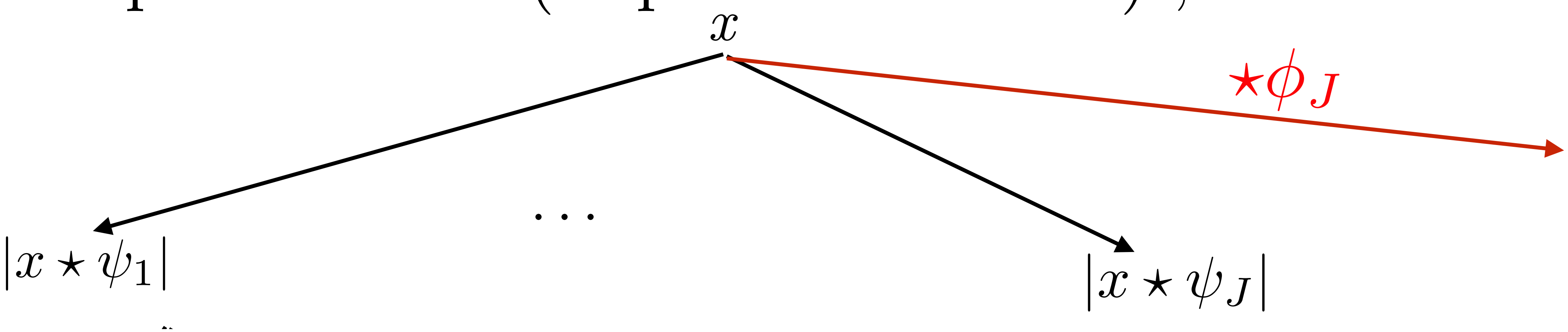
Its norm si given by:

$$\|S_J^n x\|^2 = \sum_{\lambda_1, \ldots, \lambda_k, k \leq n} \|S_J[\lambda_1, \ldots, \lambda_k]x\|^2$$

We will also write the Scattering Transform as: $S_J x \triangleq \{S_J^n x\}_{n \geq 0}$

The Scattering Transform

- Main principle: cascade wavelets AND modulus non-linearity. Depth: "order" (in practice order 2) ; J: "Scale"



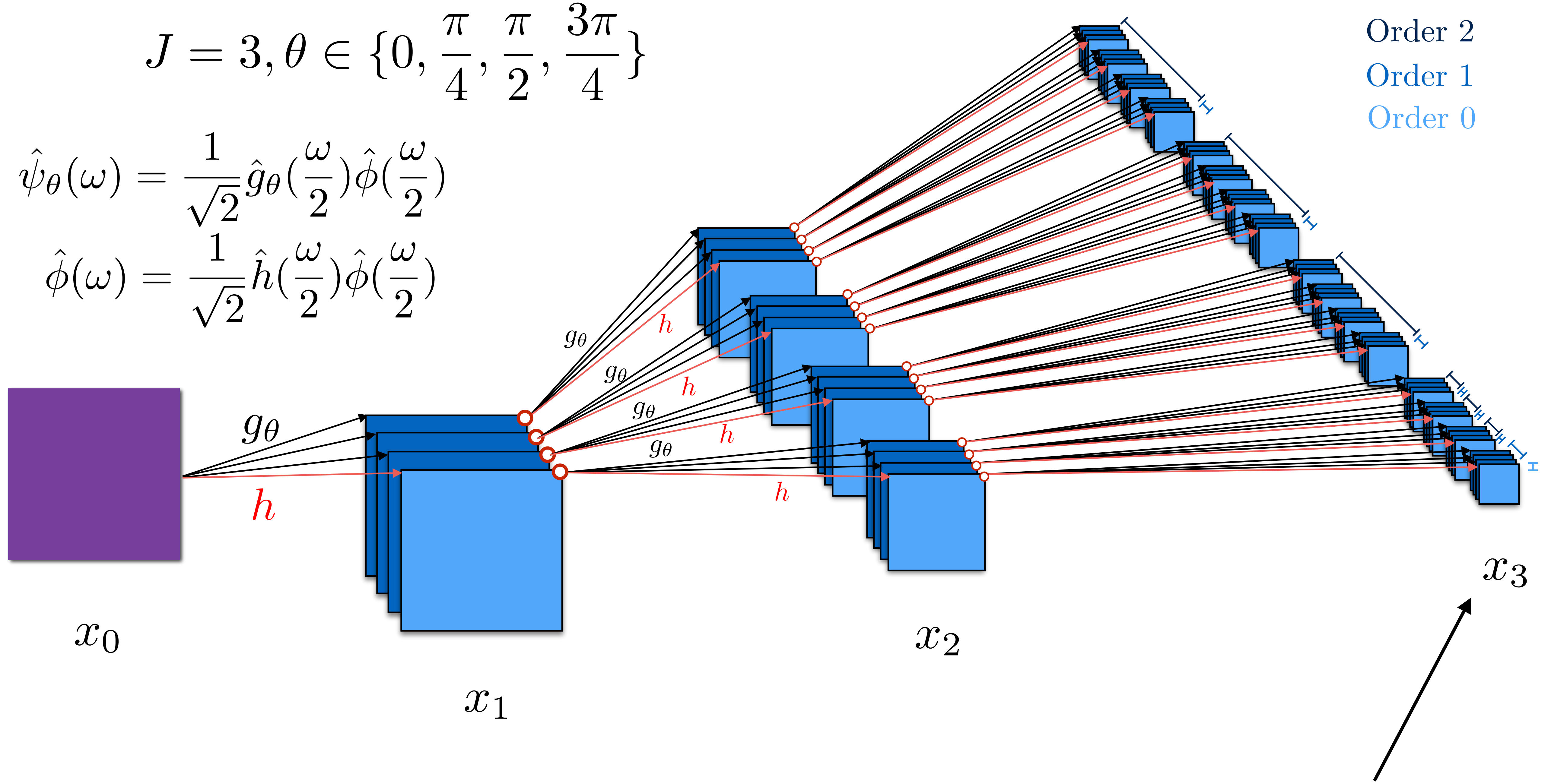
$$S_J x = \{x \star \phi_J,$$

Scattering as a CNN

$$J = 3, \theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$$

$$\hat{\psi}_\theta(\omega) = \frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$



$h \geq 0$

Scattering coefficients are only at the output

Scattering as a CNN

Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat
Recursive Interferometric Representations, S Mallat

Properties of the Scattering Transform:

- A non-linear representation which depends on an invariance parameter J *and* n wavelet transforms.

- As it cascades unitary operators, Scattering is stable:

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

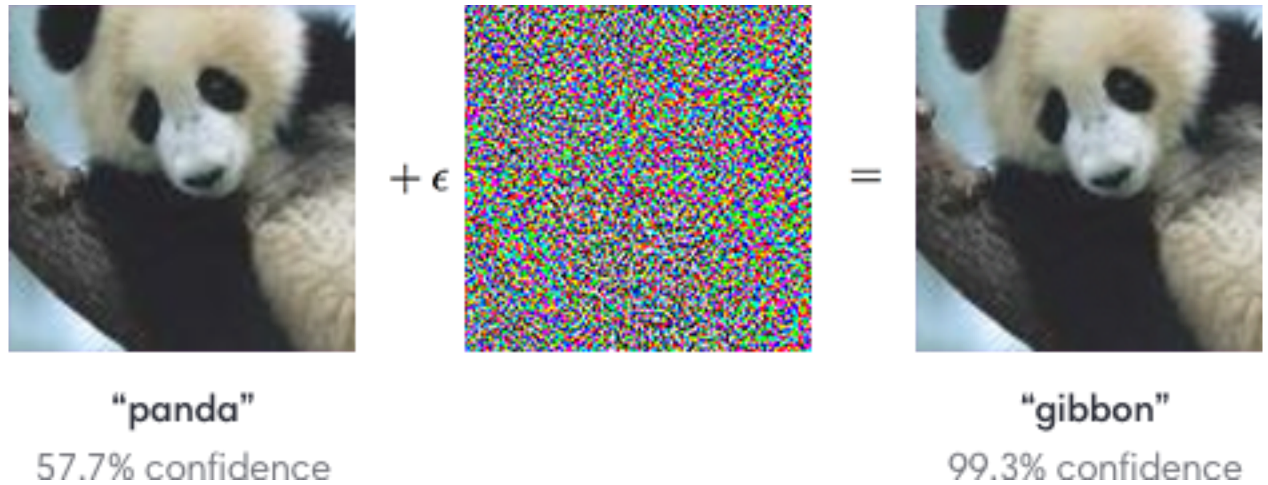
- Thanks to wavelets, it linearizes small deformations:

$$\|S_J^n x - S_J^n L_\tau x\| \leq C_n \|\nabla \tau\| \|x\|$$

- Thanks to low-pass filter, it is invariant to local translation:

$$\|a\| \ll 2^J \Rightarrow S_J^n L_a x \approx S_J^n x$$

Non-expansivity?



- NNs are super sensitive to input noise

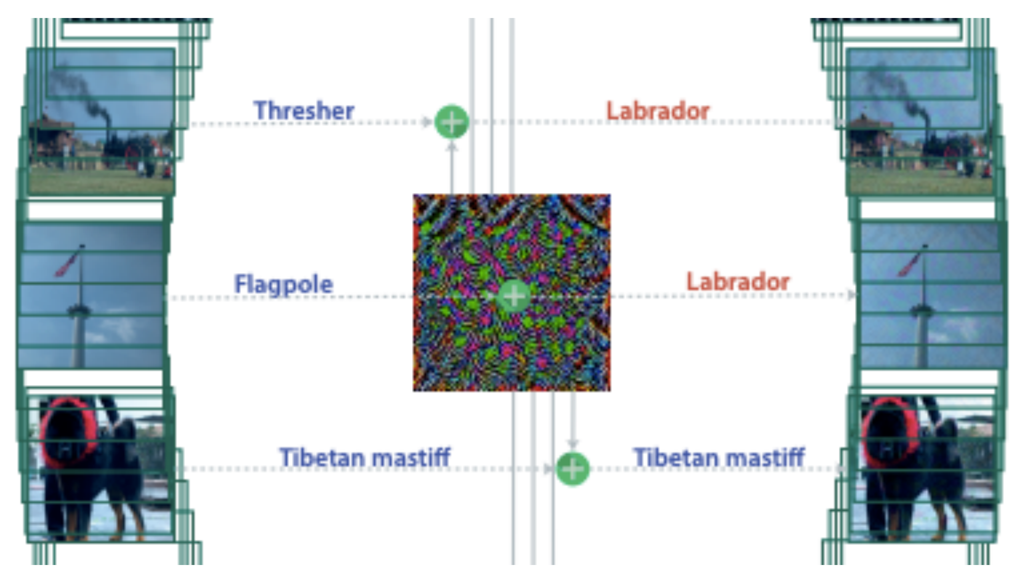
- Indeed, the NN is at most $\|W_1\|...\|W_J\|$-Lipschitz

Ref.: Lipschitz Regularity of deep neural networks, Scaman and Virmaux

$$\inf_{\Phi(x) \neq \Phi(x+\epsilon)} \|\epsilon\|$$

Or even for every class, there are algorithms with parameters (ϵ, κ) s.t.:

$$\begin{cases} \mathbb{P}\Big(\Phi(X+\delta) \neq \Phi(X)\Big) \geq 1-\kappa \\ \|\delta\| \leq \epsilon \end{cases}$$

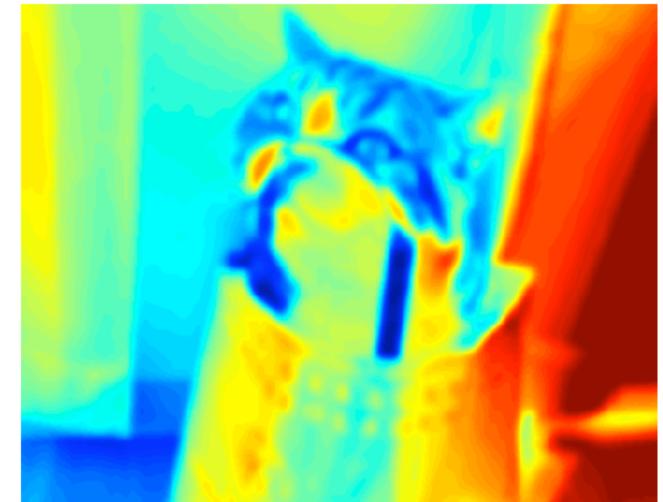


Ref.: Universal adversarial perturbations, Moosavi et al.

Feature map

x

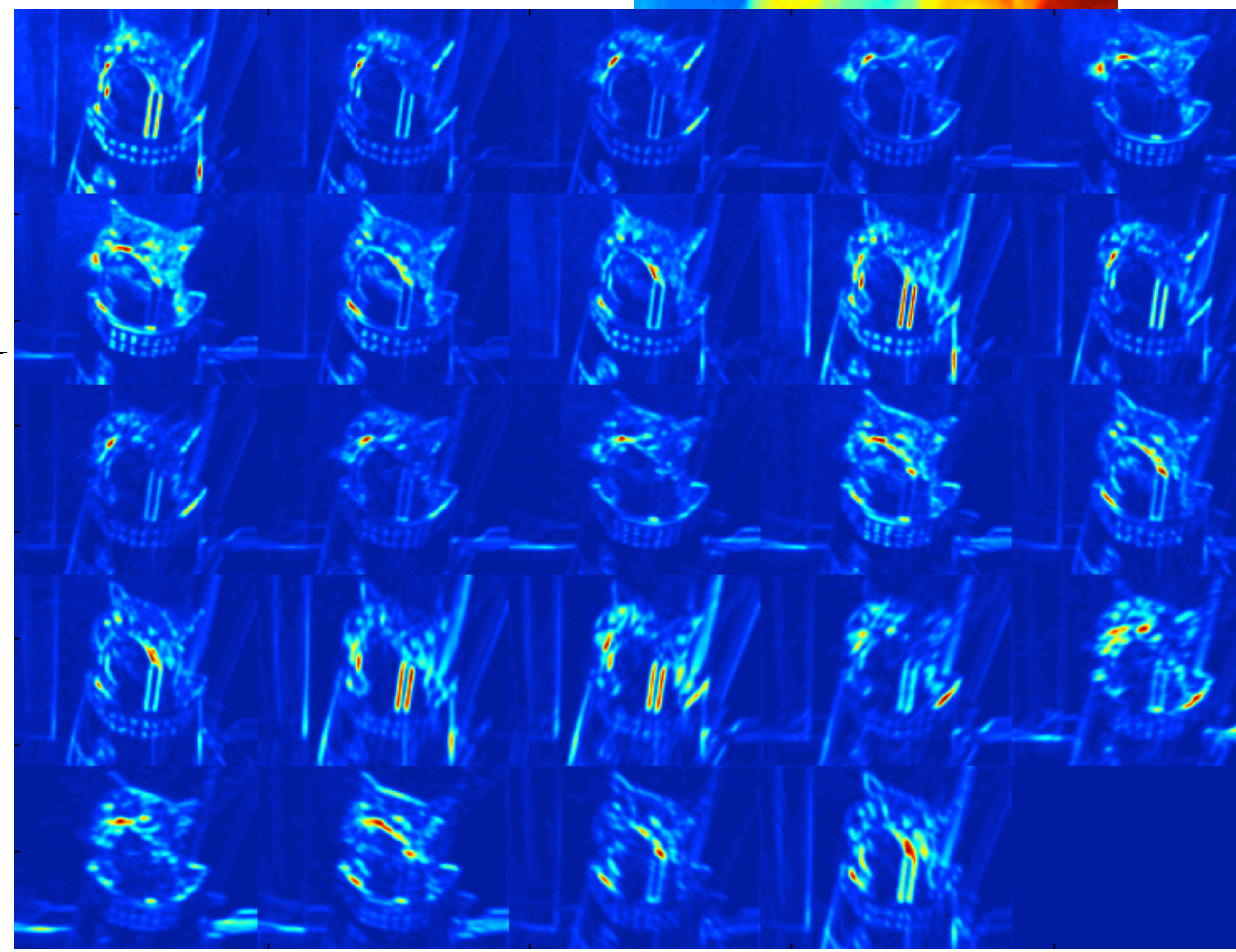
$x \star \phi$



$|x \star \psi| \star \phi$

Several features map

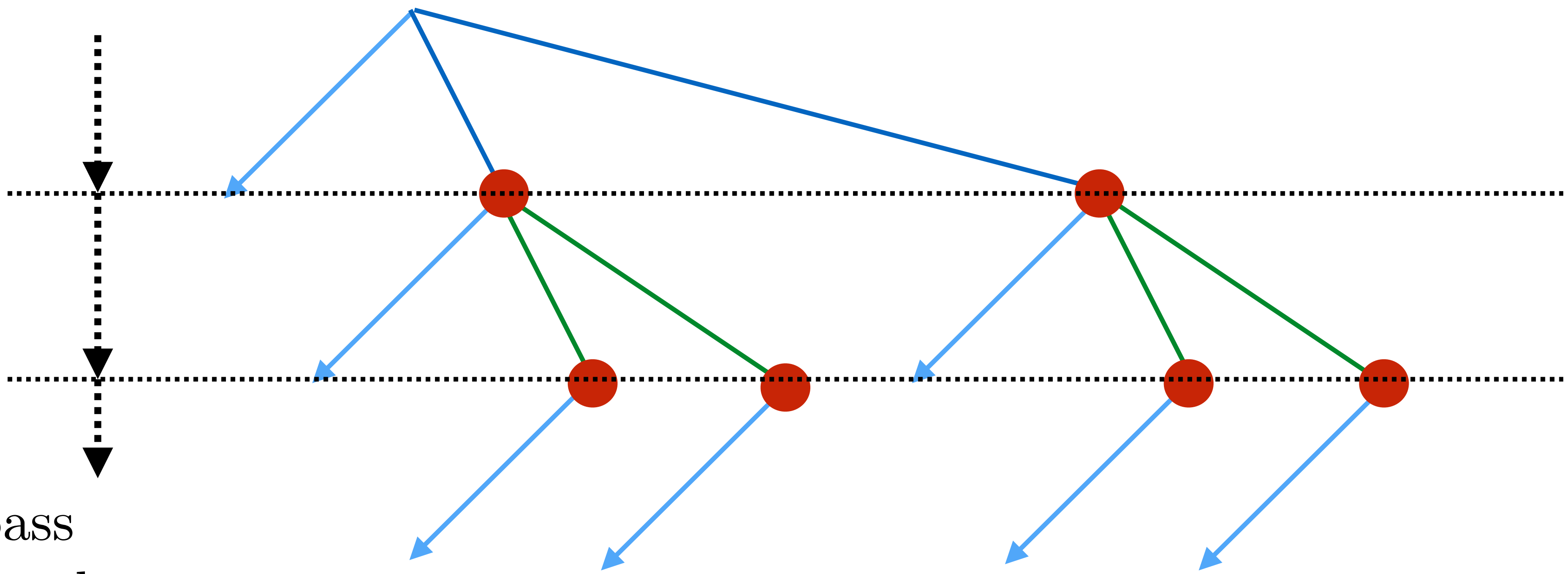
1st order coefficients

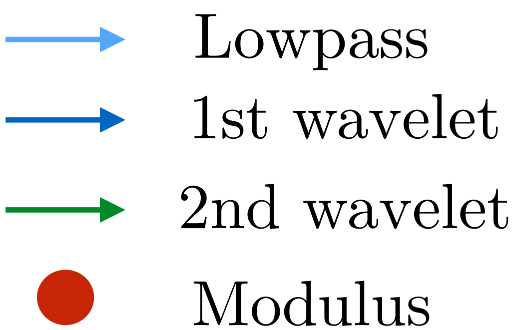


Example of Scattering coefficients

Scaling scattering on GPUs

Ref.: Thesis, EO

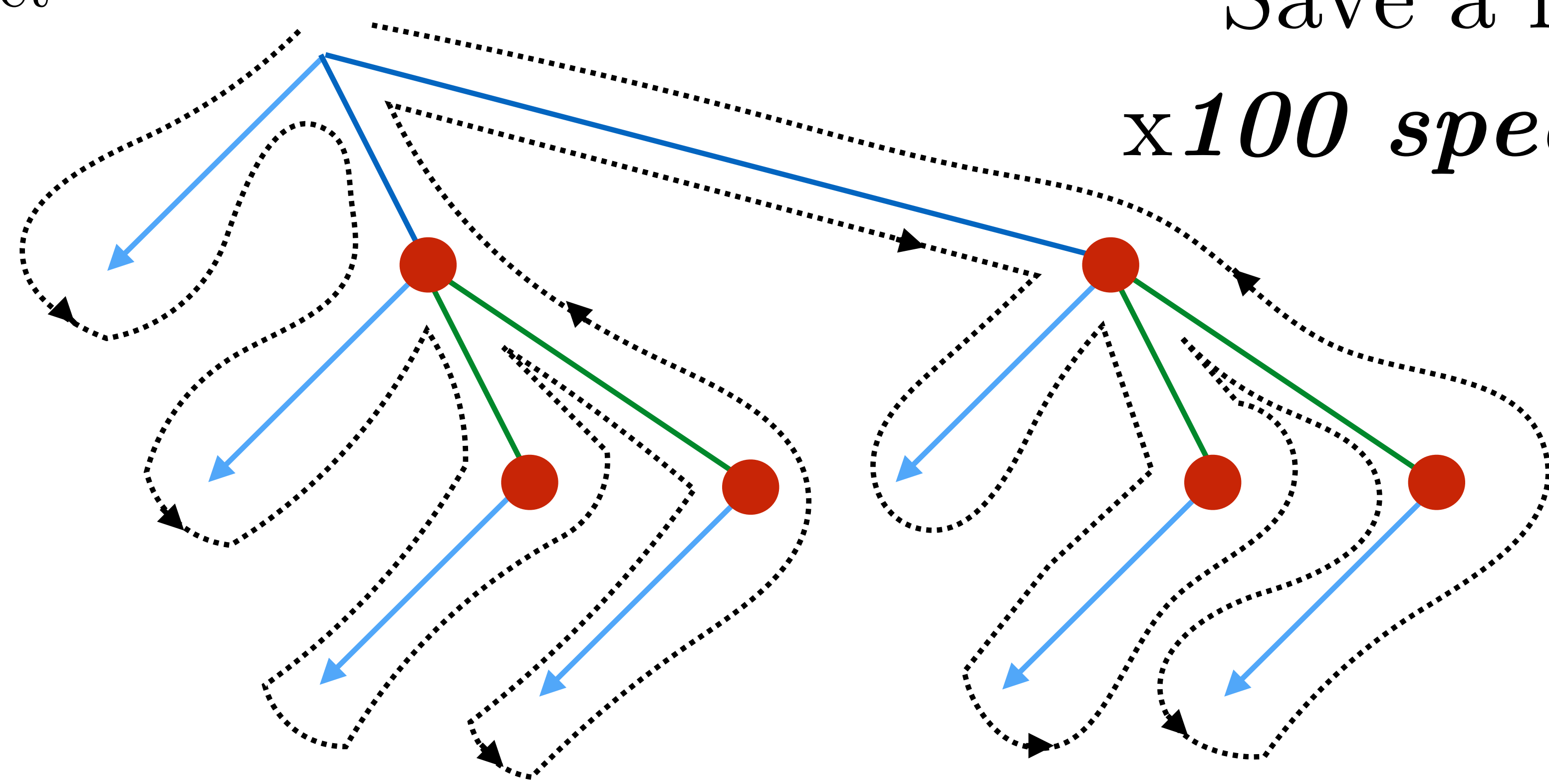




Naive algorithm

Save a lot of memory!

x***100 speed-up on GPU***



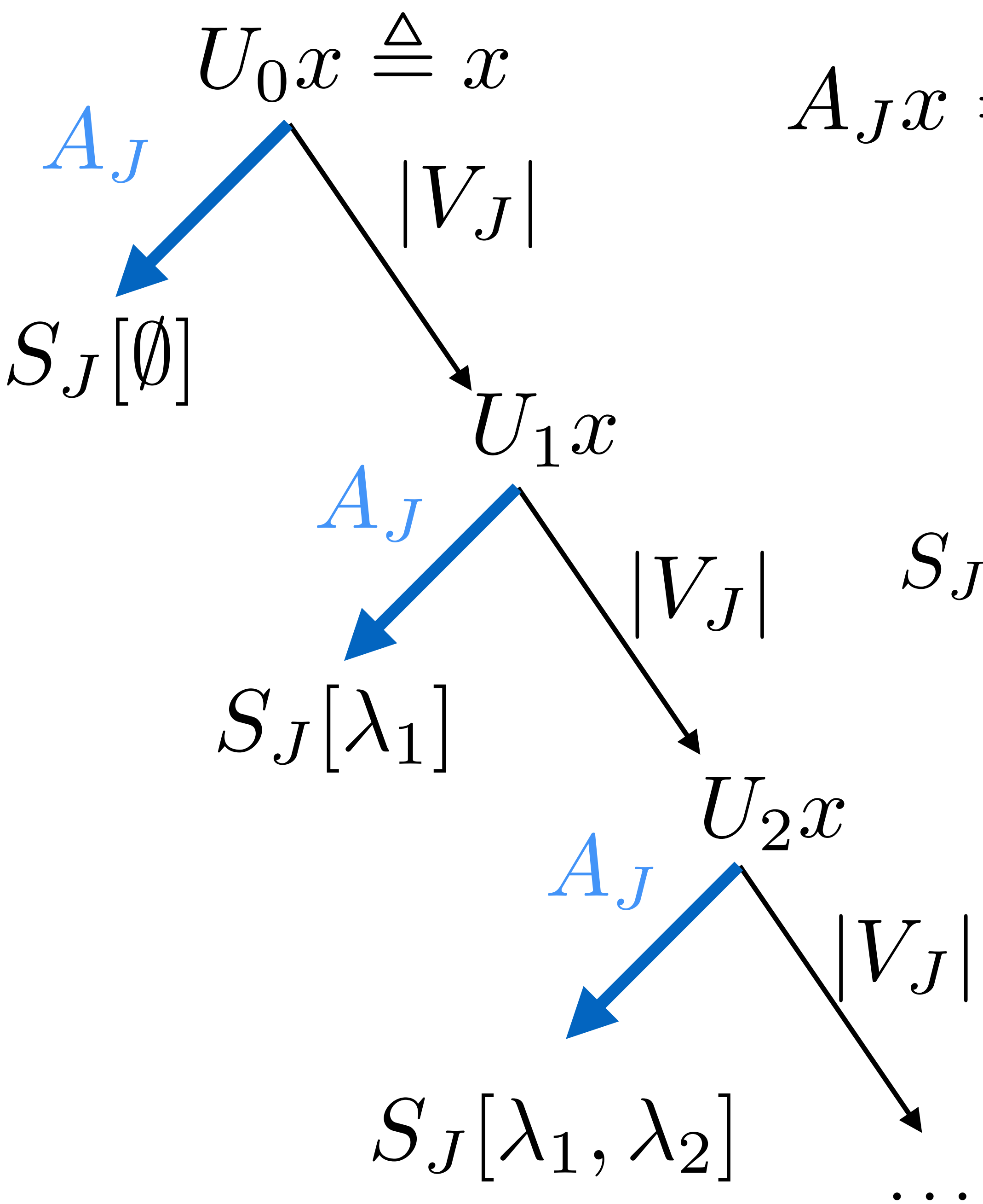
Efficient algorithm

Scattering Transform Theory

Definition & non-expansivity

Scattering Transform defined via Integral Operators

Coefficients of the scattering transform are given by:



$$A_J x = x \star \phi_J \text{ and } V_J x = \{x \star \psi_\lambda\}_{\lambda \in \Lambda}$$

$$\text{s.t.: } W_J x = V_J x \cup A_J x$$

Here:

$$S_J[\lambda_1, ..., \lambda_n] x = ||...|x \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_n}| \star \phi_J$$
$$= A_J U_n[\lambda_1, ..., \lambda_n] x$$

Scattering of order n:

$$S_J^n x = \{\cup_{\lambda_1, ..., \lambda_j \in \Lambda, j \leq n} S_J[\lambda_1, ..., \lambda_j]\}$$

Non-expansivity of Scattering Transform

- Proposition: Given $x \in L^2(\mathbb{R}^d), y \in L^2(\mathbb{R}^d)$ we have, if $\|W_J\| \leq 1$:

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

Proof:

Lemma: for $x, y \in L^2(\mathbb{R}^d)$:

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

Remark: implies boundedness

Stability to deformations

Main theorem statement

Recall that: $W_J x = \{x \star \psi_{j,\theta}, x \star \phi_J\}_{\theta, j \leq J}$

- Theorem (Adapted from Mallat, 2012): If ϕ, ψ are regular enough, $\|W_J\| \leq 1$and if $\int_u \psi(u)\, du = 0$, there exists C such that for any J, if $\|\nabla \tau\|_\infty \leq \frac{1}{2}$, then:

$$\|S_J^n L_\tau x - S_J^n x\| \leq n^{3/2} C \|x\| (\|\nabla \tau\|_\infty + \|\Delta \tau\|_\infty + \frac{\|\tau\|_\infty}{2^J})$$

In other words, the Scattering Transform is stable to small deformations.

Typical applications: n=2, J=3

Sketch of the proof.

Write: $[A, B] = AB - BA$ which measures how A, B commute.

- First, we note that:

$$\|S_J^n x - S_J^n L_\tau x\|^2 = \sum_n \|A_J U_n x - A_J U_n L_\tau x\|^2$$

- Next, we will bound each Scattering "paths"

$$\|A_J U_n x - A_J U_n L_\tau x\| \leq \left(c_1 \|A_J L_\tau - A_J\| + n c_2 \|[L_\tau, V_j]\|\right) \|x\|$$

- Finally, we will bound each operators:

$$\|A_J L_\tau - A_J\| \leq C_1 (2^{-J} \|\tau\|_\infty + \|\nabla \tau\|_\infty)$$

and

$$\|[L_\tau, V_J]\| \leq C_2 (\|\nabla \tau\|_\infty + \|\Delta \tau\|_\infty)$$

- In conclusion:

$$\|S_J^n L_\tau x - S_J^n x\| \leq n^{3/2} C \|x\| (\|\nabla \tau\|_\infty + \|\Delta \tau\|_\infty + \frac{\|\tau\|_\infty}{2^J})$$

Proof step 1

$$A_J x = x \star \phi_J \text{ and } \quad V_J x = \{x \star \psi_\lambda\}_{\lambda \in \Lambda}$$
$$\text{and } W_J = \{A_J, V_J\}$$

- Assume we proved that for ϕ, ψ regular enough, we get:

$$\|A_J L_\tau - A_J\| \leq C_1(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty)$$
and
$$\|[L_\tau, V_J]\| \leq C_2(\|\nabla\tau\|_\infty + \|\Delta\tau\|_\infty)$$

- Theorem: If ϕ, ψ are regular enough, $\|W_J\| \leq 1$ and if $\int_u \psi(u)\, du = 0$, there exists C such that for any J, if $\|\nabla\tau\|_\infty \leq \frac{1}{2}$, then:

$$\|S_J^n L_\tau x - S_J^n x\| \leq n^{3/2} C \|x\| (\|\nabla\tau\|_\infty + \|\Delta\tau\|_\infty + \frac{\|\tau\|_\infty}{2^J})$$

constants are suboptimal

Stability to deformations: low-pass filter

- Proposition: Assume: $\int_u \|\nabla\phi(u)\|\,du < \infty$ and $\int_u |\phi(u)|\,du < \infty$

  Then there exists C>0 such that for any J and $\|\nabla\tau\|_\infty \leq \frac{1}{2}$:

$$\|A_J - A_J L_\tau\| \leq C(2^{-J}\|\tau\|_\infty + \|\nabla\tau\|_\infty)$$

Commutation with deformations of high-frequencies

- Proposition: Assume that ψ is regular and $\int_u \psi(u)\, du = 0$ then there exists C such that for any J and $\|\nabla\tau\|_\infty \leq \frac{1}{2}$:

$$\|[L_\tau, V_J]\| \leq C(\|\nabla\tau\|_\infty + \|\Delta\tau\|_\infty)$$

Summary of the Scattering's properties we discussed

- Scattering is stable:
$$\|S_J x - S_J y\| \leq \|x - y\|$$
- Linearize small deformations:
$$\|S_J L_\tau x - S_J x\| \leq C\|\nabla \tau\|\|x\|$$
- Invariant to local translation:
$$\|a\| \ll 2^K \Rightarrow S_J L_a x \approx S_J x$$
- For λ, u, $S_J x(u, \lambda)$ is **covariant** with :

$$\text{if } \forall u \forall g \in SO_2(\mathbb{R}), g.x(u) \triangleq x(g^{-1}u)$$

$$S_J(g.x)(u, \lambda) = S_J x(g^{-1}u, g^{-1}\lambda) \triangleq g.S_J x(u, \lambda)$$

More Scattering

Scattering moments

Ref.: Invariant Convolutional Scattering Network, J. Bruna and S Mallat

- For a stationary process X (e.g., a texture)

$$E(X \star f) = E(X) \star f$$

- This leads to the Expected Scattering:

$$\bar{S}[\lambda_1] = \mathbb{E}|X \star \psi_{\lambda_1}|$$

Modulus is important because it can be 0!

$$\bar{S}[\lambda_1, \lambda_2] = \mathbb{E}||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$$

can be estimated via an unbiased estimator:

$$S[\lambda_1, \lambda_2]X = \int ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$$

Roto-translation scattering

Ref.: PhD of L. Sifre

$$S_0 x = \int_u x(u)\, du \quad \text{and} \quad Y^1_{j_1}(u, \theta_1) = |x \star \psi_{j_1,\theta_1}(u)|$$

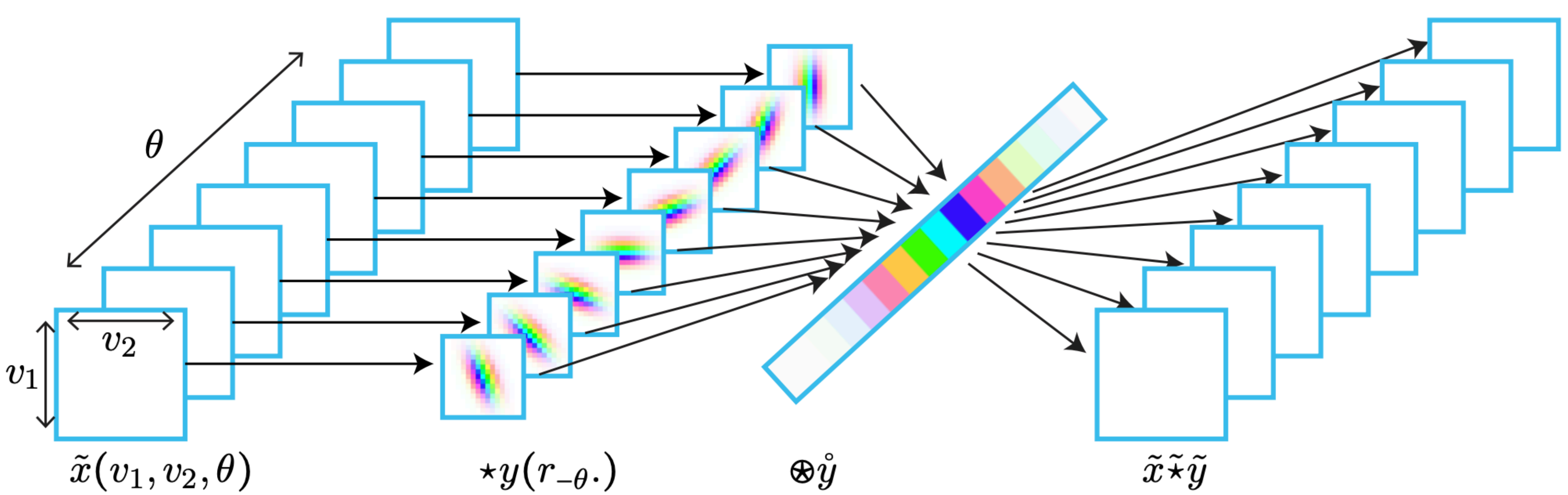
Let $S_1 x = \int_{u,\theta} Y^1(u,\theta)\, du\, d\theta$ and $\Psi(u,\theta) = \psi_{j_2,\theta_2}(u)\psi_k(\theta)$

then, we get:

$$Y^2_{j_1,j_2,\theta_2,k}(\theta, u) = \int_{\theta',u'} |x \star \psi_{j_1,\theta'}(u')|\psi_{j_2,\theta_2+\theta'}(u-u')\psi_k(\theta-\theta')du\, d\theta$$

Let $S_2 x = \int_{u,\theta} Y^2(u,\theta)\, du\, d\theta$

- Then Sx is invariant to roto-translation.

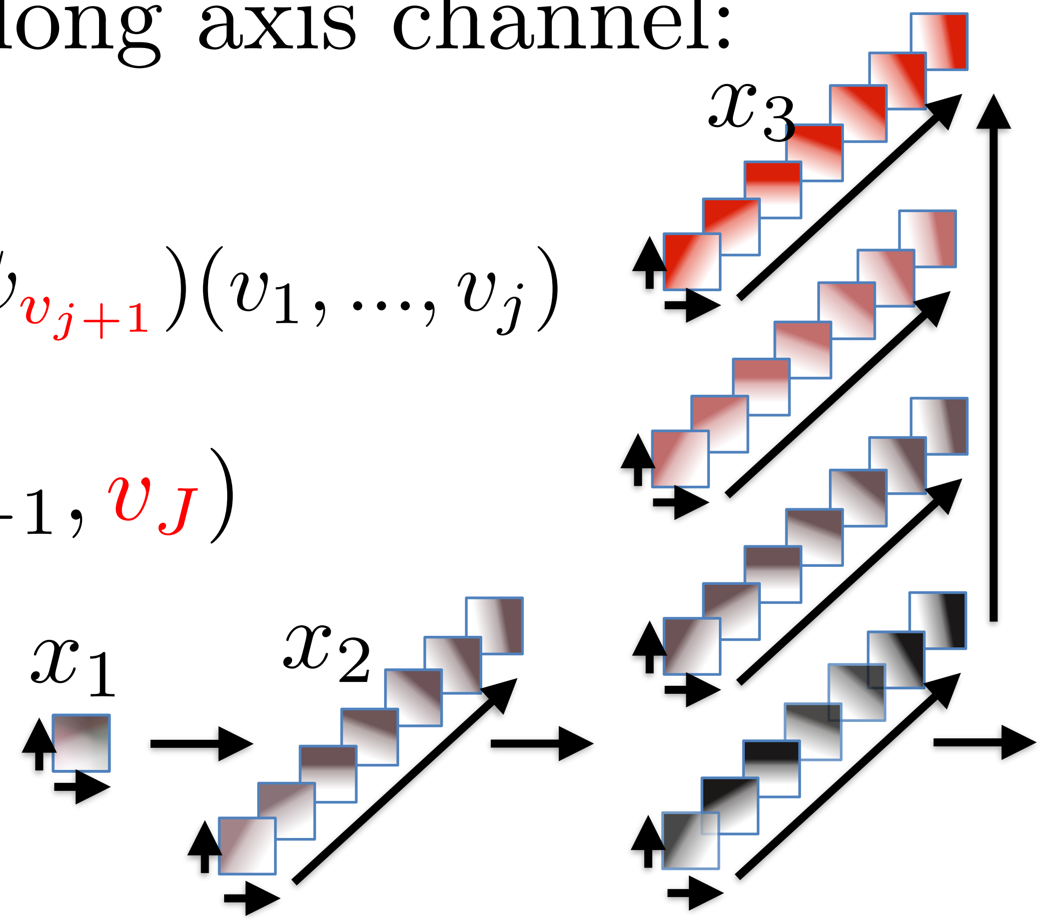


One generalization among many

- CNN that is convolutional along axis channel:

$$x_{j+1}(v_1, ..., v_j, \color{red}{v_{j+1}}) = \rho_j(x_j \star^{v_1,...,v_j} \psi_{\color{red}{v_{j+1}}})(v_1, ..., v_j)$$

$$x_J(\color{red}{v_J}) = \sum_{v_1,...,v_{J-1}} x_{J-1}(v_1, ..., v_{J-1}, \color{red}{v_J})$$



Ref.: Hiearchical CNNs, Jacobsen et al.

- For x_j, we refer to the variable v_j as an attribute that discriminates previously obtained layer.

- Representation is finally averaged: invariant along translations by v.

Very similar to equivariant CNNs

This lecture: Examples on a notebook!