Efficiency of local methods in image classification and energy regression in physics

Louis Thiry, ENS Paris
https://www.di.ens.fr/louis.thiry/

Image classification

- Predict the class of an image in a set S of classes.

$$S = \{\text{"}cat\text{"}, \text{"}dog\text{"}, \text{"}car\text{"}\} = \{(1,0,0), (0,1,0), (0,0,1)\}$$

$F : x$  $\mapsto (1,0,0)$

- Given training samples (x_i, y_i) annotated by humans, find an approximation of the classification function F

Image classification

- MNIST database, 28x28 images,
  50,000 samples, 10 classes

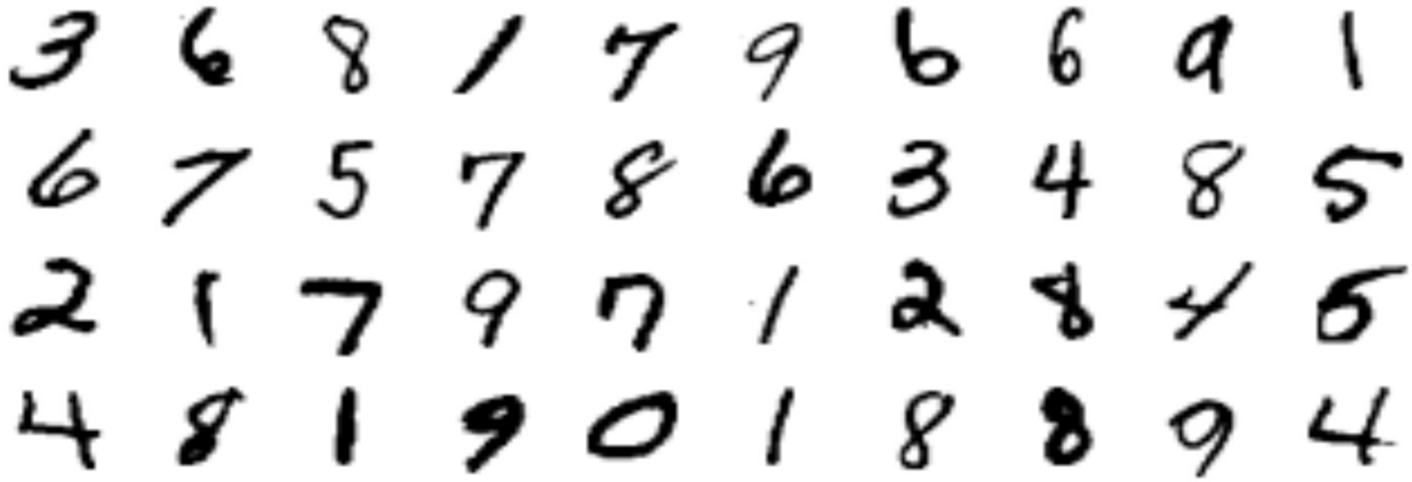


Image classification

- CIFAR-10, 50,000 samples,
  32x32 images, 10 classes

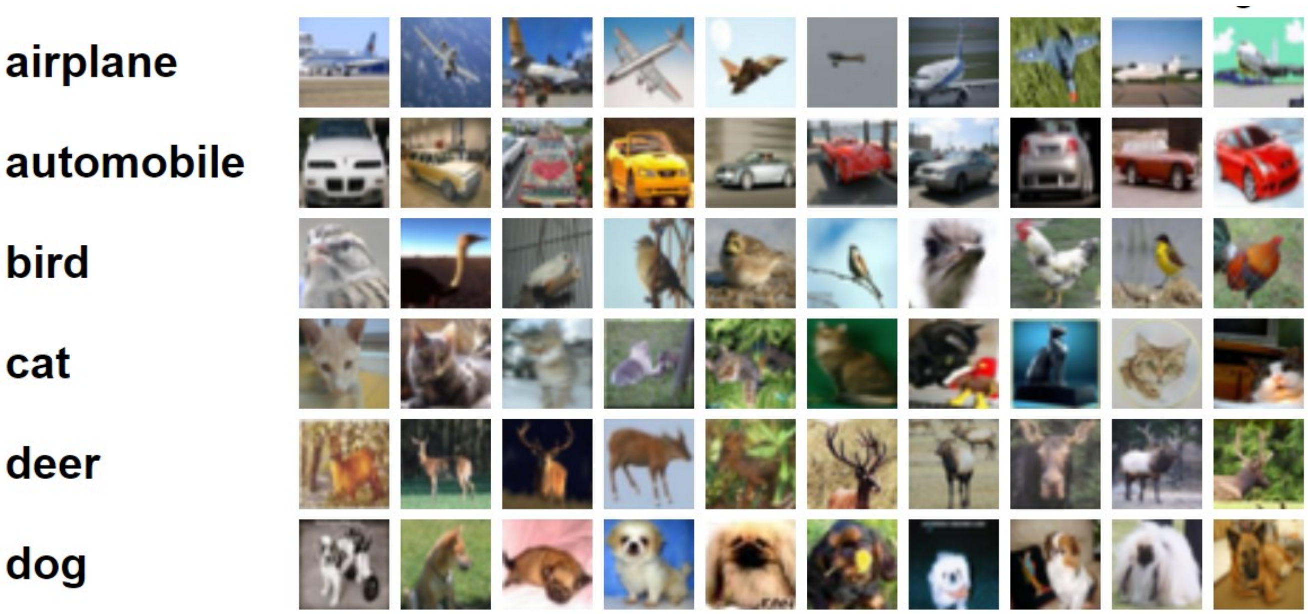


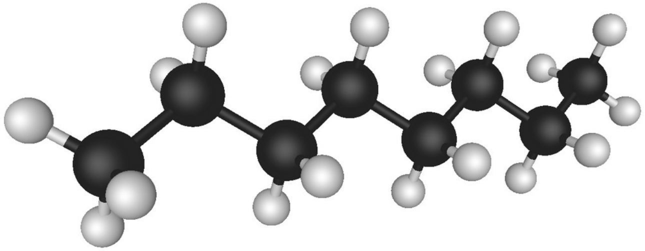
Image classification

- ImageNet, 1,3 million samples, 256x256 images, 1000 classes

Energy regression in physics

- Predict the energy of a set of interacting atoms.

$$F : (r_1, Z_1, \ldots, r_N, Z_N) \mapsto E$$



- The energy can be E_0, the ground state energy of the molecule (Born-Oppenheimer approximation)
- It can be the Free energy $F = E_0 - TS$
- Energy rules stability and chemical properties

Energy regression in physics

- Find an approximation of the energy function given training samples (x_i, y_i)

Samples

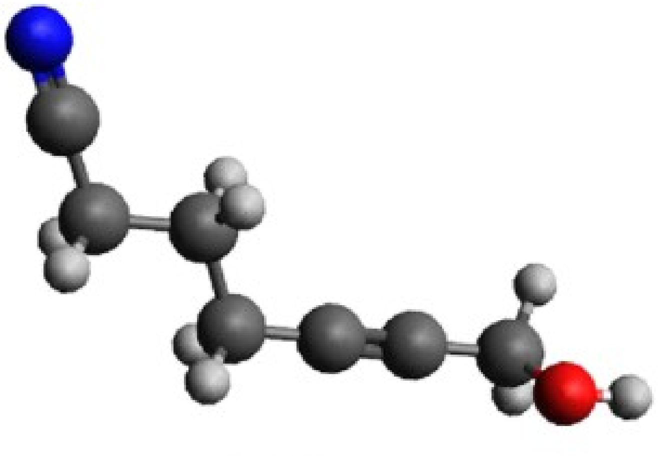
- Experimental data
- Numerical quantum mechanics computations whose cost scales likes N^2 to N^8

Goals

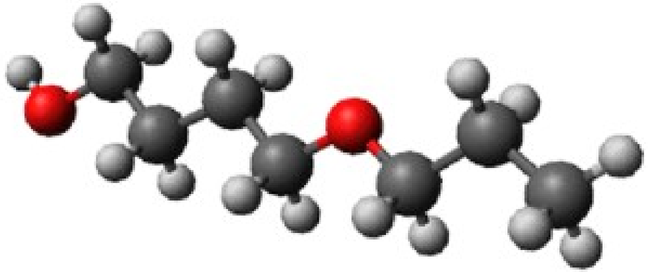
- Speed up computations and tackle large systems data

Energy regression in physics

- QM9 database
  134,000 organic molecules with up to 9 non-Hydrogen atoms
  Ground-state energies computed using DFT



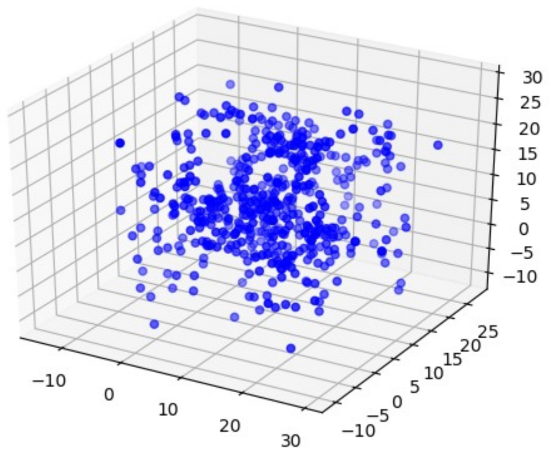
QM9

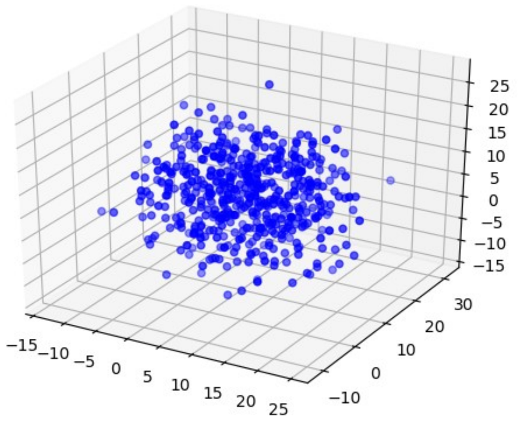


QM9

Energy regression in physics

- Graphene database
  2,500 periodic cells of carbon atoms solids (graphene)
  Ground-state energies computed using MBD





Curse of dimensionality

The input variable x is in high-dimension.

- A 256x256x3 image lies in dimension d = 196,608
- A 512 atoms cell of graphene lies in dimension d = 1536

Under usual Lipschitz regularity assumptions

$$\|F(x) - F(x')\| \leq \|x - x'\|$$

We need ϵ^{-d} samples to have a precision ϵ in the approximation

Deep Convolutional Neural Networks

ImageNet database

- 1,3 M images in dimension ~10^5
- 1000 image classes
- 84 % classification accuracy with ResNet (He et al, 2016)

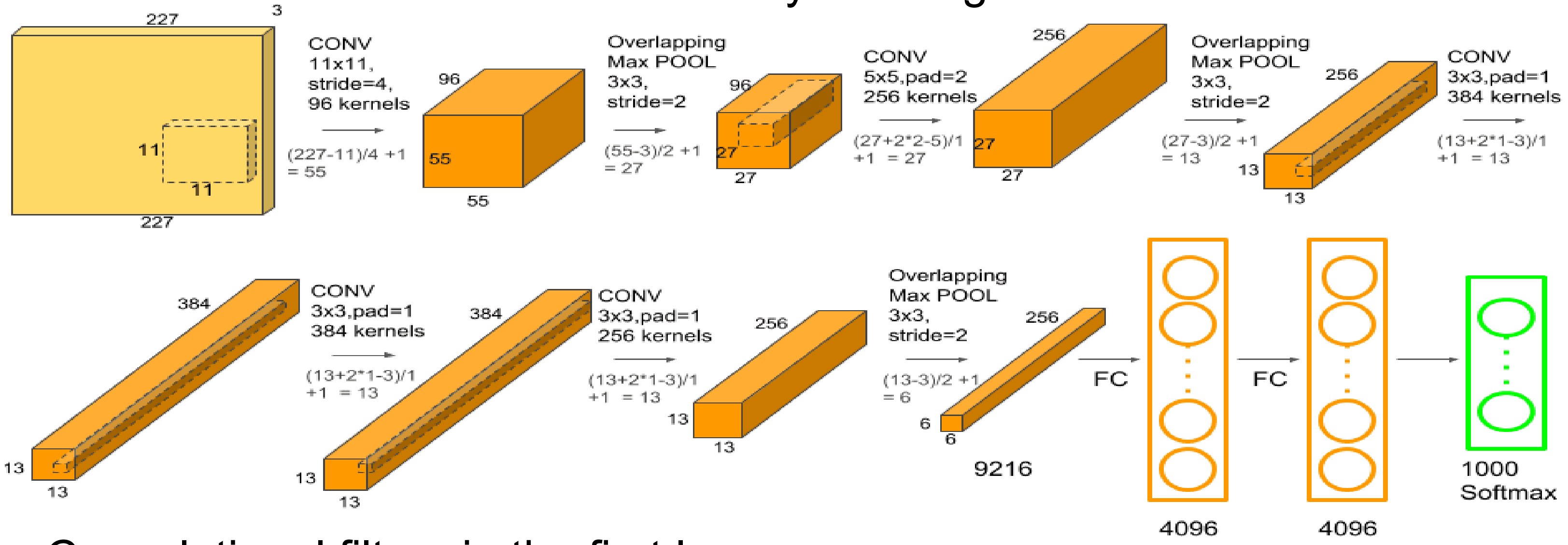
QM9 database

- 134 K molecules in dimension ~10^4
- Energies ranging from -400 to -3000 kcal/mol
- MAE of 0.3 kcal/mol with SchNet (Schutt et al, 2017)

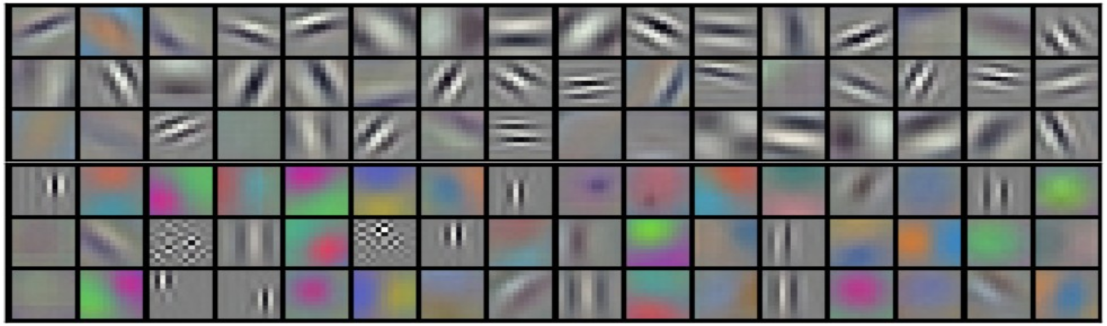
AlexNet

Krizhevsky et al. 2012

59% accuracy on ImageNet



<u>Convolutional filters in the first layer</u>

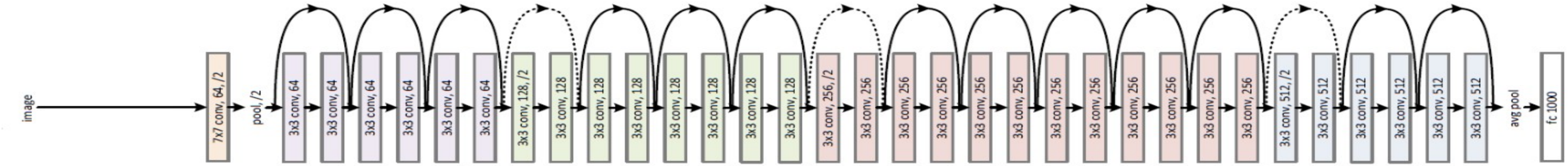


ResNet

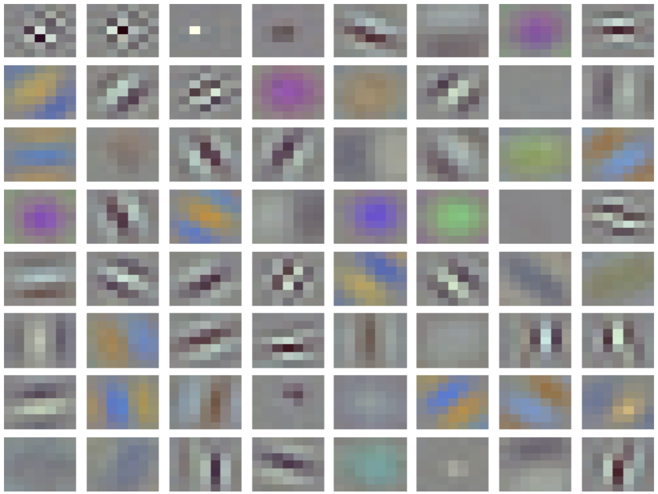
He et al. 2016

80.2 % accuracy on ImageNet

- skip connections
- up to 152 convolutional layers



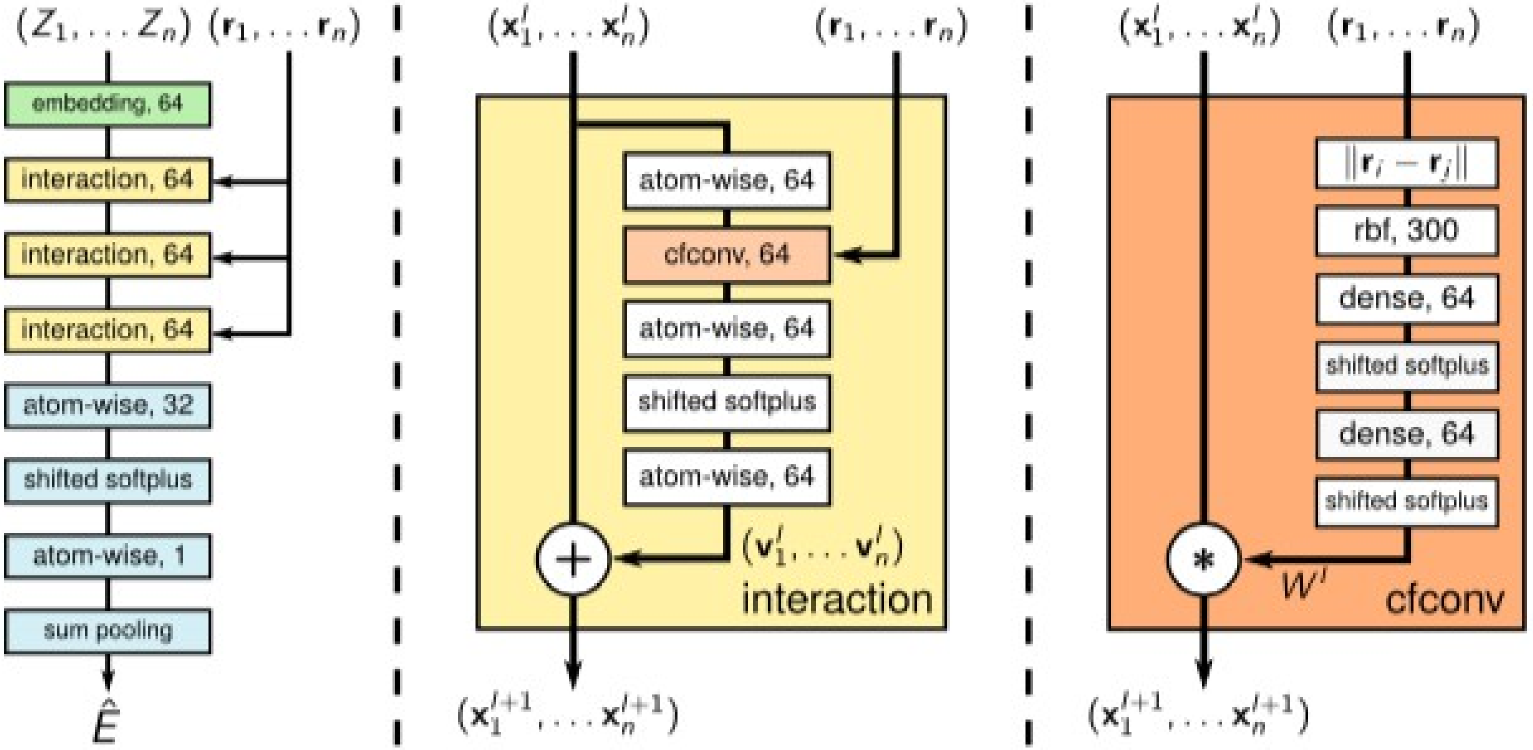
<u>Convolutional filters in the first layer</u>



SchNet

Schutt et al. 2017

MAE 0.3 kcal/mol on QM9



Deep CNNs in image classification and energy regression

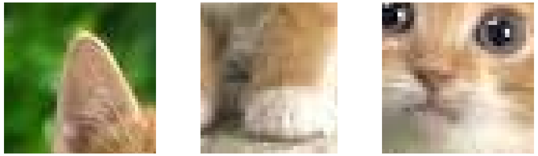
- CNN approximate well energy and classification functions
- The number of data is far below an exponential of the dimension (of the order of the dimension)

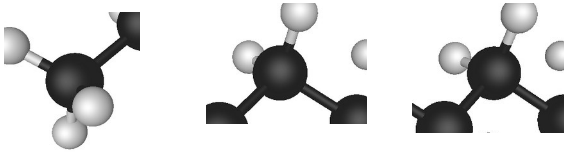
- What are these functions underlying regularity properties?
- Are there similarities between these two problems?

Image classification and energy regression

<u>Similarities</u>

- Local methods based on atomic neighborhoods and patches in images are important methods.





- Invariance properties drive atomic neighborhood or image patches :
  - rotation and translation for atoms
  - scale, lightening, and deformation for image patches.

Image classification and energy regression

Multi-scale problems

- Energy results from different scale interactions:
  - Ionic and covalent bonds at short range,
  - Van-der-Waals interactions at the mesoscale
  - Long-range Coulomb interactions.
- One can classify an image using
  - texture information at a small scale
  - pattern information at a larger scale
  - shape information at the image scale.

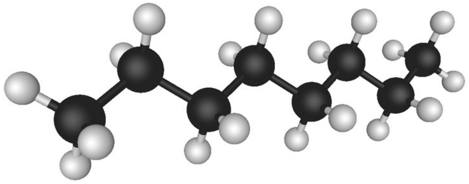


Image classification and energy regression

Differences

- Regression vs. classification.
- Continuous 3D space vs. variable sampling grid (32^2 to 2048^2)
- Absolute distances (Angstrom) vs. variable number of pixels
- Kernel methods are on par with CNNs for energy regression. CNNs far above kernel methods for image classification.

SOAP for energy regression

Bartok et al. 2013

Principle

- Energy is a sum of local energies E_l of the neighborhood x^i

$$E(r_1, Z_1, \ldots, r_N, Z_N) = \sum_{i=1}^{N_a} E_l(x^i)$$

- Local energies are computed with a Kernel Ridge Regression

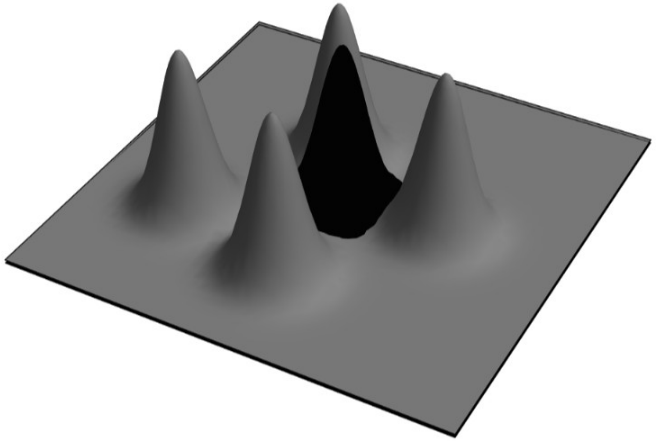
$$E_l(x^i) = \sum_{n=1}^{N_d} \alpha_n k(x^i, x_n)$$

SOAP for energy regression

Bartok et al. 2013

Atomic neighborhood representation

$$x^i = \sum_{j,\ \|r_j - r_i\| < r_c} \exp\left(-\frac{\|r_j - r_i\|^2}{2\sigma^2}\right)$$



The scalar product $\langle x^i, x \rangle$

- is invariant to global translation of the atoms
- is stable to small move of the atomic position

SOAP for energy regression

Bartok et al. 2013

<u>Atomic neighborhood similarity kernel</u>

$$k(x^i, x) = \int_{R \in SO_3} |\langle x^i, R.x \rangle|^p dR$$

This kernel is invariant to rotation of the atoms by construction.

SOAP for energy regression

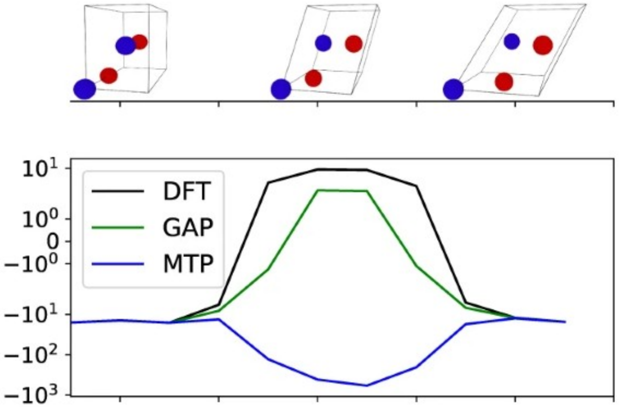
Bartok et al. 2013

QM9 Database

- MAE of 0.4 kcal/mol.
- Optimal neighborhood size is 3 A
- MAE of 0.25 kcal/mol when combining 2 SOAP

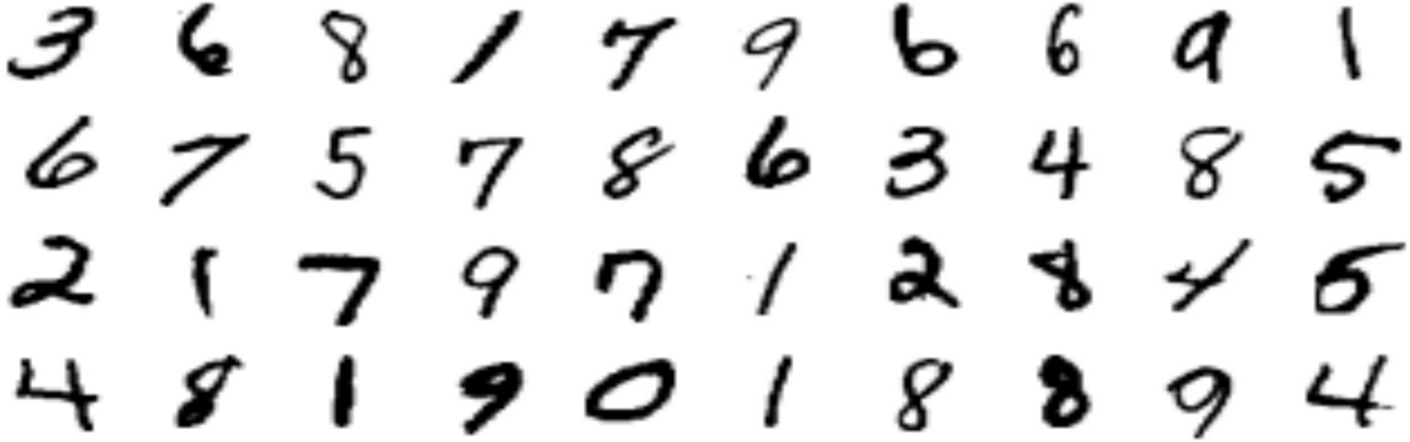
Solids database

- Graphene solids
- Silicon solids
- Ag-Pd alloys

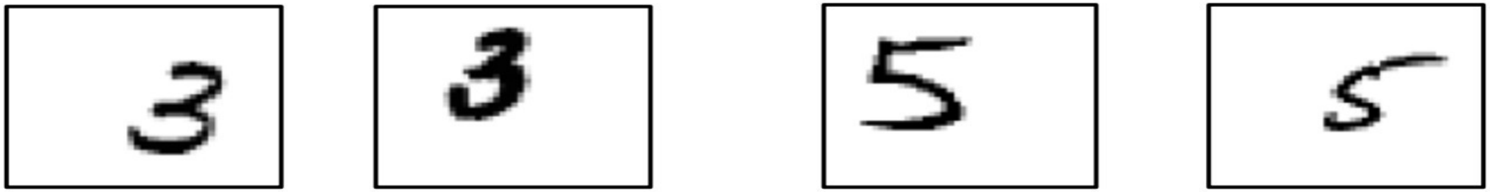


Invariant based digits classification

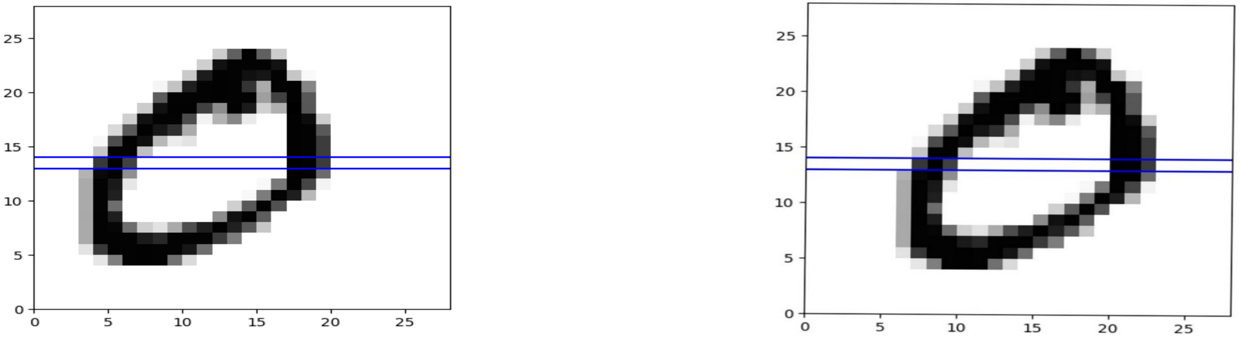
MNIST database

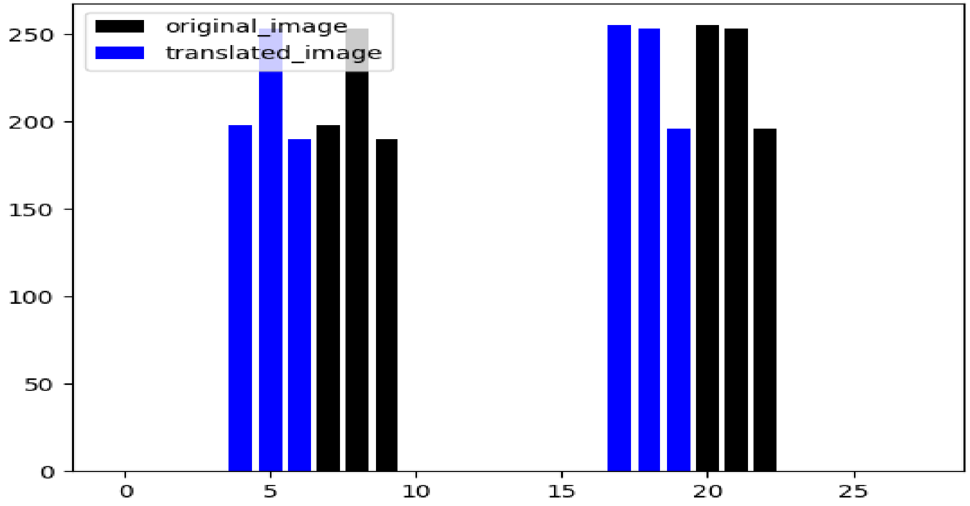


- Invariance to translations, stability to deformations

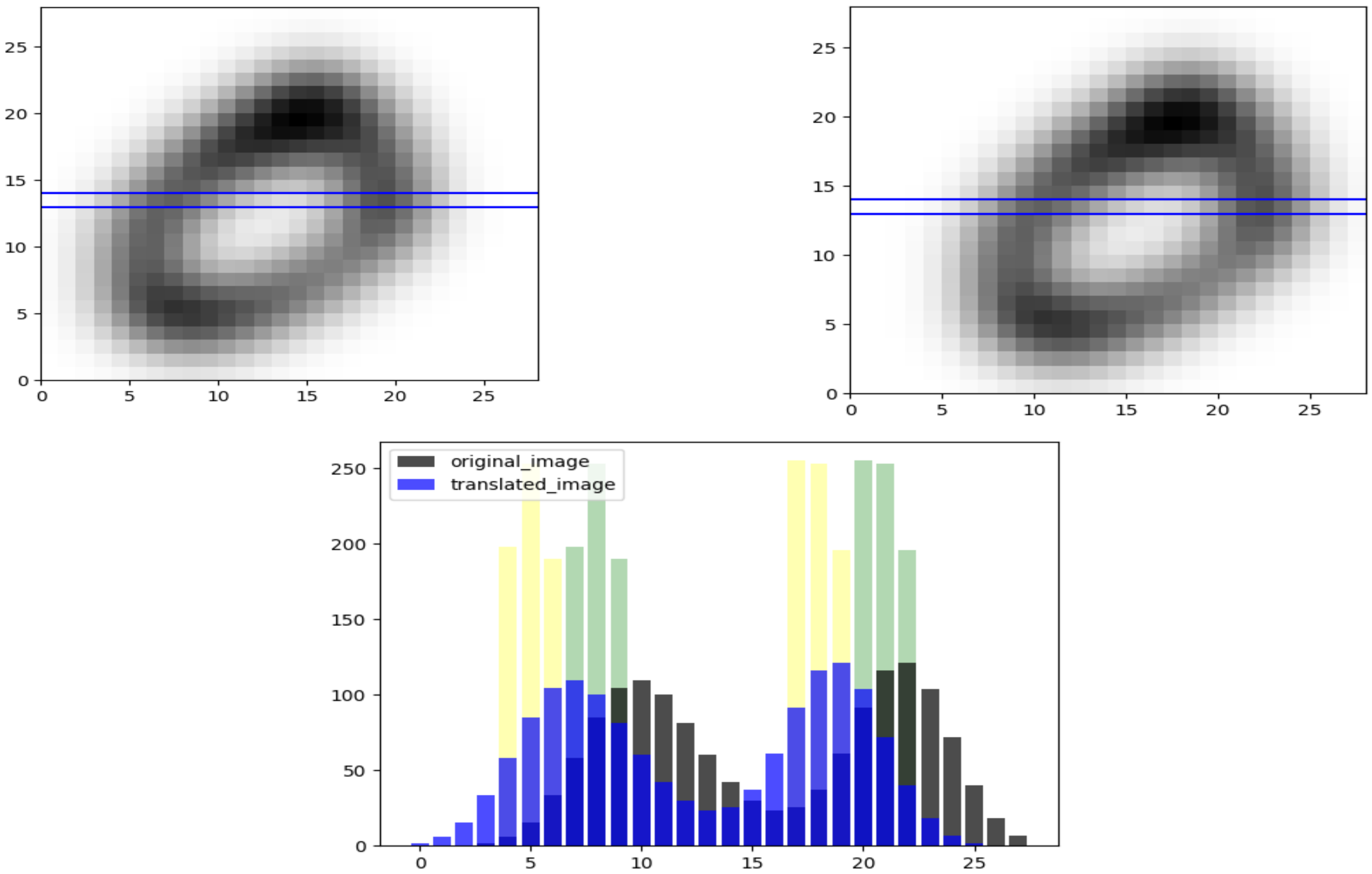


l_2 metric Instability to translations

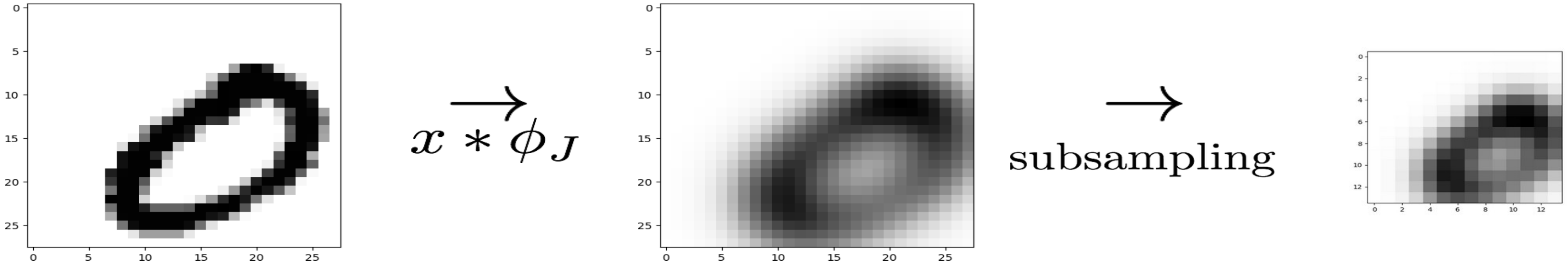




Local averaging



Stability to geometric transformations

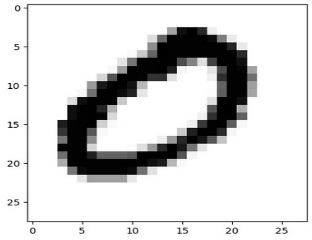


Convolution with Gaussian kernel ϕ_J :

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

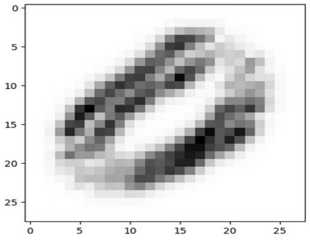
Preserving signal information

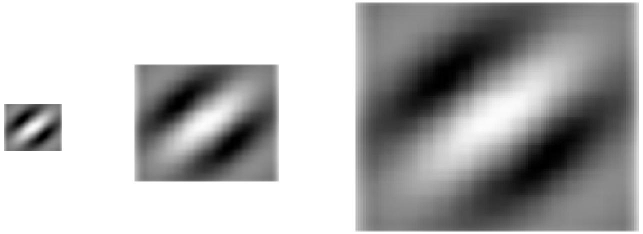
Recover information lost in averaging



$$\longrightarrow$$

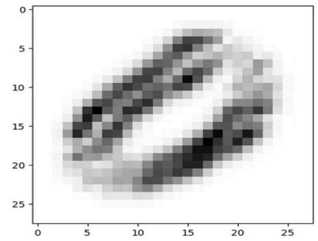
$$|x * \psi_{j,\theta}|$$





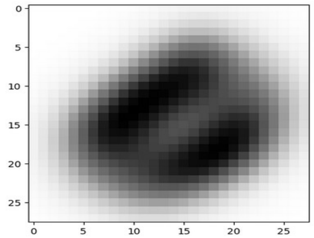
Gabor wavelets $\psi_{j,\theta}$

Stability to geometric transformations

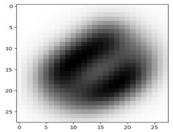


$$\longrightarrow$$

$$|x * \psi_{j,\theta}| * \phi_J$$

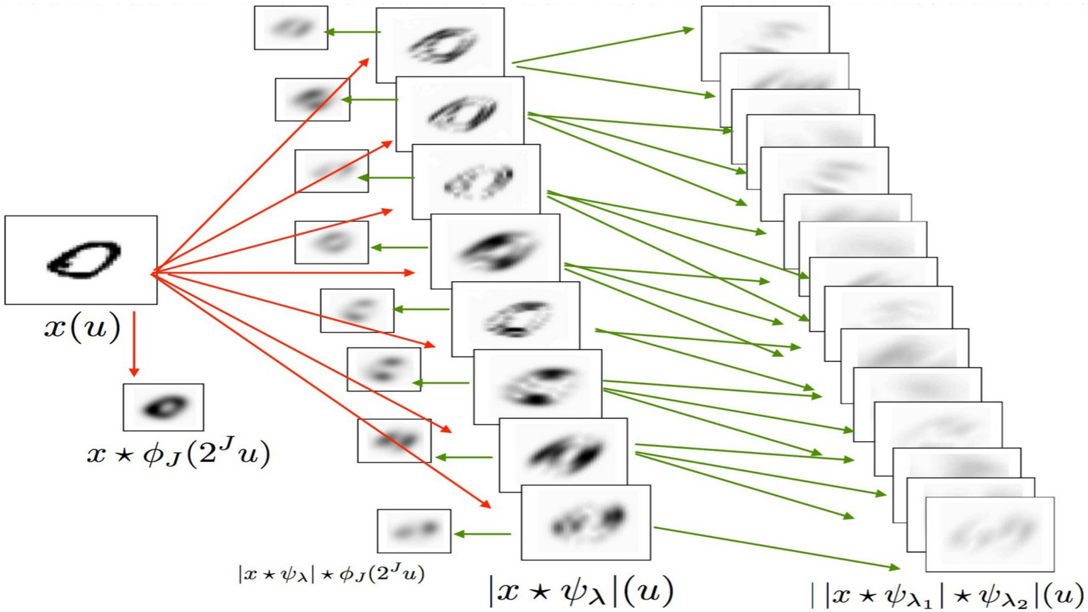


$\longrightarrow$ subsampling



Scattering transform

Mallat (2011), Mallat, Bruna (2012)

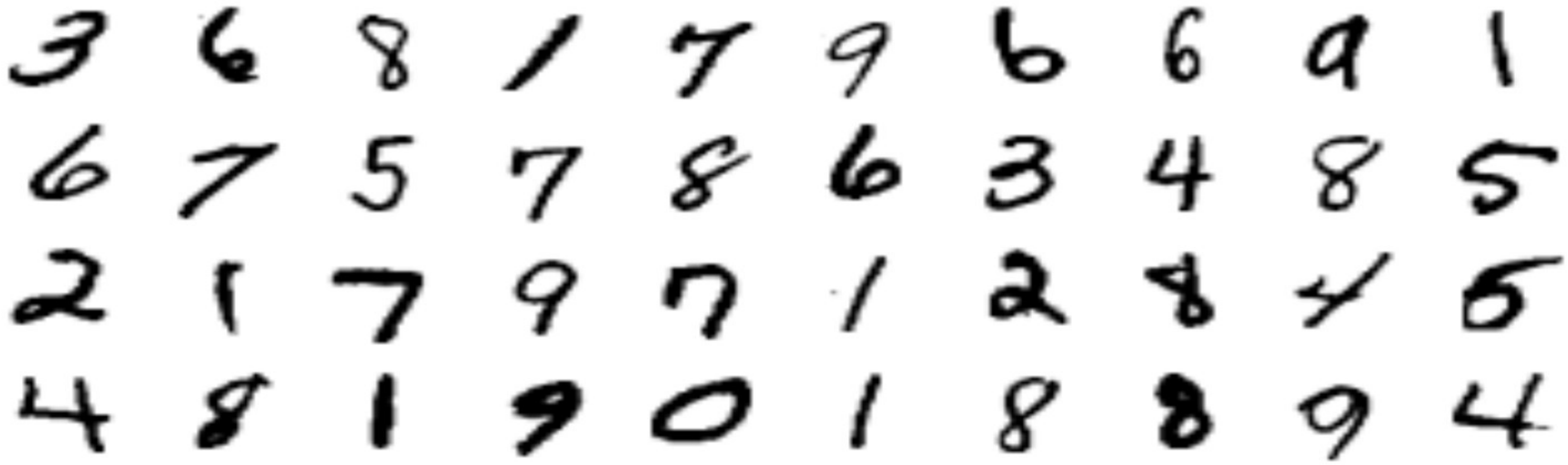


<u>Theorem</u>

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

Scattering vs Deep ConvNets

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



Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
CIFAR-10 32^2 object images 10 classes	82.3 %	89.1 %	95.5 %

Scattering vs Deep ConvNets

Dataset	Scattering Transform	AlexNet	ResNet
ImageNet 224^2 object images 1000 classes	24.3 %	58.7 %	80.2 %

Scattering vs Deep ConvNets

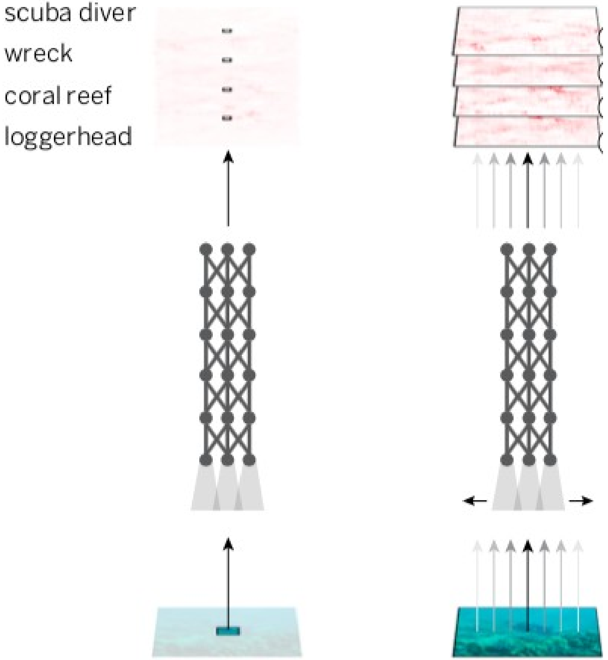
Remarks

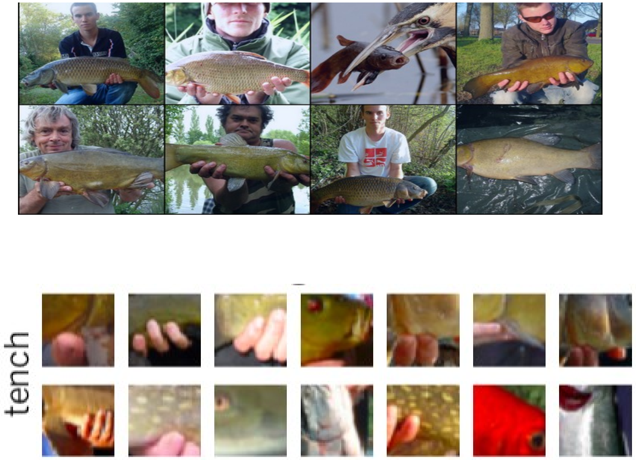
- Invariant representation is competitive for digits
- Large performance gap on ImageNet
- Energy regression:
  - Invariant properties are exact
  - Variabilities are geometry and atomic species
  - Samples are clean
- Image classification:
  - Local invariant properties
  - Huge variabilities (texture, background, noise…)
  - Samples are noisy

BagNets

Brendel et al. 2019

Patch based classification with deep CNN



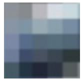


Patch based K-nearest-neighbors classifier

Ours

Dense patch extraction

6^2 patches for 32^2 CIFAR images



Mahalanobis distance

random vector X with covariance $\Sigma = P\Lambda P^T$

$$D_M(x, x') = \sqrt{(x - x')^T \Sigma^{-1} (x - x')}$$

whitening operator w

$$\mathrm{Cov}(w(\mathbf{X})) = I_n$$

$$w : \mathbf{X} \mapsto O\Lambda^{-1/2} P^T (\mathbf{X} - \mu), \quad \forall O \in O_n(\mathbb{R})$$

$$\|w(x) - w(x')\| = D_M(x, x')$$

Patch based K-nearest-neighbors classifier

Ours

Method

- Randomly select a set D of patches
- Regularized whitening operator $W = (\lambda I + \Sigma)^{-1/2}$
- For each image patch $p_{i,x}$ compute set of Mahanalobis distances

$$\mathcal{C}_{i,x} = \{\|Wp_{i,x} - Wd\| \; d \in \mathcal{D}\}$$

- K nearest neighbors encoding

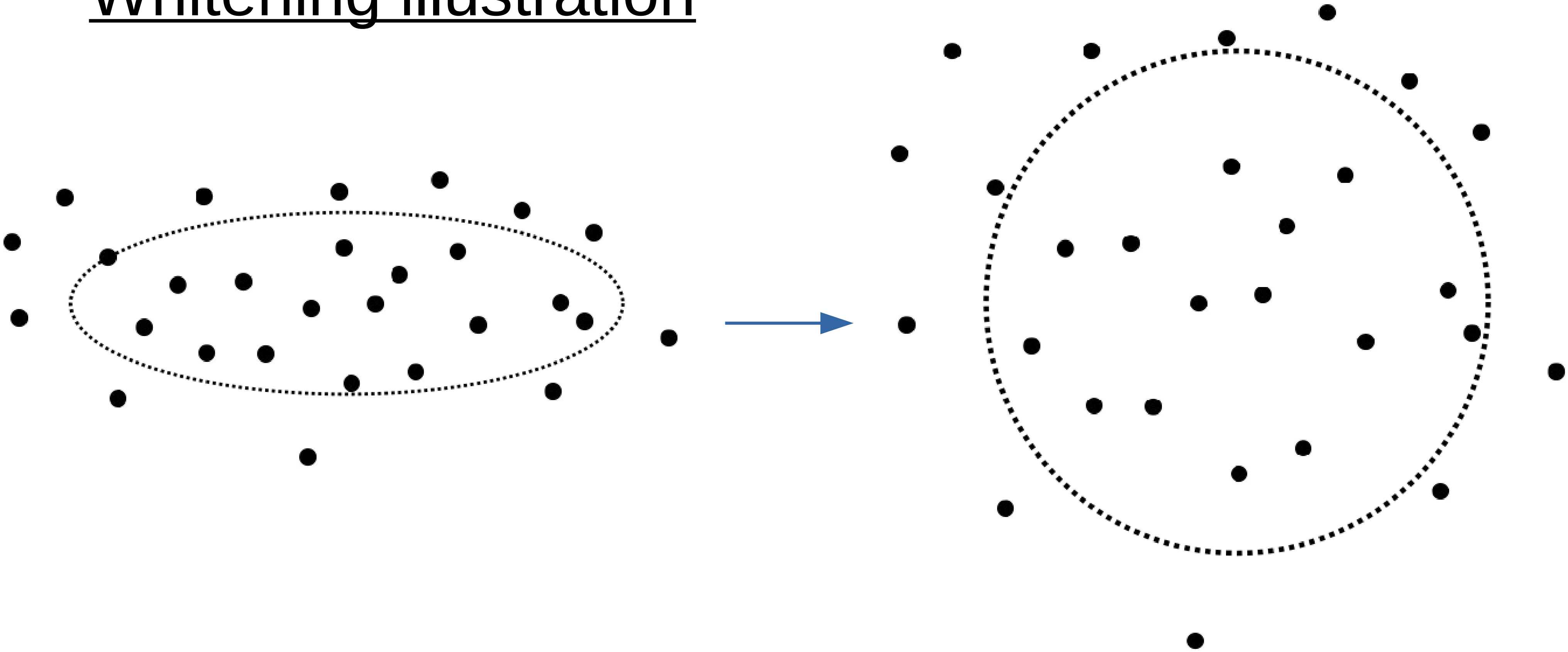
$$\tau_{i,x} \text{ the } K\text{-th smallest element of } \mathcal{C}_{i,x}$$

$$\phi(x)_{d,i} = \begin{cases} 1, & \text{if } \|p_{i,x} - d\| \leq \tau_{i,x} \\ 0, & \text{otherwise.} \end{cases}$$

Patch based K-nearest-neighbors classifier

Ours

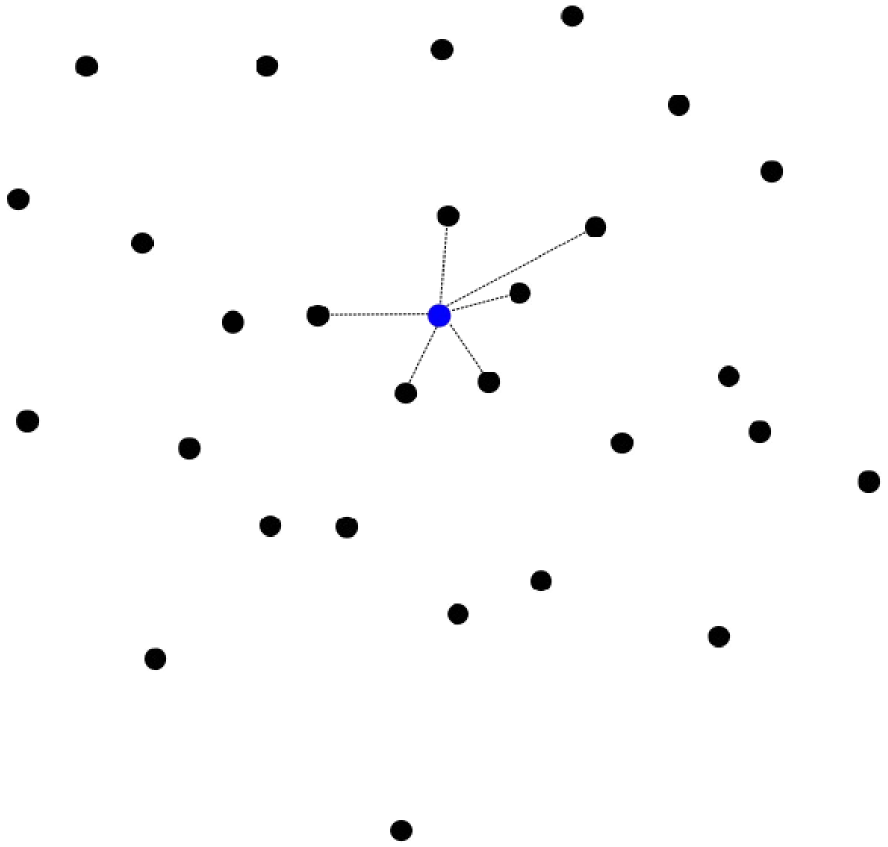
Whitening illustration



Patch based K-nearest-neighbors classifier

Ours

K nearest neighbors



Patch based K-nearest-neighbors classifier

Ours

Classification decision

- Voting system to aggregate patch evidence
- Random patch do not really have a class
- Linear classifier optimized on the training set

$$F(x) = \sum_{p \in x} \sum_{k \in \mathrm{KNN(x)}} w_k$$

Patch based K-nearest-neighbors classifier

Ours

Linear classification on CIFAR-10

Method	$\|\mathcal{D}\|$	VQ	Online	P	Acc.
Coates et al. (2011)	$1 \cdot 10^3$	✓	×	6	68.6
Ba and Caruana (2014)	$4 \cdot 10^3$	×	✓	-	81.6
Wavelets (Oyallon and Mallat, 2015)	-	×	×	8	82.2
Recht et al. (2019)	$2 \cdot 10^5$	×	×	6	85.6
SimplePatch (Ours)	$1 \cdot 10^4$	✓	✓	6	85.6
SimplePatch (Ours)	$6 \cdot 10^4$	✓	✓	6	86.7
SimplePatch (Ours)	$6 \cdot 10^4$	×	✓	6	**86.9**

Patch based K-nearest-neighbors classifier

Ours

Linear classification ImageNet

Method	$\lvert\mathcal{D}\rvert$	VQ	P	Depth	Resolution	Top1	Top5
Random (Arandjelovic et al., 2017)	-	×	-	9	224	18.9	-
Wavelets (Zarka et al., 2019)	-	×	32	2	224	26.1	44.7
SimplePatch (Ours)	2.10^3	✓	6	1	64	33.2	54.3
SimplePatch (Ours)	2.10^3	✓	12	1	128	35.9	57.4
SimplePatch (Ours)	2.10^3	×	12	1	128	36.0	**57.6**

When is nearest neighbor meaningful ?

Beyer et al. (1999)

Dimensionality and nearest-neighbors

- « *Under a broad set of conditions, for as few as 10-15 dimensions, the distance to the nearest datapoint approaches the distance to the farthest datapoin*t »

- « *Scenario where high-dimensional nearest neighbors are meaningful occurs when the underlying dimensionality of the data is much lower than the actual dimensionality* »

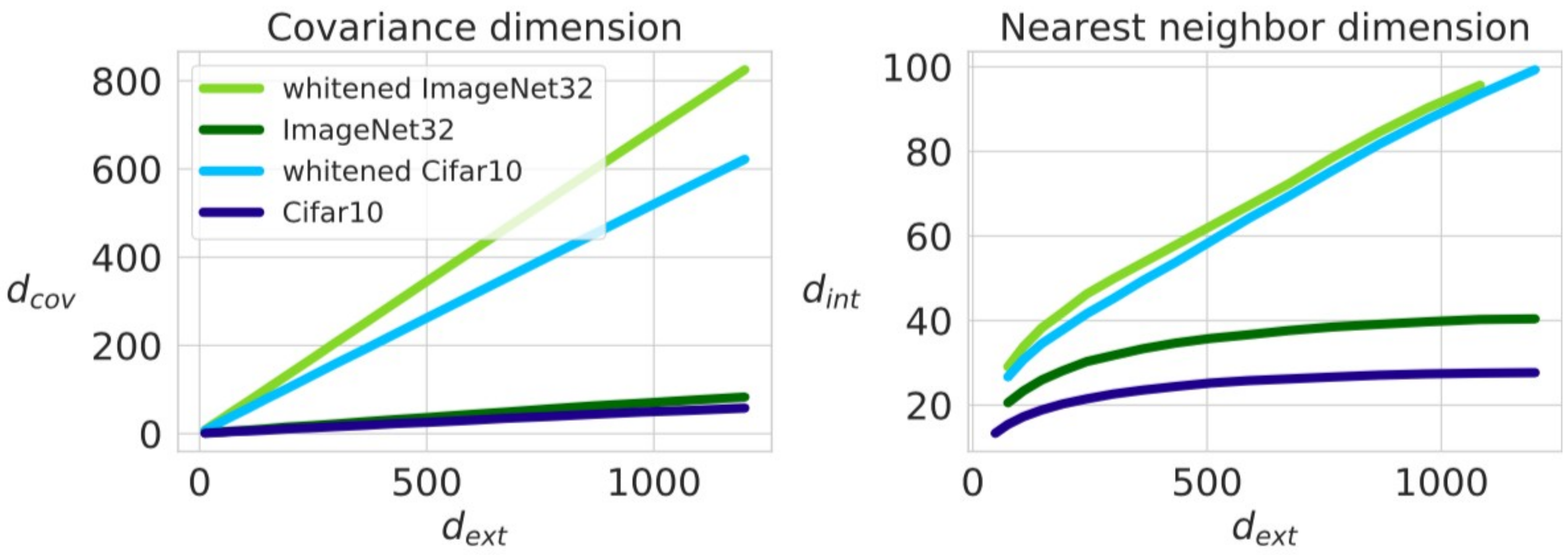
Dimensionality study

Ours

Dimensionality measures

- Covariance dimension : sum of covariance eigenvalues

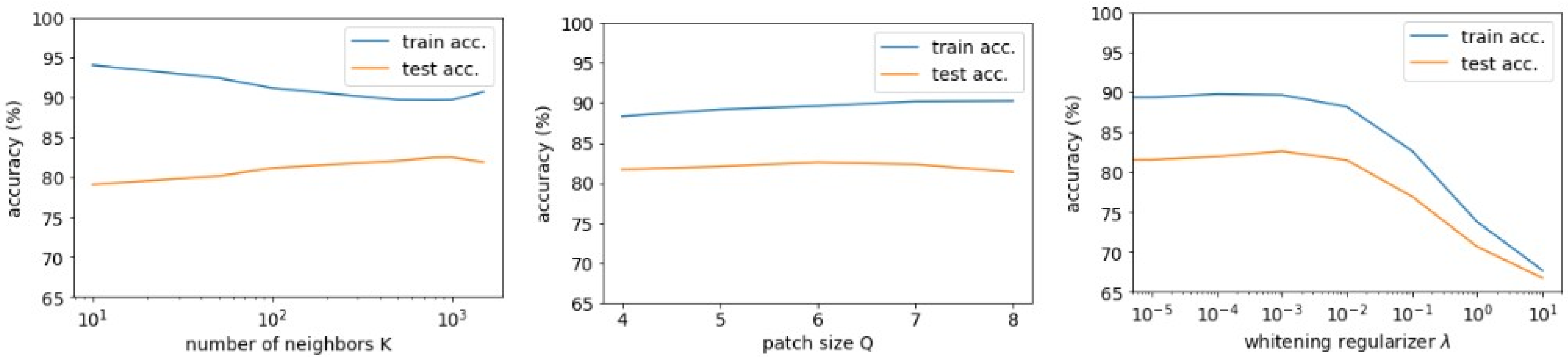
- Nearest-neighbor dimension : $$d_{\text{int}}(p) = \left(\frac{1}{K-1} \sum_{k=1}^{K-1} \log \frac{\tau_K(p)}{\tau_k(p)} \right)^{-1}$$



Patch based K-nearest-neighbors classifier

Ours

Ablation study on CIFAR 10



- Large number of neighbors reduces overfitting
- Patch size does not affect the performance
- Whitening $W = (\lambda I + \Sigma)^{-1/2}$ does not need regularization

Patch based K-nearest-neighbors classifier

Ours

Remarks

- Competitive performance
- Form of low-dimensionality in natural image patches
- Mahanalobis distance is key aspect
- Form of regularity lies in the data
- A large perfomance gap, but using 2K patches for 1,3M images

Questions ?

Paper: https://openreview.net/forum?id=aYuZO9DIdnn
Ph.D. defense in May, https://www.di.ens.fr/louis.thiry/