Sparsifying Convolutional Layers with Dual-Tree Wavelet Packets

Hubert Leterme   Kévin Polisano   Valérie Perrier   Karteek Alahari

Laboratoire Jean Kuntzmann

Université Grenoble Alpes

ORASIS 2021

Contents

Contents

CNNs vs discrete wavelet transforms

Convolutional neural networks (CNNs)[1]:

- ✓ **state-of-the-art performances in many domains – image classification, object detection, speech recognition...**
- ✗ **very resource-intensive;**
- ✗ **empirical approach; lack of mathematical understanding.**

[1] **LeCun2015**

CNNs vs discrete wavelet transforms

Convolutional neural networks (CNNs)[1]:

- ✓ **state-of-the-art performances in many domains – image classification, object detection, speech recognition...**
- ✗ **very resource-intensive;**
- ✗ **empirical approach; lack of mathematical understanding.**

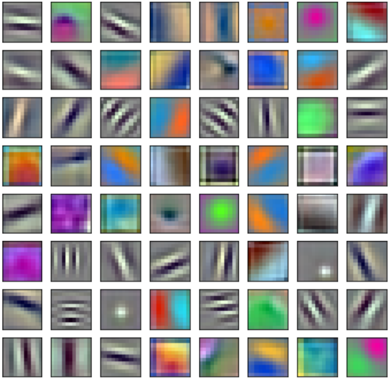
Discrete wavelet transforms[2]:

- ✓ **built on well-established mathematical framework;**
- ✓ **very efficient in tasks such as signal compression and denoising;**
- ✗ **not widely used for image classification.**

Oscillating patterns very often observed in CNN kernels[3].

[1] **LeCun2015**
[2] **Mallat2009**
[3] **Yosinski2014**

CNNs vs discrete wavelet transforms



AlexNet[4] filters (first layer) after training with ImageNet

[4] **Krizhevsky2012**

Objectives

Main objective:

- ✓ perform a **theoretical study** of CNN properties for image classification.

Objectives

Main objective:

- ✓ perform a **theoretical study** of CNN properties for image classification.

What this work is NOT about (at least not as primary objective):

- ✗ increase performance of CNNs;
- ✗ decrease training complexity.

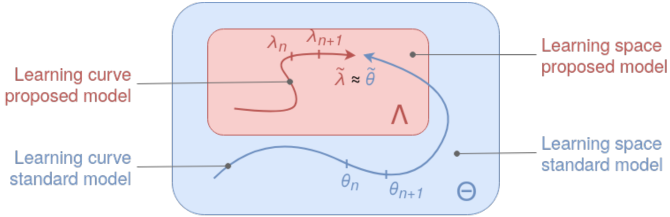
Roadmap

- **Build a sparse model** of existing CNN architectures, based on the **dual-tree wavelet packet transform** (DT-$\mathbb{C}$WPT).[5,6]

[5] **Kingsbury2001**
[6] **Bayram2008**

- **Build a sparse model** of existing CNN architectures, based on the **dual-tree wavelet packet transform** (DT-$\mathbb{C}$WPT).[5,6]

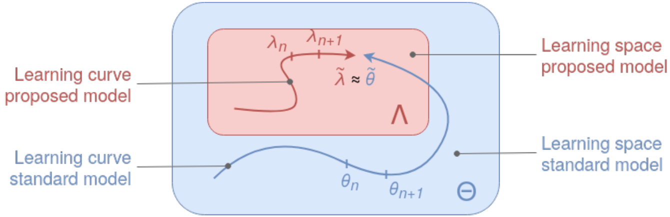
$\Longrightarrow$ **Subset selection** among all possible configurations.



[5] **Kingsbury2001**
[6] **Bayram2008**

a **Build a sparse model** of existing CNN architectures, based on the **dual-tree wavelet packet transform** (DT-$\mathbb{C}$WPT).[5,6]

$\Longrightarrow$ **Subset selection** among all possible configurations.

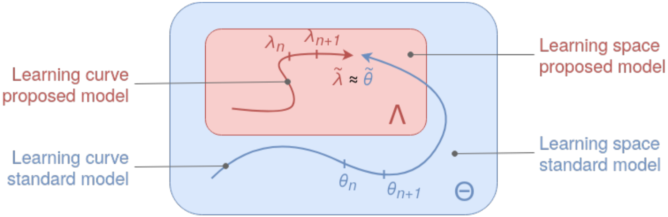


b **Assess model's accuracy** with respect to the original architecture.

[5] **Kingsbury2001**
[6] **Bayram2008**

Roadmap

a. **Build a sparse model** of existing CNN architectures, based on the **dual-tree wavelet packet transform** (DT-$\mathbb{C}$WPT).[5,6]
   $\Longrightarrow$ **Subset selection** among all possible configurations.



b. **Assess model's accuracy** with respect to the original architecture.

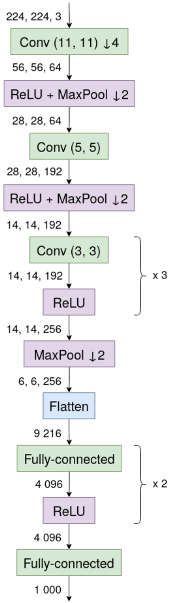
c. **Study properties of the sparse model**, such as directional selectivity, stability with respect to translations, rotations, deformation, etc.

[5] **Kingsbury2001**
[6] **Bayram2008**

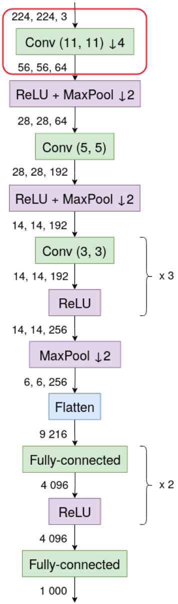
Contents

a. Sparse models of standard CNNs → b. Accuracy of the models → c. Properties of the models

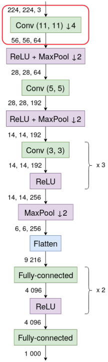
Standard AlexNet



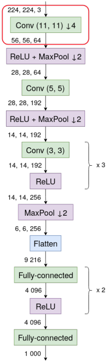
Standard AlexNet

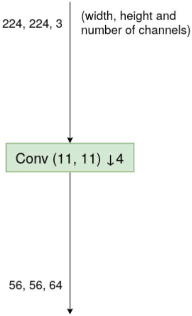


First convolution layer in standard AlexNet



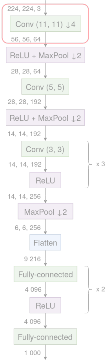
First convolution layer in standard AlexNet

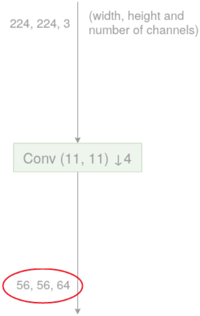




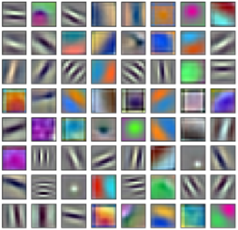
AlexNet first layer
23.3K params.

First convolution layer in standard AlexNet



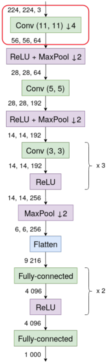


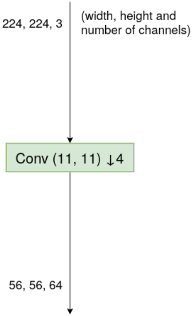
AlexNet first layer
23.3K params.



64 output channels

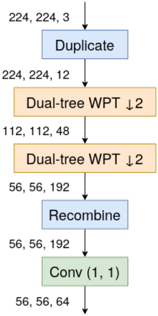
Model with 2 levels of dual-tree decomposition





AlexNet first layer

23.3K params.

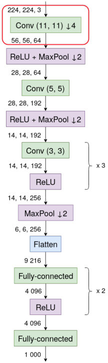


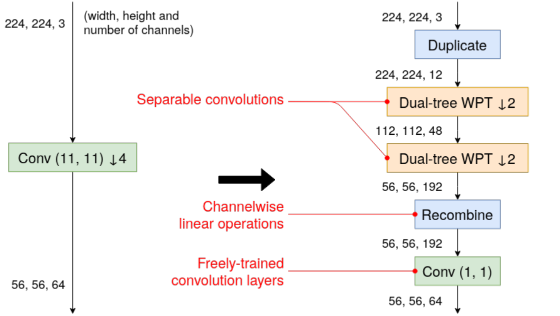
Replacement with dual-tree WPT

(2 levels of decomposition)

12.4K params.

Model with 2 levels of dual-tree decomposition





AlexNet first layer

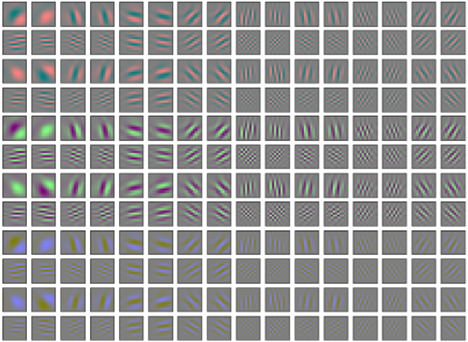
23.3K params.

Replacement with dual-tree WPT

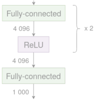
(2 levels of decomposition)

12.4K params.

Model with 2 levels of dual-tree decomposition

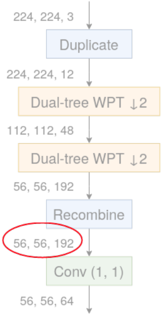


96 complex-valued feature maps



AlexNet first layer

23.3K params.

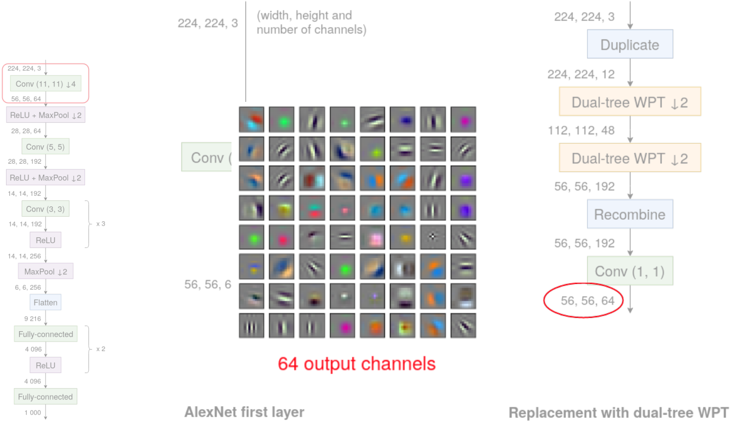


Replacement with dual-tree WPT

(2 levels of decomposition)

12.4K params.

Model with 2 levels of dual-tree decomposition



AlexNet first layer
23.3K params.

Replacement with dual-tree WPT
(2 levels of decomposition)
12.4K params.

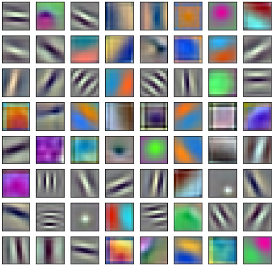
Kernel visualization

Is the model with 2 levels of dual-tree decomposition satisfactory?

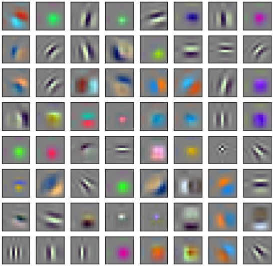
Kernel visualization

Is the model with 2 levels of dual-tree decomposition satisfactory?

- Kernel visualization **in the spatial domain**:



AlexNet first layer

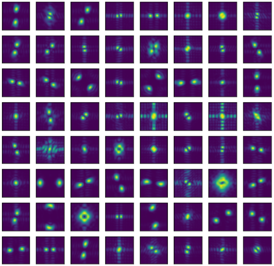


Replacement with dual-tree WPT
(2 levels of decomposition)

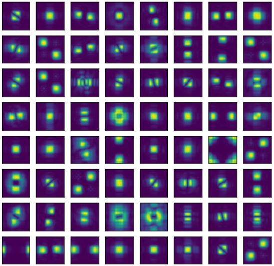
$\Longrightarrow$ **Too small spatial extent** (or too wide frequency extent), compared to standard AlexNet.

Is the model with 2 levels of dual-tree decomposition satisfactory?

- Kernel visualization **in the frequency domain**:



AlexNet first layer



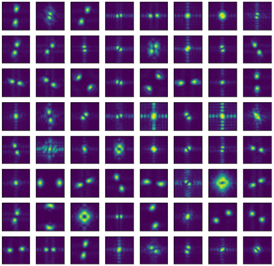
Replacement with dual-tree WPT
(2 levels of decomposition)

$\Longrightarrow$ **Too small spatial extent** (or too wide frequency extent), compared to standard AlexNet.

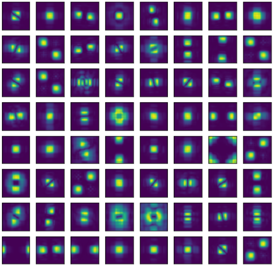
Kernel visualization

Is the model with 2 levels of dual-tree decomposition satisfactory?

- Kernel visualization **in the frequency domain**:



AlexNet first layer

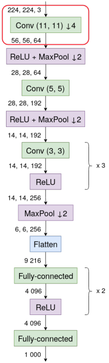


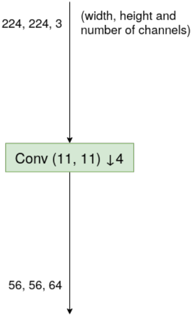
Replacement with dual-tree WPT
(2 levels of decomposition)

$\Longrightarrow$ **Too small spatial extent** (or too wide frequency extent), compared to standard AlexNet.

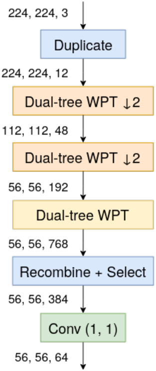
$\Longrightarrow$ Idea: **add one extra level of decomposition**.

Model with 3 levels of dual-tree decomposition



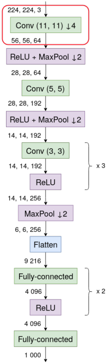


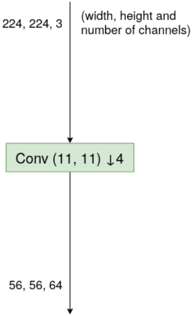
AlexNet first layer
23.3K params.



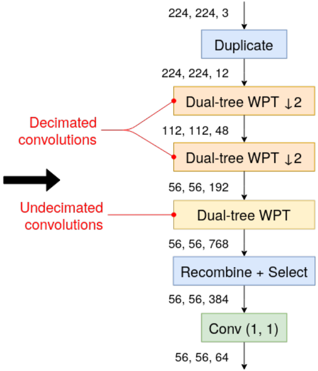
Replacement with dual-tree WPT
(3 levels of decomposition)
49.2K params. (6.1K under add. constraints)

Model with 3 levels of dual-tree decomposition



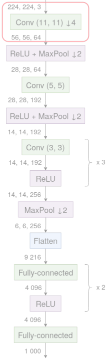


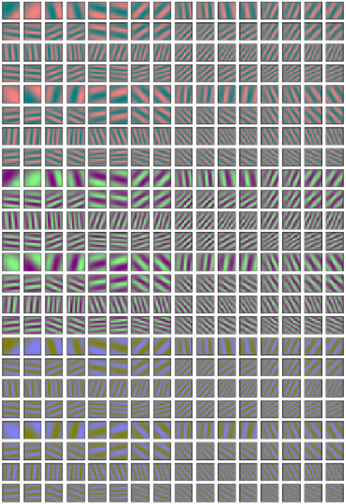
AlexNet first layer
23.3K params.



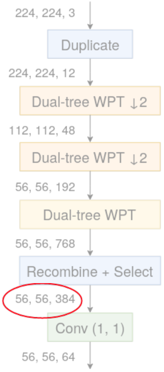
Replacement with dual-tree WPT
(3 levels of decomposition)
49.2K params. (6.1K under add. constraints)

Model with 3 levels of dual-tree decomposition



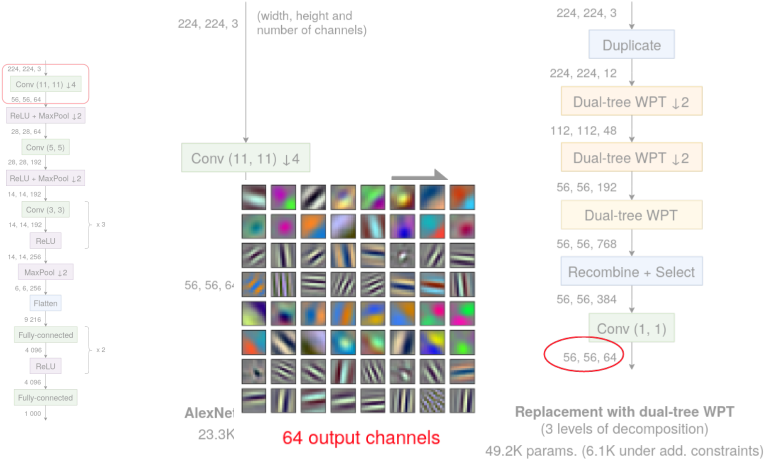


192 complex-valued feature maps



Replacement with dual-tree WPT
(3 levels of decomposition)
9.2K params. (6.1K under add. constraints)

Model with 3 levels of dual-tree decomposition

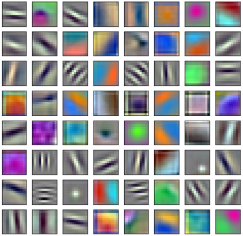


Which choice of decomposition depth?

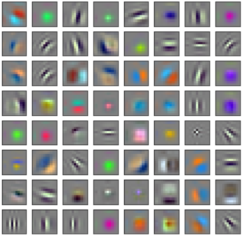
Kernel visualization

Which choice of decomposition depth?

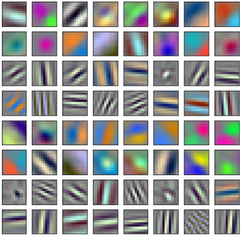
- Kernel visualization **in the spatial domain**:



AlexNet first layer



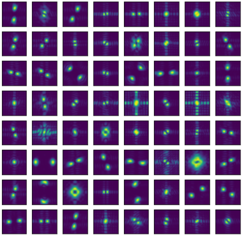
Replacement with dual-tree WPT
(2 levels of decomposition)



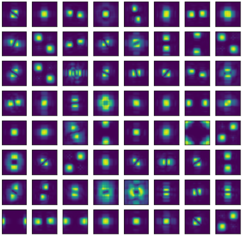
Replacement with dual-tree WPT
(3 levels of decomposition)

Which choice of decomposition depth?

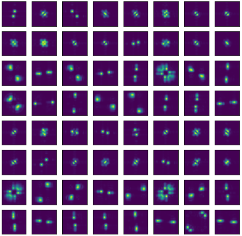
- Kernel visualization **in the frequency domain**:



AlexNet first layer



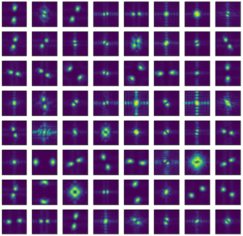
Replacement with dual-tree WPT
(2 levels of decomposition)



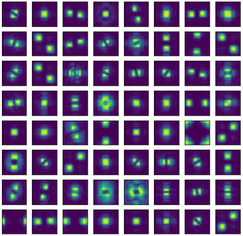
Replacement with dual-tree WPT
(3 levels of decomposition)

Which choice of decomposition depth?

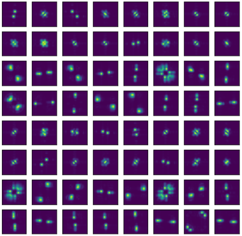
- Kernel visualization **in the frequency domain**:



AlexNet first layer



Replacement with dual-tree WPT
(2 levels of decomposition)



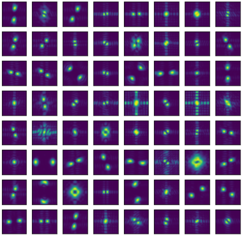
Replacement with dual-tree WPT
(3 levels of decomposition)

⟹ A model with 3 levels of decomposition seems more relevant, in both spatial and frequency domains.

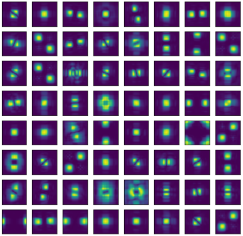
Kernel visualization

Which choice of decomposition depth?

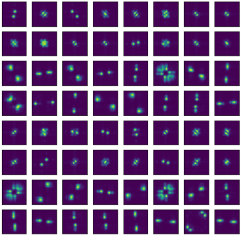
- Kernel visualization **in the frequency domain**:



AlexNet first layer



Replacement with dual-tree WPT
(2 levels of decomposition)

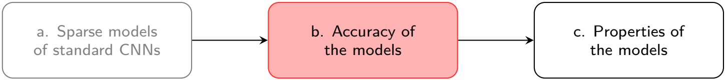


Replacement with dual-tree WPT
(3 levels of decomposition)

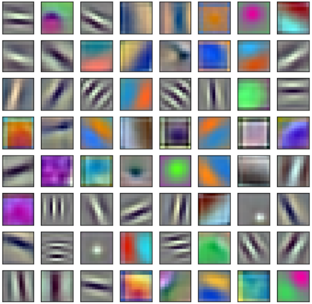
$\Longrightarrow$ A model with 3 levels of decomposition seems more relevant, in both spatial and frequency domains.

$\Longrightarrow$ **Can we find a measure of similarity between kernels?**

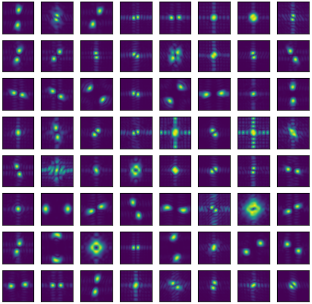
Contents


a. Sparse models of standard CNNs
b. Accuracy of the models
c. Properties of the models

Example with standard AlexNet:



Spatial representation



Frequential representation

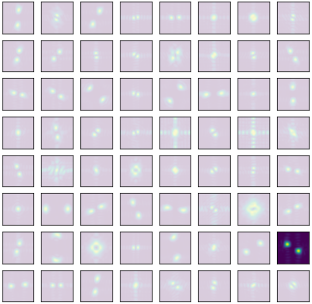
Characteristic frequencies obtained by using the 2D discrete-time Fourier transform as well as the structure tensor[7].

[7] **Jahne2004**

Similarity between convolution kernels

Example with standard AlexNet:

Spatial representation

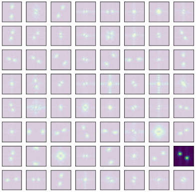


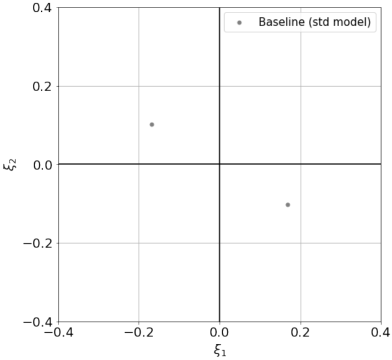
Frequential representation

Characteristic frequencies obtained by using the 2D discrete-time Fourier transform as well as the structure tensor[7].

[7] **Jahne2004**

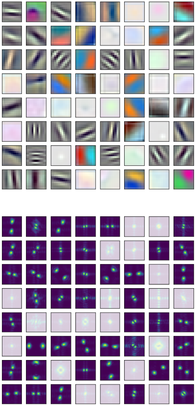
Similarity between convolution kernels

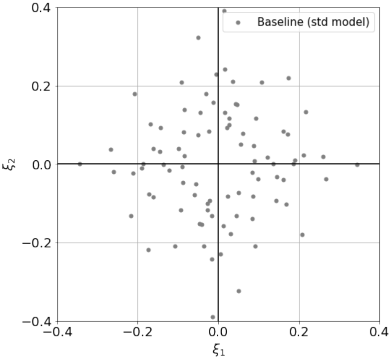




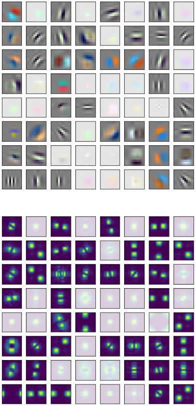
Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

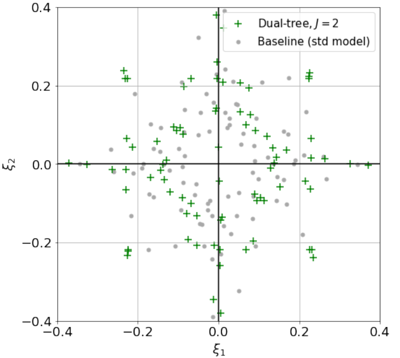
Similarity between convolution kernels





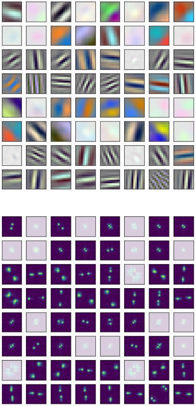
Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

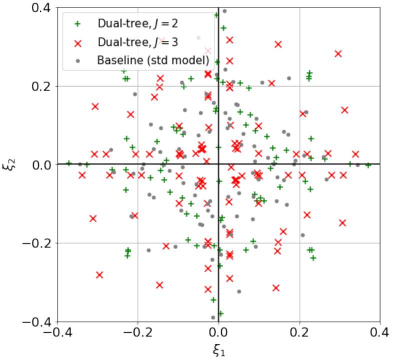




Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

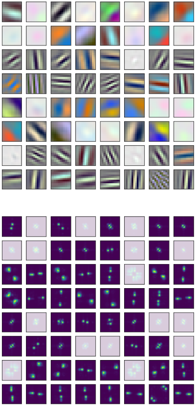
Similarity between convolution kernels

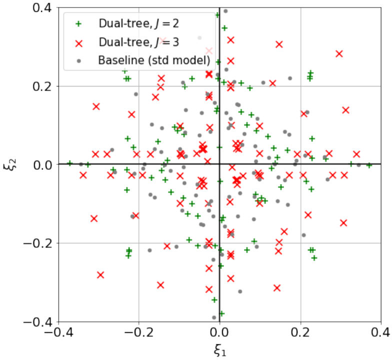




Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

Similarity between convolution kernels

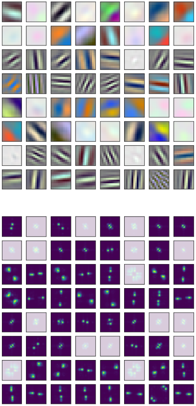


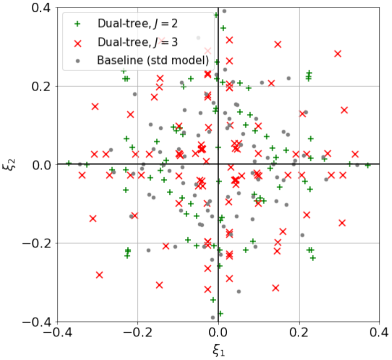


Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

$\Longrightarrow$ In addition to being more localized in the Fourier domain, the model with 3 levels of decomposition reaches lower frequencies, that we also find in the standard model.

Similarity between convolution kernels





Characteristic frequencies of AlexNet-based kernels
(J denotes the number of decomposition stages)

$\Longrightarrow$ In addition to being more localized in the Fourier domain, the model with 3 levels of decomposition reaches lower frequencies, that we also find in the standard model.

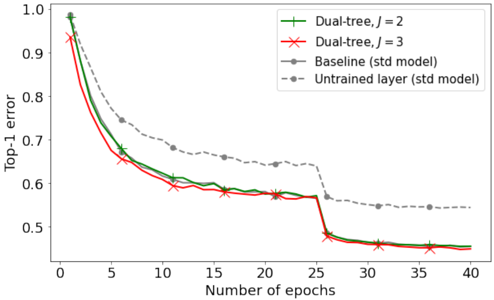
$\Longrightarrow$ **This confirms our intuition** about the choice for a "best" model.

Performance of the models

Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.

Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.

AlexNet-based models trained on ImageNet ILSVRC2012.

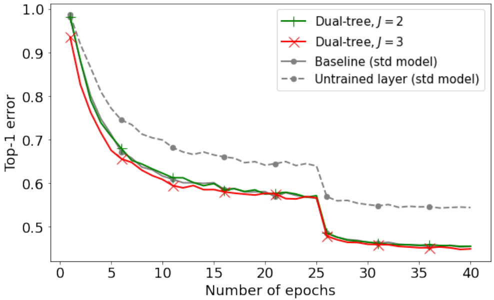


Validation error along training
(J denotes the number of decomposition stages)

Performance of the models

Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.

AlexNet-based models trained on ImageNet ILSVRC2012.



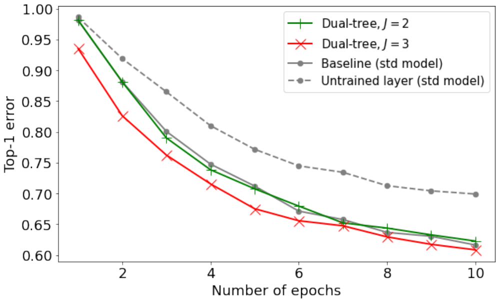
Validation error along training
(J denotes the number of decomposition stages)

$\Longrightarrow$ Our models **reach the performance** of standard AlexNet.

Performance of the models

Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.

AlexNet-based models trained on ImageNet ILSVRC2012.



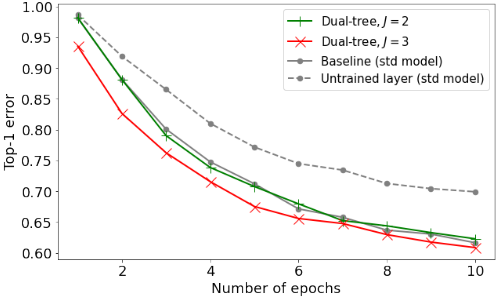
Validation error along training – focus on the first epochs
(J denotes the number of decomposition stages)

$\Longrightarrow$ Our models **reach the performance** of standard AlexNet.

Performance of the models

Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.

AlexNet-based models trained on ImageNet ILSVRC2012.



Validation error along training – focus on the first epochs
(J denotes the number of decomposition stages)

$\Longrightarrow$ Our models **reach the performance** of standard AlexNet.

$\Longrightarrow$ Validation error **decreases more rapidly** with $J = 3$. This may be **due to the reduced model complexity**, compared to Standard AlexNet.

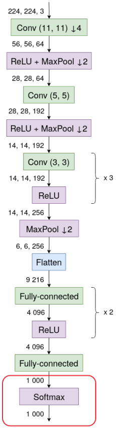
Contents

a. Sparse models of standard CNNs → b. Accuracy of the models → c. Properties of the models

Robustness with respect to small shifts

Important property of DT-$\mathbb{C}$WPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.

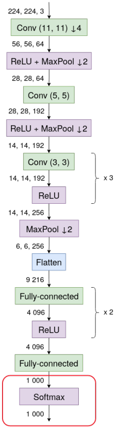
Important property of DT-ℂWPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.



- Forward-propagation of **8 shifted versions** of an image;

Robustness with respect to small shifts

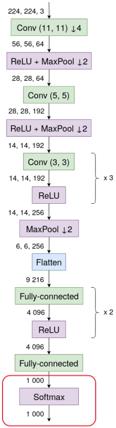
Important property of DT-ℂWPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.



- Forward-propagation of **8 shifted versions** of an image;
- **Kulback-Leibler divergence** between softmax of outputs;

Robustness with respect to small shifts

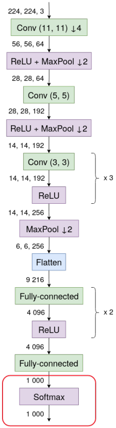
Important property of DT-ℂWPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.



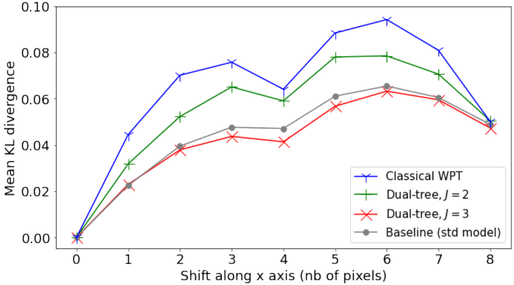
- Forward-propagation of **8 shifted versions** of an image;
- **Kulback-Leibler divergence** between softmax of outputs;
- Average values computed **over a subset of ImageNet** ($50,000$ images).

Robustness with respect to small shifts

Important property of DT-ℂWPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.

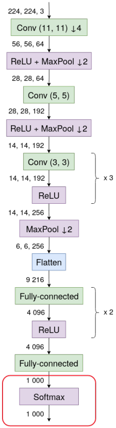


- Forward-propagation of **8 shifted versions** of an image;
- **Kulback-Leibler divergence** between softmax of outputs;
- Average values computed **over a subset of ImageNet** (50, 000 images).
- **Also included**: network implementing classical WPT, which is NOT shift-invariant.

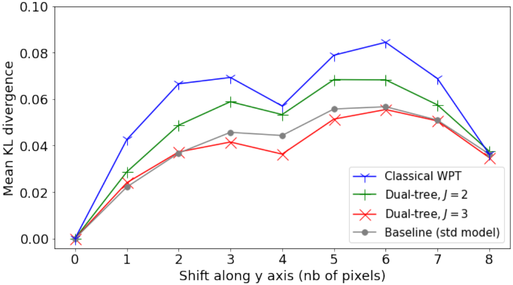


Robustness with respect to small shifts

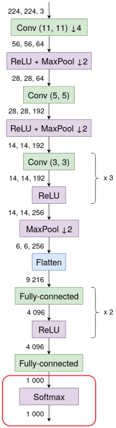
Important property of DT-ℂWPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.



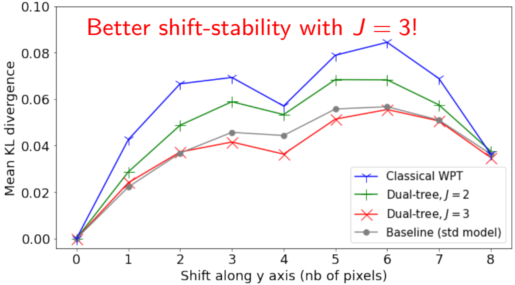
- Forward-propagation of **8 shifted versions** of an image;
- **Kulback-Leibler divergence** between softmax of outputs;
- Average values computed **over a subset of ImageNet** (50, 000 images).
- **Also included**: network implementing classical WPT, which is NOT shift-invariant.



Important property of DT-$\mathbb{C}$WPT: **near-shift invariance**, when applied to the modulus of complex coefficients.
Let's see how this property is transferred to the output of the network.



- Forward-propagation of **8 shifted versions** of an image;
- **Kulback-Leibler divergence** between softmax of outputs;
- Average values computed **over a subset of ImageNet** (50, 000 images).
- **Also included**: network implementing classical WPT, which is NOT shift-invariant.



Contents

Conclusion and future work

$\Longrightarrow$ On-going work to establish near equivalence between standard CNNs and handcrafted architectures for which theoretical properties are guaranteed.

Conclusion and future work

$\Longrightarrow$ On-going work to establish near equivalence between standard CNNs and handcrafted architectures for which theoretical properties are guaranteed.

$\Longrightarrow$ Similar study performed on ResNet architecture.

Conclusion and future work

$\Longrightarrow$ On-going work to establish near equivalence between standard CNNs and handcrafted architectures for which theoretical properties are guaranteed.

$\Longrightarrow$ Similar study performed on ResNet architecture.

$\Longrightarrow$ First step toward a more complete understanding of CNNs for computer vision.

Conclusion and future work

$\Longrightarrow$ On-going work to establish near equivalence between standard CNNs and handcrafted architectures for which theoretical properties are guaranteed.

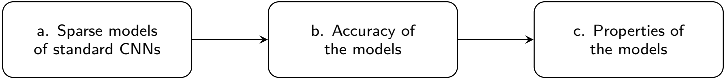
$\Longrightarrow$ Similar study performed on ResNet architecture.

$\Longrightarrow$ First step toward a more complete understanding of CNNs for computer vision.

Future work

- Consolidate analysis: other types of invariants, etc.
- Quantitative evaluation of kernel similarity.
- Focus research on deeper layers.

Thank you for your attention!



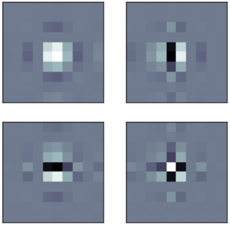
This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir, as well as the ANR grant AVENUE (ANR-18-CE23-0011).

Contents

The standard wavelet packet transform (WPT)

- h and $\mathrm{g} \in \mathbb{R}^{\mathbb{Z}}$, pair of conjugate mirror filters (CMFs)
- separable 2D filter bank

$$\mathrm{G}^{(0)} = \mathrm{h} \otimes \mathrm{h} \quad \mathrm{G}^{(1)} = \mathrm{h} \otimes \mathrm{g} \quad \mathrm{G}^{(2)} = \mathrm{g} \otimes \mathrm{h} \quad \mathrm{G}^{(3)} = \mathrm{g} \otimes \mathrm{g}.$$



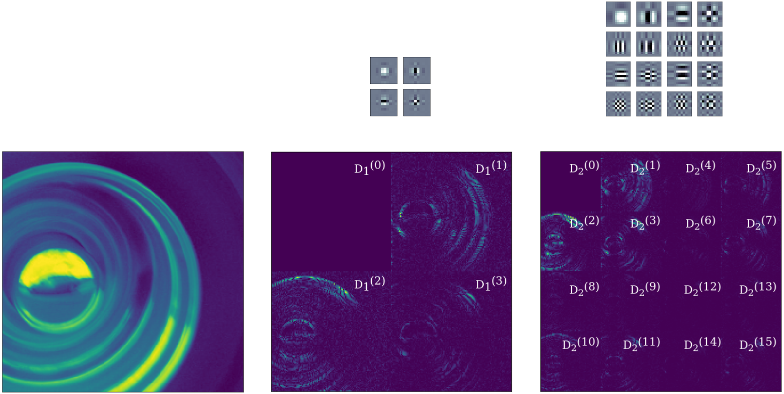
- Input image: $\mathrm{X}_0^{(0)} = \mathrm{X}$
- Successive decompositions, for each $j \in \{1 \mathinner{..} J\}$:

$$\forall l \in \{0 \mathinner{..} 3\}\,,\ \mathrm{X}_j^{(4k+l)} = \left(\mathrm{X}_{j-1}^{(k)} * \overline{\mathrm{G}^{(l)}}\right) \downarrow 2\,.$$

- $\left\{\mathrm{X}_J^{(k)}\right\}_{k \in \left\{0..4^J-1\right\}}$ is a representation of X in a wavelet packet basis.
- $J \nearrow \implies$ spatial resolution $\searrow$ and frequency resolution $\nearrow$.

The standard wavelet packet transform (WPT)

Example with 2 levels of decomposition



$j = 0$ $j = 1$ $j = 2$

The dual-tree complex wavelet packet transform (DT-$\mathbb{C}$WPT)

Properties of standard WPT:

- ✓ **sparse signal representation and vertical / horizontal feature discrimination;**
- ✗ **lack of shift invariance and a poor directional selectivity.**

The dual-tree complex wavelet packet transform (DT-$\mathbb{C}$WPT)

Properties of standard WPT:

✓ **sparse signal representation and vertical / horizontal feature discrimination;**

✗ **lack of shift invariance and a poor directional selectivity.**

Workaround: decompose images in a frame of complex oriented waveforms with minimal redundancy.

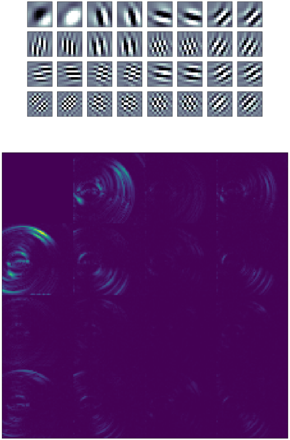
- Four WPT decompositions $\left\{\mathrm{X}_{\mathrm{a},\,J}^{(k)}\right\}$, $\left\{\mathrm{X}_{\mathrm{b},\,J}^{(k)}\right\}$, $\left\{\mathrm{X}_{\mathrm{c},\,J}^{(k)}\right\}$, $\left\{\mathrm{X}_{\mathrm{d},\,J}^{(k)}\right\}$ with suitable filter banks;
- Dual-tree coefficients $\left\{\mathrm{Z}_J^{\nearrow(k)}\right\}$ and $\left\{\mathrm{Z}_J^{\nwarrow(k)}\right\}_{k\in\left\{0..4^J-1\right\}}$:

$$\begin{pmatrix}\mathrm{Z}_J^{\nearrow(k)}\\ \mathrm{Z}_J^{\nwarrow(k)}\end{pmatrix} = \begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}\begin{pmatrix}\mathrm{X}_{\mathrm{a},\,J}^{(k)}\\ \mathrm{X}_{\mathrm{d},\,J}^{(k)}\end{pmatrix} + i\cdot\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}\mathrm{X}_{\mathrm{c},\,J}^{(k)}\\ \mathrm{X}_{\mathrm{b},\,J}^{(k)}\end{pmatrix}.$$

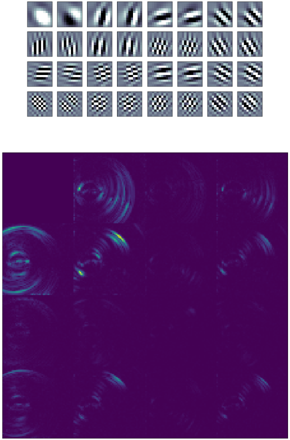
$\Rightarrow$ 6 orientations, and near shift-invariance for $\left|\mathrm{Z}_J^{\nearrow(k)}\right|$ and $\left|\mathrm{Z}_J^{\nwarrow(k)}\right|$.

The dual-tree complex wavelet packet transform (DT-$\mathbb{C}$WPT)

Example with 2 levels of decomposition



$\left|Z_2^{\nearrow(k)}\right|$



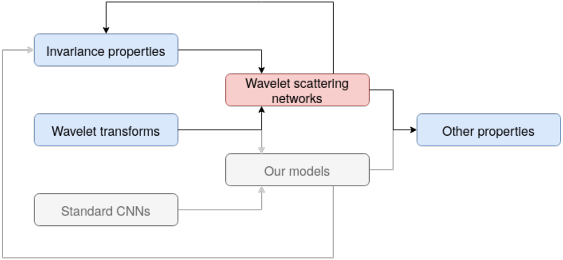
$\left|Z_2^{\nwarrow(k)}\right|$

Related work – wavelet scattering networks

Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities **as in CNNs**.[8]

[8] **Bruna2013**

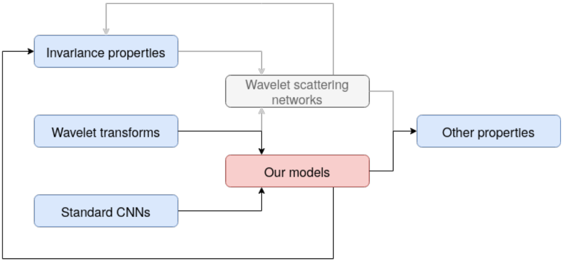
Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities **as in CNNs**.[8]



[8] **Bruna2013**

Related work – wavelet scattering networks

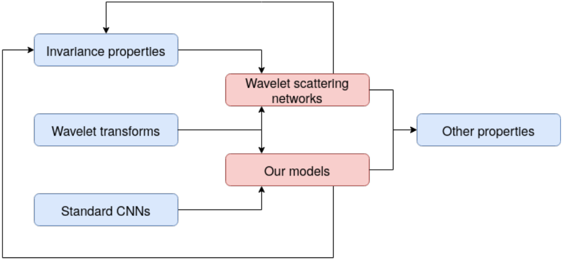
Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities **as in CNNs**.[8]



[8] **Bruna2013**

Related work – wavelet scattering networks

Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities **as in CNNs**.[8]



$\Longrightarrow$ *Our contribution: theoretical model imitating the behavior of a standard CNN, with invariance properties as in wavelet scattering networks.*

[8] **Bruna2013**

Why the dual-tree complex wavelet packet transform?

Similitudes with the **Gabor transform**:

- **complex**, **multiscale** and **oriented** filters;
- **well-localized** in the Fourier domain;
- **sparse** image representations.

Why the dual-tree complex wavelet packet transform?

Similitudes with the **Gabor transform**:

- **complex**, **multiscale** and **oriented** filters;
- **well-localized** in the Fourier domain;
- **sparse** image representations.

Differences with the Gabor transform:

- specifically designed for the **discrete world**, with **perfect reconstruction** guarantees and **minimal redundancy**;
- **decimated convolutions**, which is consistent with the CNN approach;
- **sparse description** – a single pair of out-of-phase 1D vectors is sufficient to describe the whole process.