The Mathematics of Deep Learning
Part 1: Continuous-time Theory

Helmut Bőlcskei

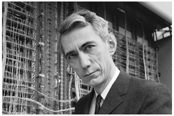
ETH zürich

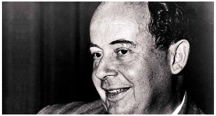
Department of Information Technology and Electrical Engineering

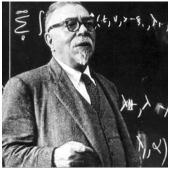
June 2016

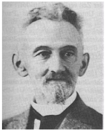
joint work with Thomas Wiatowski, Philipp Grohs, and Michael Tschannen

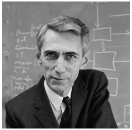
Face recognition

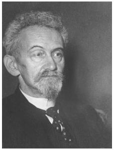


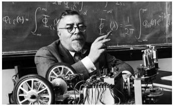


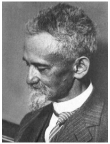




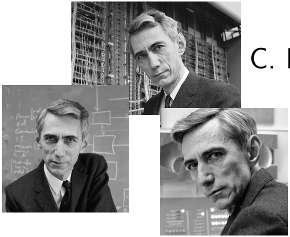






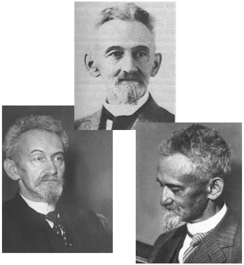


Face recognition

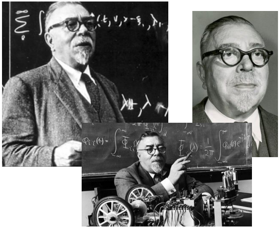


C. E. Shannon

J. von Neumann

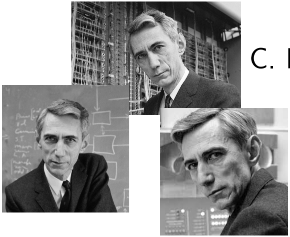


N. Wiener



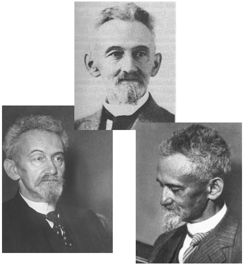
F. Hausdorff

Face recognition



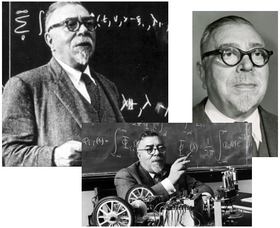
C. E. Shannon

J. von Neumann



F. Hausdorff

N. Wiener

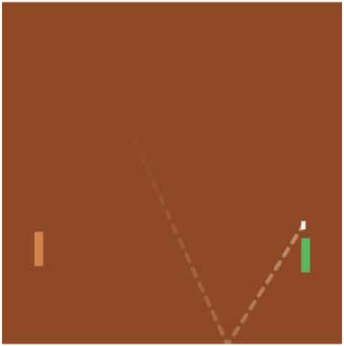


Feature extraction through deep convolutional neural networks (DCNs)

Go!

DCNs beat Go-champion Lee Sedol [*Silver et al., 2016*]

Atari games



DCNs beat professional human Atari-players [*Mnih et al., 2015*]

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos ."

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos Kleiber .”

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos Kleiber conducting the ."

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos Kleiber conducting the Vienna Philharmonic's ."

Describing the content of an image

DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos Kleiber conducting the Vienna Philharmonic's New Year's Concert .”

Describing the content of an image

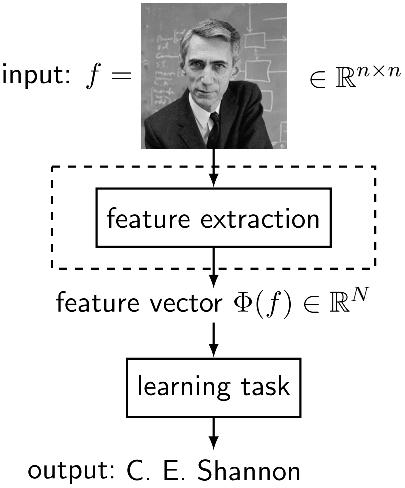
DCNs generate sentences describing the content of an image [Vinyals et al., 2015]

"Carlos Kleiber conducting the Vienna Philharmonic's New Year's Concert 1989."

Feature extraction and learning task

DCNs can be used

i) as stand-alone feature extractors [*Huang and LeCun, 2006*]

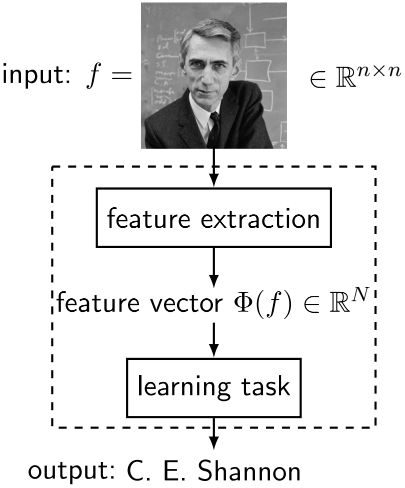


Feature extraction and learning task

DCNs can be used

i) as stand-alone feature extractors [*Huang and LeCun, 2006*]

ii) to perform feature extraction *and* the learning task directly [*LeCun et al., 1990*]



Why are DCNs so successful?

> "It is the guiding principle of many applied mathematicians that if something mathematical works really well, there must be a good underlying mathematical reason for it, and we ought to be able to understand it." [*I. Daubechies, 2015*]

Translation invariance

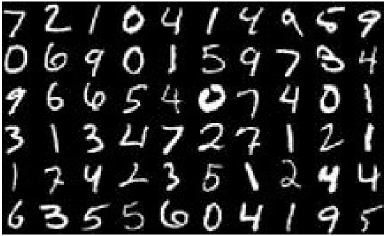
Handwritten digits from the MNIST database [*LeCun & Cortes, 1998*]

Translation invariance

Handwritten digits from the MNIST database [*LeCun & Cortes, 1998*]

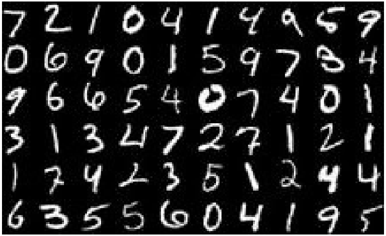
Feature vector should be invariant to spatial location
$\Rightarrow$ translation invariance

Deformation insensitivity



Handwritten digits from the MNIST database [*LeCun & Cortes, 1998*]

Deformation insensitivity

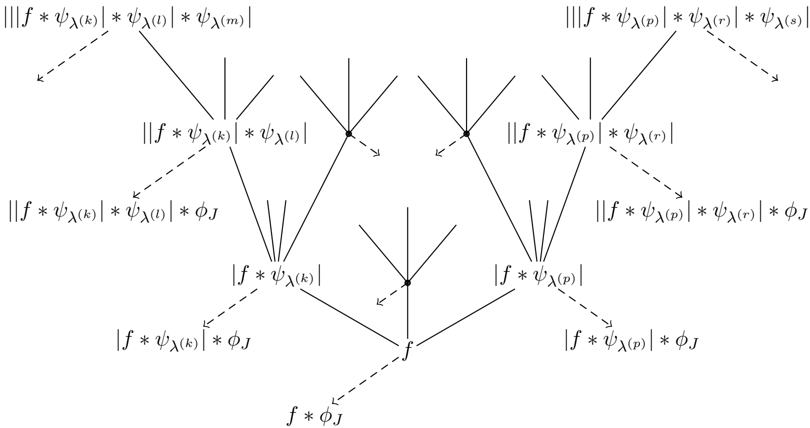


Handwritten digits from the MNIST database [*LeCun & Cortes, 1998*]

Different handwriting styles correspond to deformations of signals
$\Rightarrow$ deformation insensitivity

Mallat's wavelet-modulus DCN

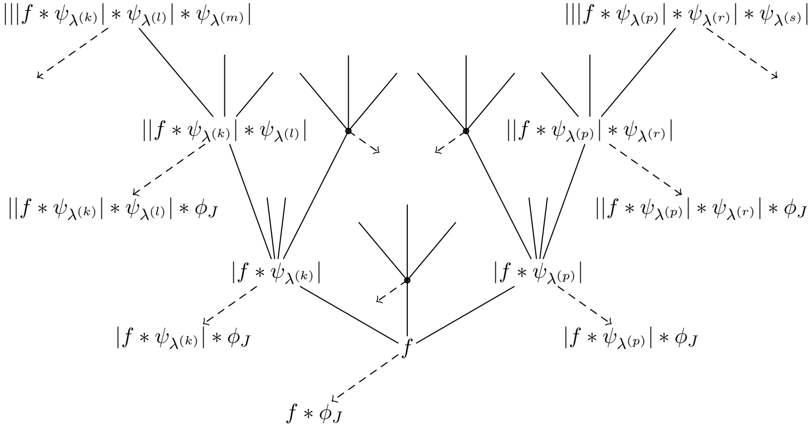
Mallat, 2012, initiated the mathematical analysis of feature extraction through DCNs



Mallat's wavelet-modulus DCN

Features generated in the n-th network layer

$$\Phi_W^n(f) := \left\{ \left| \cdots \left| \, |f * \psi_{\lambda^{(1)}}| * \psi_{\lambda^{(2)}} \right| \cdots \ * \psi_{\lambda^{(n)}} \right| * \phi_J \right\}_{\lambda^{(1)},\ldots,\lambda^{(n)} \in \Lambda_W}$$

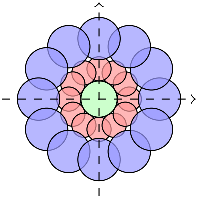


Mallat's wavelet-modulus DCN

Directional wavelet system $\{\phi_J\} \cup \{\psi_\lambda\}_{\lambda \in \Lambda_W}$,

$$\Lambda_W := \big\{\lambda = (j,k) \mid j > -J,\ k \in \{1,\dots,K\}\big\}$$

$$\|f * \phi_J\|_2^2 + \sum_{\lambda \in \Lambda_W} \|f * \psi_\lambda\|_2^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$

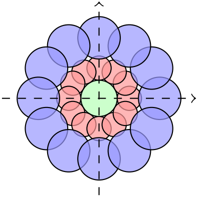


Mallat's wavelet-modulus DCN

Directional wavelet system $\{\phi_J\} \cup \{\psi_\lambda\}_{\lambda \in \Lambda_W}$,

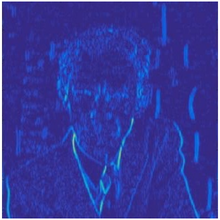
$$\Lambda_W := \{\lambda = (j,k) \mid j > -J,\ k \in \{1,\dots,K\}\}$$

$$\|f * \phi_J\|_2^2 + \sum_{\lambda \in \Lambda_W} \|f * \psi_\lambda\|_2^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$



...and its edge detection capability [*Mallat and Zhong, 1992*]

$$|f * \psi_{\lambda^{(v)}}| =$$

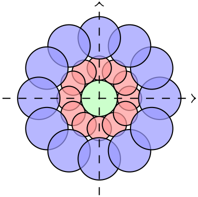


Mallat's wavelet-modulus DCN

Directional wavelet system $\{\phi_J\} \cup \{\psi_\lambda\}_{\lambda \in \Lambda_W}$,

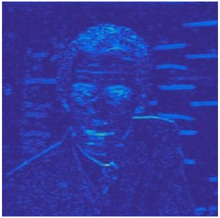
$$\Lambda_W := \{\lambda = (j,k) \mid j > -J,\ k \in \{1,\ldots,K\}\}$$

$$\|f*\phi_J\|_2^2 + \sum_{\lambda \in \Lambda_W} \|f*\psi_\lambda\|_2^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$



...and its edge detection capability [*Mallat and Zhong, 1992*]

$$|f * \psi_{\lambda^{(h)}}| =$$

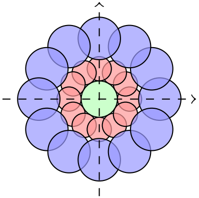


Mallat's wavelet-modulus DCN

Directional wavelet system $\{\phi_J\} \cup \{\psi_\lambda\}_{\lambda \in \Lambda_W}$,

$$\Lambda_W := \{\lambda = (j,k) \mid j > -J,\ k \in \{1,\dots,K\}\}$$

$$\|f*\phi_J\|_2^2 + \sum_{\lambda \in \Lambda_W} \|f*\psi_\lambda\|_2^2 = \|f\|_2^2, \quad \forall f \in L^2(\mathbb{R}^d)$$



...and its edge detection capability [*Mallat and Zhong, 1992*]

$$|f * \psi_{\lambda^{(d)}}| =$$

Mallat's wavelet-modulus DCN

[*Mallat, 2012*] proved that Φ_W is "horizontally" translation-invariant

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d,$$

and stable w.r.t. deformations $(F_\tau f)(x) := f(x - \tau(x))$:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

where $\|\cdot\|_W$ is a wavelet-dependent norm.

Mallat's wavelet-modulus DCN

[*Mallat, 2012*] proved that Φ_W is "horizontally" translation-invariant

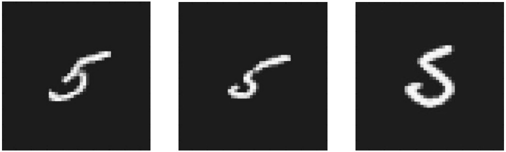
$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d,$$

and stable w.r.t. deformations $(F_\tau f)(x) := f(x - \tau(x))$:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

where $\|\cdot\|_W$ is a wavelet-dependent norm.

Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$:



Mallat's wavelet-modulus DCN

[*Mallat, 2012*] proved that Φ_W is "horizontally" translation-invariant

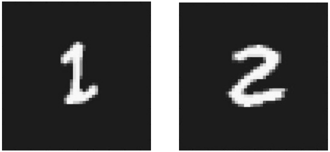
$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d,$$

and stable w.r.t. deformations $(F_\tau f)(x) := f(x - \tau(x))$:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \le C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

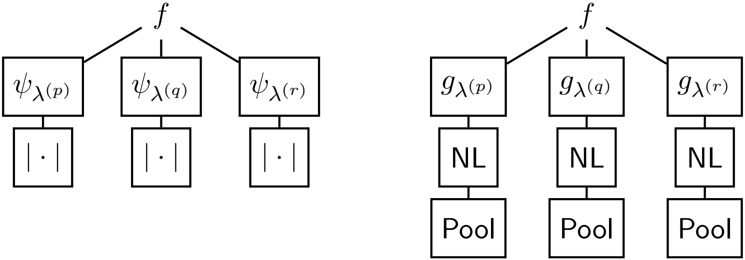
where $\|\cdot\|_W$ is a wavelet-dependent norm.

Non-linear deformation $(F_\tau f)(x) = f(x - \tau(x))$:



Generalizations

The basic operations between consecutive layers

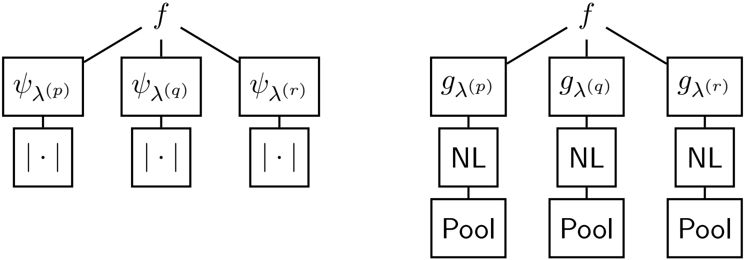


General DCNs employ a wide variety of filters g_λ

- pre-specified and structured (e.g., wavelets [*Serre et al., 2005*])
- pre-specified and unstructured (e.g., random filters [*Jarrett et al., 2009*])
- learned in a supervised [*Huang and LeCun, 2006*] or an unsupervised [*Ranzato et al., 2007*] fashion

Generalizations

The basic operations between consecutive layers

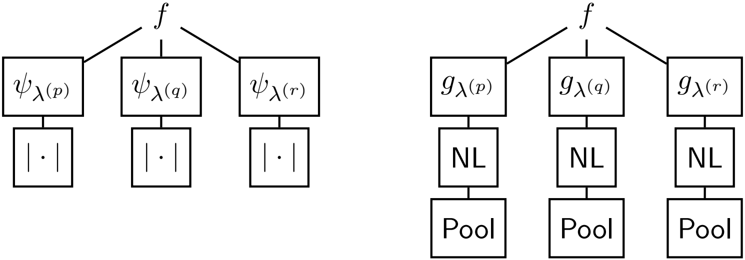


General DCNs employ a wide variety of non-linearities

- modulus [*Mutch and Lowe, 2006*]
- hyperbolic tangent [*Huang and LeCun, 2006*]
- rectified linear unit [*Nair and Hinton, 2010*]
- logistic sigmoid [*Glorot and Bengio, 2010*]

Generalizations

The basic operations between consecutive layers

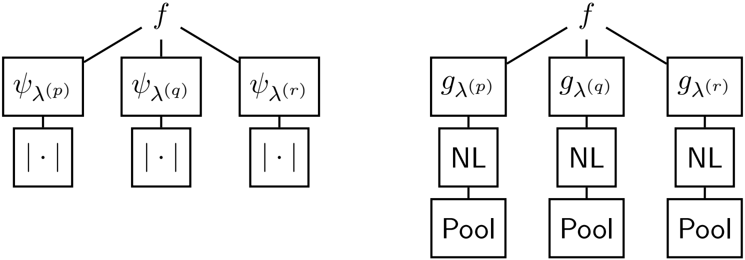


General DCNs employ intra-layer pooling

- sub-sampling [*Pinto et al., 2008*]
- average pooling [*Jarrett et al., 2009*]
- max-pooling [*Ranzato et al., 2007*]

Generalizations

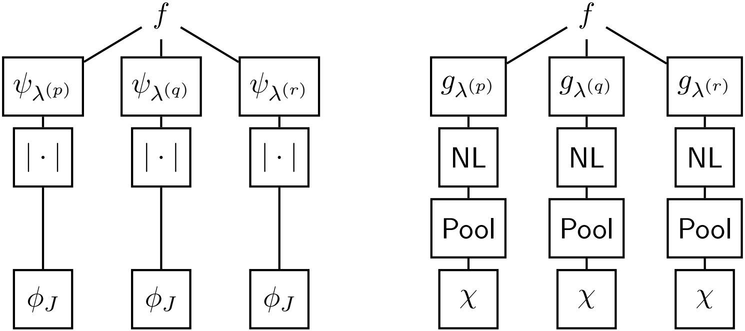
The basic operations between consecutive layers



General DCNs employ different filters, non-linearities, and pooling operations in different network layers [*LeCun et al., 2015*]

Generalizations

The basic operations between consecutive layers



General DCNs employ various output filters [*He et al., 2015*]

General filters: Semi-discrete frames

Observation: Convolutions yield semi-discrete frame coefficients

$$(f * g_\lambda)(b) = \langle f, \overline{g_\lambda(b - \cdot)} \rangle = \langle f, T_b I g_\lambda \rangle, \quad (\lambda, b) \in \Lambda \times \mathbb{R}^d$$

General filters: Semi-discrete frames

Observation: Convolutions yield semi-discrete frame coefficients

$$(f * g_\lambda)(b) = \langle f, \overline{g_\lambda(b - \cdot)} \rangle = \langle f, T_b I g_\lambda \rangle, \quad (\lambda, b) \in \Lambda \times \mathbb{R}^d$$

Definition

Let $\{g_\lambda\}_{\lambda \in \Lambda} \subseteq L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ be indexed by a countable set Λ. The collection

$$\Psi_\Lambda := \big\{T_b I g_\lambda\big\}_{(\lambda, b) \in \Lambda \times \mathbb{R}^d}$$

is a semi-discrete frame for $L^2(\mathbb{R}^d)$, if there exist constants $A, B > 0$ such that

$$A\|f\|_2^2 \leq \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda \rangle|^2 \mathrm{d}b = \sum_{\lambda \in \Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2,$$

for all $f \in L^2(\mathbb{R}^d)$.

General filters: Semi-discrete frames

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda \rangle|^2 \mathrm{d}b = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Semi-discrete frames are rooted in continuous frame theory [*Antoine et al., 1993*], [*Kaiser, 1994*]

General filters: Semi-discrete frames

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda\rangle|^2 \mathrm{d}b = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Semi-discrete frames are rooted in continuous frame theory [*Antoine et al., 1993*], [*Kaiser, 1994*]
- Sampling the translation parameter $b \in \mathbb{R}^d$ in $(T_b I g_\lambda)$ on $\mathbb{Z}^d$ leads to shift-invariant frames [*Ron and Shen, 1995*]

General filters: Semi-discrete frames

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda\rangle|^2 \mathrm{d}b = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Semi-discrete frames are rooted in continuous frame theory [*Antoine et al., 1993*], [*Kaiser, 1994*]
- Sampling the translation parameter $b \in \mathbb{R}^d$ in $(T_b I g_\lambda)$ on $\mathbb{Z}^d$ leads to shift-invariant frames [*Ron and Shen, 1995*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}(\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}^d$$

General filters: Semi-discrete frames

$$A\|f\|_2^2 \leq \sum_{\lambda \in \Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda \rangle|^2 \mathrm{d}b = \sum_{\lambda \in \Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Semi-discrete frames are rooted in continuous frame theory [*Antoine et al., 1993*], [*Kaiser, 1994*]
- Sampling the translation parameter $b \in \mathbb{R}^d$ in $(T_b I g_\lambda)$ on $\mathbb{Z}^d$ leads to shift-invariant frames [*Ron and Shen, 1995*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda \in \Lambda} |\widehat{g_\lambda}(\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}^d$$

- Structured semi-discrete frames: Weyl-Heisenberg frames, wavelets, (α)-curvelets, shearlets, and ridgelets

General filters: Semi-discrete frames

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda} \int_{\mathbb{R}^d} |\langle f, T_b I g_\lambda\rangle|^2 \mathrm{d}b = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Semi-discrete frames are rooted in continuous frame theory [*Antoine et al., 1993*], [*Kaiser, 1994*]
- Sampling the translation parameter $b \in \mathbb{R}^d$ in $(T_b I g_\lambda)$ on $\mathbb{Z}^d$ leads to shift-invariant frames [*Ron and Shen, 1995*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}(\omega)|^2 \leq B, \quad a.e.\ \omega \in \mathbb{R}^d$$

- Structured semi-discrete frames: Weyl-Heisenberg frames, wavelets, (α)-curvelets, shearlets, and ridgelets
- Λ is typically a collection of scales, directions, or frequency shifts

General non-linearities

Observation: Essentially all non-linearities $M : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ employed in the deep learning literature are

i) pointwise, i.e.,

$$(Mf)(x) = \rho(f(x)), \quad x \in \mathbb{R}^d,$$

for some $\rho : \mathbb{C} \to \mathbb{C}$,

ii) Lipschitz-continuous, i.e.,

$$\|M(f) - M(h)\| \leq L\|f - h\|, \quad \forall\, f, h, \in L^2(\mathbb{R}^d),$$

for some $L > 0$,

iii) satisfy $M(f) = 0$ for $f = 0$.

Incorporating pooling by sub-sampling

Pooling by sub-sampling can be emulated in continuous-time by the (unitary) dilation operator

$$f \mapsto R^{d/2} f(R\,\cdot), \quad f \in L^2(\mathbb{R}^d),$$

where $R \geq 1$ is the sub-sampling factor.

Different modules in different layers

Module-sequence $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$

i) in the n-th network layer, replace the wavelet-modulus convolution operation $|f * \psi_\lambda|$ by

$$U_n[\lambda_n]f := R_n^{d/2}(M_n(f * g_{\lambda_n}))(R_n\cdot)$$

Different modules in different layers

Module-sequence $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$

i) in the n-th network layer, replace the wavelet-modulus convolution operation $|f * \psi_\lambda|$ by

$$U_n[\lambda_n]f := R_n^{d/2}(M_n(f * g_{\lambda_n}))(R_n \cdot)$$

ii) extend the operator $U_n[\lambda_n]$ to paths on index sets

$$q = (\lambda_1, \lambda_2, \dots, \lambda_n) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n := \Lambda_1^n, \quad n \in \mathbb{N},$$

according to

$$U[q]f := U_n[\lambda_n] \cdots U_2[\lambda_2] U_1[\lambda_1] f$$

Output filters

- [*Mallat, 2012*] employed the same low-pass filter ϕ_J in every network layer n to generate the output according to

$$\Phi_W^n(f) := \left\{ | \cdots | \, |f * \psi_{\lambda^{(1)}}| * \psi_{\lambda^{(2)}}| \cdots * \psi_{\lambda^{(n)}}| * \phi_J \right\}_{\lambda^{(1)},\dots,\lambda^{(n)} \in \Lambda_W}$$

Output filters

- [*Mallat, 2012*] employed the same low-pass filter ϕ_J in every network layer n to generate the output according to

$$\Phi_W^n(f) := \big\{ |\cdots| \, |f * \psi_{\lambda^{(1)}}| * \psi_{\lambda^{(2)}}| \cdots * \psi_{\lambda^{(n)}}| * \phi_J \big\}_{\lambda^{(1)},\dots,\lambda^{(n)} \in \Lambda_W}$$

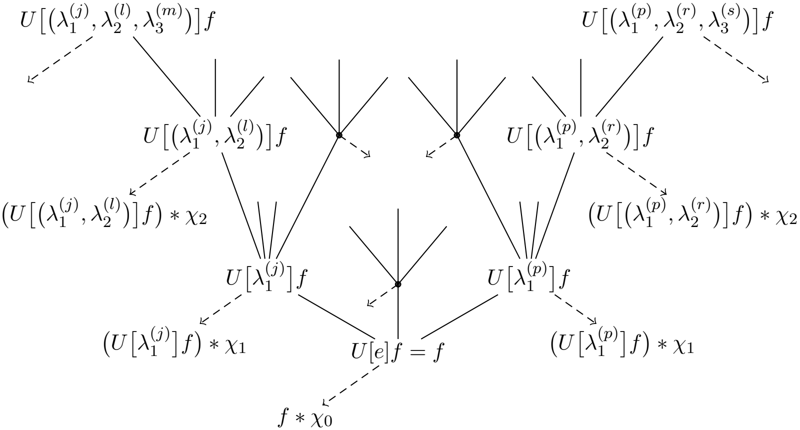
- Here, designate one of the atoms $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ as the output-generating atom $\chi_{n-1} := g_{\lambda_n^*}$, $\lambda_n^* \in \Lambda_n$, of the $(n-1)$-th layer.

  $\Rightarrow$ The atoms $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n \setminus \{\lambda_n^*\}} \cup \{\chi_{n-1}\}$ are used across two consecutive layers!

Generalized feature extractor

Features generated in the n-th network layer

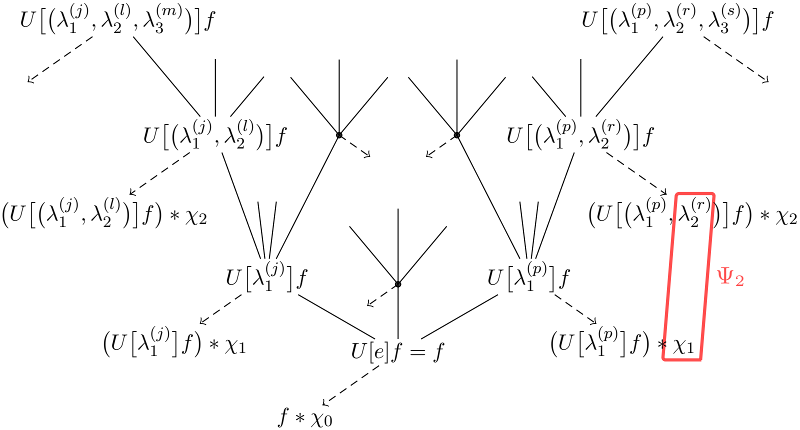
$$\Phi_\Omega^n(f) := \left\{ (U[q]f) * \chi_n \right\}_{q \in \Lambda_1^n}$$



Generalized feature extractor

Features generated in the n-th network layer

$$\Phi_\Omega^n(f) := \left\{ (U[q]f) * \chi_n \right\}_{q \in \Lambda_1^n}$$



Vertical translation invariance

Theorem (Wiatowski and HB, 2015)

Assume that $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ *satisfies the admissibility condition* $B_n \leq \min\{1, L_n^{-2}\}$, *for all* $n \in \mathbb{N}$. *If there exists a constant* $K > 0$ *such that*

$$|\widehat{\chi_n}(\omega)||\omega| \leq K, \qquad \text{a.e. } \omega \in \mathbb{R}^d, \ \forall n \in \mathbb{N}_0,$$

then

$$|||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| \leq \frac{2\pi|t|K}{R_1 \dots R_n} \|f\|_2,$$

for all $f \in L^2(\mathbb{R}^d)$, $t \in \mathbb{R}^d$, $n \in \mathbb{N}$.

Vertical translation invariance

- The admissibility condition

$$B_n \leq \min\{1, L_n^{-2}\}, \quad \forall n \in \mathbb{N},$$

is easily satisfied by normalizing Ψ_n.

Vertical translation invariance

- The admissibility condition

$$B_n \leq \min\{1, L_n^{-2}\}, \quad \forall n \in \mathbb{N},$$

is easily satisfied by normalizing Ψ_n.

- The decay condition

$$|\widehat{\chi_n}(\omega)||\omega| \leq K, \quad \text{a.e. } \omega \in \mathbb{R}^d, \ \forall n \in \mathbb{N}_0,$$

is satisfied, e.g., if $\sup_{n \in \mathbb{N}_0}\{\|\chi_n\|_1 + \|\nabla \chi_n\|_1\} < \infty$.

Vertical translation invariance

- The admissibility condition

$$B_n \leq \min\{1, L_n^{-2}\}, \quad \forall n \in \mathbb{N},$$

  is easily satisfied by normalizing Ψ_n.

- The decay condition

$$|\widehat{\chi_n}(\omega)||\omega| \leq K, \quad \text{a.e. } \omega \in \mathbb{R}^d, \ \forall n \in \mathbb{N}_0,$$

  is satisfied, e.g., if $\sup_{n\in\mathbb{N}_0}\{\|\chi_n\|_1 + \|\nabla\chi_n\|_1\} < \infty$.

- If, in addition, $\lim_{n\to\infty} R_1 \cdot R_2 \cdot \ldots \cdot R_n = \infty$, then

$$\lim_{n\to\infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d.$$

Philosophy behind invariance results

Mallat's "horizontal" translation invariance:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

Philosophy behind invariance results

Mallat's "horizontal" translation invariance:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth

Philosophy behind invariance results

Mallat's "horizontal" translation invariance:

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

- features become invariant in every network layer, but needs $J \to \infty$
- applies to wavelet transform and modulus non-linearity without pooling

"Vertical" translation invariance:

$$\lim_{n\to\infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

- features become more invariant with increasing network depth
- applies to general filters, general non-linearities, and pooling through sub-sampling

Deformation sensitivity bounds

[*Mallat, 2012*] proved that Φ_W is stable w.r.t. non-linear deformations $(F_\tau f)(x) = f(x - \tau(x))$ according to

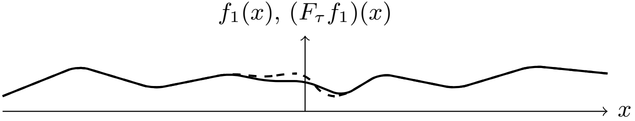
$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

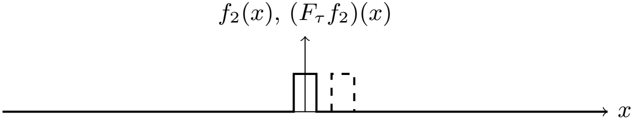
where $H_W := \{\, f \in L^2(\mathbb{R}^d) \mid \|f\|_W < \infty\}$ with

$$\|f\|_W := \sum_{n=0}^{\infty} \Big(\sum_{q \in (\Lambda_W)_1^n} \|U[q]\|_2^2 \Big)^{1/2}$$

Deformation sensitivity for signal classes

Consider $(F_\tau f)(x) = f(x - \tau(x)) = f(x - e^{-x^2})$





For given τ the amount of deformation induced can depend drastically on $f \in L^2(\mathbb{R}^d)$

Deformation sensitivity bounds: Band-limited signals

Theorem (Wiatowski and HB, 2015)

Assume that $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ *satisfies the admissibility condition* $B_n \leq \min\{1, L_n^{-2}\}$, *for all* $n \in \mathbb{N}$. *There exists a constant* $C > 0$ *(that does not depend on* Ω*) such that for all*

$$f \in \{f \in L^2(\mathbb{R}^d) \mid \operatorname{supp}(\hat{f}) \subseteq B_R(0)\}$$

and all $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ *with* $\|D\tau\|_\infty \leq \frac{1}{2d}$, *it holds that*

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq CR\|\tau\|_\infty\|f\|_2.$$

Deformation sensitivity bounds: Cartoon functions

... and what about non-band-limited signals?

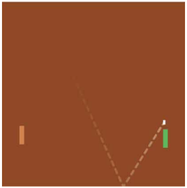


Image credit: middle [*Mnih et al., 2015*], *right* [*Silver et al., 2016*]

Deformation sensitivity bounds: Cartoon functions

... and what about non-band-limited signals?

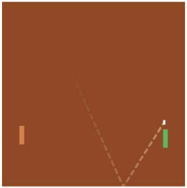
  

Image credit: middle [*Mnih et al., 2015*], *right* [*Silver et al., 2016*]

Take into account structural properties of natural images.

$\Rightarrow$ consider cartoon functions [*Donoho, 2001*]

Deformation sensitivity bounds: Cartoon functions

... and what about non-band-limited signals?

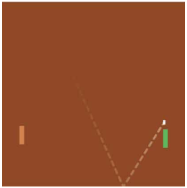


Image credit: middle [*Mnih et al., 2015*], *right* [*Silver et al., 2016*]

The class of cartoon functions of maximal size $K > 0$:

$$\mathcal{C}^{K}_{\mathrm{CART}} := \{f_1 + \mathbb{1}_B f_2 \mid f_i \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d, \mathbb{C}),\ i = 1, 2,$$
$$|\nabla f_i(x)| \le K(1 + |x|^2)^{-d/2},\ \mathrm{vol}^{d-1}(\partial B) \le K,\ \|f_2\|_\infty \le K\}$$

Deformation sensitivity bounds: Cartoon functions

Theorem (Grohs et al., 2016)

Assume that $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ *satisfies the admissibility condition* $B_n \leq \min\{1, L_n^{-2}\}$, *for all* $n \in \mathbb{N}$. *For every* $K > 0$ *there exists a constant* $C_K > 0$ *(that does not depend on* Ω*) such that for all* $f \in \mathcal{C}_{\mathrm{CART}}^{K}$ *and all* $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ *with* $\|\tau\|_\infty < \frac{1}{2}$ *and* $\|D\tau\|_\infty \leq \frac{1}{2d}$, *it holds that*

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_K \|\tau\|_\infty^{1/2}.$$

Deformation sensitivity bounds: Lipschitz functions

Cartoon functions reduce to Lipschitz functions
upon setting $f_2 = 0$ in $f_1 + \mathbb{1}_B f_2 \in \mathcal{C}_{\mathrm{CART}}^K$

Corollary (Grohs et al., 2016)

Assume that $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ *satisfies the admissibility condition* $B_n \leq \min\{1, L_n^{-2}\}$, *for all* $n \in \mathbb{N}$. *For every* $K > 0$ *there exists a constant* $C_K > 0$ *(that does not depend on* Ω*) such that for all*

$$f \in \Big\{ f \in L^2(\mathbb{R}^d) \mid f \text{ Lipschitz-continuous}, \ |\nabla f_i(x)| \leq K(1+|x|^2)^{-d/2} \Big\}$$

and all $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ *with* $\|\tau\|_\infty < \frac{1}{2}$ *and* $\|D\tau\|_\infty \leq \frac{1}{2d}$, *it holds that*

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_K \|\tau\|_\infty.$$

... and what about textures?

... and what about textures?

neither band-limited, nor a cartoon function,
nor Lipschitz-continuous

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The signal class H_W and the corresponding norm $\|\cdot\|_W$ depend on the mother wavelet (and hence the network)

Our deformation sensitivity bound:

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_\mathcal{C}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The signal class $\mathcal{C}$ (band-limited functions or cartoon functions) is independent of the network

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- Signal class description complexity implicit via norm $\|\cdot\|_W$

Our deformation sensitivity bound:

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Signal class description complexity explicit via $C_{\mathcal{C}}$
  - R-band-limited functions: $C_{\mathcal{C}} = \mathcal{O}(R)$
  - cartoon functions of maximal size K: $C_{\mathcal{C}} = \mathcal{O}(K^{3/2})$
  - K-Lipschitz functions $C_{\mathcal{C}} = \mathcal{O}(K)$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound:

$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

Our deformation sensitivity bound:

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- Decay rate $\alpha > 0$ of the deformation error is signal-class-specific (band-limited functions: $\alpha = 1$, cartoon functions: $\alpha = \frac{1}{2}$, Lipschitz functions: $\alpha = 1$)

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound depends explicitly on higher order derivatives of τ

Our deformation sensitivity bound:

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound implicitly depends on derivatives of τ via the condition $\|D\tau\|_\infty \leq \frac{1}{2d}$

Philosophy behind deformation stability/sensitivity bounds

Mallat's deformation stability bound:

$$|||\Phi_W(F_\tau f) - \Phi_W(f)||| \leq C\big(2^{-J}\|\tau\|_\infty + J\|D\tau\|_\infty + \|D^2\tau\|_\infty\big)\|f\|_W,$$

for all $f \in H_W \subseteq L^2(\mathbb{R}^d)$

- The bound is *coupled* to horizontal translation invariance

$$\lim_{J\to\infty} |||\Phi_W(T_t f) - \Phi_W(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d),\ \forall t \in \mathbb{R}^d$$

Our deformation sensitivity bound:

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_{\mathcal{C}}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

- The bound is *decoupled* from vertical translation invariance

$$\lim_{n\to\infty} |||\Phi_\Omega^n(T_t f) - \Phi_\Omega^n(f)||| = 0, \quad \forall f \in L^2(\mathbb{R}^d), \forall t \in \mathbb{R}^d$$

Proof sketch: Decoupling

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_\mathcal{C}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

1) Lipschitz continuity:

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f-h\|_2, \quad \forall f,h \in L^2(\mathbb{R}^d),$$

established through (i) frame property of Ψ_n, (ii) Lipschitz continuity of non-linearities, and (iii) admissibility condition $B_n \leq \min\{1, L_n^{-2}\}$

Proof sketch: Decoupling

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_\mathcal{C} \|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

1) Lipschitz continuity:

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d),$$

established through (i) frame property of Ψ_n, (ii) Lipschitz continuity of non-linearities, and (iii) admissibility condition $B_n \leq \min\{1, L_n^{-2}\}$

2) Signal-class-specific deformation sensitivity bound:

$$\|F_\tau f - f\|_2 \leq C_\mathcal{C} \|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

Proof sketch: Decoupling

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_\mathcal{C}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

1) Lipschitz continuity:

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d),$$

established through (i) frame property of Ψ_n, (ii) Lipschitz continuity of non-linearities, and (iii) admissibility condition $B_n \leq \min\{1, L_n^{-2}\}$

2) Signal-class-specific deformation sensitivity bound:

$$\|F_\tau f - f\|_2 \leq C_\mathcal{C}\|\tau\|_\infty^\alpha, \quad \forall f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$$

3) Combine 1) and 2) to get

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq \|F_\tau f - f\|_2 \leq C_\mathcal{C}\|\tau\|_\infty^\alpha,$$

for all $f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$

Noise robustness

Lipschitz continuity of Φ_Ω according to

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in L^2(\mathbb{R}^d),$$

also implies robustness w.r.t. additive noise $\eta \in L^2(\mathbb{R}^d)$ according to

$$|||\Phi_\Omega(f + \eta) - \Phi_\Omega(f)||| \leq \|\eta\|_2$$

Energy conservation

It is desirable to have

$$f \neq 0 \quad \Rightarrow \quad \Phi(f) \neq 0,$$

or even better

$$|||\Phi(f)||| \geq A_\Phi \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d),$$

for some $A_\Phi > 0$.

Energy conservation

It is desirable to have

$$f \neq 0 \quad \Rightarrow \quad \Phi(f) \neq 0,$$

or even better

$$|||\Phi(f)||| \geq A_\Phi \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d),$$

for some $A_\Phi > 0$.

[*Waldspurger, 2015*] proved—under analyticity assumptions on the mother wavelet—that for real-valued signals $f \in L^2(\mathbb{R}^d)$, Φ_W conserves energy according to

$$|||\Phi_W(f)||| = \|f\|_2$$

Energy conservation

Theorem (Grohs et al., 2016)

Let $\Omega = \big((\Psi_n, |\cdot|, 1)\big)_{n\in\mathbb{N}}$ *be a module-sequence employing modulus non-linearities and no sub-sampling. For every* $n \in \mathbb{N}$*, let the atoms of* Ψ_n *satisfy the following conditions:*

i) $\sum_{\lambda_n\in\Lambda_n\setminus\{\lambda_n^*\}} |\widehat{g_{\lambda_n}}(\omega)|^2 + |\widehat{\chi_{n-1}}(\omega)|^2 = 1$, *a.e.* $\omega \in \mathbb{R}^d$

ii) $\sum_{\lambda_n\in\Lambda_n\setminus\{\lambda_n^*\}} |\widehat{g_{\lambda_n}}(\omega)|^2 = 0$, *a.e.* $\omega \in B_{\delta_n}(0)$, *for some* $\delta_n > 0$

iii) *all atoms* g_{λ_n} *are analytic.*

Then,

$$|||\Phi_\Omega(f)||| = \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d)$$

Energy conservation

Theorem (Grohs et al., 2016)

Let $\Omega = \big((\Psi_n, |\cdot|, 1)\big)_{n\in\mathbb{N}}$ *be a module-sequence employing modulus non-linearities and no sub-sampling. For every* $n \in \mathbb{N}$*, let the atoms of* Ψ_n *satisfy the following conditions:*

i) $\sum_{\lambda_n\in\Lambda_n\setminus\{\lambda_n^*\}} |\widehat{g_{\lambda_n}}(\omega)|^2 + |\widehat{\chi_{n-1}}(\omega)|^2 = 1$, *a.e.* $\omega \in \mathbb{R}^d$

ii) $\sum_{\lambda_n\in\Lambda_n\setminus\{\lambda_n^*\}} |\widehat{g_{\lambda_n}}(\omega)|^2 = 0$, *a.e.* $\omega \in B_{\delta_n}(0)$, *for some* $\delta_n > 0$

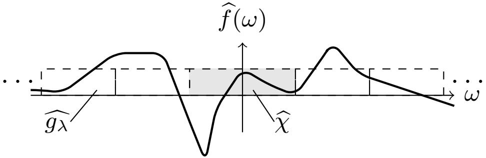
iii) *all atoms* g_{λ_n} *are analytic.*

Then,

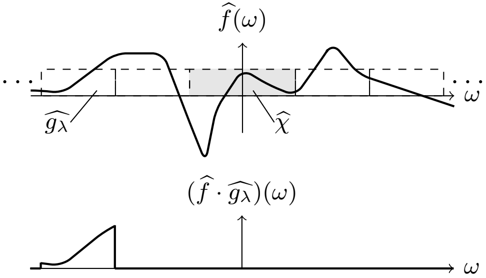
$$|||\Phi_\Omega(f)||| = \|f\|_2, \quad \forall f \in L^2(\mathbb{R}^d)$$

Various structured frames satisfy conditions i)-iii)

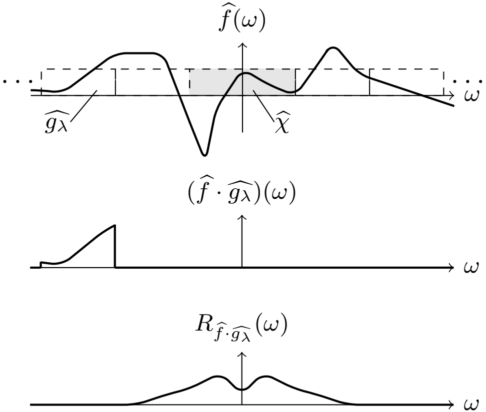
Proof sketch: Energy conservation or "What does the modulus non-linearity do?"



Proof sketch: Energy conservation or "What does the modulus non-linearity do?"

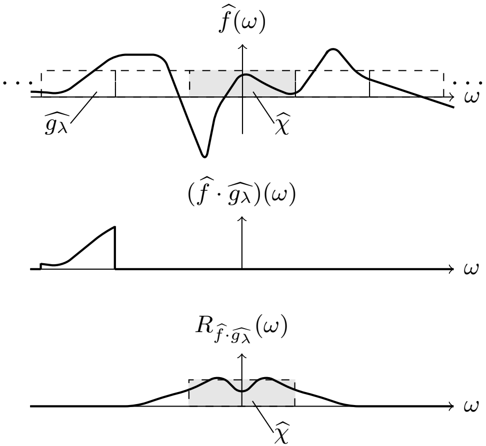


Proof sketch: Energy conservation or “What does the modulus non-linearity do?”



$$|(f * g_\lambda)(x)|^2 \quad \circ\!\!-\!\!\bullet \quad R_{\widehat{f} \cdot \widehat{g_\lambda}}(\omega)$$

Proof sketch: Energy conservation or "What does the modulus non-linearity do?"



$$|f * g_\lambda|^2 * \chi =$$

Two Meta–Theorems

Meta–Theorem

Vertical translation invariance and Lipschitz continuity (hence by decoupling also deformation insensitivity) are guaranteed by the network structure per se rather than the specific convolution kernels, non-linearities, and pooling operations.

Two Meta–Theorems

Meta–Theorem

Vertical translation invariance and Lipschitz continuity (hence by decoupling also deformation insensitivity) are guaranteed by the network structure per se rather than the specific convolution kernels, non-linearities, and pooling operations.

Meta–Theorem

For networks employing the modulus non-linearity and no intra-layer pooling, energy conservation is guaranteed for quite general convolution kernels.

Deep Frame Net

Open source software:

- MATLAB: `http://www.nari.ee.ethz.ch/commth/research`
- Python: Coming soon!

The Mathematics of Deep Learning
Part 2: Discrete-time Theory

Helmut Bőlcskei

ETH zürich

Department of Information Technology and Electrical Engineering

June 2016

joint work with Thomas Wiatowski, Michael Tschannen, and Philipp Grohs

Continuous-time theory

[*Mallat, 2012*] and [*Wiatowski and HB, 2015*] developed a continuous-time theory for feature extraction through DCNs:

- translation invariance results for $L^2(\mathbb{R}^d)$-functions
- deformation sensitivity bounds for signal classes $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$
- energy conservation for $L^2(\mathbb{R}^d)$-functions

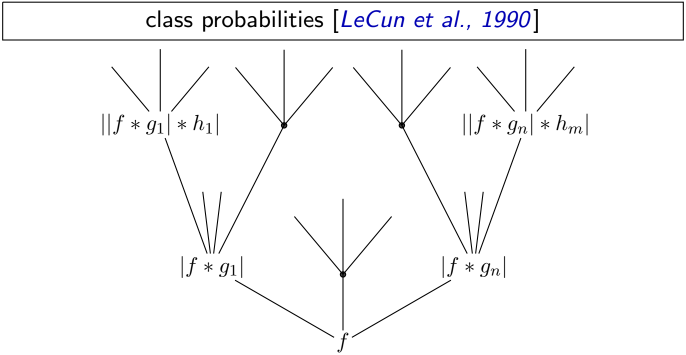
Practice is digital

In practice ... we need to handle discrete data!

$f =$  $\in \mathbb{R}^{n \times n}$

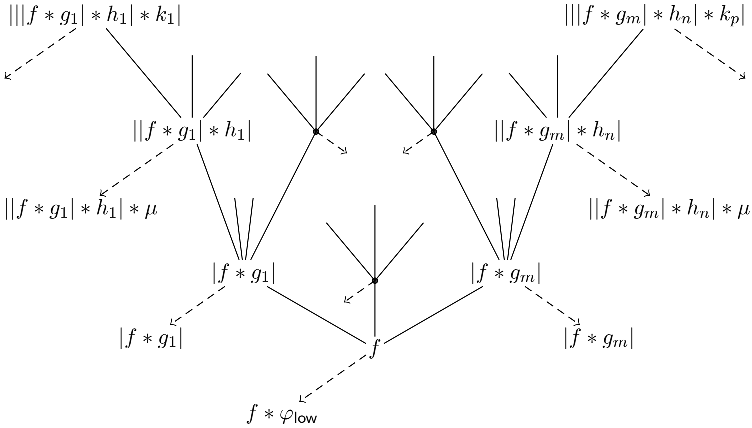
Practice is digital

In practice ... a wide variety of network architectures is used!



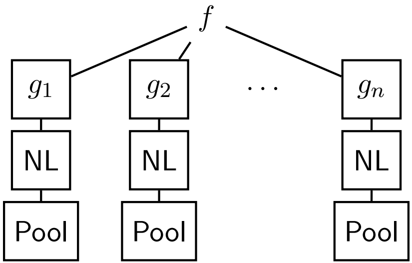
Practice is digital

In practice ... a wide variety of network architectures is used!



Architecture of general DCNs

The basic operations between consecutive layers

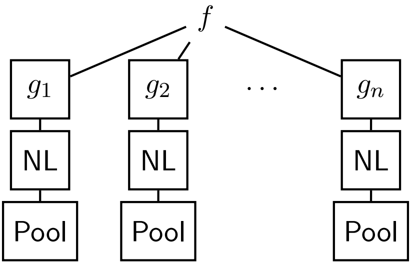


DCNs employ a wide variety of filters g_k

- pre-specified and structured (e.g., wavelets [*Serre et al., 2005*])
- pre-specified and unstructured (e.g., random filters [*Jarrett et al., 2009*])
- learned in a supervised [*Huang and LeCun, 2006*] or an unsupervised [*Ranzato et al., 2007*] fashion

Architecture of general DCNs

The basic operations between consecutive layers

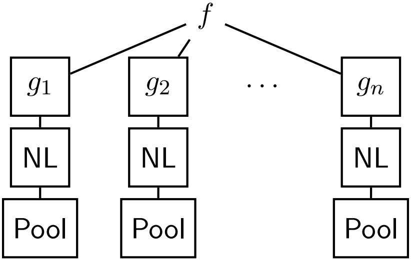


DCNs employ a wide variety of non-linearities

- modulus [*Mutch and Lowe, 2006*]
- hyperbolic tangent [*Huang and LeCun, 2006*]
- rectified linear unit [*Nair and Hinton, 2010*]
- logistic sigmoid [*Glorot and Bengio, 2010*]

Architecture of general DCNs

The basic operations between consecutive layers

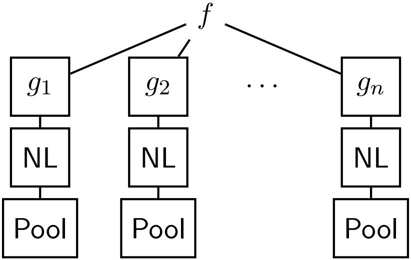


DCNs employ pooling

- sub-sampling [*Pinto et al., 2008*]
- average pooling [*Jarrett et al., 2009*]
- max-pooling [*Ranzato et al., 2007*]

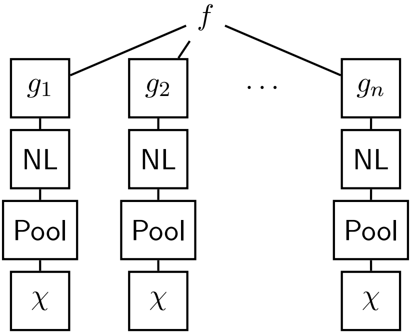
Architecture of general DCNs

The basic operations between consecutive layers



DCNs employ different filters, non-linearities, and pooling operations in different network layers [*LeCun et al., 2015*]

Architecture of general DCNs



Which layers contribute to the network's output?

- the last layer only (e.g., class probabilities [*LeCun et al., 1990*])
- subset of layers (e.g., shortcut connections [*He et al., 2015*])
- all layers (e.g., low-pass filtering [*Bruna and Mallat, 2013*])

Challenges

Challenges for discrete theory:

- flexible architectures

Challenges

Challenges for discrete theory:

- flexible architectures
- signals of varying dimensions are propagated through the network

Challenges

Challenges for discrete theory:

- flexible architectures
- signals of varying dimensions are propagated through the network
- how to incorporate general pooling operators into the theory?

Challenges

Challenges for discrete theory:

- flexible architectures
- signals of varying dimensions are propagated through the network
- how to incorporate general pooling operators into the theory?
- can not rely on asymptotics (finite network depth) to prove network properties (e.g., translation invariance)

Challenges

Challenges for discrete theory:

- flexible architectures
- signals of varying dimensions are propagated through the network
- how to incorporate general pooling operators into the theory?
- can not rely on asymptotics (finite network depth) to prove network properties (e.g., translation invariance)
- nature is analog

Challenges

Challenges for discrete theory:

- flexible architectures
- signals of varying dimensions are propagated through the network
- how to incorporate general pooling operators into the theory?
- can not rely on asymptotics (finite network depth) to prove network properties (e.g., translation invariance)
- nature is analog
- what are appropriate signal classes to be considered?

Definitions

Signal space

$$H_N := \{f : \mathbb{Z} \to \mathbb{C} \mid f[n] = f[n+N],\ \forall\, n \in \mathbb{Z}\}$$

p-Norm

$$\|f\|_p := \Big(\sum_{n \in I_N} |f[n]|^p \Big)^{1/p}, \qquad I_N := \{0, \ldots, N-1\}$$

Circular convolution

$$(f * g)[n] := \sum_{k \in I_N} f[k] g[n-k], \qquad f, g \in H_N$$

Discrete Fourier transform

$$\widehat{f}\,[k] := \sum_{n \in I_N} f[n] e^{-2\pi i k n / N}, \qquad f \in H_N$$

Filters: Shift-invariant frames for H_N

Observation: Convolutions yield shift-invariant frame coefficients

$$(f * g_\lambda)[n] = \langle f, \overline{g_\lambda(n - \cdot)} \rangle = \langle f, T_n I g_\lambda \rangle, \quad (\lambda, n) \in \Lambda \times I_N$$

Filters: Shift-invariant frames for H_N

Observation: Convolutions yield shift-invariant frame coefficients

$$(f * g_\lambda)[n] = \langle f, \overline{g_\lambda(n - \cdot)} \rangle = \langle f, T_n I g_\lambda \rangle, \quad (\lambda, n) \in \Lambda \times I_N$$

Definition

Let $\{g_\lambda\}_{\lambda \in \Lambda} \subseteq H_N$ be indexed by a finite set Λ. The collection

$$\Psi_\Lambda := \big\{T_n I g_\lambda\big\}_{(\lambda,n)\in\Lambda\times I_N}$$

is a shift-invariant frame for H_N, if there exist constants $A, B > 0$ such that

$$A\|f\|_2^2 \le \sum_{\lambda\in\Lambda} \sum_{n\in I_N} |\langle f, T_n I g_\lambda \rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \le B\|f\|_2^2,$$

for all $f \in H_N$

Filters: Shift-invariant frames for H_N

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

Filters: Shift-invariant frames for H_N

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Shift-invariant frames for $L^2(\mathbb{R}^d)$ [*Ron and Shen, 1995*], for $\ell^2(\mathbb{Z})$ [*HB et al., 1998*] and [*Cvetković and Vetterli, 1998*]

Filters: Shift-invariant frames for H_N

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Shift-invariant frames for $L^2(\mathbb{R}^d)$ [*Ron and Shen, 1995*], for $\ell^2(\mathbb{Z})$ [*HB et al., 1998*] and [*Cvetković and Vetterli, 1998*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2 \leq B, \quad \forall\, k \in I_N$$

Filters: Shift-invariant frames for H_N

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Shift-invariant frames for $L^2(\mathbb{R}^d)$ [*Ron and Shen, 1995*], for $\ell^2(\mathbb{Z})$ [*HB et al., 1998*] and [*Cvetković and Vetterli, 1998*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2 \leq B, \quad \forall\, k \in I_N$$

- Frame lower bound $A > 0$ guarantees that no essential features of f are "lost" in the network

Filters: Shift-invariant frames for H_N

$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Shift-invariant frames for $L^2(\mathbb{R}^d)$ [*Ron and Shen, 1995*], for $\ell^2(\mathbb{Z})$ [*HB et al., 1998*] and [*Cvetković and Vetterli, 1998*]
- The frame condition can equivalently be expressed as

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2 \leq B, \quad \forall\, k \in I_N$$

- Frame lower bound $A > 0$ guarantees that no essential features of f are "lost" in the network
- Structured shift-invariant frames: Weyl-Heisenberg frames, wavelets, (α)-curvelets, shearlets, and ridgelets

Filters: Shift-invariant frames for H_N

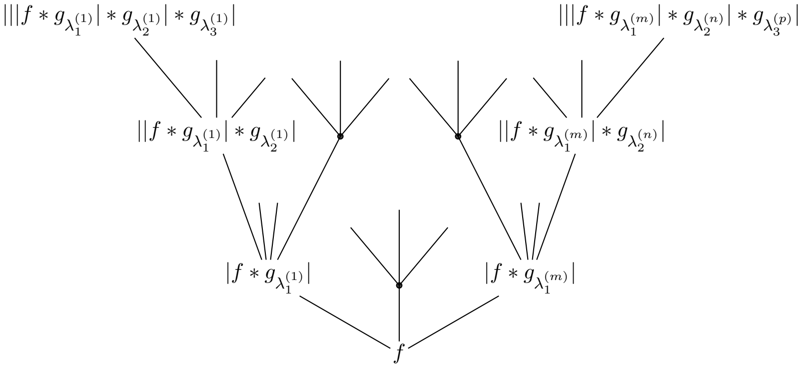
$$A\|f\|_2^2 \leq \sum_{\lambda\in\Lambda}\sum_{n\in I_N} |\langle f, T_n I g_\lambda\rangle|^2 = \sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2$$

- Shift-invariant frames for $L^2(\mathbb{R}^d)$ [*Ron and Shen, 1995*], for $\ell^2(\mathbb{Z})$ [*HB et al., 1998*] and [*Cvetković and Vetterli, 1998*]
- The frame condition can equivalently be expressed as

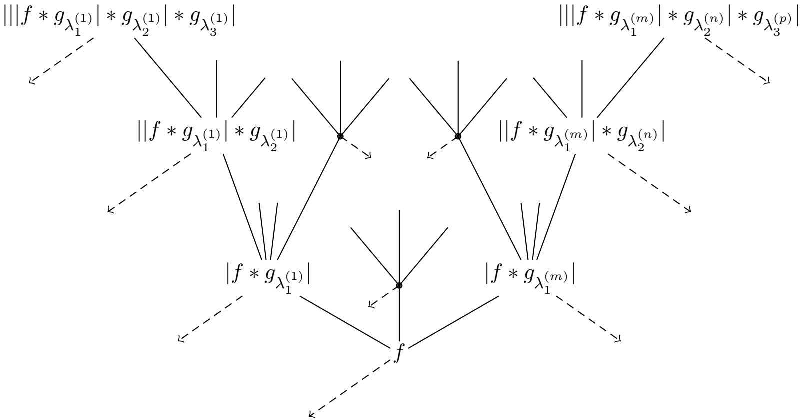
$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2 \leq B, \quad \forall\, k \in I_N$$

- Frame lower bound $A > 0$ guarantees that no essential features of f are "lost" in the network
- Structured shift-invariant frames: Weyl-Heisenberg frames, wavelets, (α)-curvelets, shearlets, and ridgelets
- Λ is typically a collection of scales, directions, or frequency shifts

Filters: Shift-invariant frames for H_N

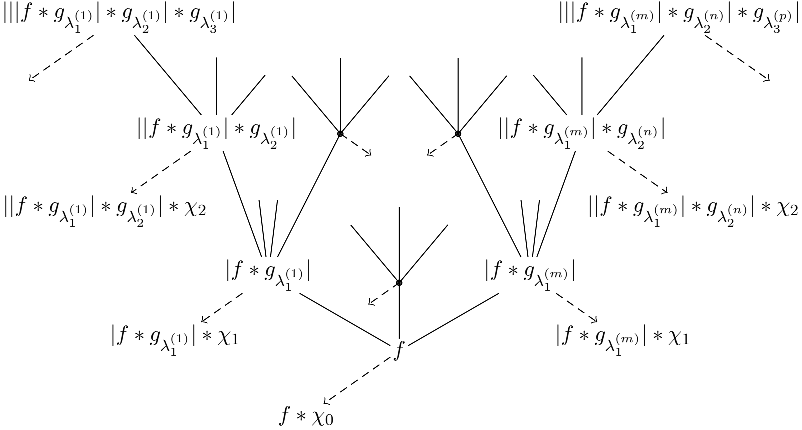


Filters: Shift-invariant frames for H_N



How to generate network output in the d-th layer?

Filters: Shift-invariant frames for H_N



How to generate network output in the d-th layer?
Convolution with general $\chi_d \in H_{N_{d+1}}$ gives flexibility!

Network output

A wide variety of architectures is encompassed, e.g.,

- output: none
  $\Rightarrow \chi_d = 0$
- output: propagated signals $|\cdots|f * g_{\lambda_1^{(m)}}| * \cdots * g_{\lambda_d^{(n)}}|$
  $\Rightarrow \chi_d = \delta$
- output: filtered signals
  $\Rightarrow \chi_d =$ filter (e.g., low-pass)

Network output

A wide variety of architectures is encompassed, e.g.,

- output: none
  $\Rightarrow \chi_d = 0$
- output: propagated signals $|\cdots|f * g_{\lambda_1^{(m)}}| * \cdots * g_{\lambda_d^{(n)}}|$
  $\Rightarrow \chi_d = \delta$
- output: filtered signals
  $\Rightarrow \chi_d =$ filter (e.g., low-pass)

$\Rightarrow \Psi_{d+1} \cup \{T_n I \chi_d\}_{n \in I_{N_{d+1}}}$ forms a shift-invariant frame for $H_{N_{d+1}}$

Start with

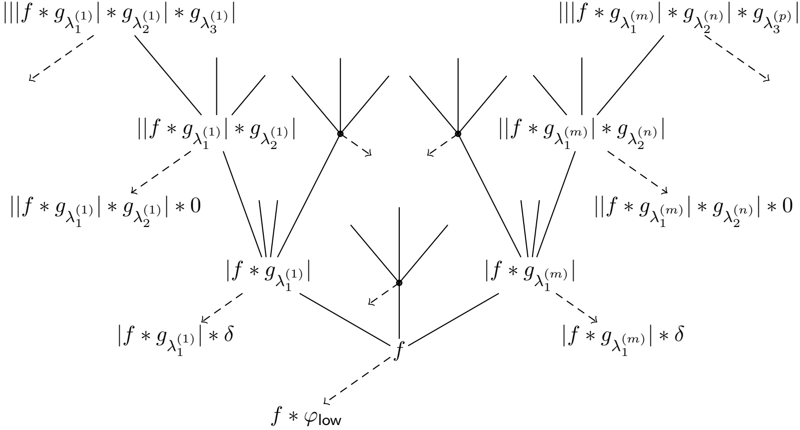
$$A_{d+1} \leq \sum_{\lambda_{d+1} \in \Lambda_{d+1}} |\widehat{g_{\lambda_{d+1}}}[k]|^2 \leq B_{d+1}, \quad \forall\, k \in I_{N_{d+1}},$$

and note that

$$A_{d+1} \leq |\widehat{\chi_d}[k]|^2 + \sum_{\lambda_{d+1} \in \Lambda_{d+1}} |\widehat{g_{\lambda_{d+1}}}[k]|^2 \leq B'_{d+1}, \quad \forall\, k \in I_{N_{d+1}}$$

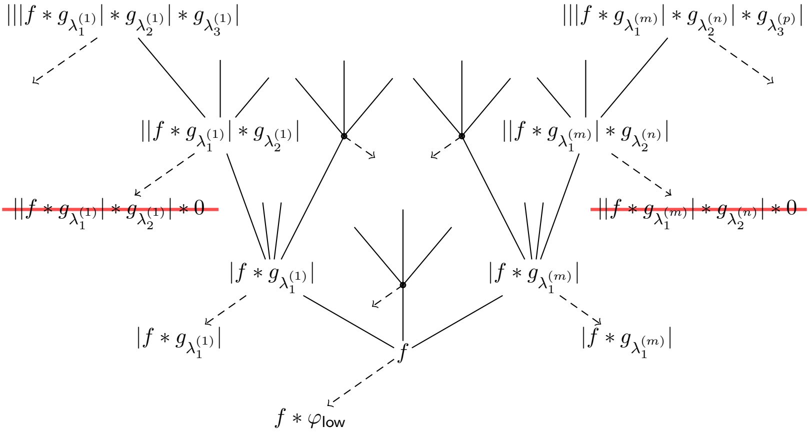
Filters: Shift-invariant frames for H_N

A wide variety of architectures is encompassed!



Filters: Shift-invariant frames for H_N

A wide variety of architectures is encompassed!



Non-linearities

Observation: Essentially all non-linearities $\rho : H_N \to H_N$ employed in the deep learning literature are

i) pointwise, i.e.,

$$(\rho f)[n] = \rho(f[n]), \quad n \in I_N,$$

ii) Lipschitz-continuous, i.e.,

$$\|\rho(f) - \rho(h)\|_2 \leq L\|f - h\|_2, \quad \forall\, f, h \in H_N,$$

for some $L > 0$

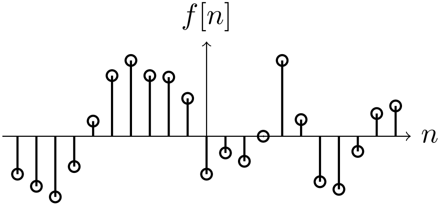
Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Averaging:

$$(Pf)[n] = \sum_{k=Sn}^{Sn+S-1} \alpha_{k-Sn} f[k]$$

- weights $\{\alpha_k\}_{k=0}^{S-1}$ can be learned [*LeCun et al., 1998*] or be pre-specified [*Pinto et al., 2008*]
- uniform averaging corresponds to $\alpha_k = \frac{1}{S}$, for $k \in \{0, \ldots, S-1\}$



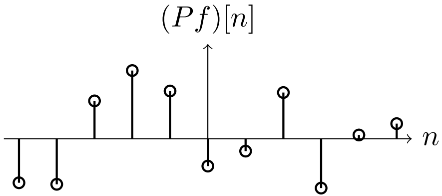
Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Averaging:

$$(Pf)[n] = \sum_{k=Sn}^{Sn+S-1} \alpha_{k-Sn} f[k]$$

- weights $\{\alpha_k\}_{k=0}^{S-1}$ can be learned [*LeCun et al., 1998*] or be pre-specified [*Pinto et al., 2008*]
- uniform averaging corresponds to $\alpha_k = \frac{1}{S}$, for $k \in \{0, \ldots, S-1\}$

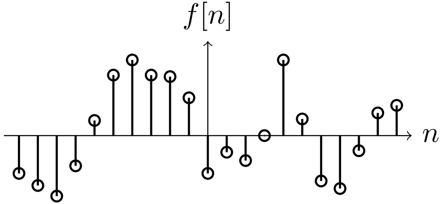


Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Maximization:

$$(Pf)[n] = \max_{k \in \{nS, \ldots, nS+S-1\}} |f[k]|$$

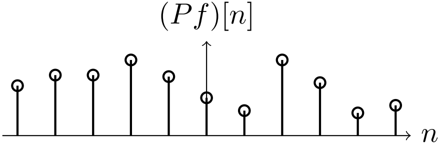


Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Maximization:

$$(Pf)[n] = \max_{k \in \{nS, \dots, nS+S-1\}} |f[k]|$$



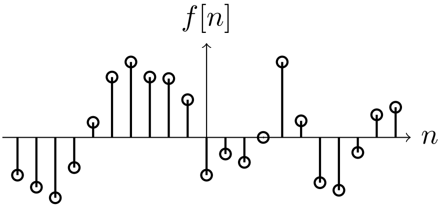
Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Sub-sampling:

$$(Pf)[n] = f[Sn]$$

- $S = 1$ corresponds to "no pooling"



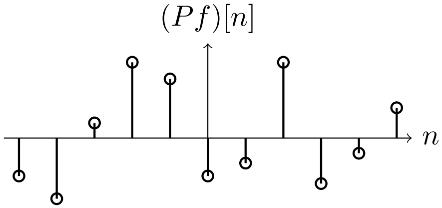
Pooling $P : H_N \to H_{N/S}$

Pooling: Combining nearby values / picking one representative value

Sub-sampling:

$$(Pf)[n] = f[Sn]$$

- $S = 1$ corresponds to "no pooling"



Pooling

Common to all pooling operators P_d:

- Lipschitz continuity with Lipschitz constant R_d:
  - averaging: $R_d = S_d^{1/2} \max_{k \in \{0,\dots,S_d-1\}} |\alpha_k^d|$
  - maximization: $R_d = 1$
  - sub-sampling: $R_d = 1$

Pooling

Common to all pooling operators P_d:

- Lipschitz continuity with Lipschitz constant R_d:
  - averaging: $R_d = S_d^{1/2} \max_{k \in \{0,\dots,S_d-1\}} |\alpha_k^d|$
  - maximization: $R_d = 1$
  - sub-sampling: $R_d = 1$
- Pooling factor S_d:
  - "size" of the neighborhood values are combined in
  - dimensionality-reduction from d-th to $(d+1)$-th layer, i.e., $N_{d+1} = \frac{N_d}{S_d}$

Different modules in different layers

Module-sequence $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$

i) in the d-th network layer, we compute

$$U_d[\lambda_d]f := P_d(\rho_d(f * g_{\lambda_d}))$$

Different modules in different layers

Module-sequence $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$

i) in the d-th network layer, we compute

$$U_d[\lambda_d]f := P_d(\rho_d(f * g_{\lambda_d}))$$

ii) extend the operator $U_d[\lambda_d]$ to paths on index sets

$$q = (\lambda_1, \lambda_2, \dots, \lambda_d) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_d := \Lambda_1^d, \quad d \in \{1, \dots, D\},$$

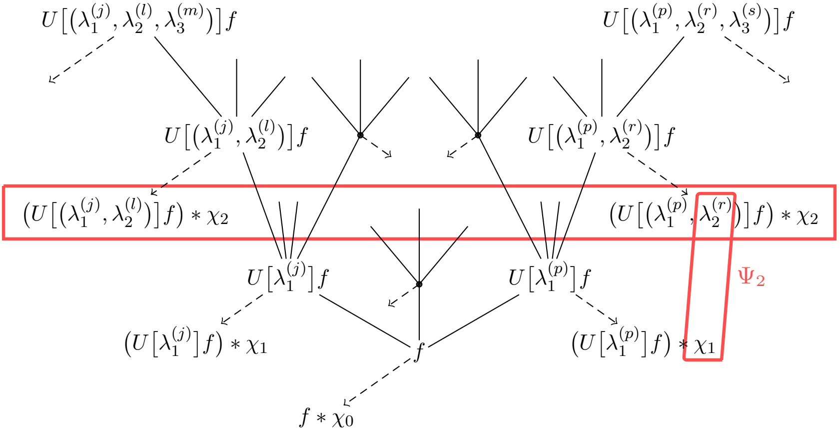
according to

$$U[q]f := U_d[\lambda_d] \cdots U_2[\lambda_2] U_1[\lambda_1] f$$

Local and global properties

Features generated in the d-th network layer

$$\Phi_\Omega^d(f) := \left\{(U[q]f) * \chi_d\right\}_{q\in\Lambda_1^d}$$



Global properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

Assume that $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$ *satisfies the admissibility condition* $B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}$, *for all* $d \in \{1, \ldots, D\}$. *Then, the feature extractor is Lipschitz-continuous, i.e.,*

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

Global properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

Assume that $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$ *satisfies the admissibility condition* $B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}$, *for all* $d \in \{1, \ldots, D\}$. *Then, the feature extractor is Lipschitz-continuous, i.e.,*

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

... this implies ...

- robustness w.r.t. additive noise $\eta \in L^2(\mathbb{R}^d)$ according to

$$|||\Phi_\Omega(f + \eta) - \Phi_\Omega(f)||| \leq \|\eta\|_2, \quad \forall f \in H_{N_1}$$

Global properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

Assume that $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$ *satisfies the admissibility condition* $B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}$, *for all* $d \in \{1, \dots, D\}$. *Then, the feature extractor is Lipschitz-continuous, i.e.,*

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

... this implies ...

- robustness w.r.t. additive noise $\eta \in L^2(\mathbb{R}^d)$ according to

$$|||\Phi_\Omega(f + \eta) - \Phi_\Omega(f)||| \leq \|\eta\|_2, \quad \forall f \in H_{N_1}$$

- an upper bound on the feature vector's energy according to

$$|||\Phi_\Omega(f)||| \leq \|f\|_2, \quad \forall f \in H_{N_1}$$

Global properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

Assume that $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$ *satisfies the admissibility condition* $B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}$, *for all* $d \in \{1, \dots, D\}$. *Then, the feature extractor is Lipschitz-continuous, i.e.,*

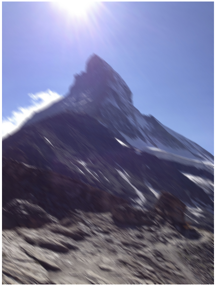
$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \leq \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

The admissibility condition

$$B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}, \quad \forall d \in \{1, \dots, D\},$$

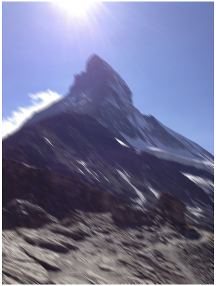
is easily satisfied by normalizing the frame elements in Ψ_d

Global properties: Deformation sensitivity bounds



- Network output should be independent of cameras (of different resolutions), and insensitive to small acquisition jitters

Global properties: Deformation sensitivity bounds



- Network output should be independent of cameras (of different resolutions), and insensitive to small acquisition jitters
- $\Rightarrow$ Want to analyze sensitivity w.r.t. continuous-time deformations

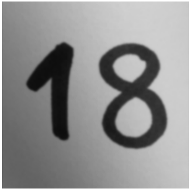
$$(F_\tau f)(x) = f(x - \tau(x)), \quad x \in \mathbb{R},$$

and hence consider

$$(F_\tau f)[n] = f(n/N - \tau(n/N)), \quad n \in I_N$$

Global properties: Deformation sensitivity bounds

Goal: Deformation sensitivity bounds for practically relevant signal classes



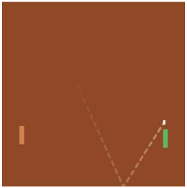
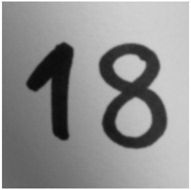


Image credit: middle [Mnih et al., 2015], right [Silver et al., 2016]

Global properties: Deformation sensitivity bounds

Goal: Deformation sensitivity bounds for practically relevant signal classes



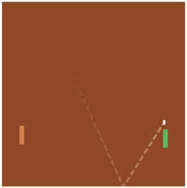


Image credit: middle [*Mnih et al., 2015*], *right* [*Silver et al., 2016*]

Take into account structural properties of natural images

$\Rightarrow$ consider cartoon functions [*Donoho, 2001*]

Global properties: Deformation sensitivity bounds

Goal: Deformation sensitivity bounds for practically relevant signal classes

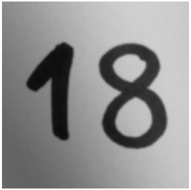
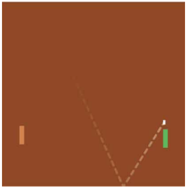




Image credit: middle [*Mnih et al., 2015*], *right* [*Silver et al., 2016*]

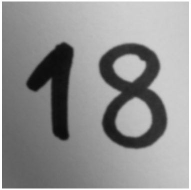
Continuous-time [*Donoho, 2001*]:

Cartoon functions of maximal variation $K > 0$:

$$\mathcal{C}^K_{\mathrm{CART}} := \{c_1 + \mathbb{1}_{[a,b]} c_2 \mid |c_i(x) - c_i(y)| \leq K|x - y|, \\ \forall\, x, y \in \mathbb{R},\ i = 1, 2,\ \|c_2\|_\infty \leq K\}$$

Global properties: Deformation sensitivity bounds

Goal: Deformation sensitivity bounds for practically relevant signal classes



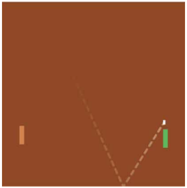


Image credit: middle [Mnih et al., 2015], right [Silver et al., 2016]

Discrete-time [*Wiatowski et al., 2016*]:

Sampled cartoon functions of length N and maximal variation $K > 0$:

$$\mathcal{C}_{\mathrm{CART}}^{N,K} := \left\{ f[n] = c(n/N),\ n \in I_N \ \middle|\ c = (c_1 + \mathbb{1}_{[a,b]} c_2) \in \mathcal{C}_{\mathrm{CART}}^{K} \right\}$$

Global properties: Deformation sensitivity bounds

Theorem (Wiatowski et al., 2016)

Assume that $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{d=1}^{D}$ *satisfies the admissibility condition* $B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}$, *for all* $d \in \{1, \ldots, D\}$. *For every* $N_1 \in \mathbb{N}$, *every* $K > 0$, *and every* $\tau : [0,1] \to [-1,1]$, *it holds that*

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq 4 K N_1^{1/2} \|\tau\|_\infty^{1/2},$$

for all $f \in \mathcal{C}_{\mathrm{CART}}^{N_1, K}$.

Philosophy behind deformation sensitivity bounds

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq 4KN_1^{1/2}\|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

- Bound depends explicitly on the analog signal's description complexity via K and N_1

Philosophy behind deformation sensitivity bounds

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq 4KN_1^{1/2}\|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

- Bound depends explicitly on the analog signal's description complexity via K and N_1
- Lipschitz exponent $\alpha = \frac{1}{2}$ for $\|\tau\|_\infty$ is signal-class-specific (*larger* Lipschitz exponents for *smoother* functions)

Philosophy behind deformation sensitivity bounds

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq 4KN_1^{1/2}\|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

- Bound depends explicitly on the analog signal's description complexity via K and N_1
- Lipschitz exponent $\alpha = \frac{1}{2}$ for $\|\tau\|_\infty$ is signal-class-specific (*larger* Lipschitz exponents for *smoother* functions)
- Particularizing to translations: $\tau_t(x) = t$, $x \in [0,1]$, results in *translation sensitivity* bound according to

$$|||\Phi_\Omega(F_{\tau_t} f) - \Phi_\Omega(f)||| \leq 4KN_1^{1/2}|t|^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

Global properties: Energy conservation

Theorem (Wiatowski et al., 2016)

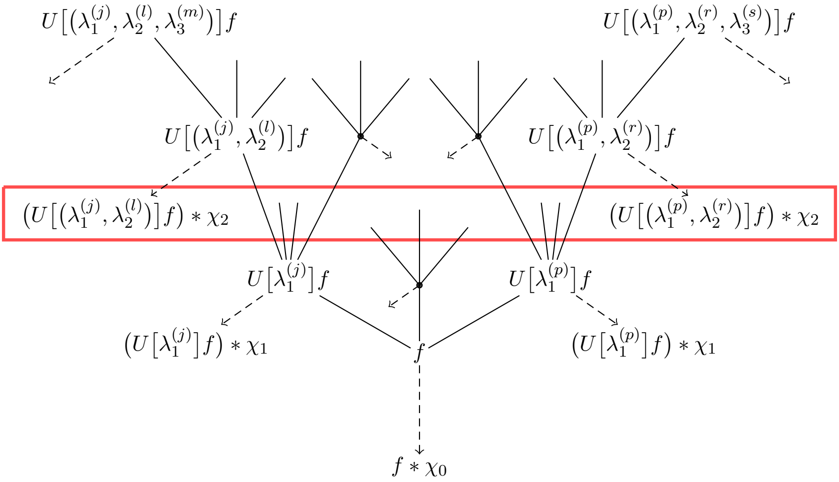
Let $\Omega = \big((\Psi_n, |\cdot|, P_{S=1}^{sub})\big)_{n\in\mathbb{N}}$ *be a module-sequence employing modulus non-linearities and no pooling. For every* $d \in \{1,\dots,D\}$*, let the atoms of* Ψ_d *satisfy*

$$\sum_{\lambda_d\in\Lambda_d} |\widehat{g_{\lambda_d}}[k]|^2 + |\widehat{\chi_{d-1}}[k]|^2 = 1, \quad \forall k \in I_{N_d}.$$

Let the output-generating atom of the last layer be the delta function, i.e., $\chi_{D-1} = \delta$*, then*

$$|||\Phi_\Omega(f)||| = \|f\|_2, \quad \forall f \in H_{N_1}.$$

Local properties



Local properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^d B_k L_k^2 R_k^2 \Big)^{1/2},$$

i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \leq L_\Omega^d \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

Local properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^d B_k L_k^2 R_k^2 \Big)^{1/2},$$

i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \le L_\Omega^d \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

The Lipschitz constant L_Ω^d

- determines the noise sensitivity of $\Phi_\Omega^d(f)$ according to

$$|||\Phi_\Omega^d(f+\eta) - \Phi_\Omega^d(f)||| \le L_\Omega^d \|\eta\|_2, \quad \forall f \in H_{N_1}$$

Local properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^d B_k L_k^2 R_k^2 \Big)^{1/2},$$

i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \le L_\Omega^d \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

The Lipschitz constant L_Ω^d

- impacts the energy of $\Phi_\Omega^d(f)$ according to

$$|||\Phi_\Omega^d(f)||| \le L_\Omega^d \|f\|_2, \quad \forall f \in H_{N_1}$$

Local properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^d B_k L_k^2 R_k^2 \Big)^{1/2},$$

i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \leq L_\Omega^d \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

The Lipschitz constant L_Ω^d

- quantifies the impact of deformations τ according to

$$|||\Phi_\Omega^d(F_\tau f) - \Phi_\Omega^d(f)||| \leq 4 L_\Omega^d K N_1^{1/2} \|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1, K}$$

Local properties: Lipschitz continuity

Theorem (Wiatowski et al., 2016)

The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^{d} B_k L_k^2 R_k^2 \Big)^{1/2},$$

i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \le L_\Omega^d \|f - h\|_2, \quad \forall f, h \in H_{N_1}.$$

The Lipschitz constant L_Ω^d

- is hence a characteristic constant for the features $\Phi_\Omega^d(f)$ generated in the d-th network layer

Local properties: Lipschitz continuity

$$L_\Omega^d = \frac{\|\chi_d\|_1 B_d^{1/2} L_d R_d}{\|\chi_{d-1}\|_1} L_\Omega^{d-1}$$

If $\|\chi_d\|_1 < \frac{\|\chi_{d-1}\|_1}{B_d^{1/2} L_d R_d}$, then $L_\Omega^d < L_\Omega^{d-1}$, and hence

- the features $\Phi_\Omega^d(f)$ are less deformation-sensitive than $\Phi_\Omega^{d-1}(f)$, thanks to

$$|||\Phi_\Omega^d(F_\tau f) - \Phi_\Omega^d(f)||| \leq 4 L_\Omega^d K N_1^{1/2} \|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

Local properties: Lipschitz continuity

$$L_\Omega^d = \frac{\|\chi_d\|_1 B_d^{1/2} L_d R_d}{\|\chi_{d-1}\|_1} L_\Omega^{d-1}$$

If $\|\chi_d\|_1 < \frac{\|\chi_{d-1}\|_1}{B_d^{1/2} L_d R_d}$, then $L_\Omega^d < L_\Omega^{d-1}$, and hence

- the features $\Phi_\Omega^d(f)$ are less deformation-sensitive than $\Phi_\Omega^{d-1}(f)$, thanks to

$$|||\Phi_\Omega^d(F_\tau f) - \Phi_\Omega^d(f)||| \leq 4 L_\Omega^d K N_1^{1/2} \|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

- the features $\Phi_\Omega^d(f)$ contain less energy than $\Phi_\Omega^{d-1}(f)$, owing to

$$|||\Phi_\Omega^d(f)||| \leq L_\Omega^d \|f\|_2, \quad \forall f \in H_{N_1}$$

Local properties: Lipschitz continuity

$$L_\Omega^d = \frac{\|\chi_d\|_1 B_d^{1/2} L_d R_d}{\|\chi_{d-1}\|_1} L_\Omega^{d-1}$$

If $\|\chi_d\|_1 < \frac{\|\chi_{d-1}\|_1}{B_d^{1/2} L_d R_d}$, then $L_\Omega^d < L_\Omega^{d-1}$, and hence

- the features $\Phi_\Omega^d(f)$ are less deformation-sensitive than $\Phi_\Omega^{d-1}(f)$, thanks to

$$|||\Phi_\Omega^d(F_\tau f) - \Phi_\Omega^d(f)||| \leq 4 L_\Omega^d K N_1^{1/2} \|\tau\|_\infty^{1/2}, \quad \forall f \in \mathcal{C}_{\mathrm{CART}}^{N_1,K}$$

- the features $\Phi_\Omega^d(f)$ contain less energy than $\Phi_\Omega^{d-1}(f)$, owing to

$$|||\Phi_\Omega^d(f)||| \leq L_\Omega^d \|f\|_2, \quad \forall f \in H_{N_1}$$

$\Rightarrow$ Tradeoff between deformation sensitivity and energy preservation!

Local properties: Covariance-Invariance

Theorem (Wiatowski et al., 2016)

Let $\{S_k\}_{k=1}^{d}$ *be pooling factors. The features generated in the* d*-th network layer are translation-covariant according to*

$$\Phi_\Omega^d(T_m f) = T_{\frac{m}{S_1 \ldots S_d}} \Phi_\Omega^d(f),$$

for all $f \in H_{N_1}$ *and all* $m \in \mathbb{Z}$ *with* $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$*.*

Local properties: Covariance-Invariance

Theorem (Wiatowski et al., 2016)

Let $\{S_k\}_{k=1}^{d}$ be pooling factors. The features generated in the d-th network layer are translation-covariant according to

$$\Phi^d_\Omega(T_m f) = T_{\frac{m}{S_1 \ldots S_d}} \Phi^d_\Omega(f),$$

for all $f \in H_{N_1}$ and all $m \in \mathbb{Z}$ with $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$.

- Translation covariance on signal grid induced by the pooling factors

Local properties: Covariance-Invariance

Theorem (Wiatowski et al., 2016)

Let $\{S_k\}_{k=1}^{d}$ be pooling factors. The features generated in the d-th network layer are translation-covariant according to

$$\Phi_\Omega^d(T_m f) = T_{\frac{m}{S_1 \ldots S_d}} \Phi_\Omega^d(f),$$

for all $f \in H_{N_1}$ and all $m \in \mathbb{Z}$ with $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$.

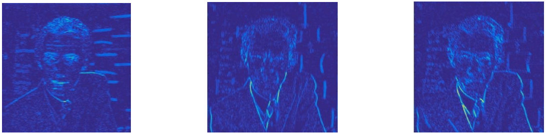
- Translation covariance on signal grid induced by the pooling factors
- In the absence of pooling, i.e., $S_k = 1$, for $k \in \{1, \ldots, d\}$, we get translation covariance w.r.t. the fine grid the input signal $f \in H_{N_1}$ lives on

Experiments

Experiments

The implementation in a nutshell

- Filters: Tensorized wavelets
  - extract *visual features* w.r.t. 3 directions (horizontal, vertical, diagonal)

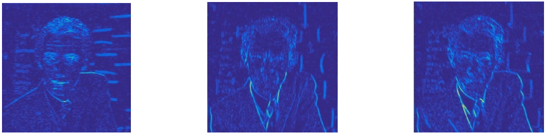


  - efficiently implemented using the *algorithme à trous* [*Holschneider et al., 1989*]

Experiments

The implementation in a nutshell

- Filters: Tensorized wavelets
  - extract *visual features* w.r.t. 3 directions (horizontal, vertical, diagonal)

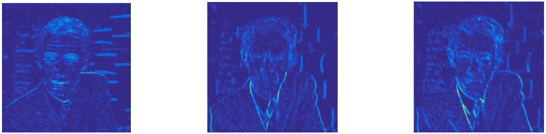


  - efficiently implemented using the *algorithme à trous* [*Holschneider et al., 1989*]
- Non-linearities: Modulus, rectified linear unit, hyperbolic tangent, logistic sigmoid

Experiments

The implementation in a nutshell

- Filters: Tensorized wavelets
  - extract *visual features* w.r.t. 3 directions (horizontal, vertical, diagonal)

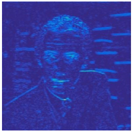
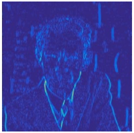


  - efficiently implemented using the *algorithme à trous* [*Holschneider et al., 1989*]
- Non-linearities: Modulus, rectified linear unit, hyperbolic tangent, logistic sigmoid
- Pooling: no pooling, sub-sampling, max-pooling, average-pooling

Experiments

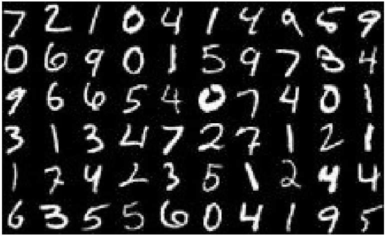
The implementation in a nutshell

- Filters: Tensorized wavelets
  - extract *visual features* w.r.t. 3 directions (horizontal, vertical, diagonal)

  - efficiently implemented using the *algorithme à trous* [*Holschneider et al., 1989*]
- Non-linearities: Modulus, rectified linear unit, hyperbolic tangent, logistic sigmoid
- Pooling: no pooling, sub-sampling, max-pooling, average-pooling
- Output-generating atoms: Low-pass filters

Experiment: Handwritten digit classification



- Dataset: MNIST database of handwritten digits [*LeCun & Cortes, 1998*]; 60,000 training and 10,000 test images
- Setup for Φ_Ω: $D = 3$ layers; same filters, non-linearities, and pooling operators in all layers
- Classifier: SVM with radial basis function kernel [*Vapnik, 1995*]
- Dimensionality reduction: Supervised orthogonal least squares scheme [*Chen et al., 1991*]

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig
- pooling-results ($S = 2$) are competitive with those without pooling at significantly lower computational cost

Experiment: Handwritten digit classification

Classification error in percent:

	Haar wavelet				Bi-orthogonal wavelet			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.57	0.57	1.35	1.49	0.51	0.57	1.12	1.22
sub.	0.69	0.66	1.25	1.46	0.61	0.61	1.20	1.18
max.	0.58	0.65	0.75	0.74	0.52	0.64	0.78	0.73
avg.	0.55	0.60	1.27	1.35	0.58	0.59	1.07	1.26

- modulus and ReLU perform better than tanh and LogSig
- pooling-results ($S = 2$) are competitive with those without pooling at significantly lower computational cost
- State-of-the-art: 0.43 [*Bruna and Mallat, 2013*]
  - similar feature extraction network with directional, but non-separable, wavelets and no pooling
  - significantly higher computational complexity

Experiment: Feature importance evaluation

Question: Which features are important in

- handwritten digit classification?

- detection of facial landmarks (eyes, nose, mouth) through regression?

Compare importance of features corresponding to (i) different layers, (ii) wavelet scales, and (iii) wavelet directions.

Experiment: Feature importance evaluation

Setup for Φ_Ω:

- $D = 4$ layers; Haar wavelets with $J = 3$ scales and modulus non-linearity in every network layer
- no pooling in the first layer, average pooling with uniform weights in the second and third layer $(S = 2)$

Experiment: Feature importance evaluation

Handwritten digit classification:

- Dataset: MNIST database (10,000 training and 10,000 test images)
- Random forest classifier [*Breiman, 2001*] with 30 trees
- Feature importance: Gini importance [*Breiman, 1984*]

Experiment: Feature importance evaluation

Handwritten digit classification:

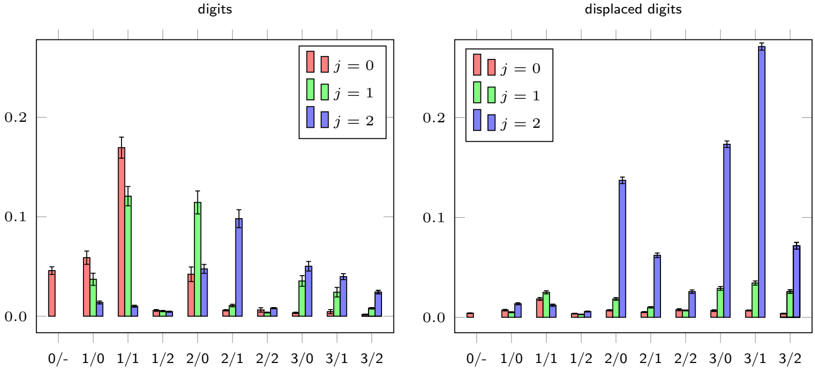
- Dataset: MNIST database (10,000 training and 10,000 test images)
- Random forest classifier [*Breiman, 2001*] with 30 trees
- Feature importance: Gini importance [*Breiman, 1984*]

Facial landmark detection:

- Dataset: Caltech Web Faces database (7092 images; 80% for training, 20% for testing)
- Random forest regressor [*Breiman, 2001*] with 30 trees
- Feature importance: Gini importance [*Breiman, 1984*]

Experiment: Feature importance evaluation

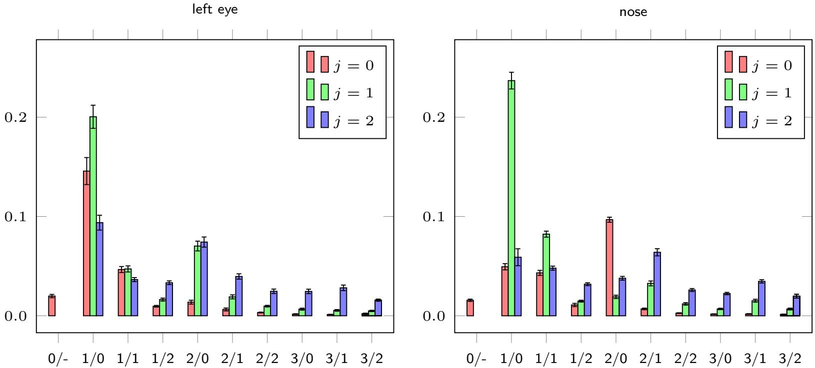
Average cumulative feature importance: Digit classification



- triplet of bars $[d/r]$ corresponds to horizontal $r = 0$, vertical $r = 1$, and diagonal $r = 2$ features in layer d

Experiment: Feature importance evaluation

Average cumulative feature importance: Facial landmarks



- triplet of bars $[d/r]$ corresponds to horizontal $r = 0$, vertical $r = 1$, and diagonal $r = 2$ features in layer d

Experiment: Feature importance evaluation

Average cumulative feature importance per layer:

	left eye	right eye	nose	mouth	digits	disp. digits
Layer 0	0.020	0.023	0.016	0.014	0.046	0.004
Layer 1	0.629	0.646	0.576	0.490	0.426	0.094
Layer 2	0.261	0.236	0.298	0.388	0.337	0.280
Layer 3	0.090	0.095	0.110	0.108	0.192	0.622

Experiment: Feature importance evaluation

Average cumulative feature importance per layer:

	left eye	right eye	nose	mouth	digits	disp. digits
Layer 0	0.020	0.023	0.016	0.014	0.046	0.004
Layer 1	0.629	0.646	0.576	0.490	0.426	0.094
Layer 2	0.261	0.236	0.298	0.388	0.337	0.280
Layer 3	0.090	0.095	0.110	0.108	0.192	0.622

- Digit classification: Features in deeper layers have higher importance

$\Rightarrow$ exploit vertical reduction in translation / deformation sensitivity

Experiment: Feature importance evaluation

Average cumulative feature importance per layer:

	left eye	right eye	nose	mouth	digits	disp. digits
Layer 0	0.020	0.023	0.016	0.014	0.046	0.004
Layer 1	0.629	0.646	0.576	0.490	0.426	0.094
Layer 2	0.261	0.236	0.298	0.388	0.337	0.280
Layer 3	0.090	0.095	0.110	0.108	0.192	0.622

- Facial landmark detection: Features in shallower layers have higher importance as they are translation-covariant on a finer grid

Experiment: Feature importance evaluation

Average cumulative feature importance per layer:

	left eye	right eye	nose	mouth	digits	disp. digits
Layer 0	0.020	0.023	0.016	0.014	0.046	0.004
Layer 1	0.629	0.646	0.576	0.490	0.426	0.094
Layer 2	0.261	0.236	0.298	0.388	0.337	0.280
Layer 3	0.090	0.095	0.110	0.108	0.192	0.622

- Facial landmark detection: Features in shallower layers have higher importance as they are translation-covariant on a finer grid

Given a particular machine learning task, it may be attractive to generate output in individual layers only!

Deep Frame Net

Open source software:

- MATLAB: `http://www.nari.ee.ethz.ch/commth/research`
- Python: Coming soon!

Thank you

"If you ask me anything I don't know, I'm not going to answer."

Y. Berra