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

1 Introduction

2 Proposed models

3 Accuracy of the models

4 Experimental properties

5 Conclusion and future work
Contents

1 Introduction

2 Proposed models

3 Accuracy of the models

4 Experimental properties

5 Conclusion and future work
Convolutional neural networks (CNNs)

✓ 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)\textsuperscript{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\textsuperscript{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\textsuperscript{3}.

\textsuperscript{1}LeCun2015
\textsuperscript{2}Mallat2009
\textsuperscript{3}Yosinski2014
CNNs vs discrete wavelet transforms

AlexNet\textsuperscript{4} filters (first layer) after training with ImageNet

\textsuperscript{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-CWPT).\textsuperscript{5,6}

\textsuperscript{5} Kingsbury2001
\textsuperscript{6} Bayram2008
a **Build a sparse model** of existing CNN architectures, based on the **dual-tree wavelet packet transform** (DT-CWPT).\(^5,\!^6\)

\[ \implies \text{Subset selection among all possible configurations.} \]

\(^5\)Kingsbury2001

\(^6\)Bayram2008
Roadmap

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

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

---

\textsuperscript{5} Kingsbury2001  
\textsuperscript{6} Bayram2008
Roadmap

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

- **Assess model’s accuracy** with respect to the original architecture.

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

\(^5\) Kingsbury2001
\(^6\) Bayram2008
Standard AlexNet

- **Introduction**
- **Proposed models**
- **Accuracy of the models**
- **Experimental properties**
- **Conclusion and future work**
Standard AlexNet

- **Conv (11, 11) ↓ 4**
- **ReLU + MaxPool ↓ 2**
- **Conv (5, 5)**
- **ReLU + MaxPool ↓ 2**
- **Conv (3, 3)**
- **ReLU**
- **MaxPool ↓ 2**
- **Flatten**
- **Fully-connected**
- **ReLU**
- **Fully-connected**
First convolution layer in standard AlexNet
First convolution layer in standard AlexNet

224, 224, 3

Conv (11, 11) ↓4

56, 56, 64

ReLU + MaxPool ↓2

28, 28, 64

Conv (5, 5)

28, 28, 192

ReLU + MaxPool ↓2

14, 14, 192

Conv (3, 3)

14, 14, 192

ReLU

14, 14, 256

MaxPool ↓2

6, 6, 256

Flatten

9 216

x 2

Fully-connected

4 096

ReLU

4 096

Fully-connected

1 000

56, 56, 64

Conv (11, 11) ↓4

(width, height and number of channels)

AlexNet first layer

23.3K params.
First convolution layer in standard AlexNet

AlexNet first layer
23.3K 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

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

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

  ![Kernel visualization](image)

  - **AlexNet first layer**
  - **Replacement with dual-tree WPT (2 levels of decomposition)**

  **⇒ 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)

⇒ **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:

  ➞ Too small spatial extent (or too wide frequency extent), compared to standard AlexNet.

  ➞ 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

Duplicate

224, 224, 3
224, 224, 12
112, 112, 48
56, 56, 192
56, 56, 768
56, 56, 384

Conv (1, 1)

Recombine + Select

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

**Proposed models**

- Model with 3 levels of dual-tree decomposition

**Accuracy of the models**

- 64 output channels

**Experimental properties**

- 49.2K params. (6.1K under add. constraints)

**Conclusion and future work**

- 23.3K
Kernel visualization

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)

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

=⇒ Can we find a measure of similarity between kernels?
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)

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

⇒ Can we find a measure of similarity between kernels?
Contents

1 Introduction

2 Proposed models

3 Accuracy of the models

4 Experimental properties

5 Conclusion and future work

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

Example with standard AlexNet:

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

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. This confirms our intuition about the choice for a "best" model.

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

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.

This confirms our intuition about the choice for a "best" model.

Feature frequencies of AlexNet-based kernels

(J denotes the number of decomposition stages)
Similarity between convolution kernels

\[
\text{Characteristic frequencies of AlexNet-based kernels}\quad (J \text{ denotes the number of decomposition stages})
\]
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

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.

This confirms our intuition about the choice for a “best" model.
Another way of assessing the accuracy of our models is to compare their performances with respect to standard CNNs.
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](image)

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

*Our models reach the performance of standard AlexNet.*
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.**

![Graph showing validation error along training](image)

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

⇒ Our models reach the performance of standard AlexNet.
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.

Our models reach the performance of standard AlexNet.

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

1 Introduction

2 Proposed models

3 Accuracy of the models

4 Experimental properties

5 Conclusion and future work

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-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.
Robustness with respect to small shifts

Important property of DT-ĆWPT: **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-\(\mathcal{C}\text{WPT}^{\prime}\): **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;

![Diagram]
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.

![Diagram](image_url)
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.
Conclusion and future work

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

Similar study performed on ResNet architecture.
Conclusion and future work

- On-going work to establish near equivalence between standard CNNs and handcrafted architectures for which theoretical properties are guaranteed.
- Similar study performed on ResNet architecture.
- First step toward a more complete understanding of CNNs for computer vision.
Conclusion and future work

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

Similar study performed on ResNet architecture.

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.
1 Introduction

2 Proposed models

3 Accuracy of the models

4 Experimental properties

5 Conclusion and future work

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).
Appendix

Background on discrete wavelet transform

Related work – wavelet scattering networks
The standard wavelet packet transform (WPT)

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

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

Input image: \( X_0^{(0)} = X \)

Successive decompositions, for each \( j \in \{1 \ldots J\} \):

\[
\forall l \in \{0 \ldots 3\}, \quad X_j^{(4k+l)} = \left( X_{j-1}^{(k)} \ast \overline{G(l)} \right) \downarrow 2.
\]

\( \left\{ X_j^{(k)} \right\}_{k \in \{0 \ldots 4^J-1\}} \) is a representation of \( X \) in a wavelet packet basis.

\( J \uparrow \quad \Longrightarrow \) spatial resolution \( \downarrow \) and frequency resolution \( \uparrow \).
The standard wavelet packet transform (WPT)

Example with 2 levels of decomposition

\[ j = 0 \quad j = 1 \quad j = 2 \]
The dual-tree complex wavelet packet transform (DT-CWPT)

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-CWPT)

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 \( \left\{ X_{a, J}^{(k)} \right\}, \left\{ X_{b, J}^{(k)} \right\}, \left\{ X_{c, J}^{(k)} \right\}, \left\{ X_{d, J}^{(k)} \right\} \) with suitable filter banks;
- Dual-tree coefficients \( \left\{ Z_{J}^{\uparrow(k)} \right\} \) and \( \left\{ Z_{J}^{\downarrow(k)} \right\} \):

\[
\begin{pmatrix}
Z_{J}^{\uparrow(k)} \\
Z_{J}^{\downarrow(k)}
\end{pmatrix} =
\begin{pmatrix}
1 & -1 \\
1 & 1
\end{pmatrix}
\begin{pmatrix}
X_{a, J}^{(k)} \\
X_{c, J}^{(k)}
\end{pmatrix} +
i \cdot
\begin{pmatrix}
1 & 1 \\
1 & -1
\end{pmatrix}
\begin{pmatrix}
X_{b, J}^{(k)} \\
X_{d, J}^{(k)}
\end{pmatrix}.
\]

⇒ 6 orientations, and near shift-invariance for \( \left| Z_{J}^{\uparrow(k)} \right| \) and \( \left| Z_{J}^{\downarrow(k)} \right| \).
The dual-tree complex wavelet packet transform (DT-CWPT)

Example with 2 levels of decomposition

\[ |Z_2^{\rightarrow(k)}| \]

\[ |Z_2^{\leftarrow(k)}| \]
Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities **as in CNNs**.\(^8\)
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.

\[ \Rightarrow \]

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

---

\[ ^8 \text{Bruna2013} \]
Related work – wavelet scattering networks

Image representation based on the **continuous wavelet transform**, involving convolutions and non-linearities as in CNNs.\(^8\)

\[ \Rightarrow \text{Our contribution: theoretical model imitating the behavior of a standard CNN, with invariance properties as in wavelet scattering networks.} \]

---

\(^8\) Bruna2013
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.