Efficiency of patch representations for image classification

Louis Thiry,
IRMAR, INRIA, Fluminance Team
Previously: DATA Team, Computer Science Department, ENS, PSL.

Image classification and Energy regression

High-dimensional learning problems.

Set of samples (x_i, y_i).

- Learn a classification function F

$$F : x = \text{  } \mapsto \text{"cat"}$$

- Learn an energy function E (potential)

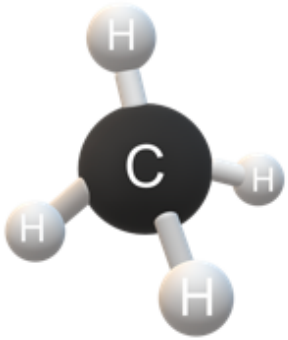
$$E : x = \text{  } \mapsto -417.96 \text{ kcal/mol}$$

Image classification benchmark

- MNIST, 28^2 grayscale images, 10 digits classes.

- CIFAR-10, 32^2 RGB grayscale images, 10 classes.

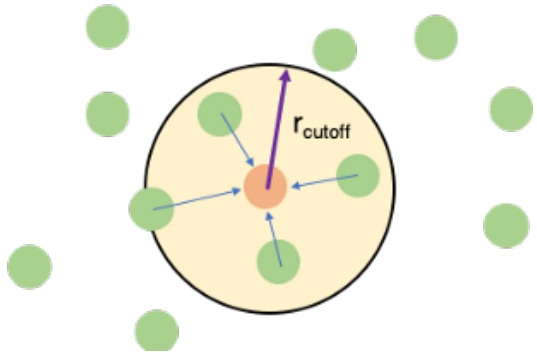
- ImageNet, 256^2 RGB images, 1000 classes (among which 50 dog classes).

Similarities between Image classification and energy regression

- F **invariance properties :**
  - Energy invariant to atoms rotations and translations
  - Image class invariant to scale, lightening and translations
- **Multi-scale aspects**.
  - Physics : small-scale ionic and covalent bonds, medium-scale Van-der-Waals interactions, large-scale Coulomb interactions.
  - Image: small-scale texture information, medium-scale pattern information, large-scale shape information.
- Local methods: Atomic neighborhoods $\Longleftrightarrow$ image patches.

Local methods

- Energy regression: Separation into atomic neighborhoods $\mathcal{N}_i$



Energy sum of local contributions

$$E(x) = \sum_i E\left(\mathcal{N}_i\right)$$

- Image Classification: separation of the image into patches

Sum over patch evidences

$$F(x) = \sum_{p \in x} f(p)$$

Local vs non-local methods

Physics

- Local methods give state-of-the-art results: SOAP (Bartók et al., 2013).
- Simple non-local terms can be added when necessary.

Image classification (ImageNet)

- Before deep-learning era: 70 % top5 accuracy local methods (SIFT + Fisher Vectors, Sanchez et al. 2013)
- AlexNet, 2012: 85 % top 5 with **non-local** Convolutional Neural Networks.
- ResNet, 2016: 96 % top 5 with **non-local** very deep Convolutional Neural Networks.

Hypothesis: Importance of CNN's multi-scale/hierarchical structure.

Plan

1. Hierarchical hypothesis for image classification
2. Local/patch methods in Image classification
3. A simple patch-based classifier

Plan

1. Hierarchical hypothesis for image classification
2. Local/patch methods in Image classification
3. A simple patch-based classifier

CNNs hierarchical structure

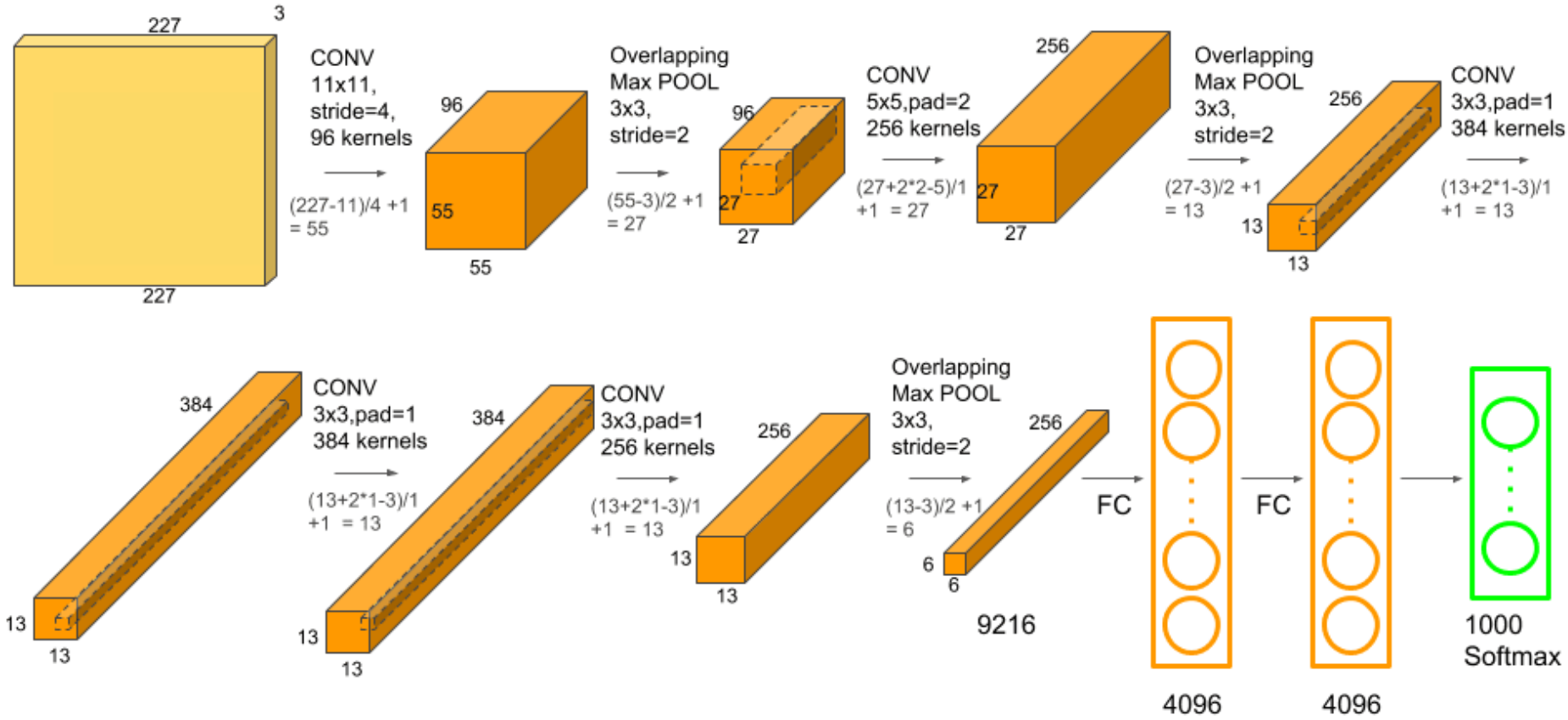


Figure: Architecture AlexNet (Krizhevsky et al., 2012). The receptive field is equal to the whole image.

CNNs receptive field

Defintion: Receptive field at layer ℓ. Patch's size encoded by 1 pixel in the convolution layer ℓ.

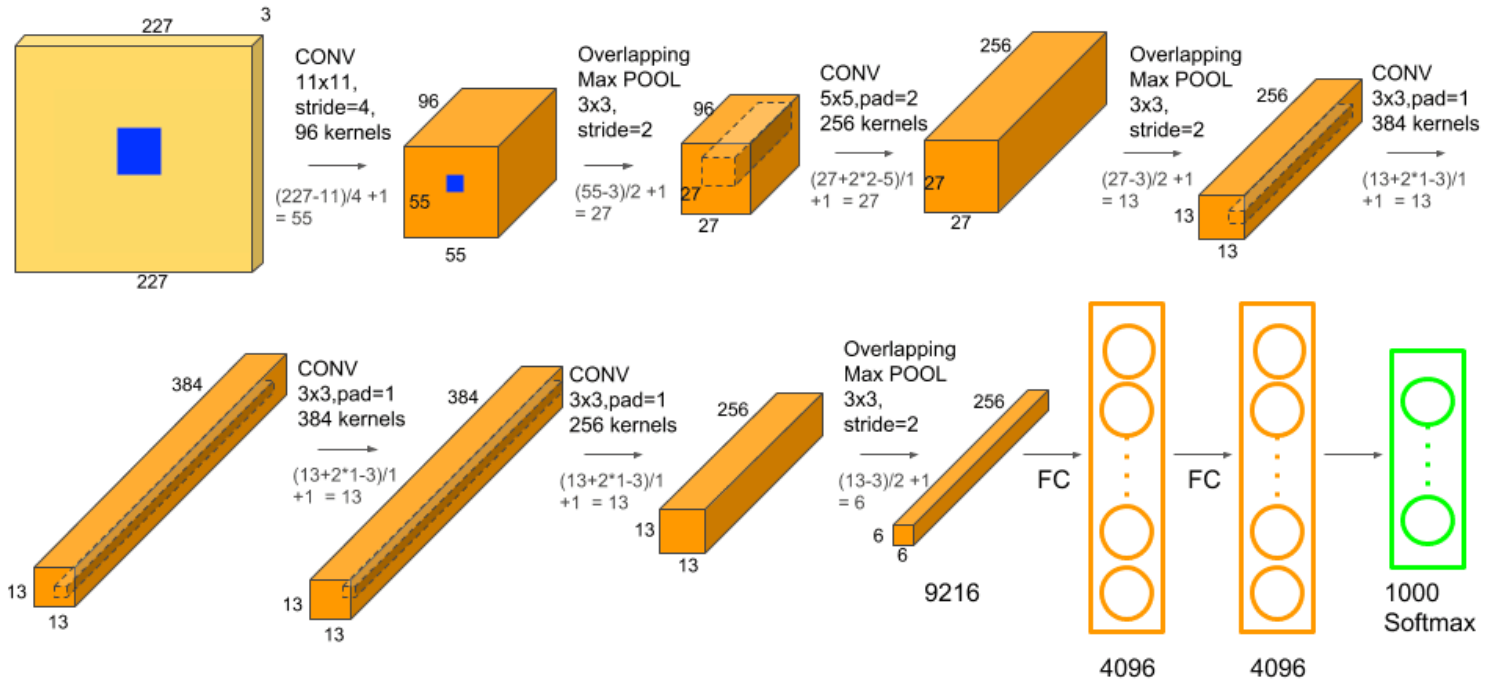


Figure: Architecture AlexNet (Krizhevsky et al., 2012). The receptive field is equal to the whole image.

CNNs receptive field

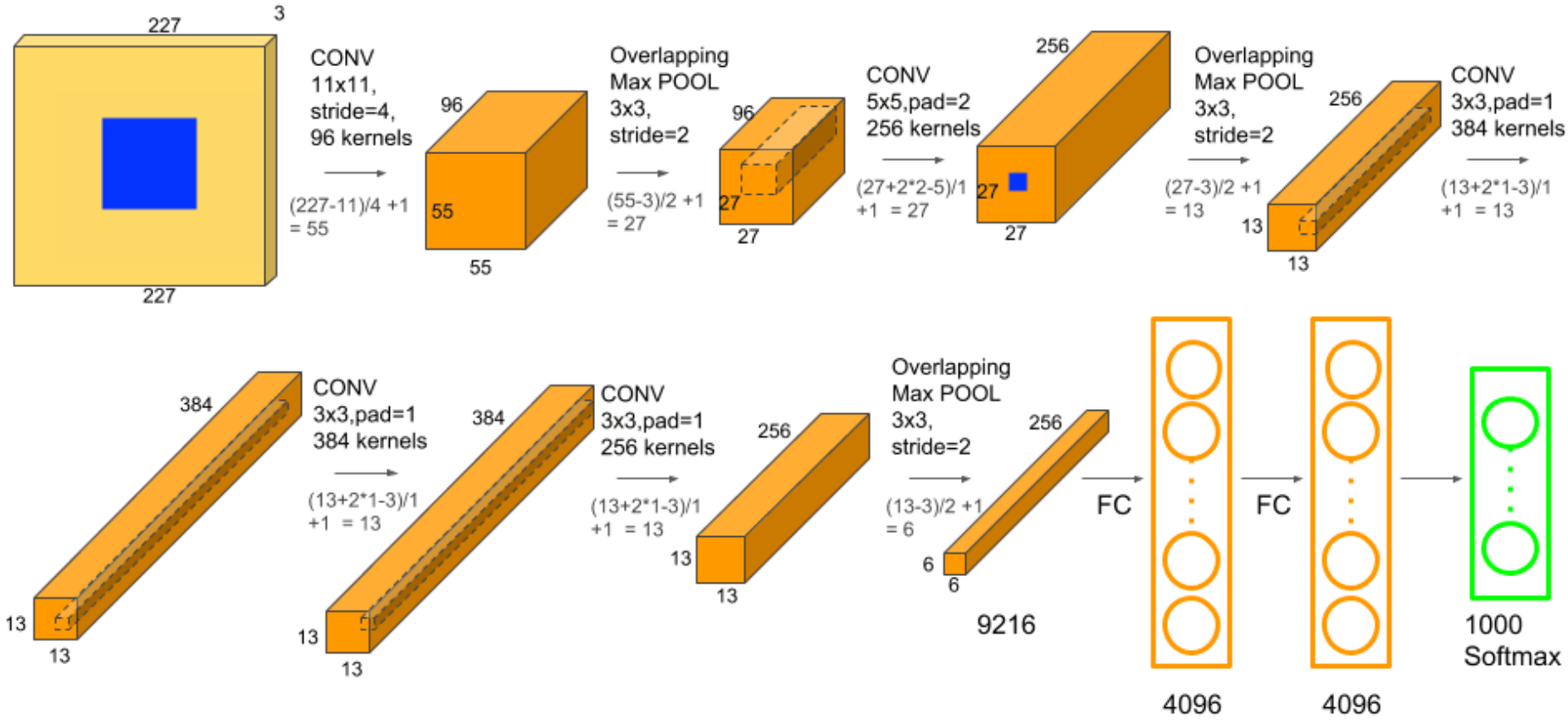


Figure: Architecture AlexNet (Krizhevsky et al., 2012). The receptive field is equal to the whole image.

CNNs receptive field

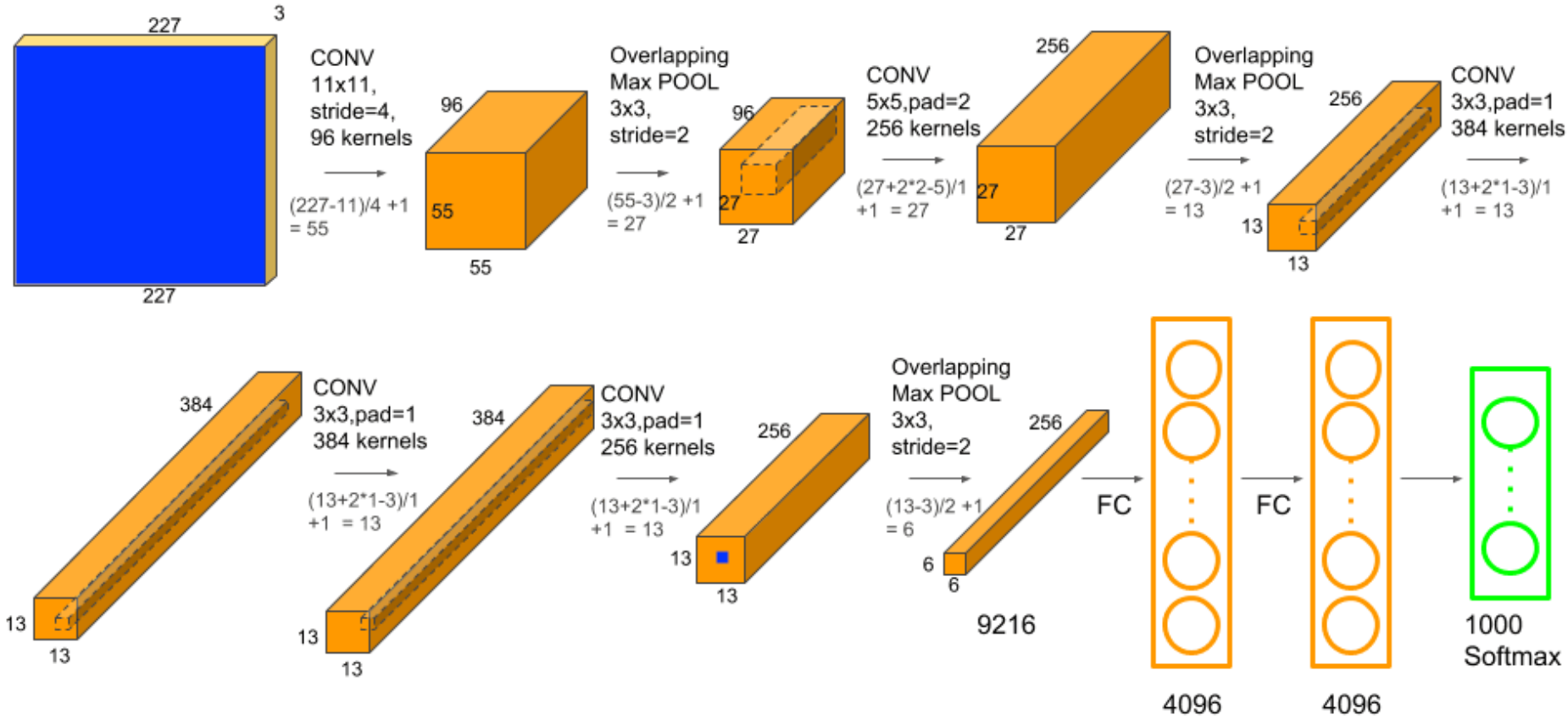


Figure: Architecture AlexNet (Krizhevsky et al., 2012). The receptive field is equal to the whole image.

CNN Visualization techniques

Support the multi-scale/hierarchical hypothesis.

- Visualizing and Understanding Convolutional Networks, Zeiler and Fergus, 2014
- Deep visualization Toolbox, Yosinski et al., 2015

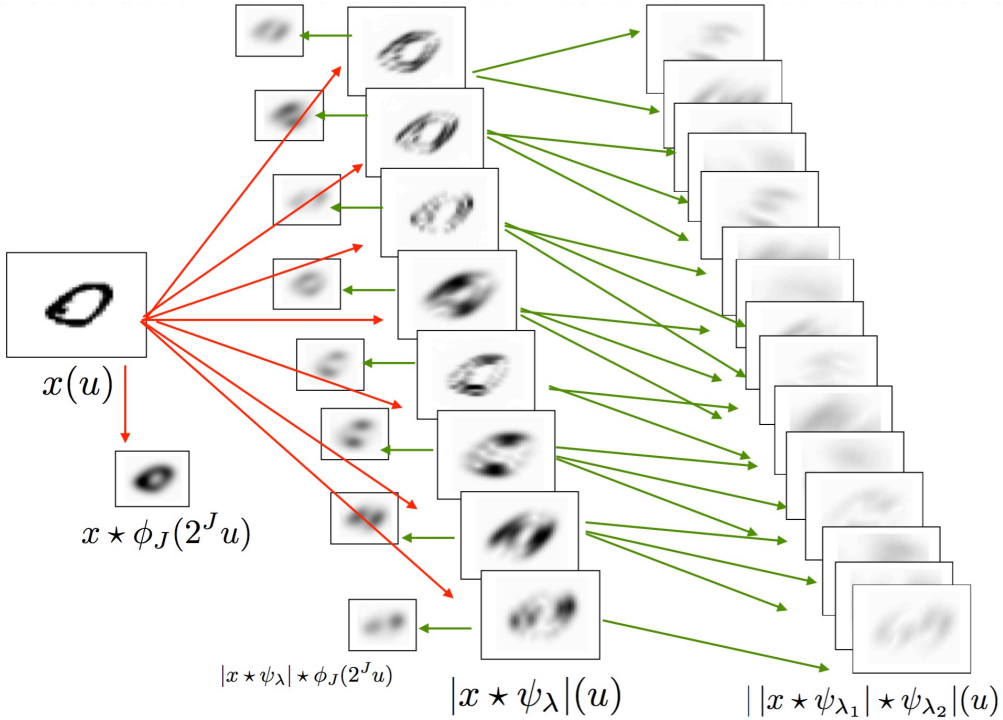
Multi-scale/hierarchical hypothesis

Deep Learning, Lecun, Bengio, Hinton 2015:

- *" The first layer [of the convolutional network] represents the presence or absence of edges in the image. "*
- *" The second layer typically detects motifs"*
- *" The third layer may assemble motifs into [...] parts of familiar objects."*
- *" Subsequent layers would detect objects as combinations of these parts."*

$\rightarrow$ **Importance of the multi-scale/hierarchical hypothesis.**

Scattering Transform: a simple hierarchical model



Scattering transform (Mallat, 2012; Bruna and Mallat, 2013)

- $Sx(u)$: multi-scale image descriptor.
- Stable to small geometric deformations.
- $>$99.5 % accuracy for handwritten digits recognition.

Scattering Transform: a simple hierarchical model

CIFAR-10

- Large variability: pose, texture, background...
- Scattering Transform: 82 % accuracy.

ImageNet

- Huge variability
- Scattering Transform: 42 % accuracy.

Plan

1 Hierarchical hypothesis for image classification

2 Local/patch methods in Image classification

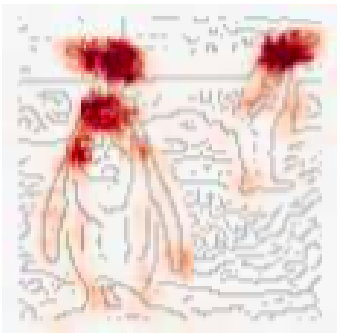
3 A simple patch-based classifier

Evidences against hierarchical hypothesis

- CNN are biased towards textures, Geihros et al 2018.
- *BagNet*, Brendel Wieland, 2019:
  - Local method with f convolutional neural network

$$F(x) = \sum_{p \in x} f(p)$$

  - Accuracy: 88 % on ImageNet.
  - Explainability of the classification decision: patch evidence

Patch-based deep neural networks

- Visual Transformes, Dosovitskiy et al, 2021
- ResMLP, Touvron et al, 2021
- ConvMixer, Trockman et Kalter, 2021
- Adapative Fourier Neural Operators, Guibas et al., 2021, with an application to weather forecast.

$\rightarrow$ patches are back in the game !

Plan

1. Hierarchical hypothesis for image classification
2. Local/patch methods in Image classification
3. A simple patch-based classifier

Patch K-nearest-neighbors

Motivations

- Patches are good input representation for classification.
- 16×16 patch, $D = 768 \rightarrow$ **they still live high dimension**.

"*As dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point. Empirical results on both real and synthetic data sets demonstrate that this effect can occur for as few as 10–15 dimensions.*", Beyer et al. 1999.

- Are there low-dimensional properties of natural image patches in spite of this seeming high dimension?
- What is the performance of a patch K-nearest-neighbor-based classifier?

The unreasonable effectiveness of patches in Convolutional Kernel Methods, (Thiry et al., 2021).

Naive K-nearest-neighbors

1. Image Level KNN: 58 % accuracy on CIFAR-10 with mahanalobis distance.

2. Voting system at the Patch Level:

$$F(x) = \sum_{p \in x} \sum_{n \in \mathrm{KNN}(p)} 1_{\mathrm{class(n)}}$$

- Performs poorly: $\sim$ 30% with CIFAR-10 subset.
- Heavy nearest-neighbor search (millions of patches)
- Does not ignore non-informative patches

Informative patches in BagNet (Brendel and Bethge, 2019)

Our Patch K-nearest-neighbors

Goals:

- As close as possible of nearest neighbor classifier
- Reduce the nearest-neighbor search computational cost
- Filter non-informative patches

$\rightarrow$ Learn the class evidence w_n of the patches:

$$F(x) = \sum_{p \in x} \left(\sum_{n \in \mathrm{KNN}(p)} w_n \right)$$

Our Patch K-nearest-neighbors

Algorithm

- Select N patches of size P^2 randomly in the training set
- Mahanalobis Euclidean distance: patches whitening operation
- Patches nearest-neighbors one-hot encoding spatial map

$$\Phi(x) = \left(1_{\mathrm{KNN}(p[i,j])}\right)_{i,j}$$

- linear regression

$$F(x) = \langle W, \Phi(x) \rangle = \sum_{p[i,j] \in x} \left(\sum_{n \in \mathrm{KNN}(p[i,j])} w_n^{i,j} \right)$$

Classification results

CIFAR-10 linear classification

Method	N patches	P	Acc.
Scattering (Oyallon et al. 2015)	-	8	82.2
SimplePatch ℓ^2 (Ours)	$10k$	6	65.4
SimplePatch Mahanalobis (Ours)	$10k$	6	85.6
SimplePatch Mahanalobis (Ours)	$60k$	6	**86.9**

- Mahanalobis distance is key aspect
- Surprisingly good accuracy

ImageNet linear classification

Method	N patches	P	Res.	Top5
Scattering	-	16	224	42.3
Ours	$2k$	12	128	**57.6**

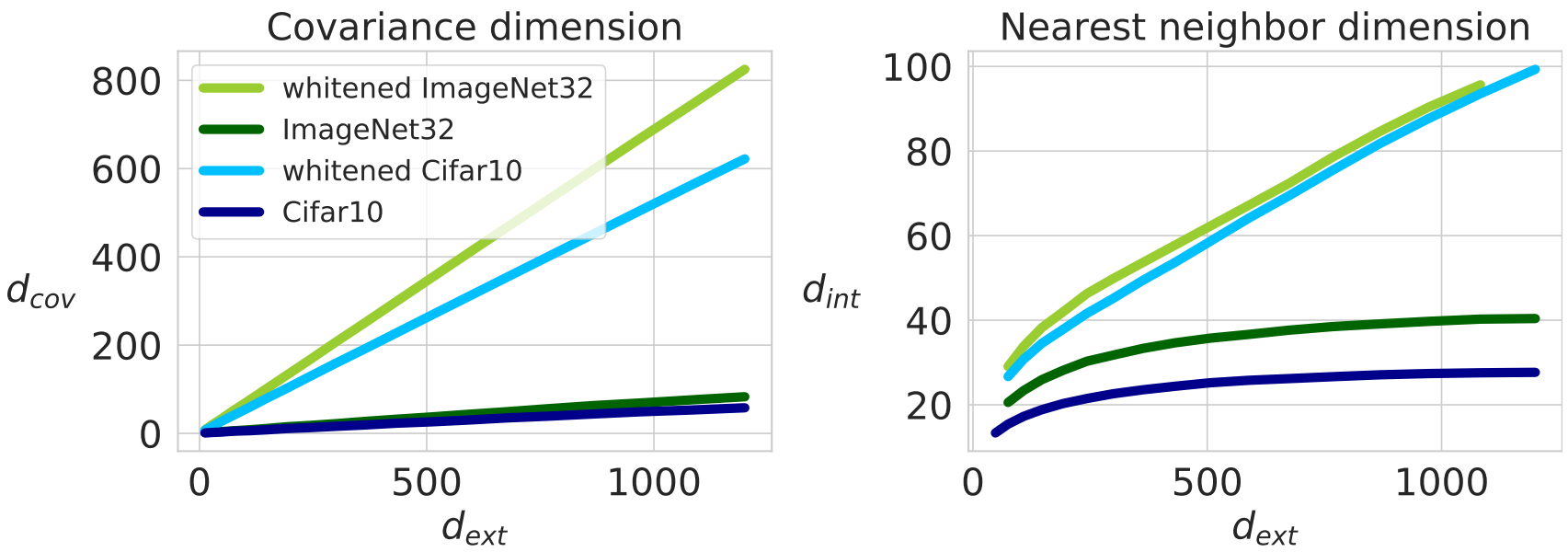
Classification results

CIFAR-10 Convolutional Kernel Classification		
Method	**Classifier**	**Acc.**
SimplePatch (Ours)	linear	86.9
SimplePatch (Ours)	1-hidden-layer	88.5
NKWT (Li et al. 2019)	kernel	89.1
NK (Shankar et al. 2020)	kernel	89.8
CKN (Mairal et al. 2016)	kernel	89.8

- Competitive accuracy with convolutional kernel methods
- Possible line of explanation of their success

Low-dimensionality anaylsis

"Scenario where high-dimensional nearest neighbors are meaningful occurs when the underlying dimensionality of the data is much lower than the actual dimensionality.", (Beyer et al., 1999).



Dimensionality measures w.r.t. patch extrinsic dimensionality d_{ext}

Our Patch K-nearest-neighbors

- Nearest neighbors classifier works much better at patch level.
- State-of-the-art performance as **non-learned** (i.e. non-optimized) representation
- Competitive Convolutional Kernel method
- Random patch subsets: tiny fraction of the training set:
  - $60,000$ out of 35 millions CIFAR-10 patches
  - $2,000$ out of 10 billions ImageNet patches
- Patches low-dimensional properties
- Line of explanation for the use of patches in deep networks ?

Questions ?

References I

Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, August 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

References II

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Louis Thiry, Michael Arbel, Eugene Belilovsky, and Edouard Oyallon. The unreasonable effectiveness of patches in deep convolutional kernels methods. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=aYuZ09DIdnn`.