Interpretability and Signal Processing in Deep Learning

Edouard Oyallon

edouard.oyallon@lip6.fr

CNRS, LIP6

Interpretability?

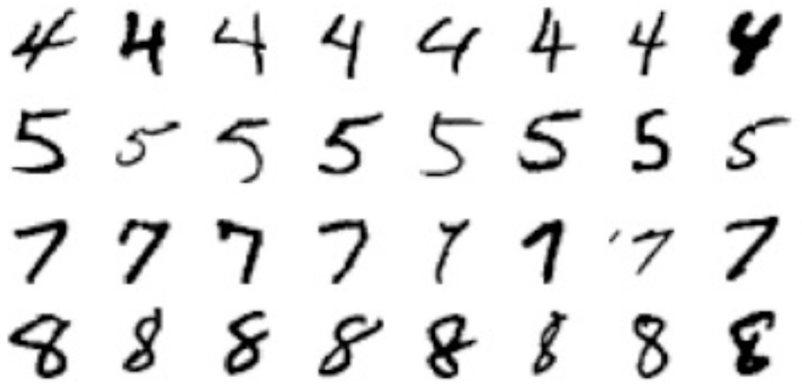
- A notion that can be vague...
- "*Ability to provide explanations that are humanly understandable*"
- Trustable model, learning guarantees and stability lead to a good interpretable model...
- Structure is desirable!

Example of an interpretable AI

A supervised classification task

1. Propose a model of your data.

   Ex.: MNIST (60k samples)



Small deformations
+Translation

2. Design a representation.

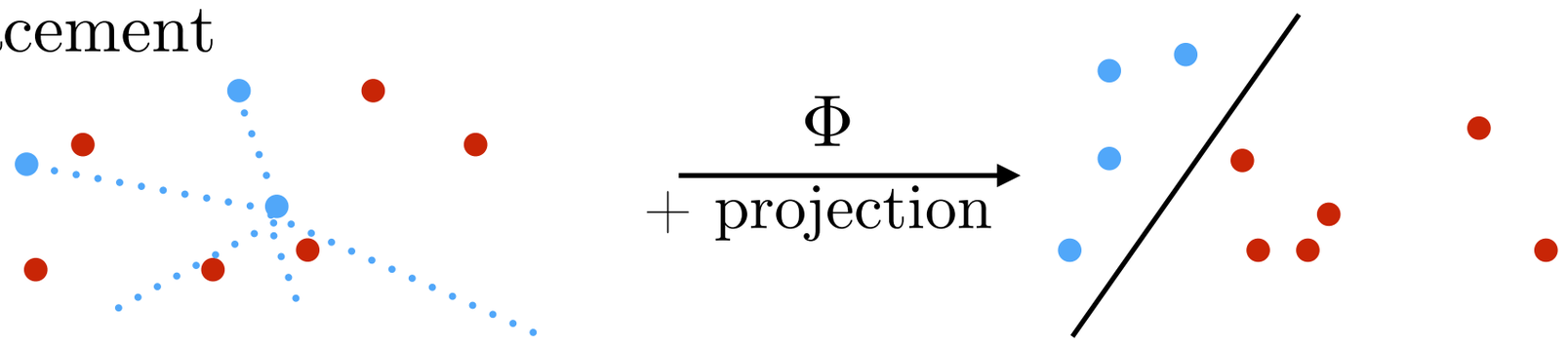
   Ex.: Scattering Transform.

   Achieves translation invariance, linearises deformation.

3. Propose a classifier.

   Ex.: Linear SVM.

Displacement



4. Obtain reasonable performances.

In the following...

1. No model known on real images
2. Limited *a priori*, except translation invariance
3. Learn each parameters...
4. Obtain the best performances

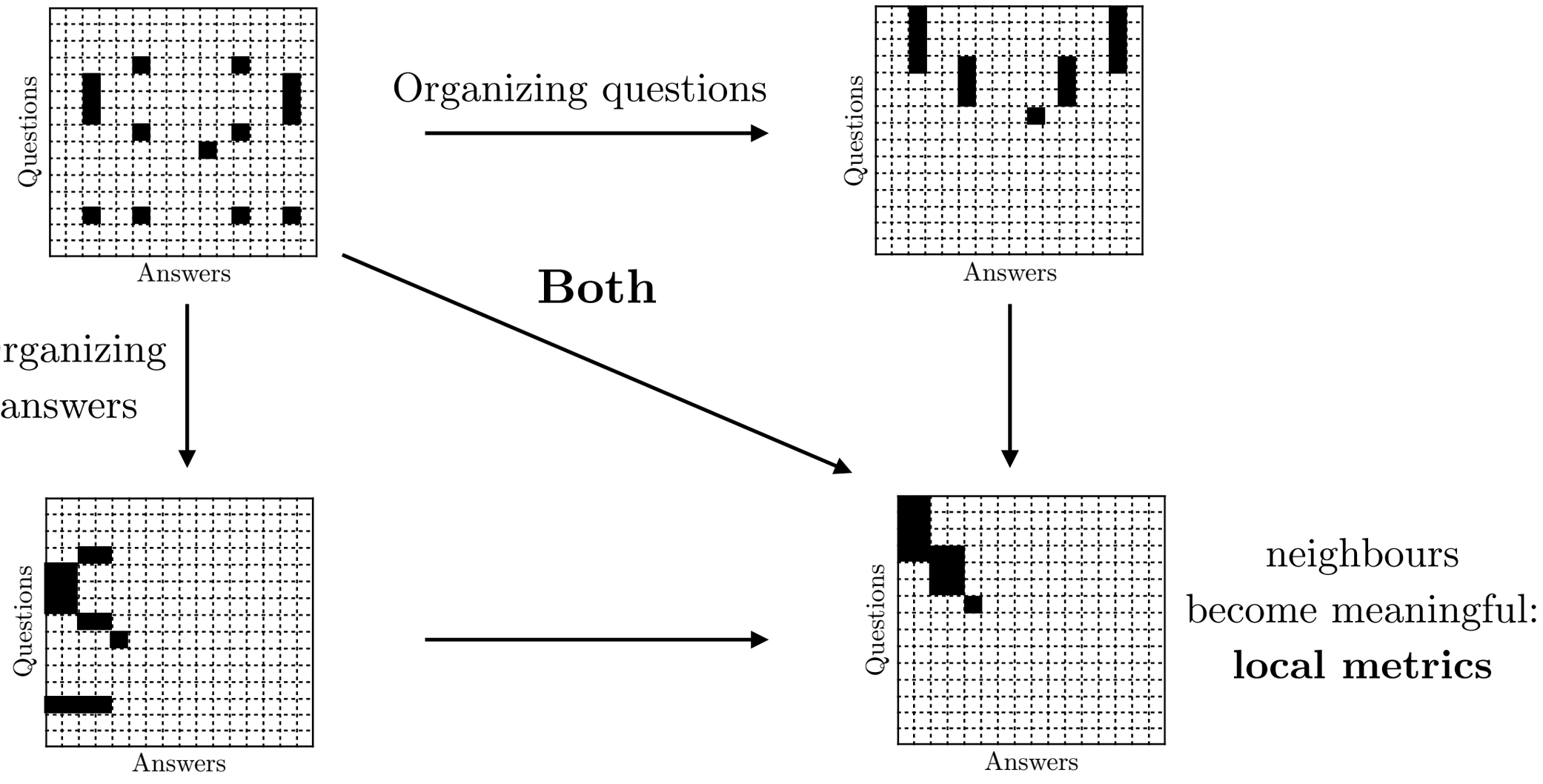
The reason of their success is unclear...

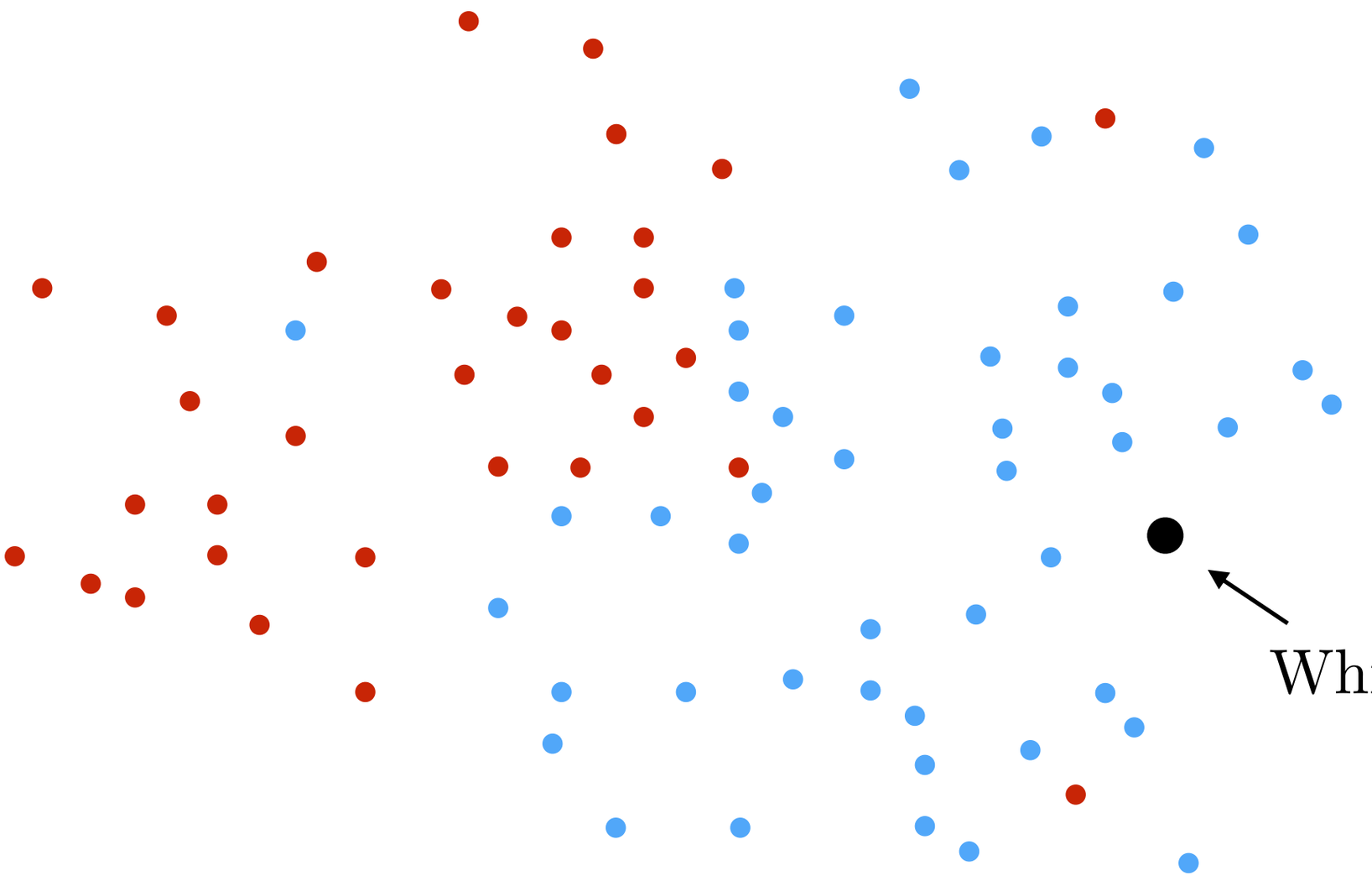
Statistical issues

Organization is a key

- Consider a problem of questionnaires: people answer to 0 or 1 to some question. What does interpretability mean?

Ref.: Harmonic Analysis of Digital Data Bases
Coifman R. et al.





An example of supervised task: classification

Supervised task

$\mathcal{X} = \mathbb{R}^2$ Samples space

$\mathcal{Y} = \{$$\}$ Labels

Input data $x \in \mathcal{X}$ $\xrightarrow{\Phi ?}$ Output data $\Phi(x) \approx y \in \mathcal{Y}$

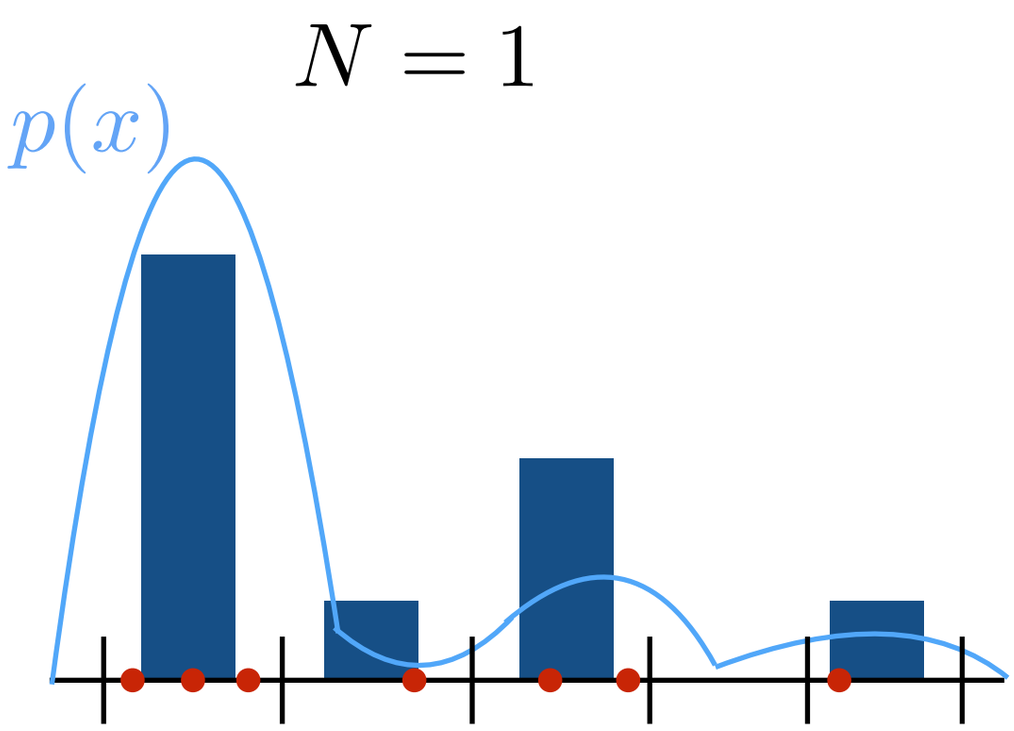
- Estimating a label $\boldsymbol{y}$ from a sample x, by training a model Φ on a training set. Validation of the model is done on a different test set.

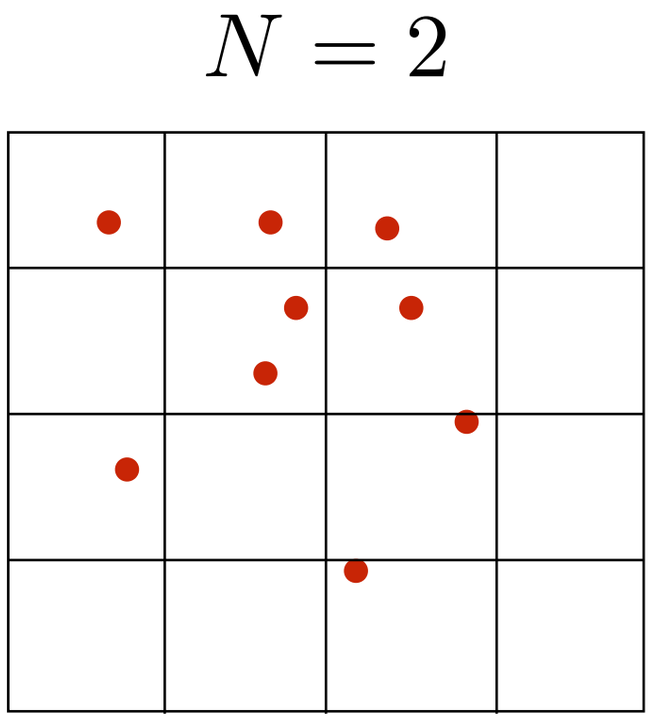
- Examples: prediction, regression, classification,...

- Best setting: dimensions of x and y is small, $\mathcal{X}$ large

High dimensional images

- PdFs are difficult to estimate in high dimension.





- For a fixed number of points and bin size, as N increases, the bins are likely to be empty.

Curse of dimensionality:
occurs in many machine learning problems

Very high-dimensional images

- Curse of dimensionality!

$$(x_i, y_i) \in \mathbb{R}^{224^2} \times \{1, ..., 1000\}, i < 10^6 \longrightarrow \hat{y}(x)?$$

Estimation problem

Training set to predict labels

"Rhino"

Large datasets. . .

- ImageNet 2012: (350GB)
  1 million training images, 1 000 classes
  400 000 test images
  Large coloured images of various sizes

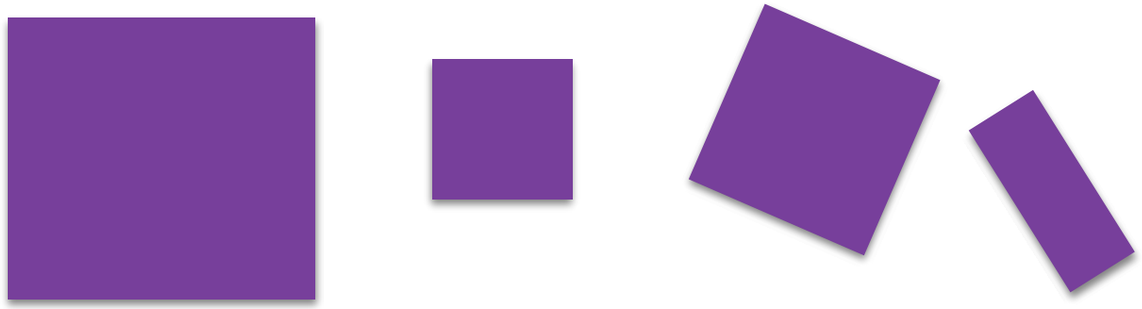
  Ref.: image-net.org

- Labels obtained via Amazon Turk (complex process that requires human labelling)

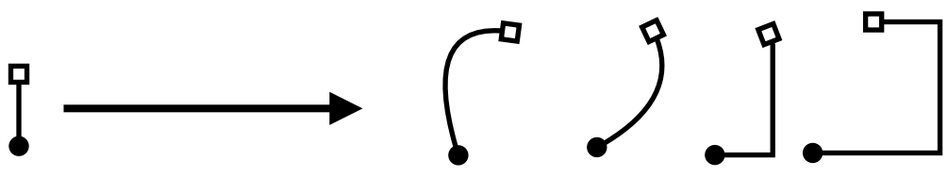
Difficult problems due to Image variabilities

Geometric variability

Groups acting on images:
translation, rotation, scaling



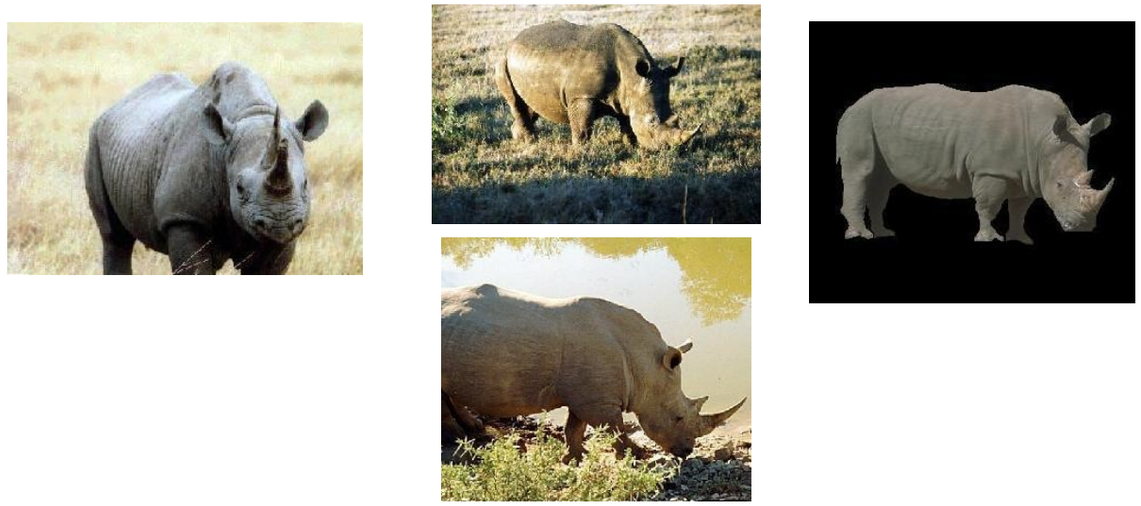
Other sources : luminosity, occlusion,
small deformations



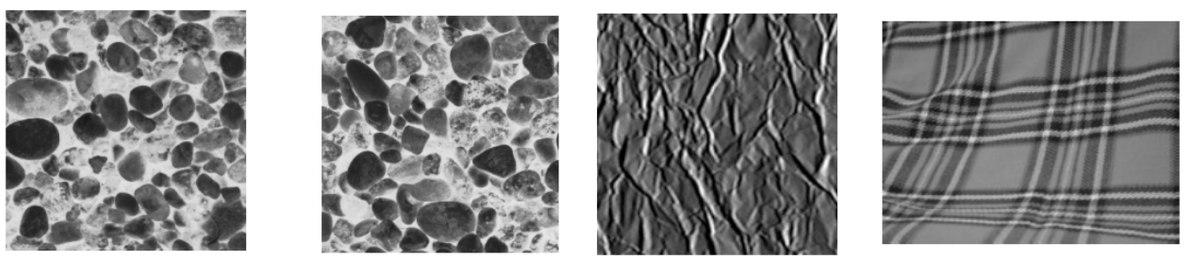
Class variability

Intraclass variability

Not informative



Extraclass variability



High variance: hard to reduce!

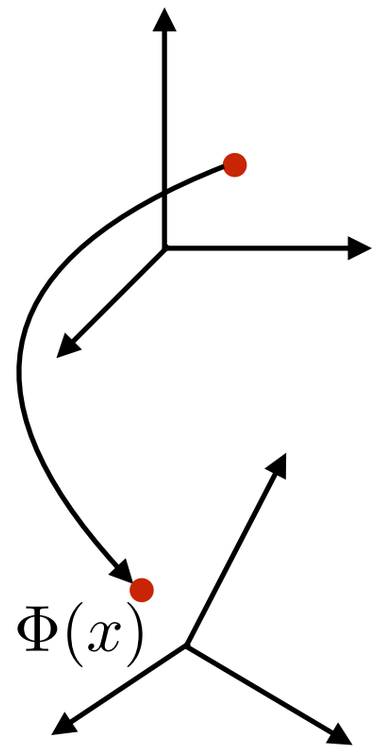
Desirable properties of a representation

- **Invariance** to group G of transformation (e.g. roto-translation):

$$\forall x, \forall g \in G, \Phi(g.x) = \Phi(x)$$

- **Stability** to noise

$$\forall x, y, \|\Phi(x) - \Phi(y)\|_2 \leq \|x - y\|_2$$



- **Reconstruction** properties

$$y = \Phi(x) \Longleftrightarrow x = \Phi^{-1}(y)$$

- **Linear separation** of the different classes

$$\forall i \neq j, \|E(\Phi(X_i)) - E(\Phi(X_j))\|_2 \gg 1$$

$$\forall i, \sigma(\Phi(X_i)) \ll 1$$

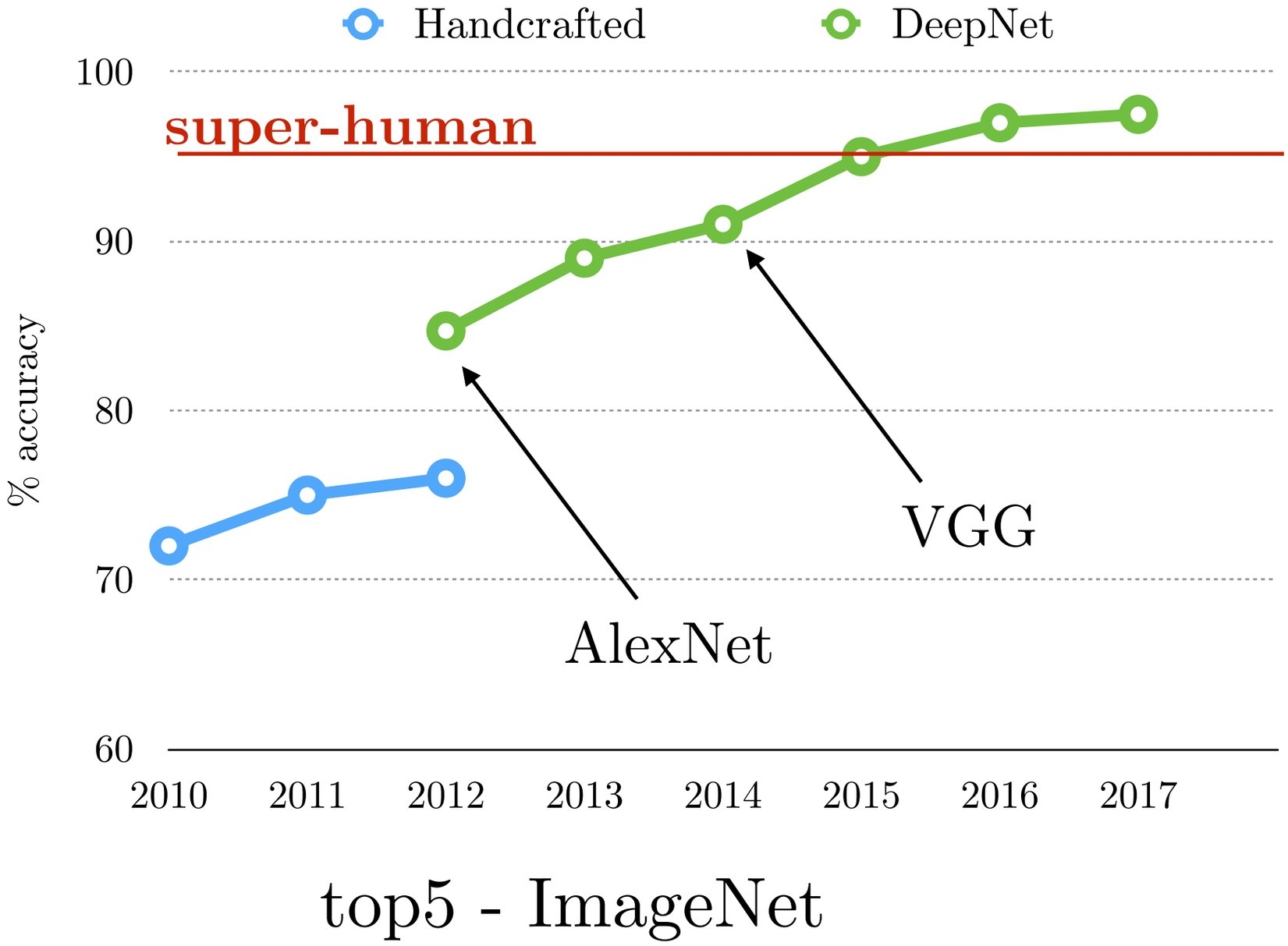
Can be difficult to handcraft..

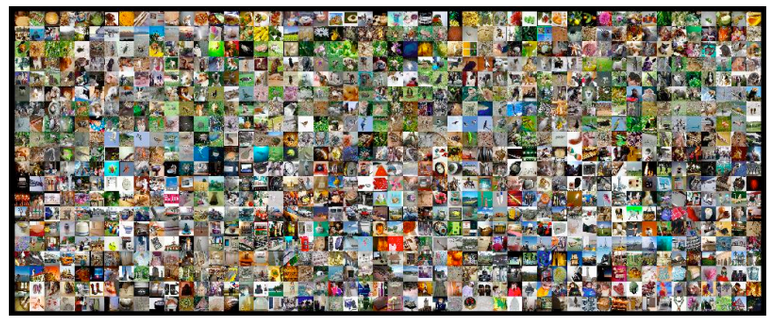
Is this solvable?

Years of research...

Of course?

- Huge gap thanks to deep neural networks.





ImageNet:

1 million training images, 1 000 classes

400 000 test images

Large coloured images of various sizes

Theory for good performances?

Interpreting subsequent layers

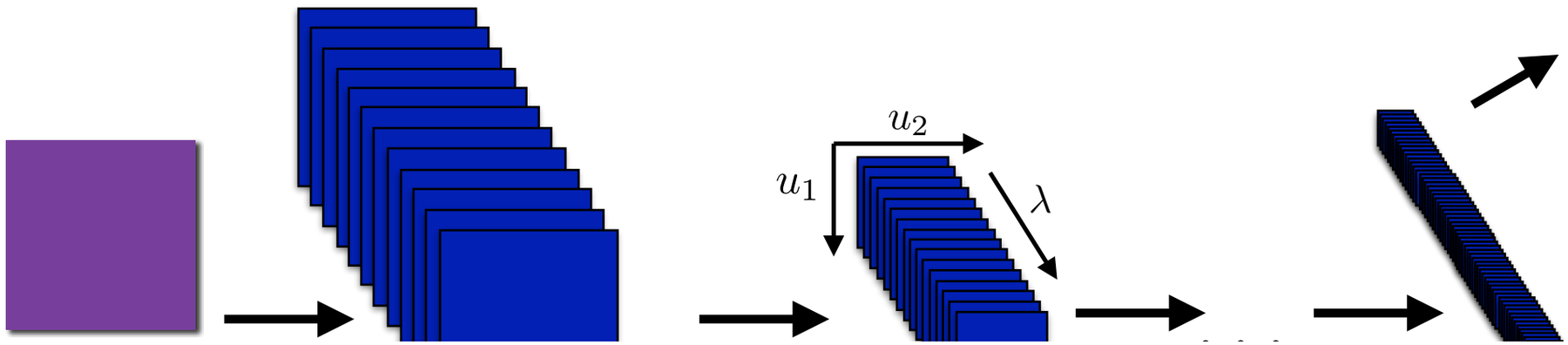
Convolutional Neural Networks

input signal

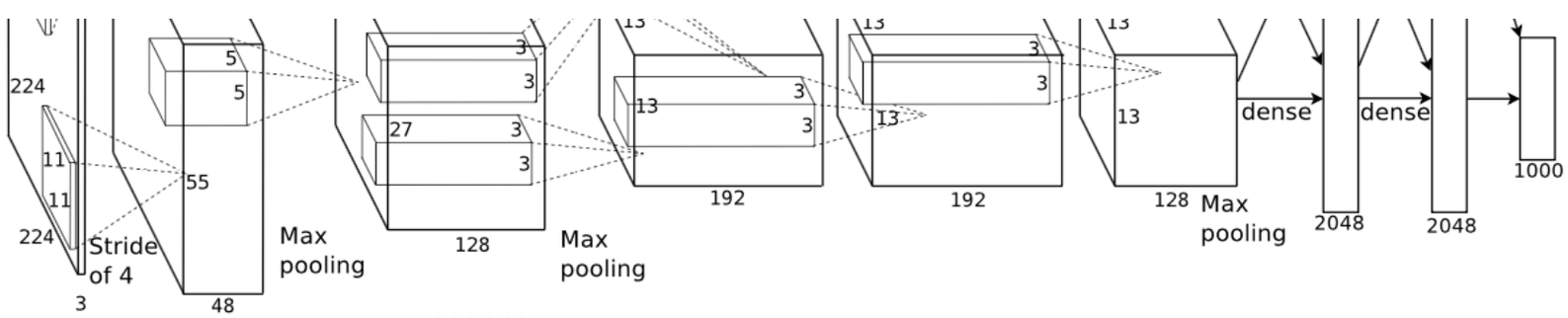
output signal

$$x \longrightarrow W_1 \longrightarrow \rho \longrightarrow W_2 \longrightarrow \dots \longrightarrow \rho \longrightarrow W_J \longrightarrow \Phi(x)$$

Schematic



Engineering



Each layer: $x_{j+1} = \rho W_j x_j$

learned kernel

that leads to: $x_{j+1}(u, \lambda_{j+1}) = \rho\left(\sum_{\lambda_j} \left(x_j(., \lambda_j) \star w_{\lambda_j, \lambda_{j+1}} \right)(u) \right)$

where: $\rho(x) = \max(0, x)$ s.t. $|\rho(x) - \rho(y)| \leq |x - y|$

Fourier?

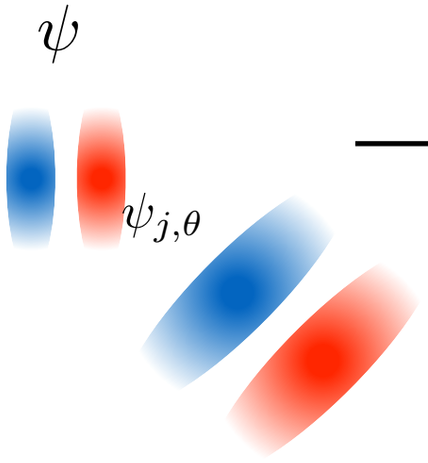
Fourier transform:

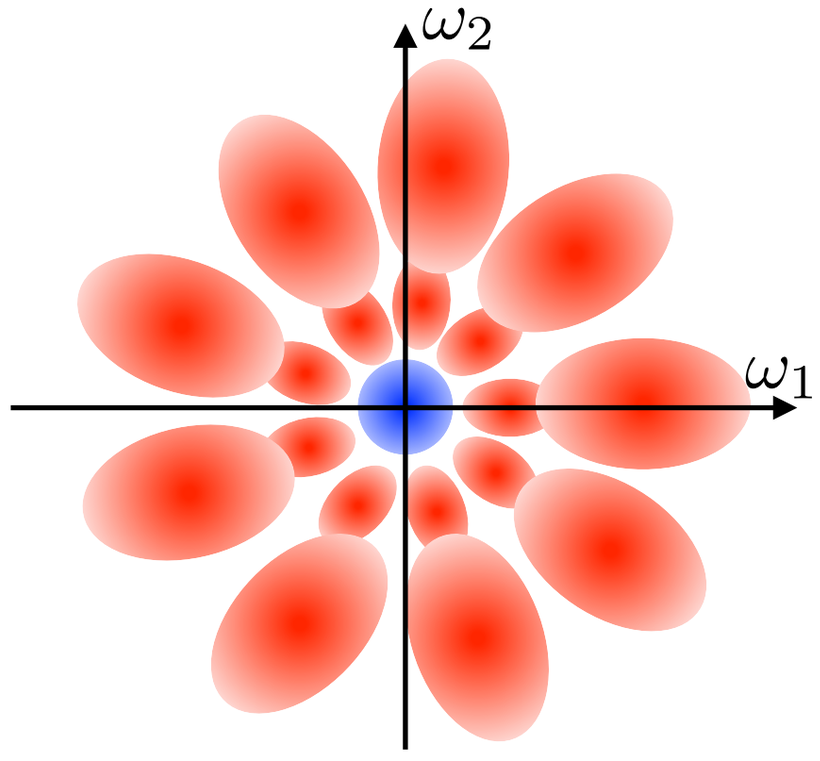
$$\mathcal{F}x(\omega) \triangleq \int_{\mathbb{R}^2} x(u) e^{-i\omega^T u} du \quad \text{is an isometry of } L^2$$

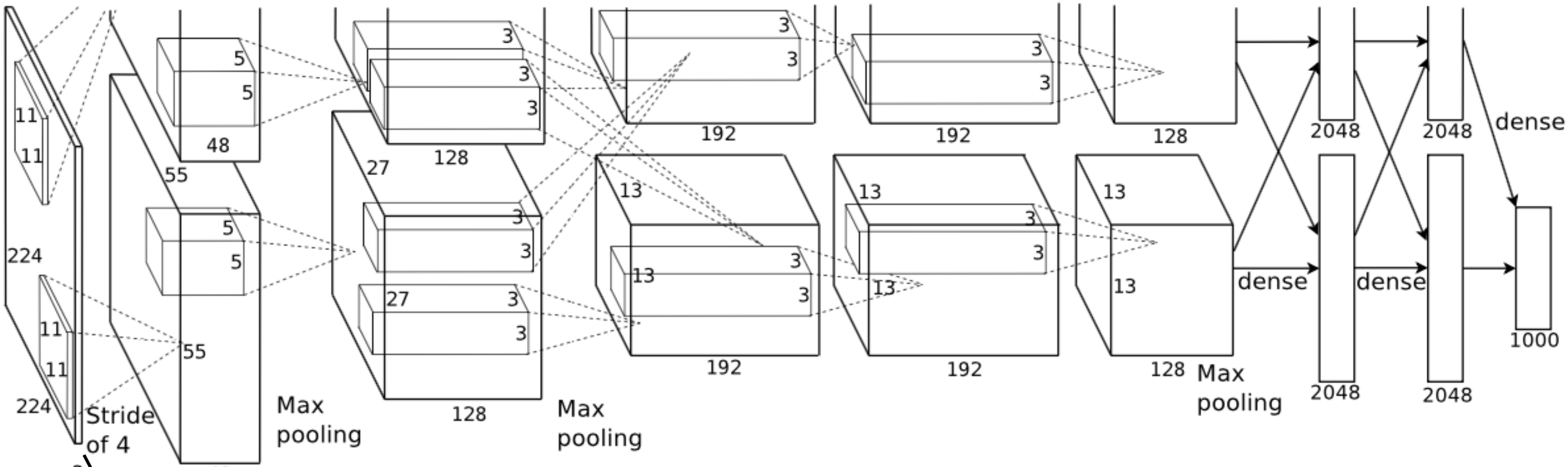
Diagonalizes convolution: $\mathcal{F}(g \star h)(\omega) = \mathcal{F}g(\omega).\mathcal{F}h(\omega)$

Ex.: wavelets

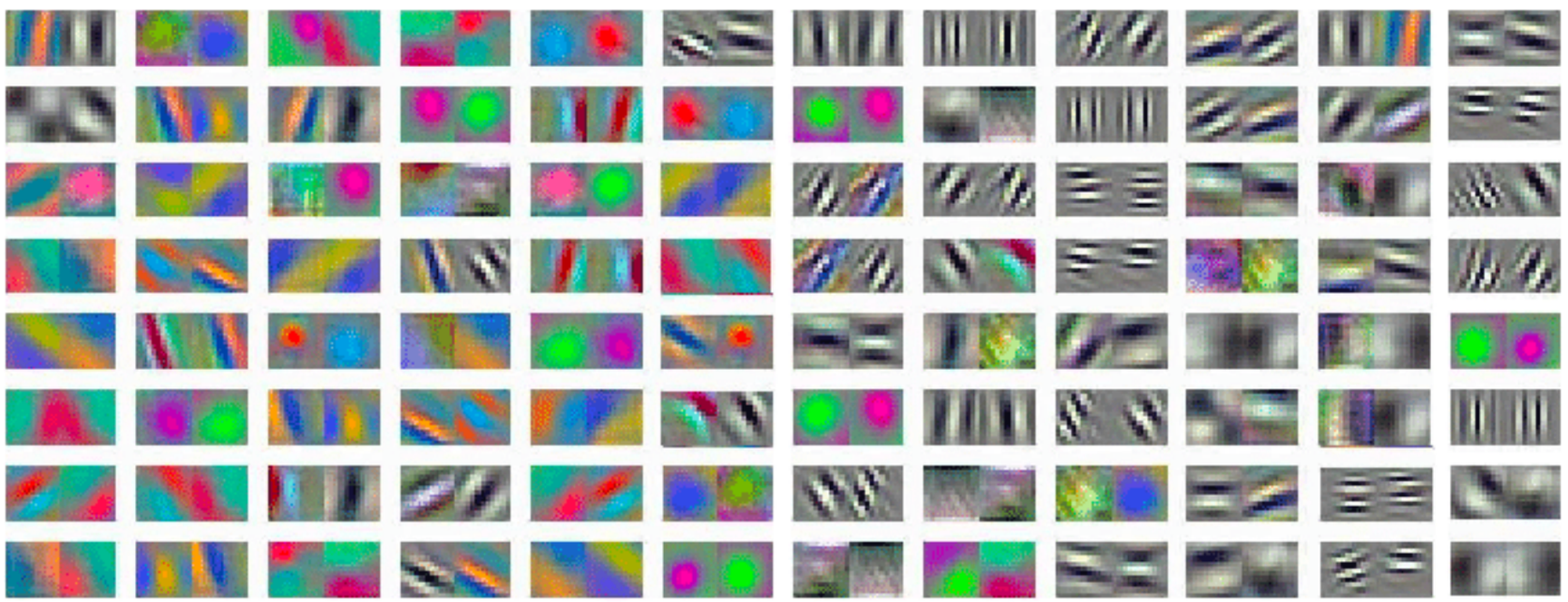
$$\psi_{j,\theta} = \frac{1}{2^{2j}} \psi(\frac{r_\theta(u)}{2^j})$$







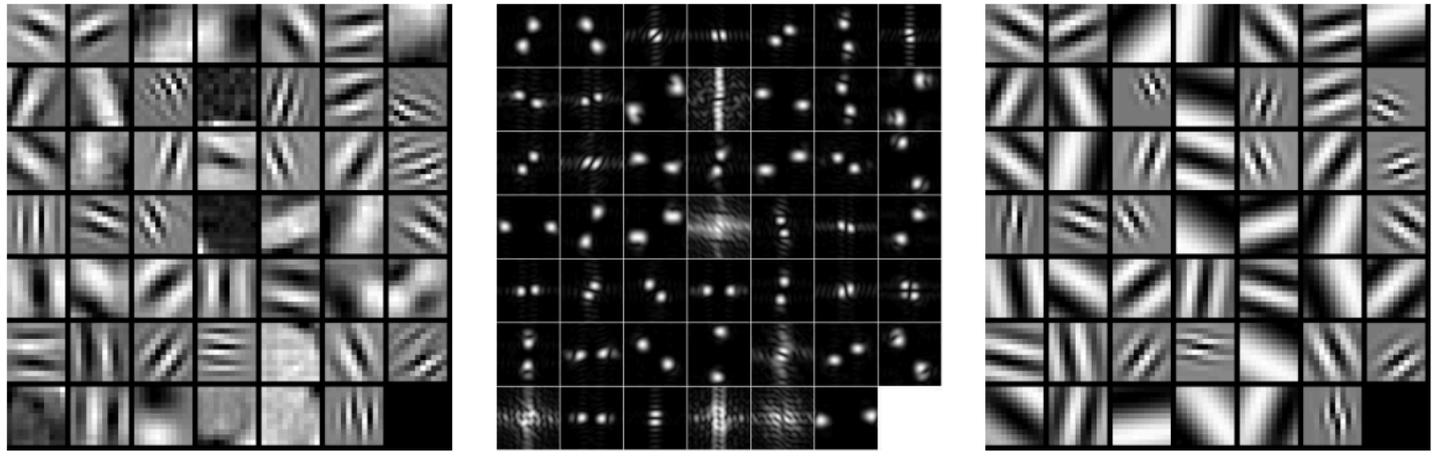
Color, scales, rotation, smoothness...



One detailed example: the AlexNet

Model for the first layer

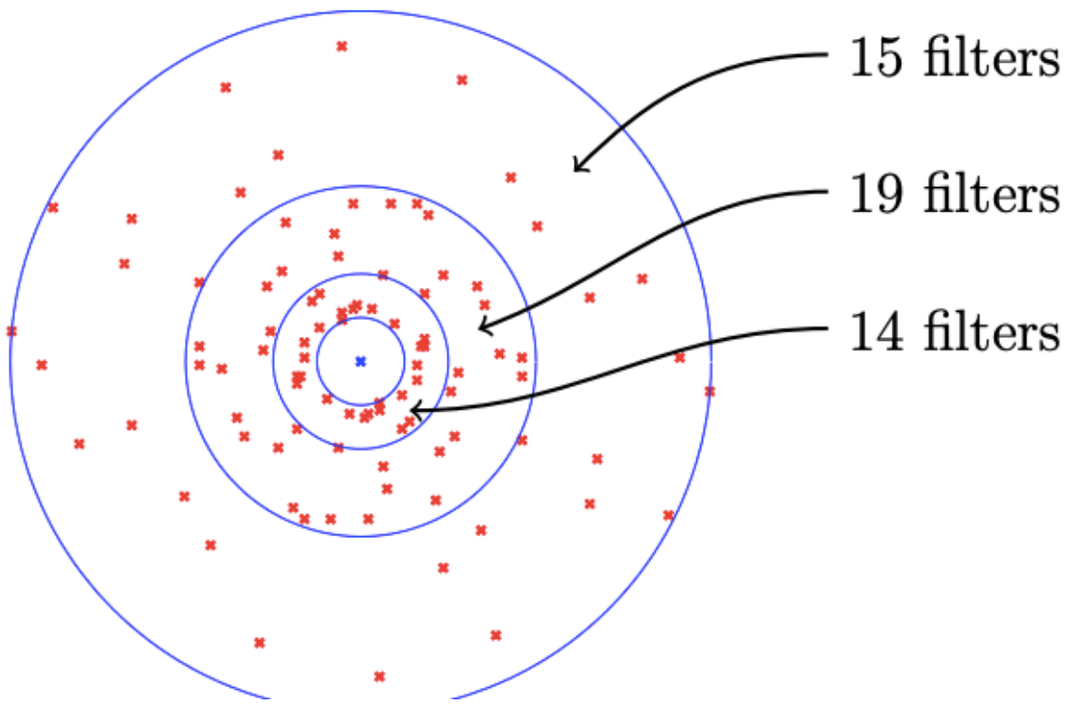
$$\psi_{C,D,\xi}(u) = Ce^{-u^T Du} e^{iu^T \xi}$$



Ref.: I Waldspurger's phd

- Consider Gabor filters and fit the model.

This principle is core in many models (V1, Scattering,...)



Model for the second layer

Ref.: I Waldspurger's phd

First layer:

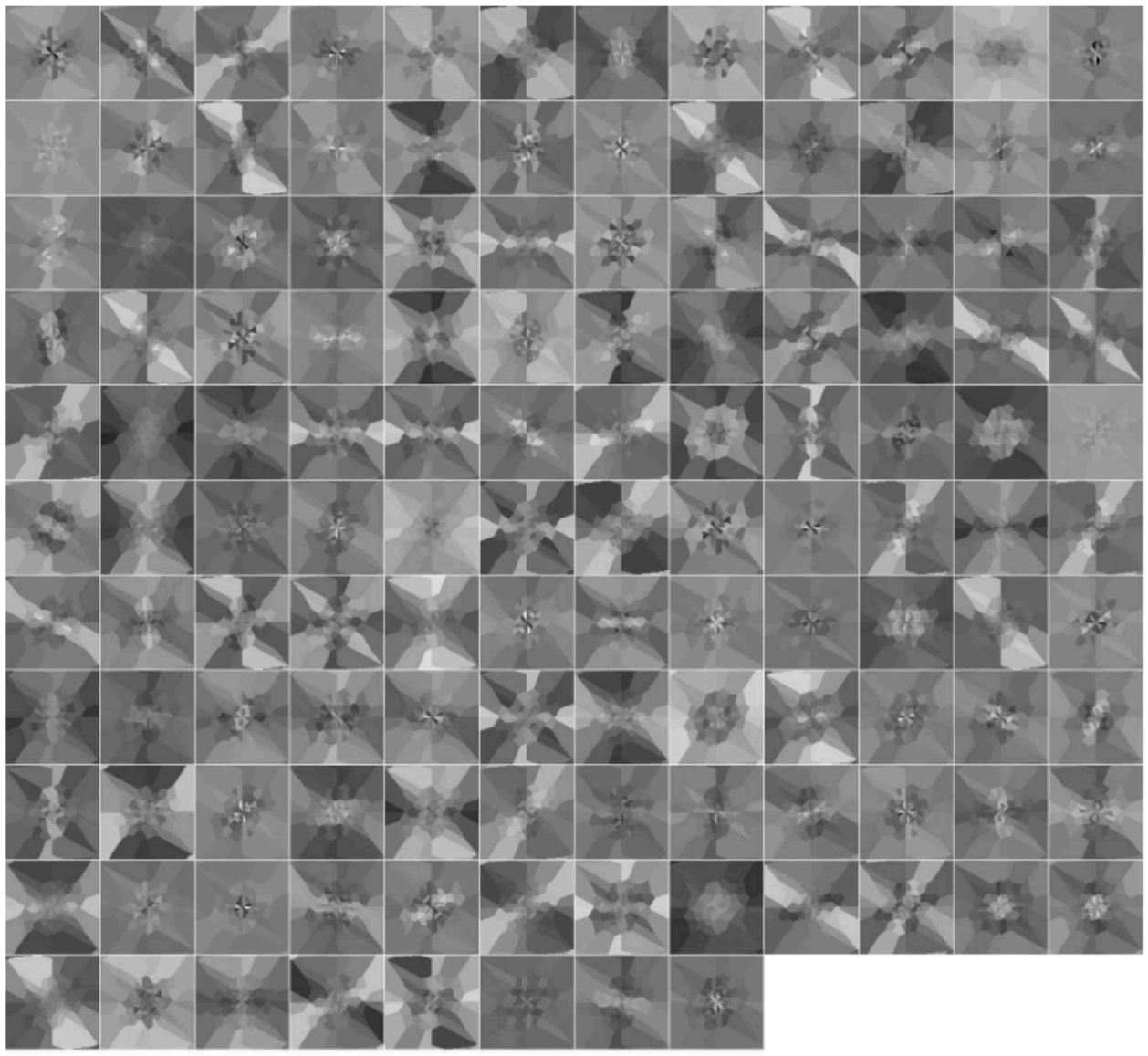
$$\psi_\lambda(u)$$

Second layer:

$$\psi(u,\lambda) \approx \phi^1(u) \times \phi^2(\lambda)$$

Recombines along λ

Why was this possible?
We were aware of the topology
of the previous layer!



Visualisation of ϕ^2
in the frequency plane

Visualization methods

- Given a representation of a CNN (*specific activation, output probes)*, find a relevant input (*patches, full image*)

- Many successful methods! Yet, quite heuristic...

- It has lead to techniques like "deep dreams"

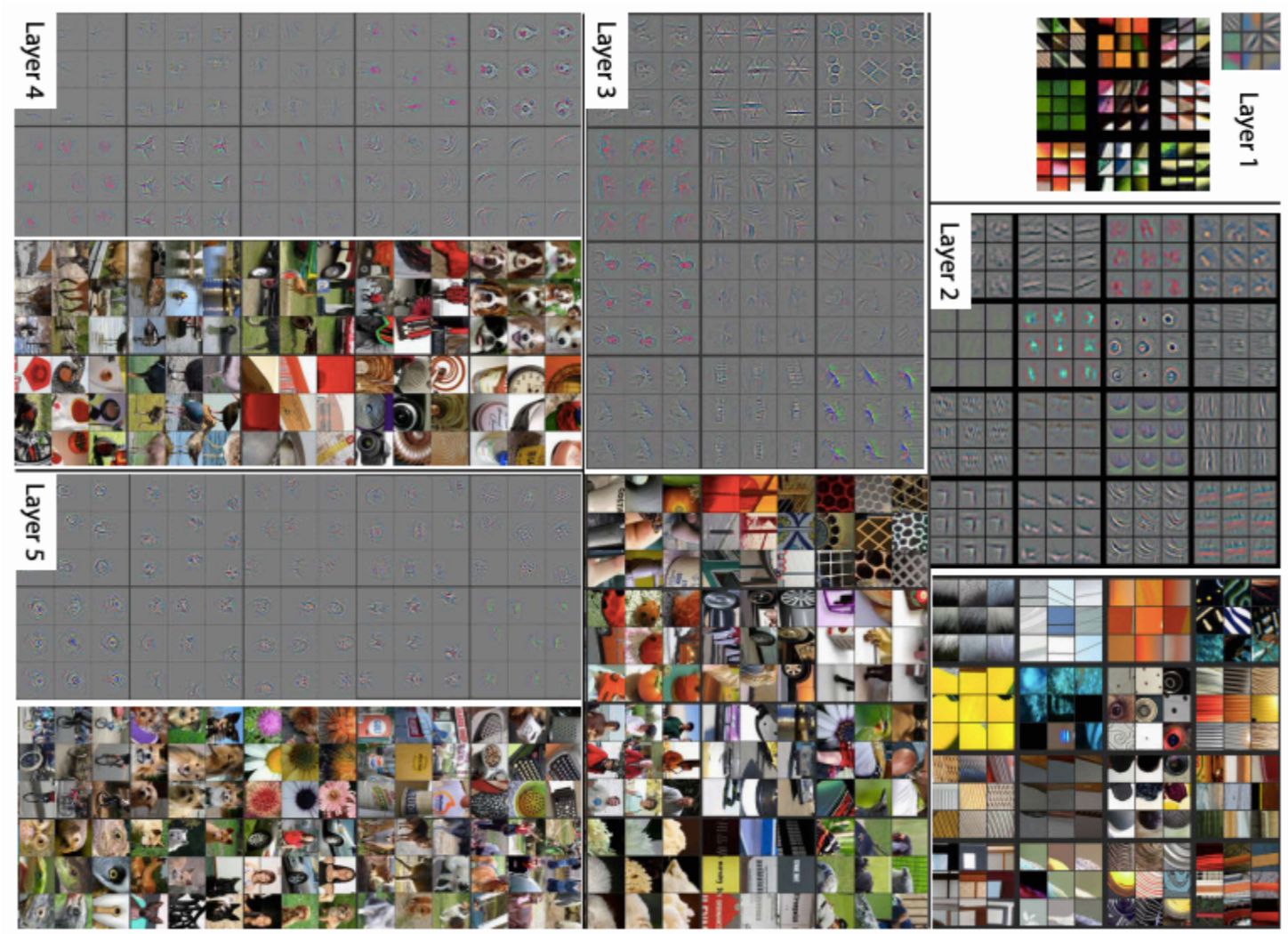
DeconvNet

$$X_0 \longrightarrow \boxed{W_1} \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_2} \longrightarrow \dots \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_J} \longrightarrow \Phi(x)$$

- Consider a CNN:

Ref.: Visualizing and Understanding Convolutional Networks, Zeiler and Fergus

$$\varphi_j \longrightarrow \boxed{W_j^T} \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_{j-1}^T} \longrightarrow \dots \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_1^T} \longrightarrow X(\varphi_j)$$

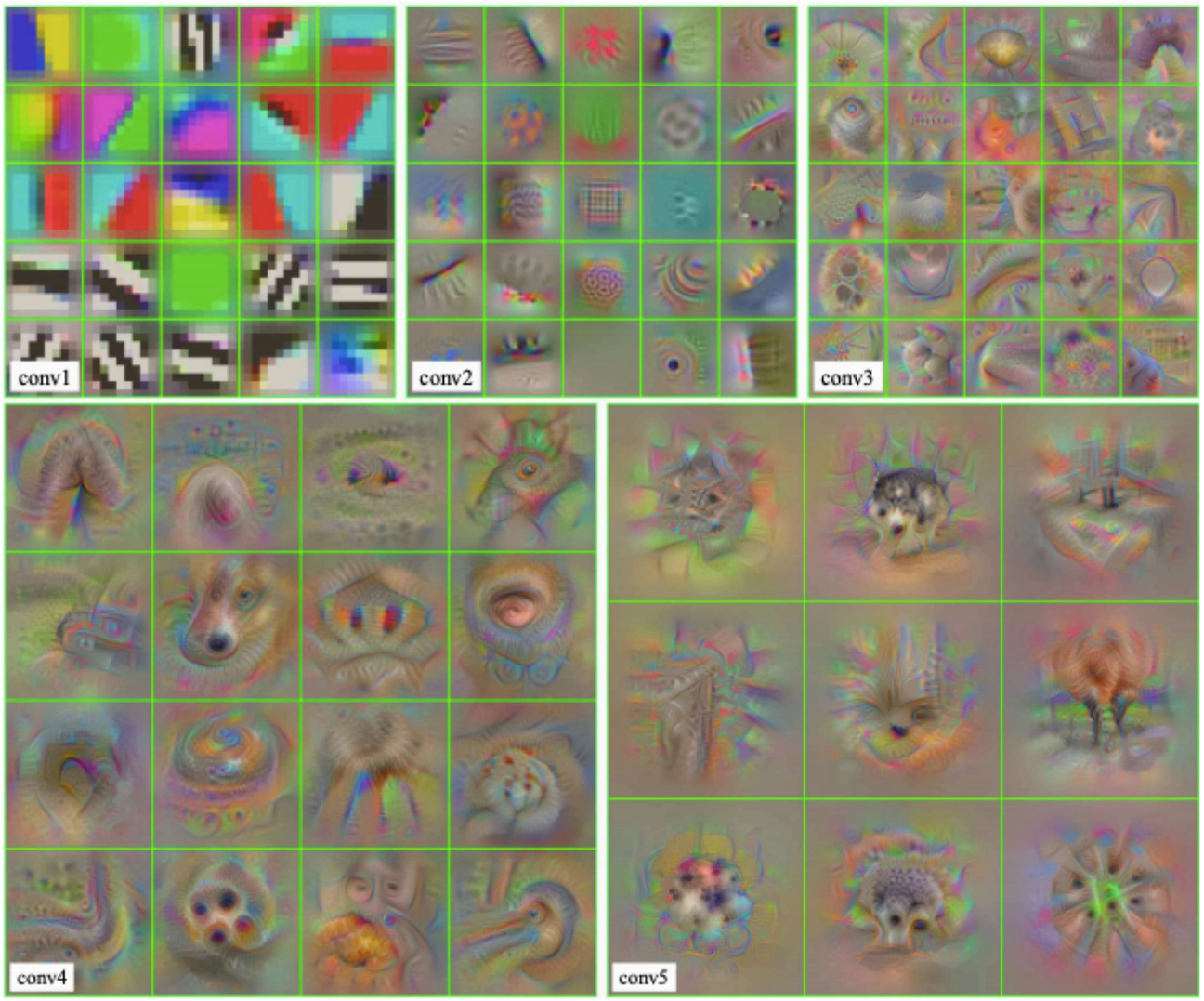


"sort of" dual network

$$\nabla_x (W \circ f) = W^T \nabla_x f$$

No learning!

Maximization of the activation



Ref.: Visualizing Deep Convolutional Neural Networks Using Natural Pre-images, mahendran and vedaldi

Fix an index i of a representation, set:

$$\Phi_i = (0, ..., 0, 1, 0, ..., 0)$$

and consider

$$x = \arg\min_x \mathcal{R}(x) + \lambda \langle \Phi x, \Phi_i \rangle$$

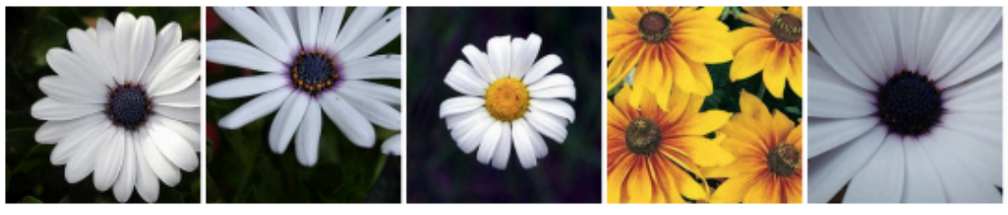
image prior (TV norm)

It consists in "studying" a specific neuron.

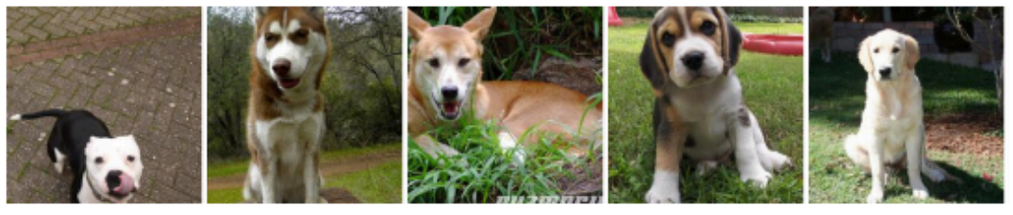
Concept of neuron?

Ref.: Intriguing properties of Deep Neural Networks, Szegedy et al.

- Consider: $v \in \mathbb{R}^{1000}, \quad x_v = \arg\max_{x \in \mathcal{D}} \langle \Phi x, v \rangle$ ← dataset
- Claim 1: $v = (0, ..., 0, 1, 0, ..., 0)$ has a semantic meaning
- Claim 2: any unit norm v has a semantic meaning.

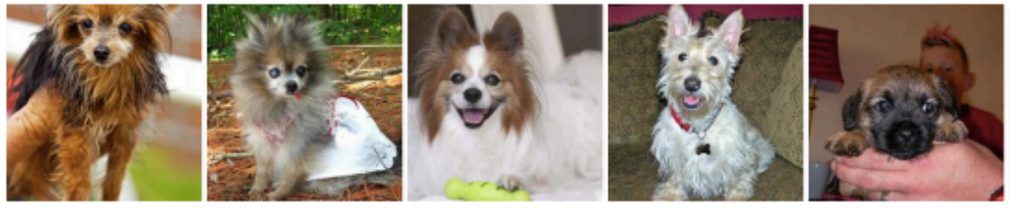


(a) Direction sensitive to white, spread flowers.



(b) Direction sensitive to white dogs.

(c) Direction sensitive to spread shapes.

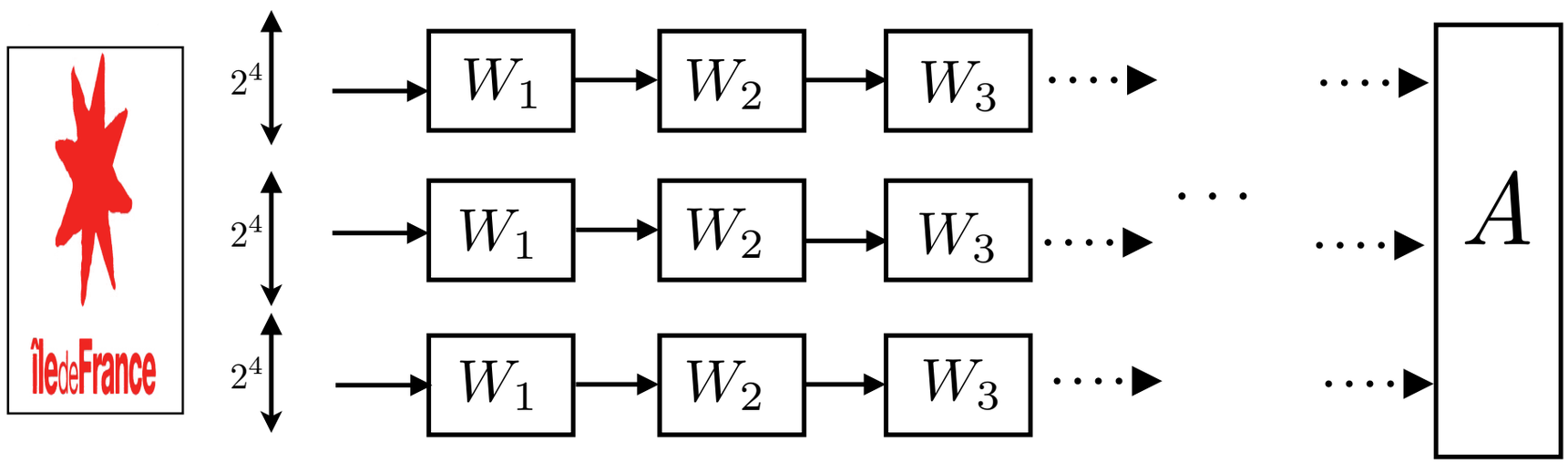


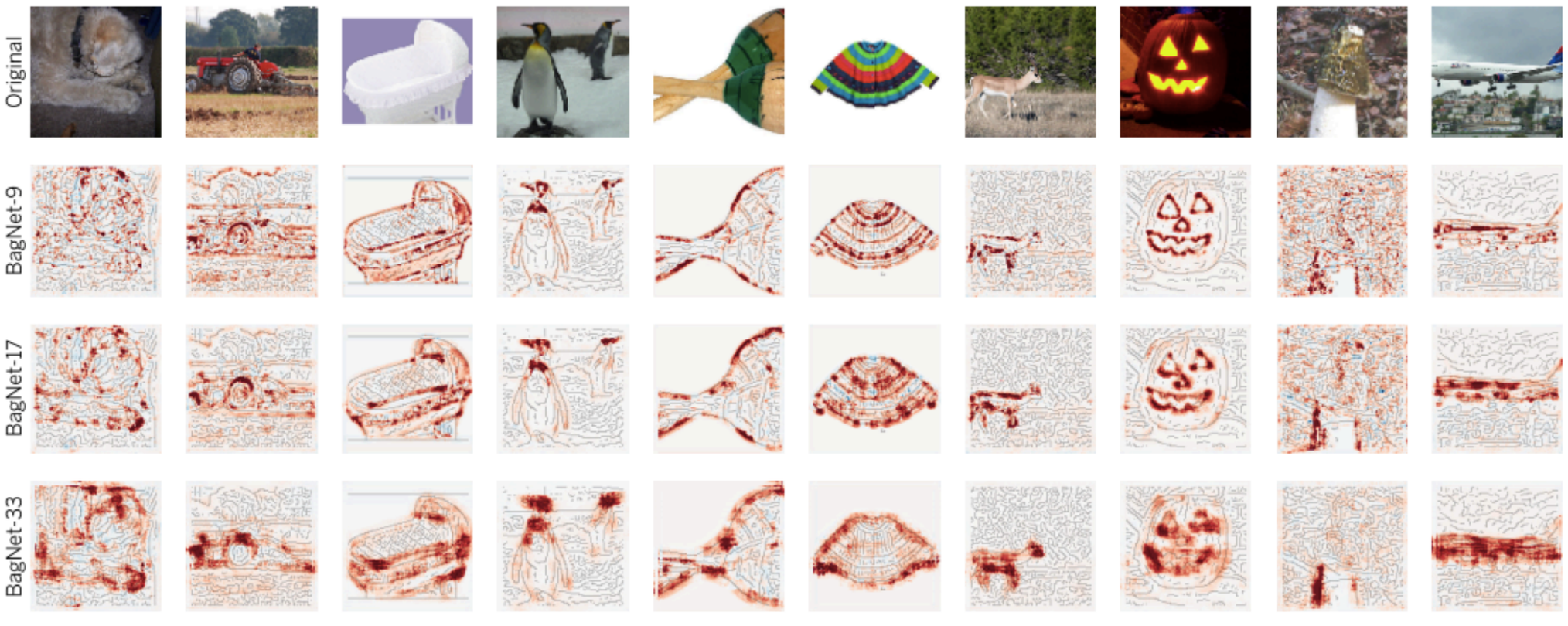
(d) Direction sensitive to dogs with brown heads.

Surprising BagNet Spatial distribution

"BagNet"

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



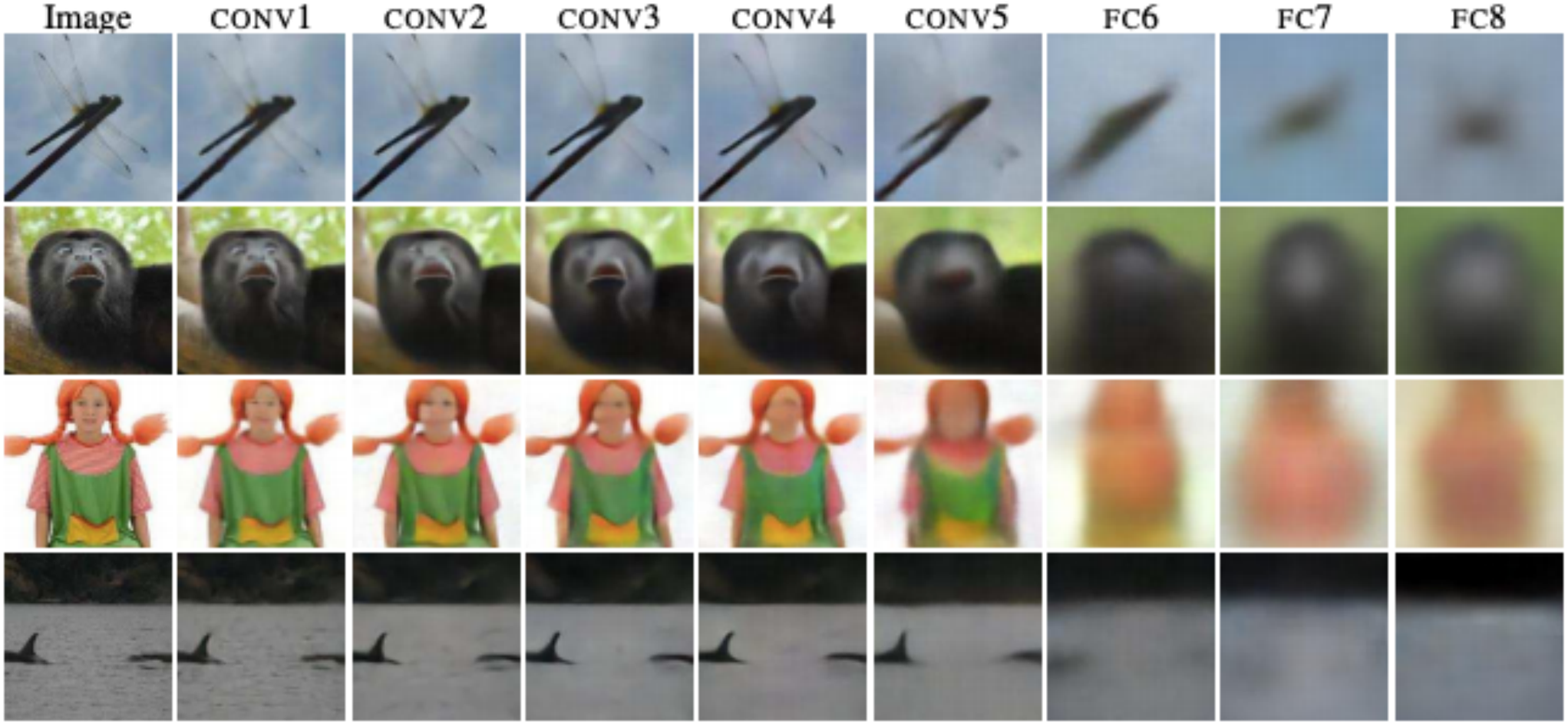


Reconstruction from a given layer?

$$X_0 \longrightarrow \boxed{W_1} \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_2} \longrightarrow \cdots \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_J} \longrightarrow \Phi(x)$$

Learn the operators!

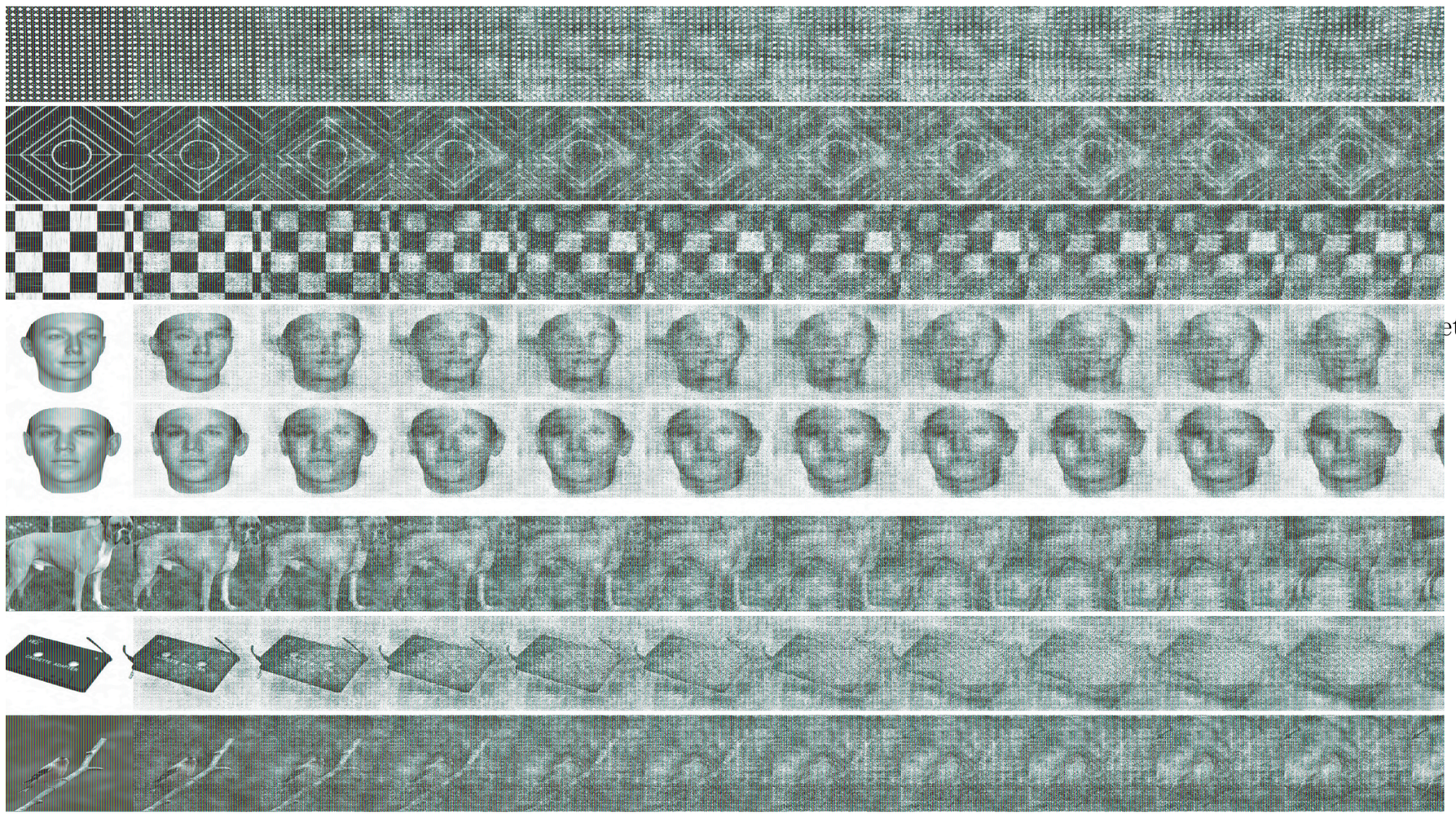
$$\varphi_j \longrightarrow \boxed{\tilde{W}_j} \longrightarrow \boxed{\rho} \longrightarrow \boxed{\tilde{W}_{j-1}} \longrightarrow \cdots \longrightarrow \boxed{\rho} \longrightarrow \boxed{\tilde{W}_1} \longrightarrow X(\varphi_j)$$



Ref.: Inverting Visual Representations with Convolutional Networks, Dodovistky et al.

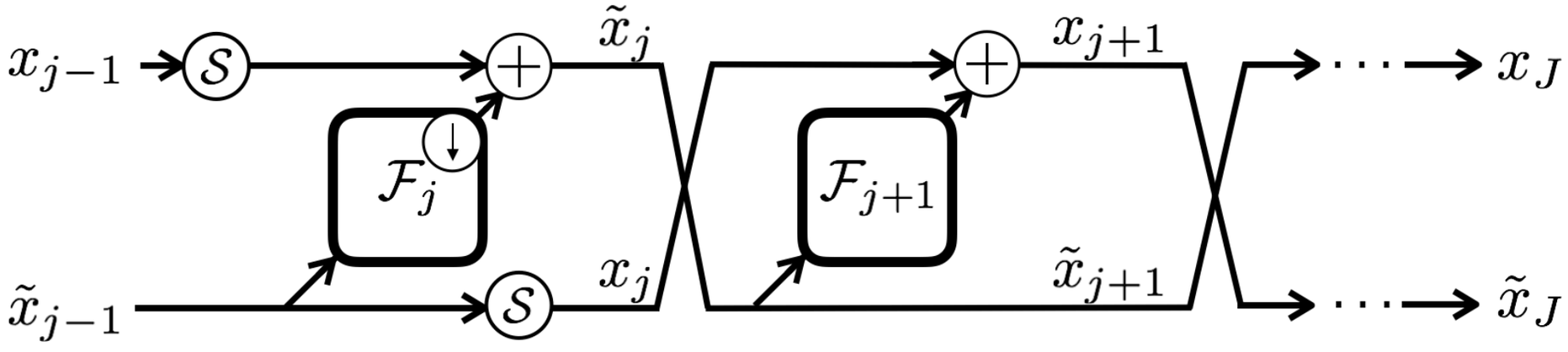
Even better: invertibility

Ref.: i-Revnet, depp invertible networks Jacobsen, Smeulder and EO



et al.

Lifting scheme



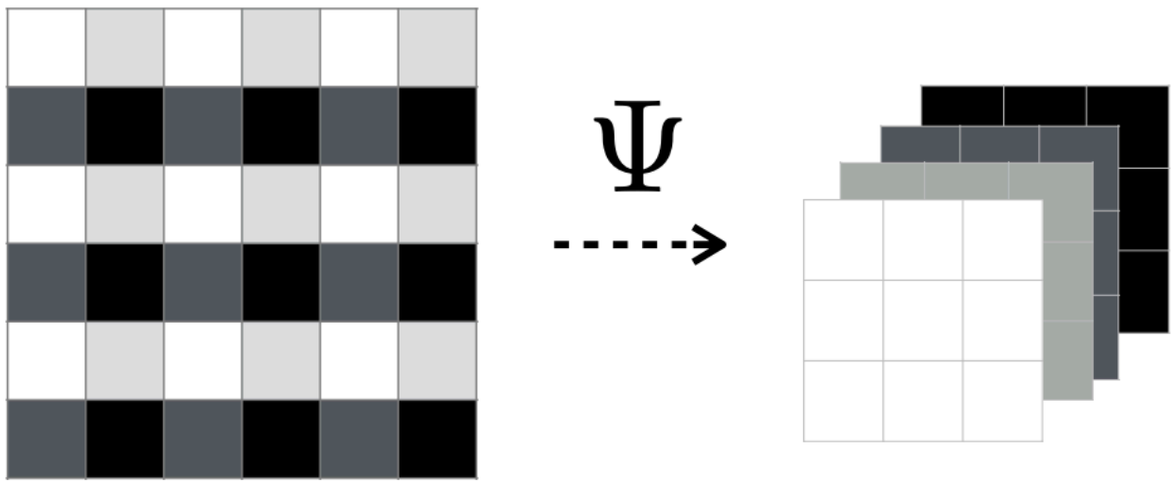
Invertibility?

- It's for free!

$$x_{j+1} = \tilde{x}_j \qquad \longleftrightarrow \qquad \tilde{x}_j = x_{j+1}$$

$$\tilde{x}_{j+1} = x_j + F_j \tilde{x}_j \qquad x_j = \tilde{x}_{j+1} - F_j \tilde{x}_j = \tilde{x}_{j+1} - F_j x_{j+1}$$

- Output dimension is the same as the input dimension...



input: $3 \times N \times N$

output: $3K^2 \times \frac{N}{K} \times \frac{N}{K}$

Reducing mutual information (Information bottleneck)

- Reducing the information sounds relevant:

$$I(X;Y) = \int_{\mathbb{R}^2} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\mathrm{d}y = H(X) - H(X|Y)$$

Ref.: Opening the Black Box of Deep Neural Networks via Information, R Shwartz-Ziv and N Tishby

Measures the dependancy between variables

$$I(X;\Phi_1 X) \geq I(X;\Phi_2 X) \geq ... \geq I(X;\Phi_J X)$$

"Compress" X

$$I(X;Y) \geq I(\Phi_1 X;Y) \geq ... \geq I(\Phi_J X;Y)$$

... but "reveal" Y

They propose to introduce:

$$\Phi_{j,\lambda} = \arg\inf_{\Phi} I(\Phi_{j-1}X, \Phi_j X) - \lambda I(\Phi_j X, Y)$$

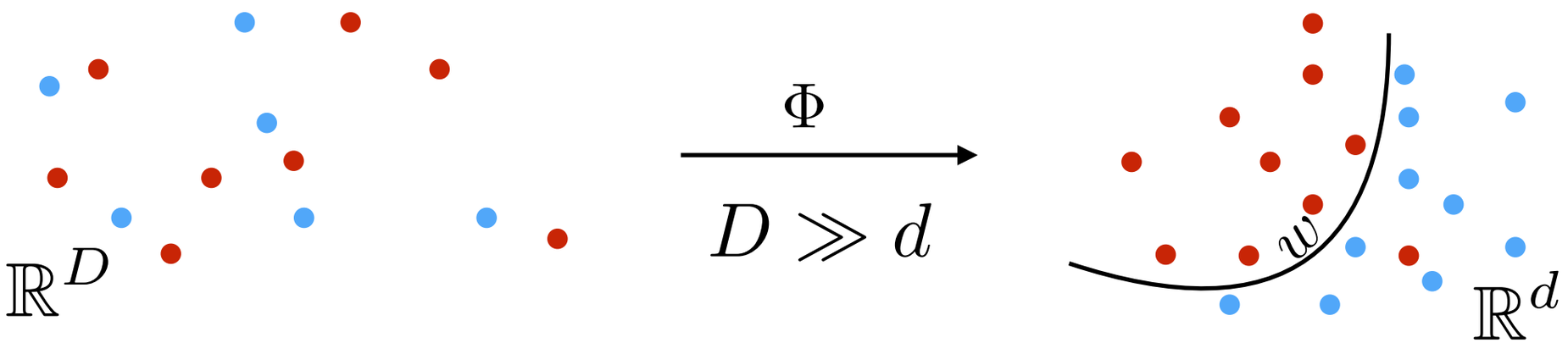
- But one can easily build invertible CNNs...

Ref.: i-RevNet: Deep Invertible Networks, J Jacobsen et al.

Data model

Fighting the curse of dimensionality

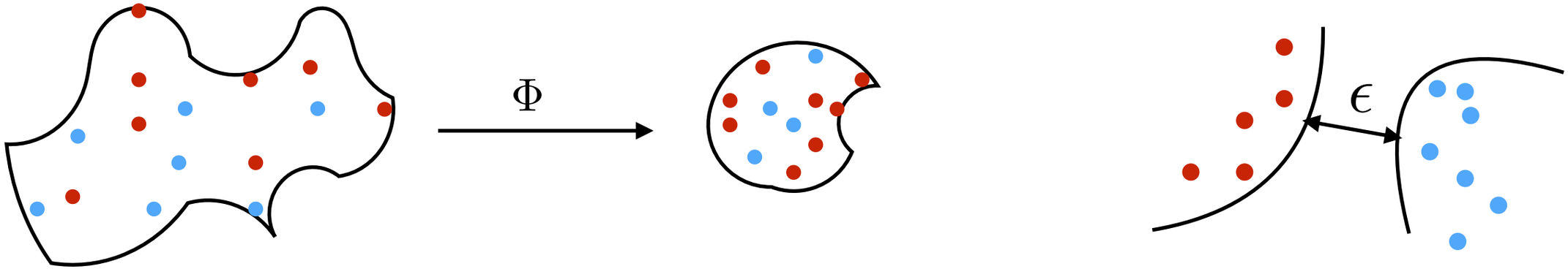
- **Objective**: building a representation Φx of x such that a simple (say euclidean) classifier $\hat{y}$ can estimate the label y:



- Designing Φ: must be regular with respect to the class:

$$\|\Phi x - \Phi x'\| \lll 1 \Rightarrow \hat{y}(x) = \hat{y}(x')$$

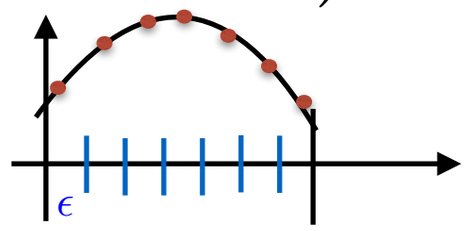
- **Necessary** dimensionality reduction and separation to break the curse of dimensionality:



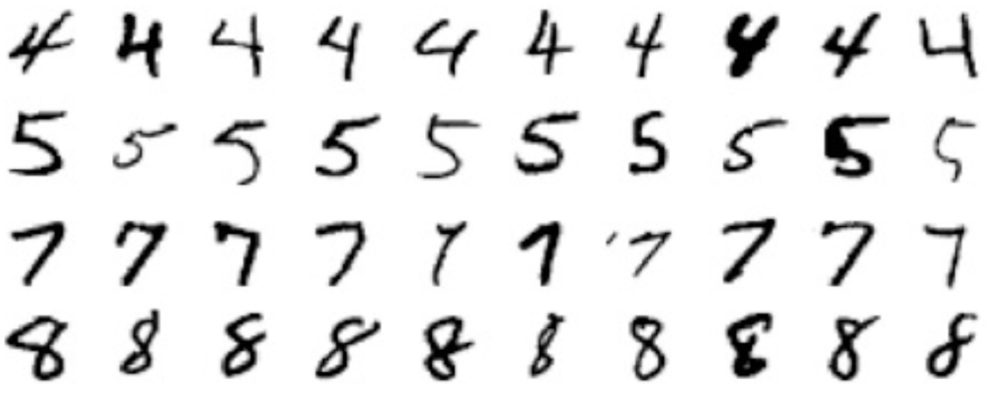
Model on the data: low dimensional manifold hypothesis?

- Low dimensional manifold: dimension up to 6. Not higher:

Property: if $f : \mathbb{R}^D \to [0,1]$ is 1-Lipschitz, then let $N_\epsilon = \arg\inf_N \sup_{i \leq N} \left(|f(x) - f(x_i)| < \epsilon \right)$.
Then $N_\epsilon = \mathcal{O}(\epsilon^{-D})$



- Can be true for MNIST...



All variabilities are known

Small "limited" deformations
+Translation

- Yet high dimensional deformations are an issue in the general case!

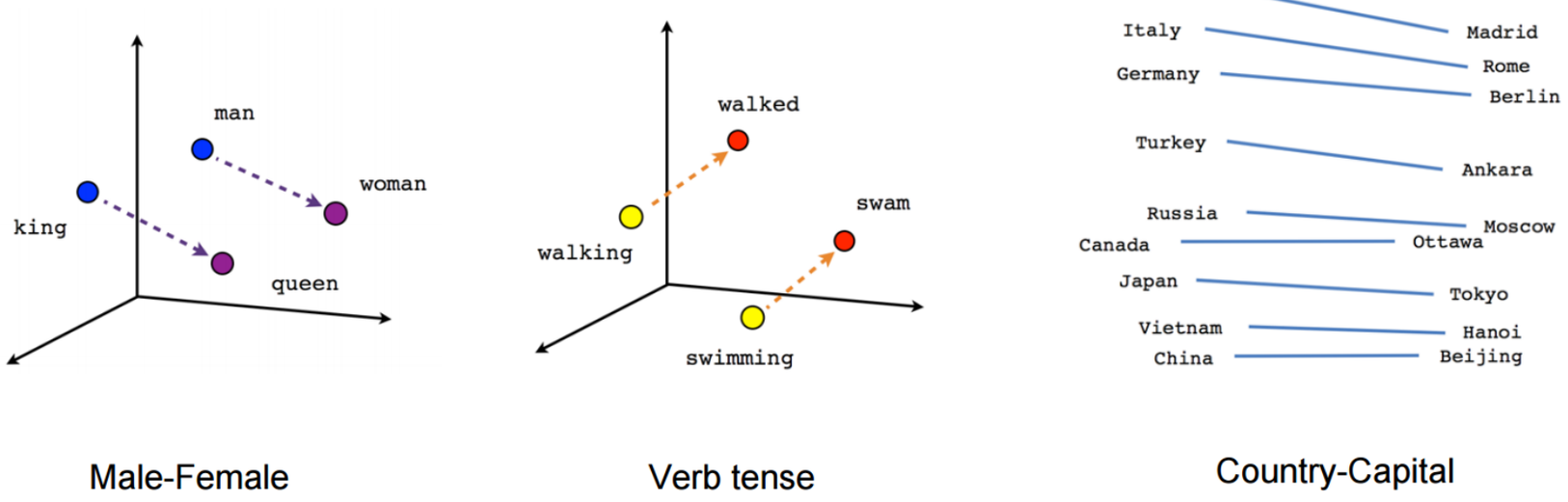
Flattening the space: progressive manifold?

- Parametrize variability on synthetic data: $L_\theta, \theta \in \mathbb{R}^d$ and observe it after PCA

Ref.: Understanding deep features with computer-generated imagery, M Aubry, B Russel

(c) Object color (d) Background color (a) Lighting (b) Scale

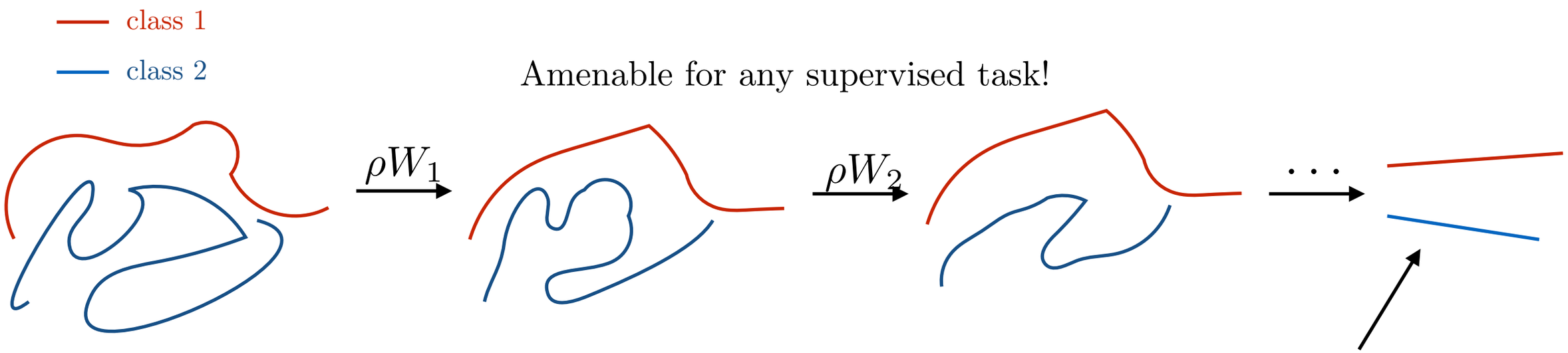
- Data tends to live on flattened space. Tangent space?



Male-Female Verb tense Country-Capital

Difficult to find evidences of such phenomen

Mechanism proposal: Flattening the level sets



Ref.: Understanding Deep Convolutional Networks, Mallat, 2016

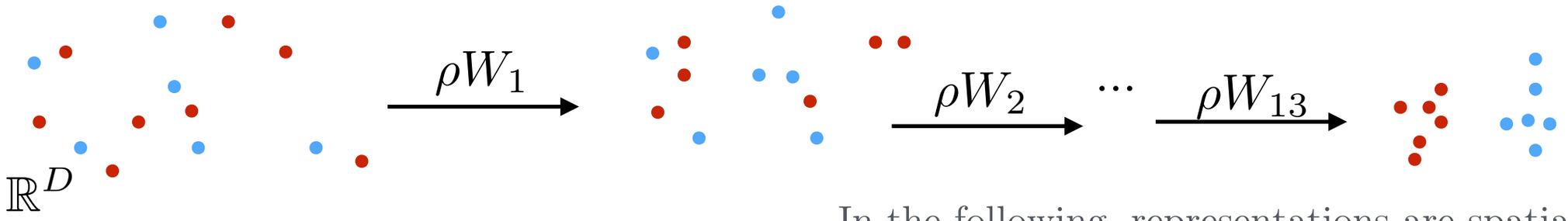
- How to linearize? Ex.: Gâteaux differentiability

$$\exists C_x,\ \sup_{\mathcal{T}} \frac{\|\Phi x - \Phi \mathcal{T} x\|}{\|\mathcal{T}\|} < C_x \quad \Rightarrow \exists \partial\Phi_x :\ \Phi\mathcal{T}x \approx \Phi x + \partial\Phi_x.\mathcal{T}$$

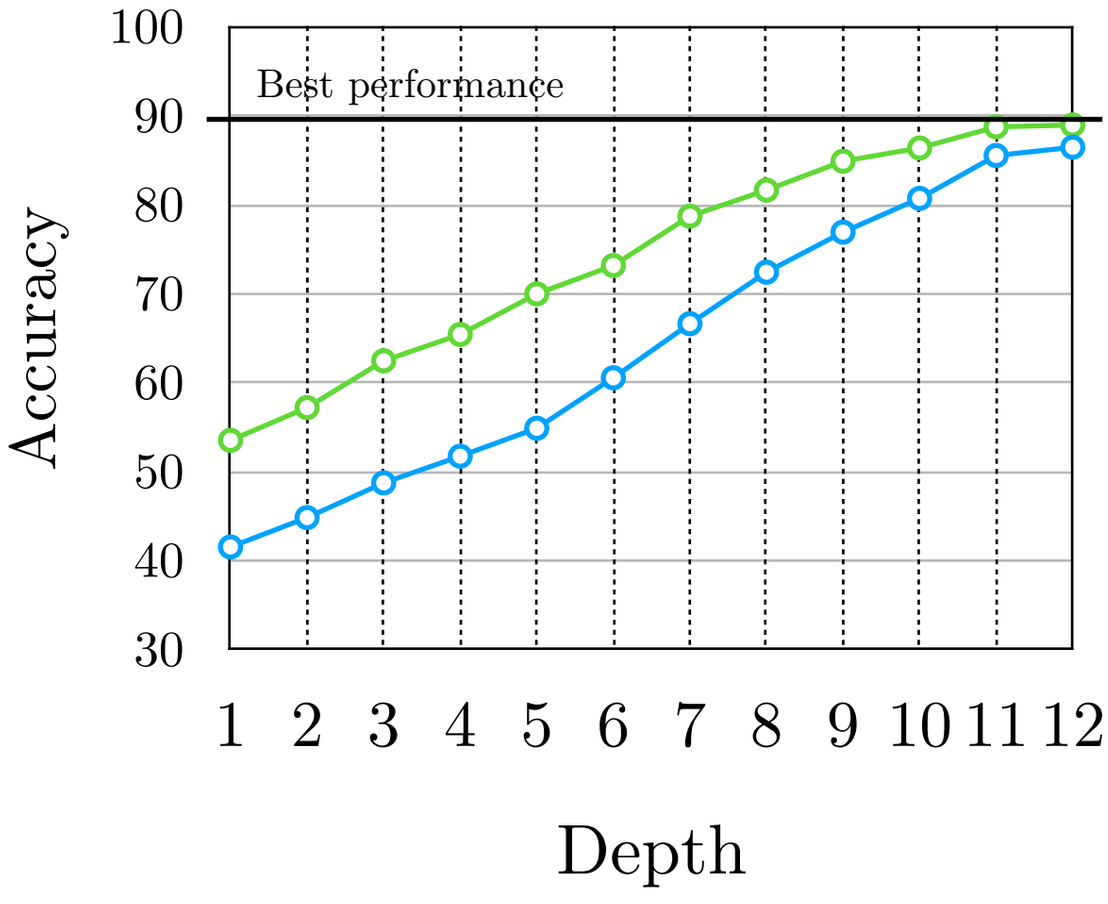
- However, exhibiting $\mathcal{T}$ can be difficult. (*curse of dimensionality*)
  Ex.: linear translations $\mathcal{T}_a\big(x\big)(u) \triangleq x(u+a)$, yet non linear case?

Empirical observation: Progressive separability

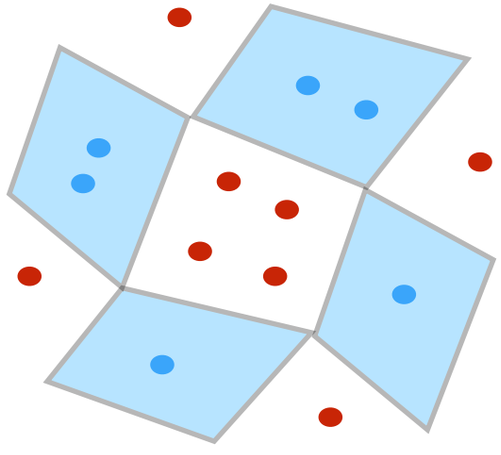
- **Typical CNN exhibits a progressive contraction & separation, w.r.t. the depth:**



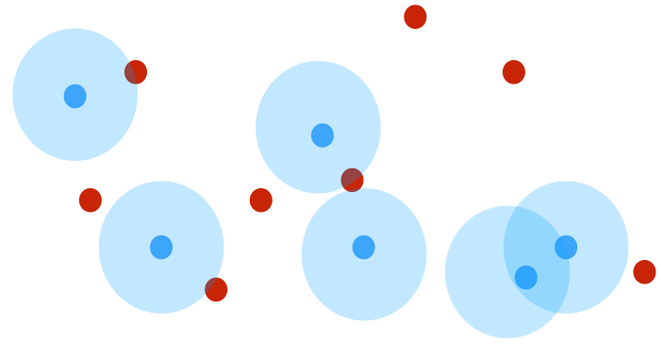
In the following, representations are spatially averaged.



Nearest Neighbor (NN)



Gaussian SVM

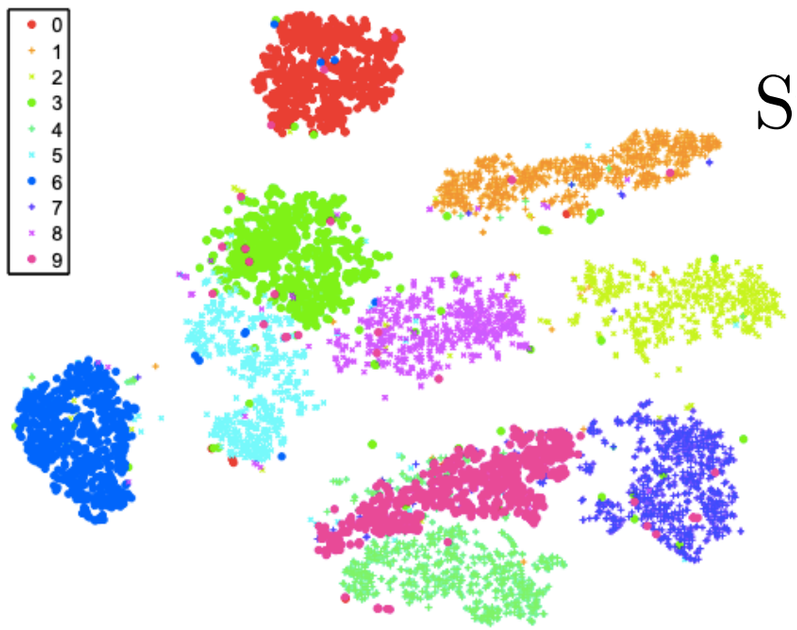


Localised classifiers

Ref.: Building a Regular Decision Boundary with Deep Networks, EO

- **How can we explain it?**

t-SNE



Step 1: get a localised distribution from $x_i \in \mathbb{R}^N$

$$p_{i,j} = \frac{\exp(-\|x_i - x_j\|)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|)}$$

Step 2: consider $y_i \in \mathbb{R}^n$

$$q_{i,j} = \frac{1 + \|y_i - y_j\|}{\sum_{k \neq l} 1 + \|y_k - y_l\|}$$

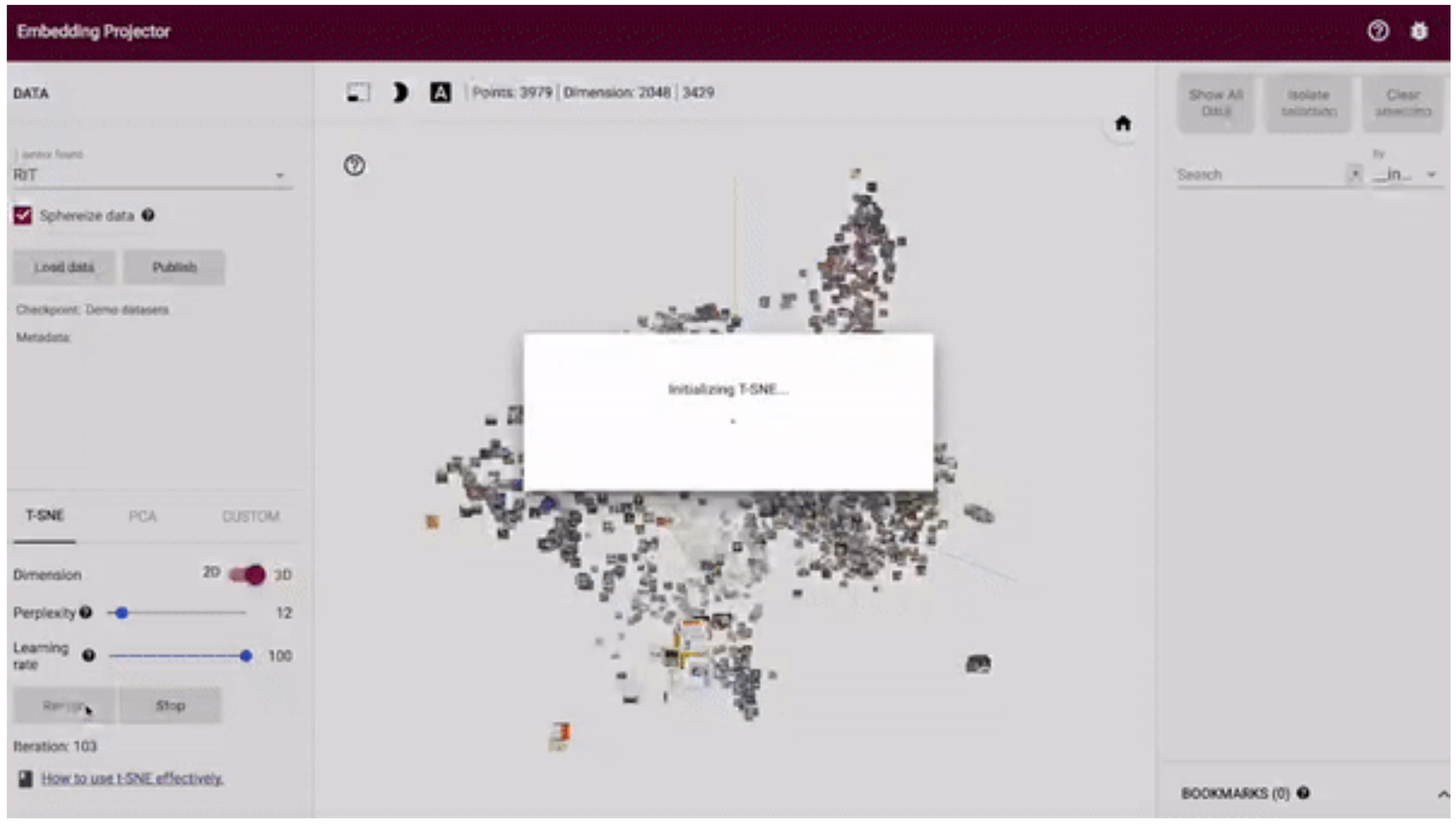
$n \ll N$

Step 3: optimise

$$\inf_{\{y_i\}} \sum_{i,j} p_{i,j} \frac{\log p_{i,j}}{\log q_{i,j}}$$

Issues: curse of dimensionality...

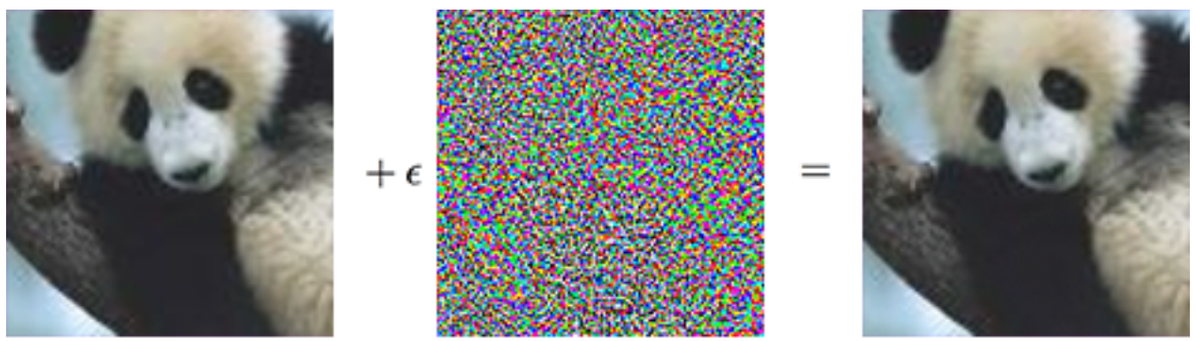
t-SNE on DNN



Allows to studying neighbours of a point.

Stability of NN

Adversarial examples



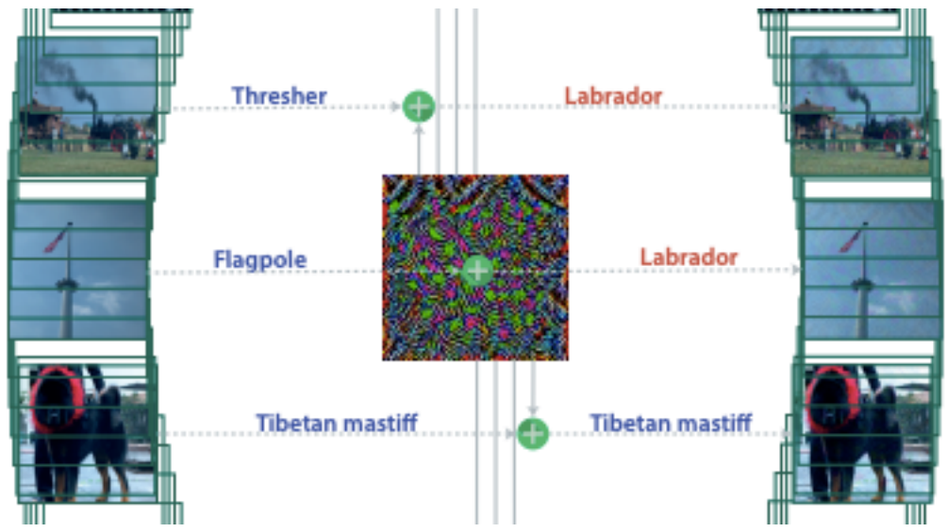
- How can we obtain adversarial examples? For ex:

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

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.



Why instabilities?

- Lipschitz constant!

$$\|f\|_L = \sup \frac{\|f(x) - f(y)\|}{\|x - y\|} \longrightarrow \|f \circ g\|_L \leq \|f\|_L \|g\|_L$$

For any f, large $\|f\|_L$ induce instabilities !

- Techniques to remove them: reducing individual lipschitz constant, adapting learning, controlling norms...

Lipschitz Constant

$$\|\rho W_j\|_L \leq \|W_j\|$$

Ref.: Lipschitz regularity of deep neural networks: analysis and efficient estimation, Scaman and Virmaux

- Those b refined.

Layer	Size	Stride	Upper bound
Conv. 1	$3 \times 11 \times 11 \times 96$	4	2.75
Conv. 2	$96 \times 5 \times 5 \times 256$	1	10
Conv. 3	$256 \times 3 \times 3 \times 384$	1	7
Conv. 4	$384 \times 3 \times 3 \times 384$	1	7.5
Conv. 5	$384 \times 3 \times 3 \times 256$	1	11
FC. 1	9216×4096	N/A	3.12
FC. 2	4096×4096	N/A	4
FC. 3	4096×1000	N/A	4

How to compute norms?

FC

$A_{k,k'}$

$\downarrow$ SVD on A

$\sigma_1, ..., \sigma_K$

$\boxed{\mathrm{Sp}(AA^T)}$

Conv

$$\tilde{x}(u, \tilde{\lambda}) = \sum_{\lambda} w_{\tilde{\lambda},\lambda} \star^u x(., \lambda)$$

$\downarrow$ Fourier

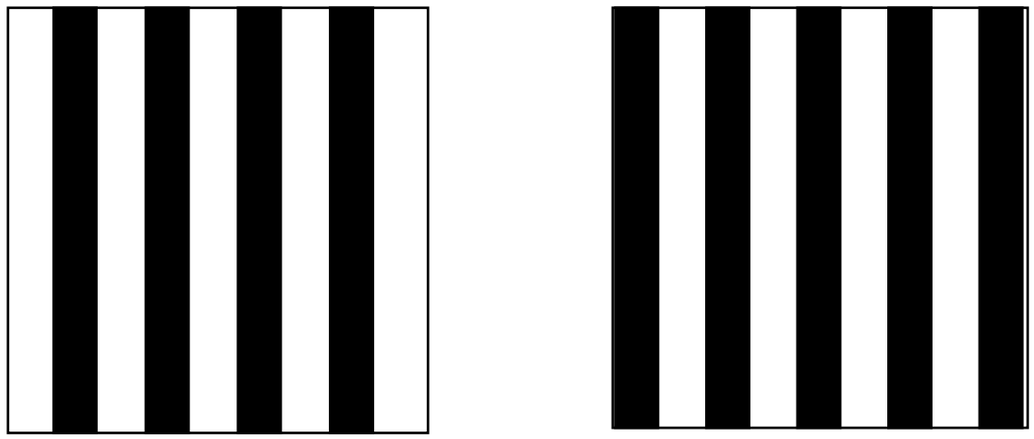
$$\hat{\tilde{x}}(\omega, \tilde{\lambda}) = \sum_{\lambda} \hat{w}_{\tilde{\lambda},\lambda}(\omega) \hat{x}(\omega, \lambda)$$

$\downarrow$ SVD on $\hat{w}_{\tilde{\lambda},\lambda}$

$\sigma_{\omega,1}, ..., \sigma_{\omega,\Lambda}$

Invariance Mechanisms

Translation



x y

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

Rotation

x y

Averaging is the key to get invariants

High dimensionality issues

Non-informative Variability

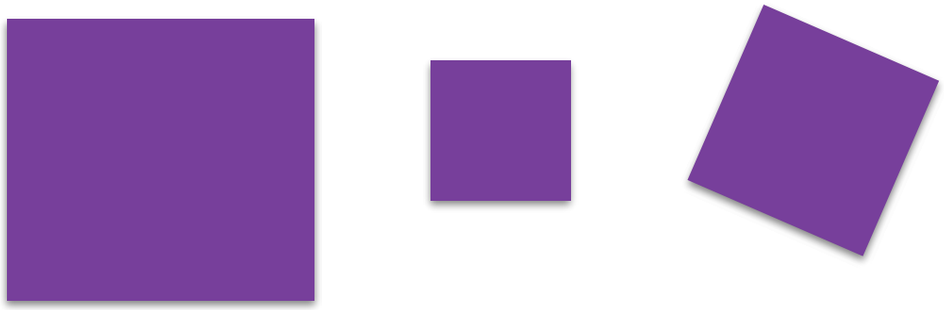
Consider a target function f (e.g. the class of a sample x)

- An invertible operator is said to be a symmetry if:

$$f(\mathcal{T}x) = f(x)$$

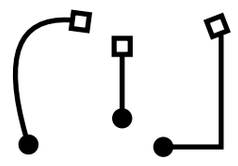
- Example of linear-geometric variability for classification:

$$r_\theta.x(u) \triangleq x(r_\theta u)$$



$$\tau_a.x(u) \triangleq x(u+a)$$

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

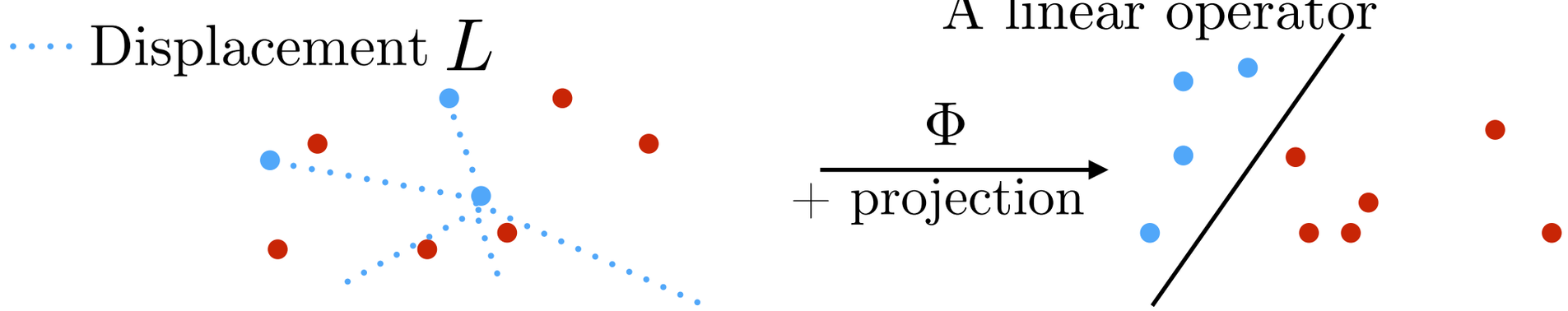


How to build invariance?

(1)

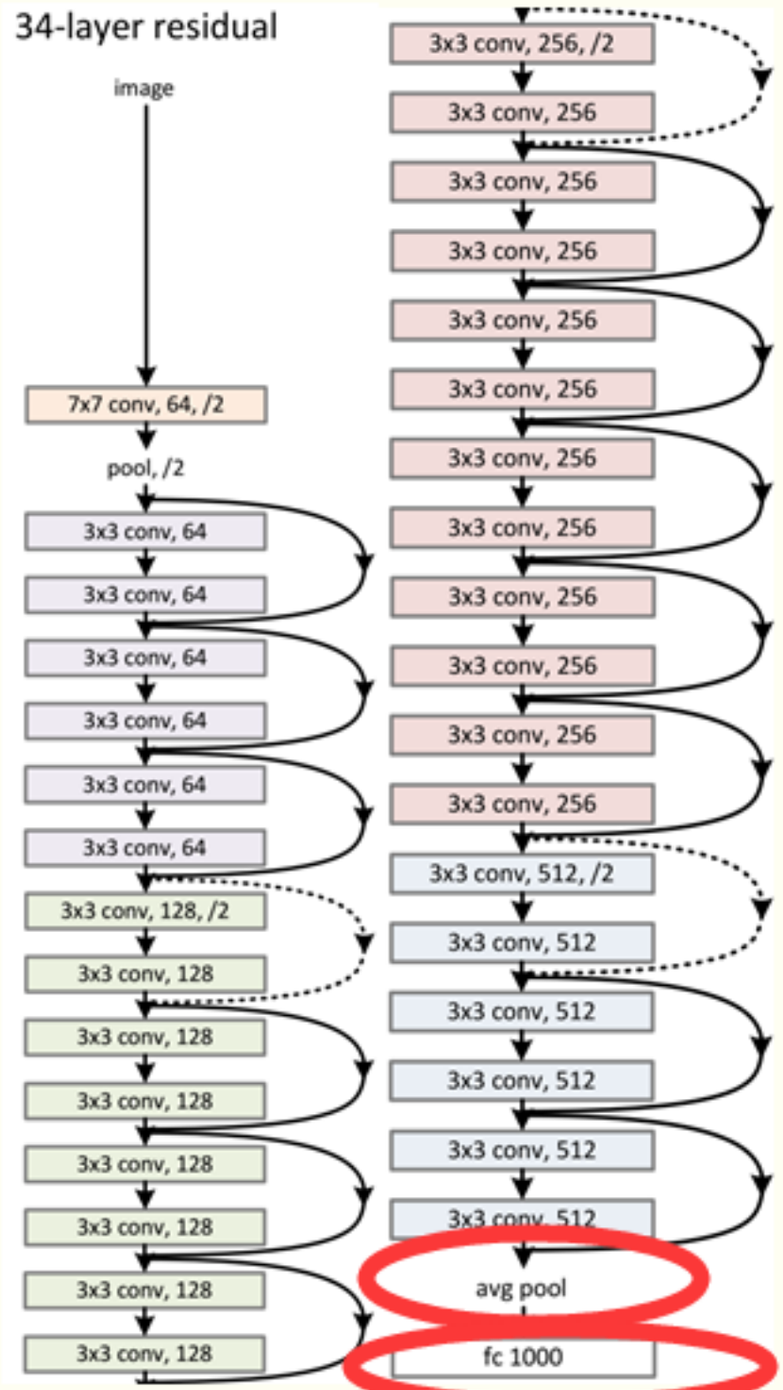
$$\sup_L \frac{\|\Phi Lx - \Phi x\|}{\|Lx - x\|} < \infty \Rightarrow \exists \text{ "weak" } \partial_x \Phi$$

$$\Rightarrow \Phi Lx \approx \Phi x + \underbrace{\partial_x \Phi L}_{\text{A linear operator}} + o(\|L\|)$$



Example: The scattering transform with deformations...

How to build invariance? (2)



- Obtain a covariant representation, and then, average! Ex.: convolutions.

covariance

$$W_j \mathcal{L}_a = \mathcal{L}_a W_j \quad \text{and} \quad \rho \mathcal{L}_a = \mathcal{L}_a \rho$$

and

$$A \mathcal{L}_a = A$$

this implies:

$$\begin{aligned} \Phi \mathcal{L}_a x &= A \rho W_J \ldots W_2 \rho W_1 \mathcal{L}_a x \\ &= A \mathcal{L}_a \rho W_J \ldots W_2 \rho W_1 x \\ &= \Phi x \end{aligned}$$

More complex covariance in CNNs?

$$\mathrm{X}_0 \longrightarrow \boxed{W_1} \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_2} \quad \dots \quad \longrightarrow \boxed{\rho} \longrightarrow \boxed{W_j} \longrightarrow \Phi_j x$$

Let $\mathcal{T}$ be a symmetry and estimate the (linear) inverse mapping:

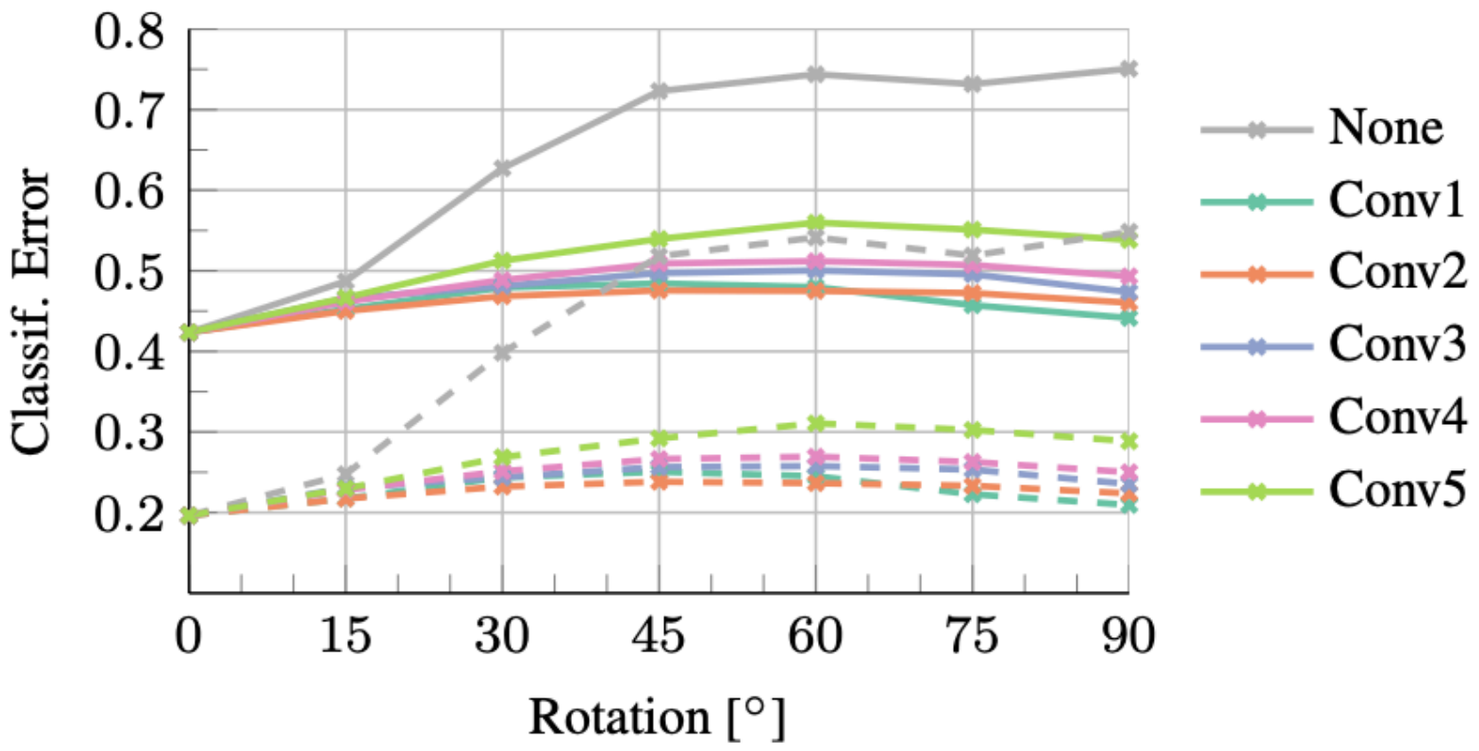
Ref.: Xxxx, Machin et al.

$$\inf_{T_j} \sum_n \|T_j^{-1} \Phi_j \mathcal{T} x^n - \Phi_j x^n\|$$

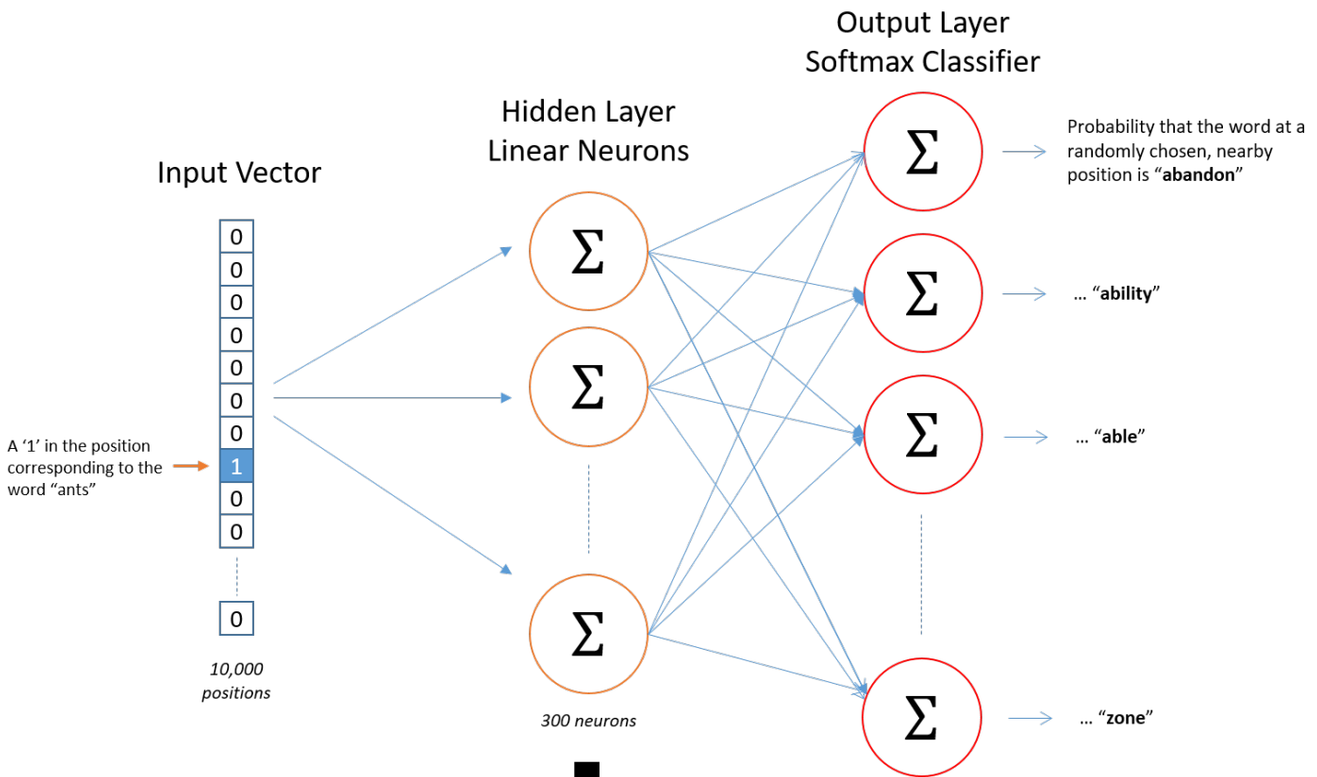
Ex.:

CNNs still classify correctly when switching the terms.

Ref.: Understanding image representations by measuring their equivariance and equivalence, Lenc et al.



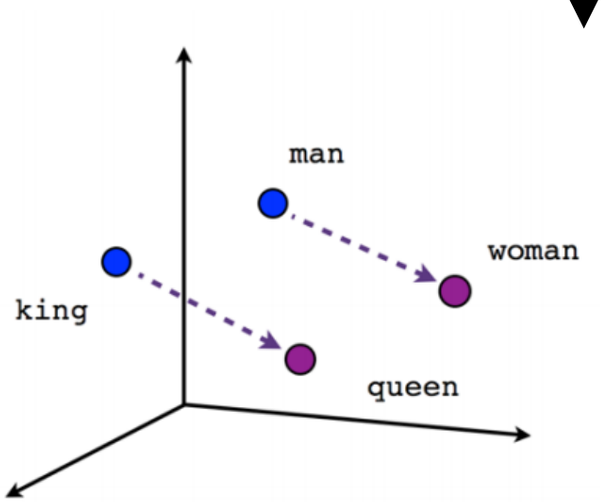
Word2vec covariance in text



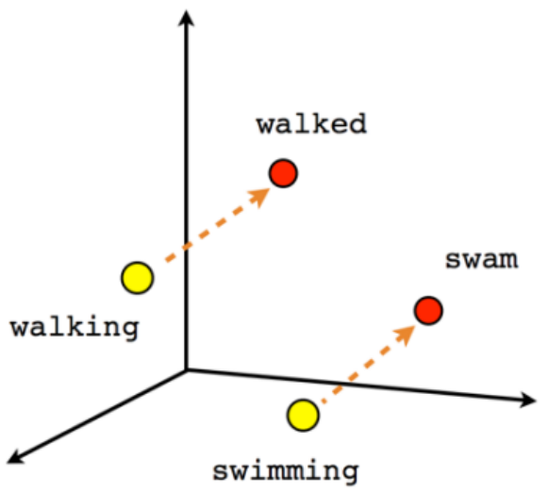
Ref.: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Given a word
output the probability
a word is nearby

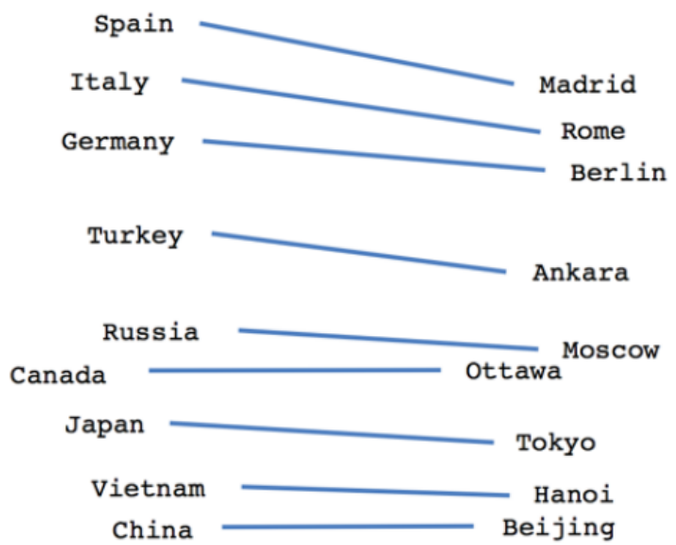
Suggest linearization!



Male-Female



Verb tense



Country-Capital

The Scattering Transform

An example: translation

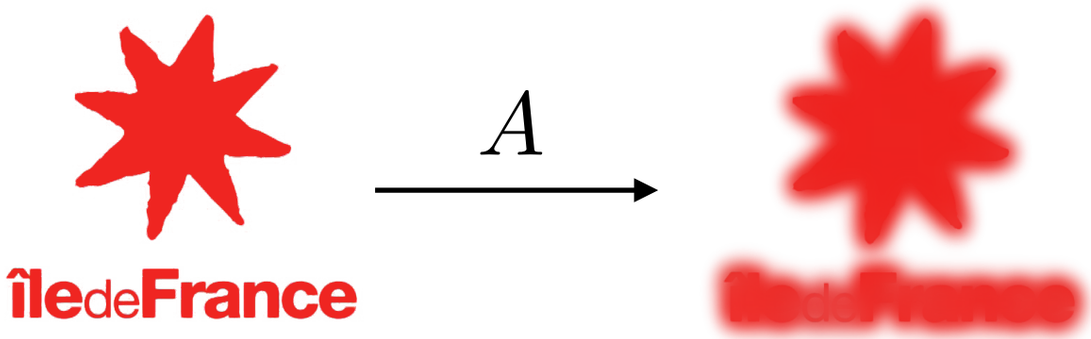
- Translation is a linear action:

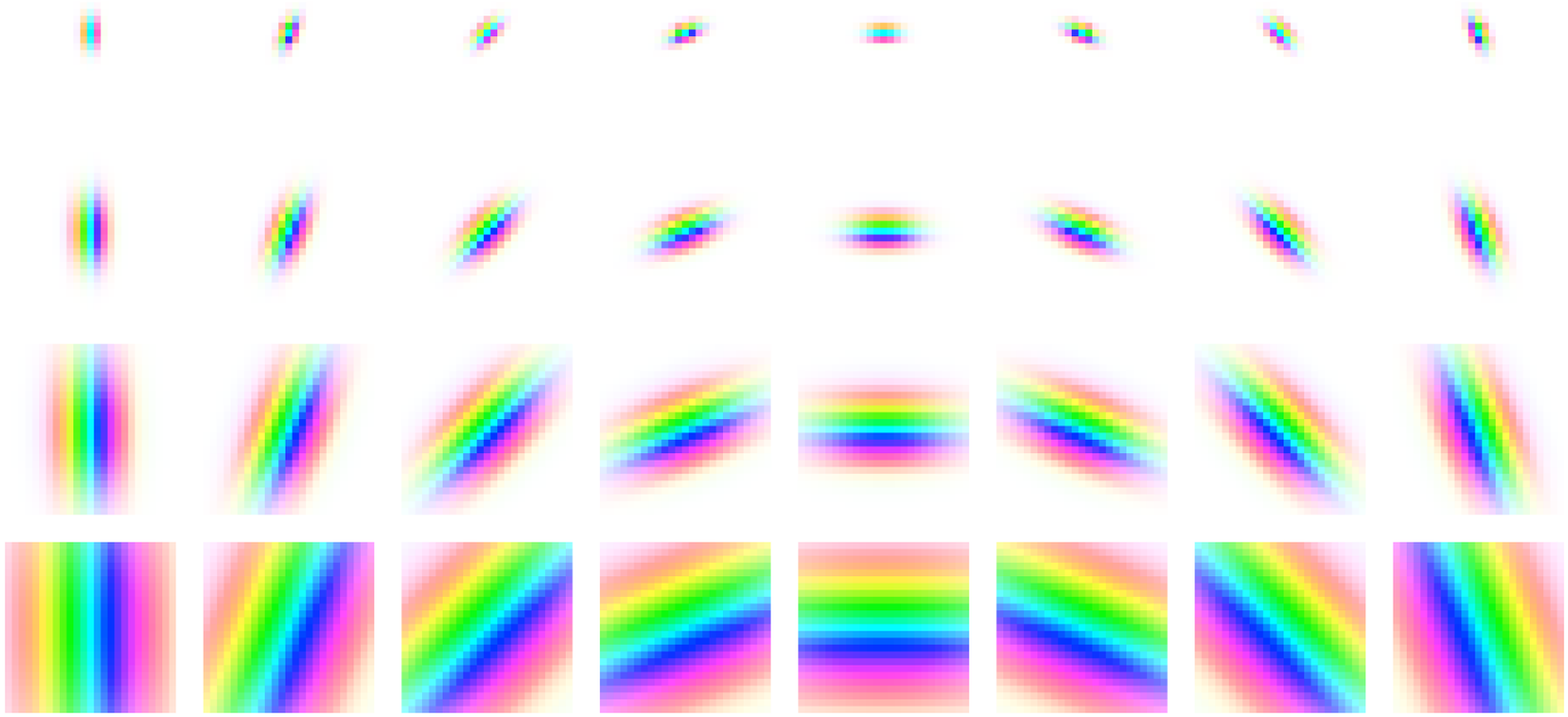
$$\forall u \in \mathbb{R}^2, L_a x(u) = x(u - a)$$

- 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!





$$\psi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}} (e^{i\xi.u} - \kappa)$$

$$\phi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}}$$

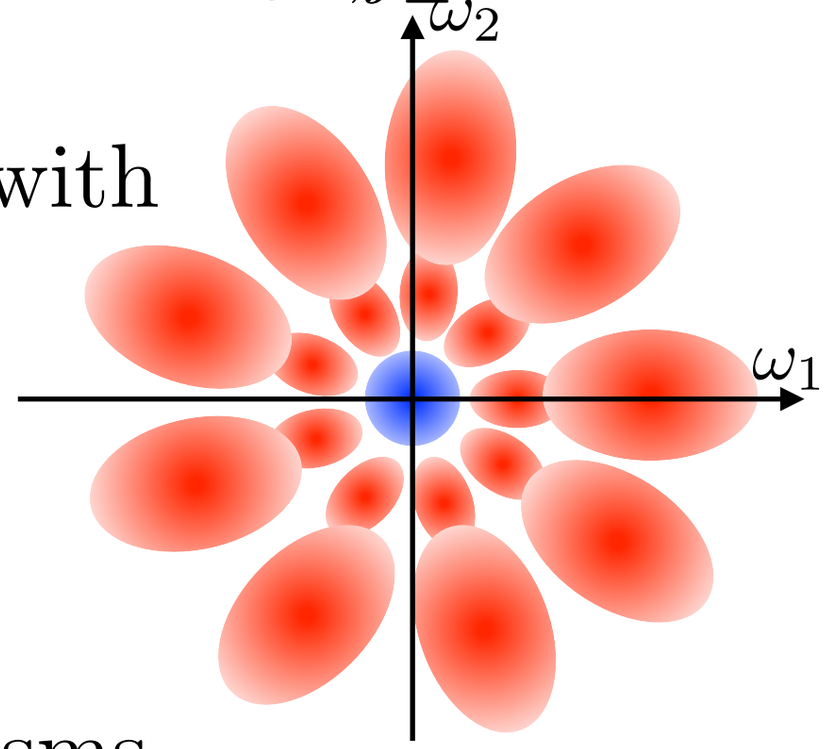
(for sake of simplicity, formula are given in the isotropic case)

The Gabor wavelet

Wavelet Transform

- Wavelet transform: $Wx = \{x \star \psi_{j,\theta}, x \star \phi_J\}_{\theta, j \leq J}$
- Isometric and linear operator of L^2, with

$$\|Wx\|^2 = \sum_{\theta, j \leq J} \int |x \star \psi_{j,\theta}|^2 + \int x \star \phi_J^2$$



- Covariant with translation L_a:

$$WL_a = L_a W$$

- Nearly commutes with diffeomorphisms

$$\|[W, L_\tau]\| \leq C\|\nabla \tau\|$$

Ref.: Group Invariant Scattering, Mallat S

- A good baseline to describe an image!

Filter bank implementation of a Fast WT

Ref.: Fast WT, Mallat S, 89

- Assume it is possible to find h and g such that
$$\hat{\psi}_\theta(\omega) = \frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \quad \textbf{and} \quad \hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$
- Set:

$$x_j(u,0) = x \star \phi_j(u) = h \star (x \star \phi_{j-1})(2u) \quad \textbf{and}$$
$$x_j(u,\theta) = x \star \psi_{j,\theta}(u) = g_\theta \star (x \star \phi_{j-1})(2u)$$

- The WT is then given by $Wx = \{x_j(.,\theta), x_J(.,0)\}_{j \leq J,\theta}$

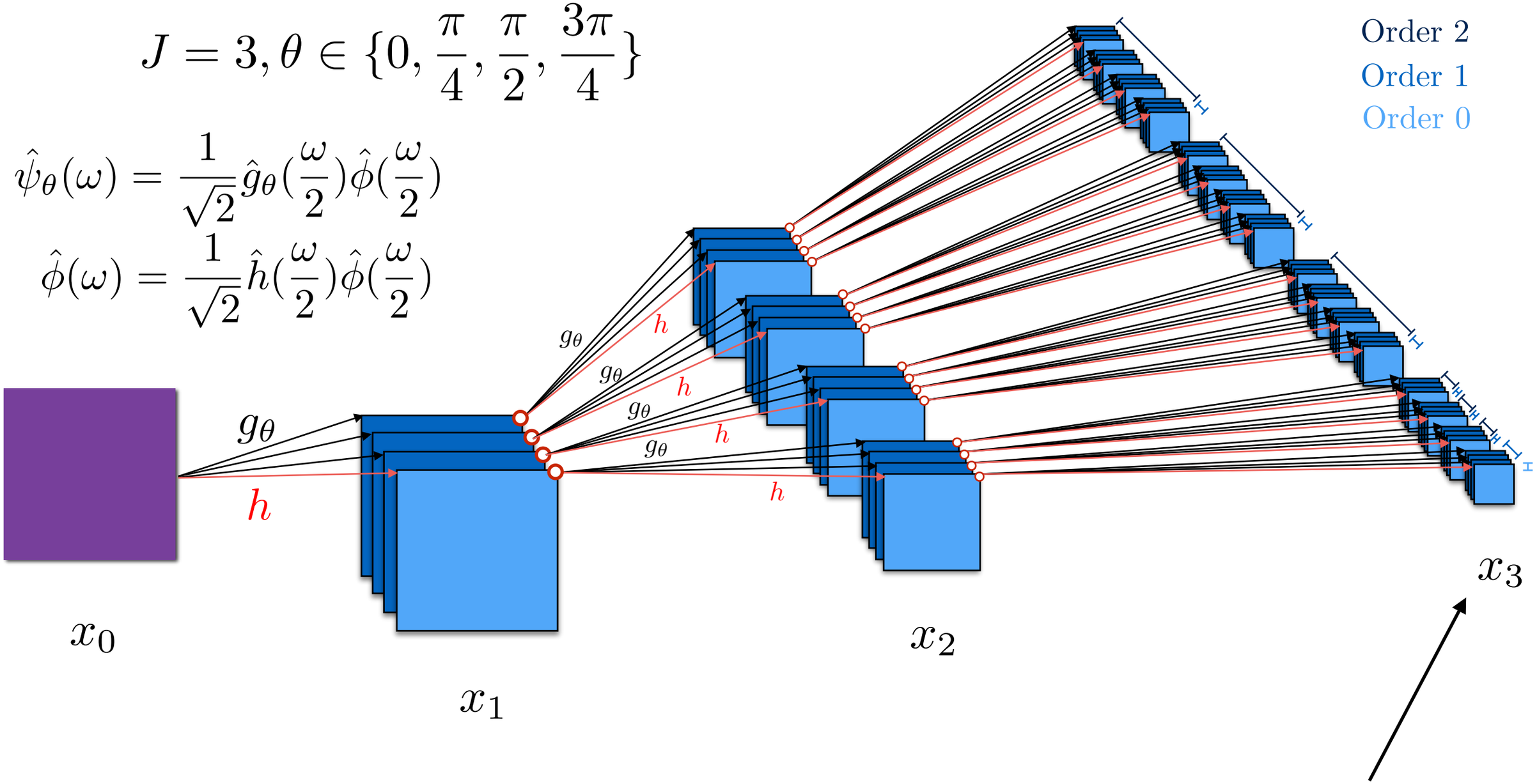
- A WT can be interpreted as a **deep cascade** of linear operator, which is approximatively verified for the Gabor Wavelets.

Scattering as a CNN

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

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

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



Modulus

$h \geq 0$

Scattering coefficients are only at the output

Scattering as a CNN

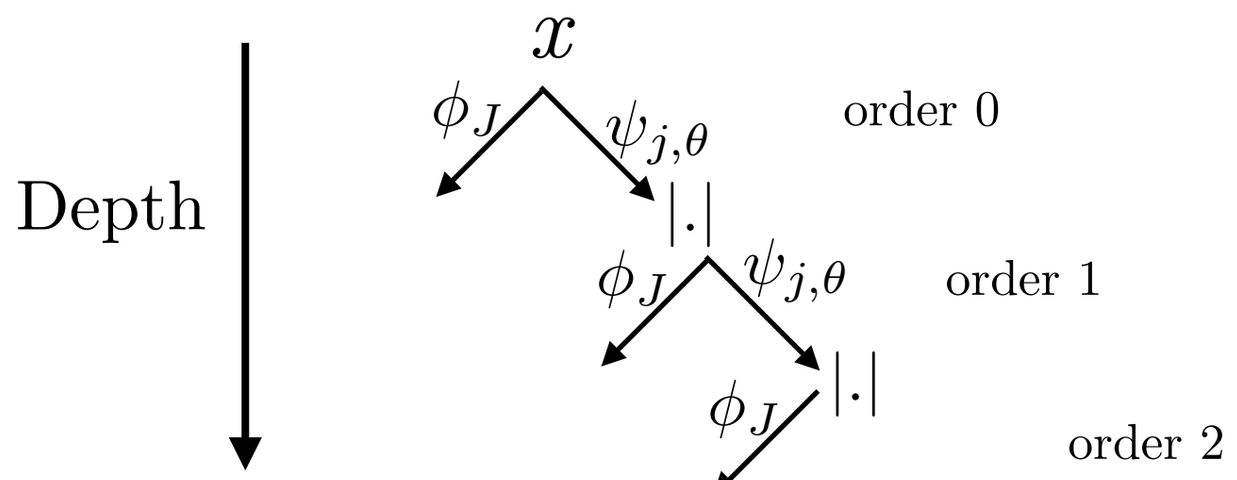
Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat
Recursive Interferometric Representations, S Mallat

Scattering Transform

- Scattering transform at scale J is the cascading of complex WT with modulus non-linearity, followed by a low pass-filtering:

Ref.: Group Invariant Scattering, Mallat S

$$S_J x = \{x \star \phi_J,$$
$$|x \star \psi_{j_1,\theta_1}| \star \phi_J,$$
$$||x \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2}| \star \phi_J\}$$



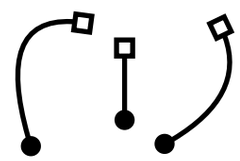
- **Mathematically** well defined for a large class of wavelets.

Properties of a Scattering Transform

- Scattering is stable:
$$\|S_J x - S_J y\| \leq \|x - y\|$$
- Linearize small deformations:
$$\|S_J L_\tau x - S_J x\| \leq C\|\nabla\tau\|\|x\|$$

Deformations

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



- Invariant to local translation:
$$|a| \ll 2^J \Rightarrow S_J L_a x \approx S_J$$

Ref.: Group Invariant Scattering, Mallat S

- For λ, u, $S_J x(u, \lambda)$ is **covariant** with $SO_2(\mathbb{R})$:

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

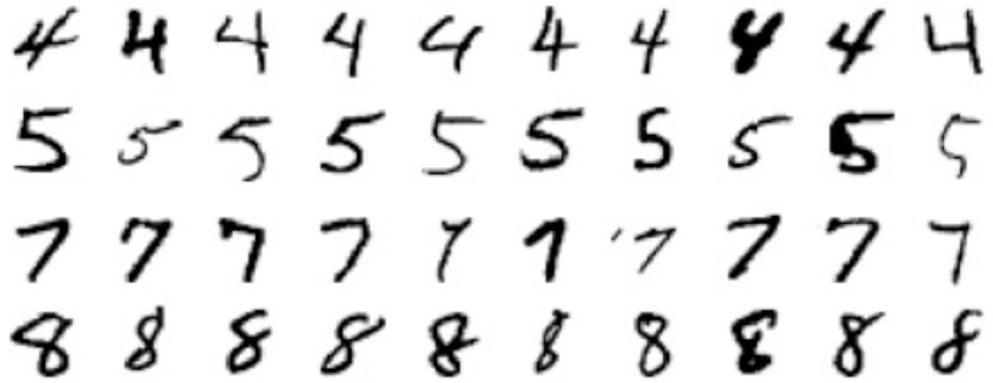
$$S_J(g.x)(u, \lambda) = S_J x(g^{-1}u, g^{-1}\lambda) \triangleq g.S_J x(u, \lambda)$$

A successful representation in image classification

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

- Successfully used in several applications:
  - Digits
  - Textures

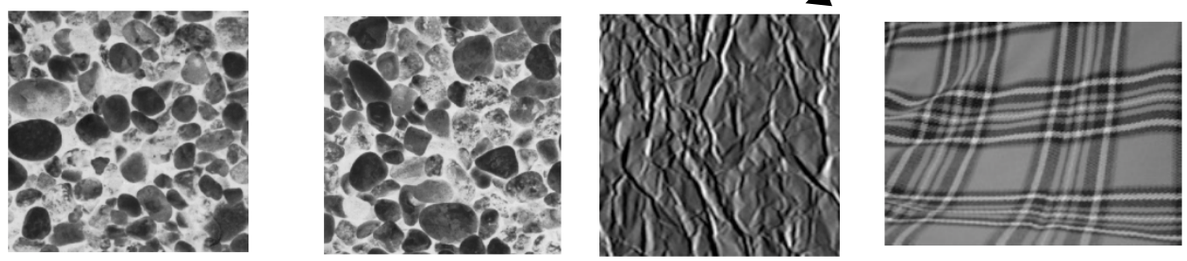
Ref.: Rotation, Scaling and Deformation Invariant Scattering for texture discrimination, Sifre L and Mallat S.



All variabilities are known

Small deformations +Translation

Rotation+Scale



- The design of the scattering transform is guided by the euclidean group
- To which extent can we compete with other architectures on more complex problems (e.g. variabilities are more complex)?

Conclusion

- Theory and practice shouldn't be opposed

- They should be combined together to improve our understanding of a method.

- (e.g., the scattering transform helps to understand invariance even if no learning is involved)