ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Discrete deep feature extraction: A theory and new architectures

T. Wiatowski and M. Tschannen and A. Stanic and P. Grohs and

H. Bölcskei

Research Report No. 2016-29
May 2016

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Discrete Deep Feature Extraction: A Theory and New Architectures

Thomas Wiatowski[1] WITHOMAS@NARI.EE.ETHZ.CH
Michael Tschannen[1] MICHAELT@NARI.EE.ETHZ.CH
Aleksandar Stanić[1] ASTANIC@STUDENT.ETHZ.CH
Philipp Grohs[2] PHILIPP.GROHS@UNIVIE.AC.AT
Helmut Bölcskei[1] BOELCSKEI@NARI.EE.ETHZ.CH
[1]Dept. IT & EE, ETH Zurich, Switzerland
[2]Dept. Math., University of Vienna, Austria

Abstract

First steps towards a mathematical theory of deep convolutional neural networks for feature extraction were made—for the continuous-time case—in Mallat, 2012, and Wiatowski and Bölcskei, 2015. This paper considers the discrete case, introduces new convolutional neural network architectures, and proposes a mathematical framework for their analysis. Specifically, we establish deformation and translation sensitivity results of local and global nature, and we investigate how certain structural properties of the input signal are reflected in the corresponding feature vectors. Our theory applies to general filters and general Lipschitz-continuous non-linearities and pooling operators. Experiments on handwritten digit classification and facial landmark detection—including feature importance evaluation—complement the theoretical findings.

1. Introduction

Deep convolutional neural networks (DCNNs) have proven tremendously successful in a wide range of machine learning tasks (Bengio et al., 2013; LeCun et al., 2015). Such networks are composed of multiple layers, each of which computes convolutional transforms followed by the application of non-linearities and pooling operators.

DCNNs are typically distinguished according to (i) whether the filters employed are learned (in a supervised (LeCun et al., 1998; Huang & LeCun, 2006; Jarrett et al., 2009) or unsupervised (Ranzato et al., 2006; 2007; Jarrett et al., 2009) fashion) or pre-specified (and structured, such as, e.g., wavelets (Serre et al., 2005; Mutch & Lowe, 2006; Mallat, 2012), or unstructured, such as random filters (Ranzato et al., 2007; Jarrett et al., 2009)), (ii) the non-linearities used (e.g., logistic sigmoid, hyperbolic tangent, modulus, or rectified linear unit), and (iii) the pooling operator employed (e.g., sub-sampling, average pooling, or max-pooling). While a given choice of filters, non-linearities, and pooling operators will lead to vastly different performance results across datasets, it is remarkable that the overall DCNN architecture allows for impressive classification results across an extraordinarily broad range of applications. It is therefore of significant interest to understand the mechanisms underlying this universality.

First steps towards addressing this question and developing a mathematical theory of DCNNs for feature extraction were made—for the continuous-time case—in (Mallat, 2012; Wiatowski & Bölcskei, 2015). Specifically, (Mallat, 2012) analyzed so-called scattering networks, where signals are propagated through layers that employ directional wavelet filters and modulus non-linearities but no intra-layer pooling. The resulting wavelet-modulus feature extractor is horizontally (i.e., in every network layer) translation-invariant (accomplished by letting the wavelet scale parameter go to infinity) and deformation-stable, both properties of significance in practical feature extraction applications. Recently, (Wiatowski & Bölcskei, 2015) considered Mallat-type networks with arbitrary filters (that may be learned or pre-specified), general Lipschitz-continuous non-linearities (e.g., rectified linear unit, shifted logistic sigmoid, hyperbolic tangent, and the modulus function), and a continuous-time pooling operator that amounts to a dilation. The essence of the results in (Wiatowski & Bölcskei, 2015) is that vertical (i.e., asymptotically in the network depth) translation invariance and Lipschitz continuity of the feature extractor are induced by the network structure per se rather than the specific choice of filters and non-linearities. For band-limited signals

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

(Wiatowski & Bölcskei, 2015), Lipschitz-continuous functions (Grohs et al., 2016), and cartoon functions (Grohs et al., 2016), Lipschitz continuity of the feature extractor automatically leads to bounds on deformation sensitivity.

A discrete-time setup for wavelet-modulus scattering networks (referred to as ScatNets) was considered in (Bruna & Mallat, 2013).

Contributions. The purpose of the present paper is to develop a theory of discrete DCNNs for feature extraction. Specifically, we follow the philosophy put forward in (Wiatowski & Bölcskei, 2015; Grohs et al., 2016). Our theory incorporates general filters, Lipschitz non-linearities, and Lipschitz pooling operators. In addition, we introduce and analyze a wide variety of new network architectures which build the feature vector from subsets of the layers. This leads us to the notions of local and global feature vector properties with globality pertaining to characteristics brought out by the union of features across all network layers, and locality identifying attributes made explicit in individual layers.

Besides providing analytical performance results of general validity, we also investigate how certain structural properties of the input signal are reflected in the corresponding feature vectors. Specifically, we analyze the (local and global) deformation and translation sensitivity properties of feature vectors corresponding to sampled cartoon functions (Donoho, 2001). For simplicity of exposition, we focus on the 1-D case throughout the paper, noting that the extension to the higher-dimensional case does not pose any significant difficulties.

Our theoretical results are complemented by extensive numerical studies on facial landmark detection and handwritten digit classification. Specifically, we elucidate the role of local feature vector properties through a feature relevance study.

Notation. The complex conjugate of $z \in \mathbb{C}$ is denoted by $\overline{z}$. We write $\mathrm{Re}(z)$ for the real, and $\mathrm{Im}(z)$ for the imaginary part of $z \in \mathbb{C}$. We let $H_N := \{f : \mathbb{Z} \to \mathbb{C} \mid f[n] = f[n+N],\ \forall n \in \mathbb{Z}\}$ be the set of N-periodic discrete-time signals[1], and set $I_N := \{0, 1, \ldots, N-1\}$. The delta function $\delta \in H_N$ is $\delta[n] := 1$, for $n = kN$, $k \in \mathbb{Z}$, and $\delta[n] := 0$, else. For $f, g \in H_N$, we set $\langle f, g\rangle := \sum_{k\in I_N} f[k]\overline{g[k]}$, $\|f\|_1 := \sum_{n\in I_N} |f[n]|$, $\|f\|_2 := (\sum_{n\in I_N} |f[n]|^2)^{1/2}$, and $\|f\|_\infty := \sup_{n\in I_N} |f[n]|$. We denote the discrete Fourier transform (DFT) of $f \in H_N$ by $\hat{f}[k] := \sum_{n\in I_N} f[n]e^{-2\pi ikn/N}$. The circular convolution of $f \in H_N$ and $g \in H_N$ is $(f * g)[n] := \sum_{k\in I_N} f[k]g[n-k]$. We write $(T_m f)[n] := f[n-m]$, $m \in \mathbb{Z}$, for the cyclic translation operator. The supremum norm of a continuous-time function $c : \mathbb{R} \to \mathbb{C}$ is $\|c\|_\infty := \sup_{x\in\mathbb{R}} |c(x)|$. The indicator function of an interval $[a, b] \subseteq \mathbb{R}$ is defined as $\mathbb{1}_{[a,b]}(x) := 1$, for $x \in [a, b]$, and $\mathbb{1}_{[a,b]}(x) := 0$, for $x \in \mathbb{R}\backslash[a, b]$. The cardinality of the set $\mathcal{A}$ is denoted by $\mathrm{card}(\mathcal{A})$.

[1] We note that H_N is isometrically isomorphic to $\mathbb{C}^N$, but we prefer to work with H_N for the sake of expositional simplicity.

2. The basic building block

The basic building block of a DCNN, described in this section, consists of a convolutional transform followed by a non-linearity and a pooling operation.

2.1. Convolutional transform

A convolutional transform is made up of a set of filters $\Psi_\Lambda = \{g_\lambda\}_{\lambda\in\Lambda}$. The finite index set Λ can be thought of as labeling a collection of scales, directions, or frequency-shifts. The filters g_λ—referred to as atoms—may be learned (in a supervised or unsupervised fashion), pre-specified and unstructured such as random filters, or pre-specified and structured such as wavelets, curvelets, shearlets, or Weyl-Heisenberg functions.

Definition 1. *Let Λ be a finite index set. The collection $\Psi_\Lambda = \{g_\lambda\}_{\lambda\in\Lambda} \subseteq H_N$ is called a convolutional set with Bessel bound $B \geq 0$ if*

$$\sum_{\lambda\in\Lambda} \|f * g_\lambda\|_2^2 \leq B\|f\|_2^2, \qquad \forall f \in H_N. \tag{1}$$

Condition (1) is equivalent to

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

and hence, every finite set $\{g_\lambda\}_{\lambda\in\Lambda}$ is a convolutional set with Bessel bound $B^* := \max_{k\in I_N} \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2$. As $(f * g_\lambda)[n] = \langle f, \overline{g_\lambda[n-\cdot]}\rangle$, $n \in I_N$, $\lambda \in \Lambda$, the outputs of the filters g_λ may be interpreted as inner products of the input signal f with translates of the atoms g_λ. Frame theory (Daubechies, 1992) therefore tells us that the existence of a lower bound $A > 0$ in (2) according to

$$A \leq \sum_{\lambda\in\Lambda} |\widehat{g_\lambda}[k]|^2 \leq B, \qquad \forall k \in I_N, \tag{3}$$

implies that every element in H_N can be written as a linear combination of elements in the set $\{\overline{g_\lambda[n-\cdot]}\}_{n\in I_N,\lambda\in\Lambda}$ (or in more technical parlance, the set $\{\overline{g_\lambda[n-\cdot]}\}_{n\in I_N,\lambda\in\Lambda}$ is complete for H_N). The absence of a lower bound $A > 0$ may therefore result in Ψ_Λ failing to extract essential features of the signal f. We note, however, that even learned filters are likely to satisfy (3) as all that is needed is, for each $k \in I_N$, to have $\widehat{g_\lambda}[k] \neq 0$ for at least one $\lambda \in \Lambda$. As we shall see below, the existence of a lower bound $A > 0$ in (3) is, however, not needed for our theory to apply.

Examples of structured convolutional sets with $A = B = 1$ include, in the 1-D case, wavelets (Daubechies, 1992) and Weyl-Heisenberg functions (Bölcskei & Hlawatsch, 1997), and in the 2-D case, tensorized wavelets (Mallat, 2009), curvelets (Candès et al., 2006), and shearlets (Kutyniok & Labate, 2012a).

2.2. Non-linearities

The non-linearities $\rho : \mathbb{C} \to \mathbb{C}$ we consider are all point-wise and satisfy the Lipschitz property $|\rho(x) - \rho(y)| \leq L|x - y|$, $\forall x, y \in \mathbb{C}$, for some $L > 0$.

2.2.1. Example non-linearities

- The *hyperbolic tangent* non-linearity, defined as $\rho(x) = \tanh(\mathrm{Re}(x)) + i\tanh(\mathrm{Im}(x))$, where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, has Lipschitz constant $L = 2$.
- The *rectified linear unit* non-linearity is given by $\rho(x) = \max\{0, \mathrm{Re}(x)\} + i\max\{0, \mathrm{Im}(x)\}$, and has Lipschitz constant $L = 2$.
- The *modulus* non-linearity is $\rho(x) = |x|$, and has Lipschitz constant $L = 1$.
- The logistic sigmoid non-linearity is defined as $\rho(x) = \mathrm{sig}(\mathrm{Re}(x)) + i\,\mathrm{sig}(\mathrm{Im}(x))$, where $\mathrm{sig}(x) = \frac{1}{1+e^{-x}}$, and has Lipschitz constant $L = 1/2$.

We refer the reader to (Wiatowski & Bölcskei, 2015) for proofs of the Lipschitz properties of these example non-linearities.

2.3. Pooling operators

The essence of pooling is to reduce signal dimensionality in the individual network layers and to ensure robustness of the feature vector w.r.t. deformations and translations.

The theory developed in this paper applies to general pooling operators $P : H_N \to H_{N/S}$, where $N, S \in \mathbb{N}$ with $N/S \in \mathbb{N}$, that satisfy the Lipschitz property $\|Pf - Pg\|_2 \leq R\|f - g\|$, $\forall f, g \in H_N$, for some $R > 0$. The integer S will be referred to as pooling factor, and determines the "size" of the neighborhood values are combined in.

2.3.1. Example pooling operators

- *Sub-sampling*, defined as $P : H_N \to H_{N/S}$, $(Pf)[n] = f[Sn]$, $n \in I_{N/S}$, has Lipschitz constant $R = 1$. For $S = 1$, P is the identity operator which amounts to "no pooling".
- *Averaging*, defined as $P : H_N \to H_{N/S}$, $(Pf)[n] = \sum_{k=Sn}^{Sn+S-1} \alpha_{k-Sn} f[k]$, $n \in I_{N/S}$, has Lipschitz constant $R = S^{1/2} \max_{k\in\{0,\ldots,S-1\}} |\alpha_k|$. The weights $\{\alpha_k\}_{k=0}^{S-1}$ can be learned (LeCun et al., 1998) or pre-specified (Pinto et al., 2008) (e.g., uniform pooling corresponds to $\alpha_k = \frac{1}{S}$, for $k \in \{0, \ldots, S-1\}$).
- *Maximization*, defined as $P : H_N \to H_{N/S}$, $(Pf)[n] = \max_{k\in\{Sn,\ldots,Sn+S-1\}} |f[k]|$, $n \in I_{N/S}$, has Lipschitz constant $R = 1$.

We refer to Appendix B in the Supplement for proofs of the Lipschitz property of these three example pooling operators along with the derivations of the corresponding Lipschitz constants.

3. The network architecture

The architecture we consider is flexible in the following sense. In each layer, we can feed into the feature vector either the signals propagated down to that layer (i.e., the feature maps), filtered versions thereof, or we can decide not to have that layer contribute to the feature vector.

The basic building blocks of our network are the triplets (Ψ_d, ρ_d, P_d) of filters, non-linearities, and pooling operators associated with the d-th network layer and referred to as *modules*. We emphasize that these triplets are allowed to be different across layers.

Definition 2. *For network layers d, $1 \leq d \leq D$, let $\Psi_d = \{g_{\lambda_d}\}_{\lambda_d\in\Lambda_d} \subseteq H_{N_d}$ be a convolutional set, $\rho_d : \mathbb{C} \to \mathbb{C}$ a point-wise Lipschitz-continuous non-linearity, and $P_d : H_{N_d} \to H_{N_{d+1}}$ a Lipschitz-continuous pooling operator with $N_{d+1} = \frac{N_d}{S_d}$, where $S_d \in \mathbb{N}$ denotes the pooling factor in the d-th layer. Then, the sequence of triplets*

$$\Omega := \Big((\Psi_d, \rho_d, P_d)\Big)_{1\leq d\leq D}$$

is called a module-sequence.

Note that the dimensions of the spaces H_{N_d} satisfy $N_1 \geq N_2 \geq \ldots \geq N_D$. Associated with the module (Ψ_d, ρ_d, P_d), we define the operator

$$(U_d[\lambda_d]f) := P_d(\rho_d(f * g_{\lambda_d})) \tag{4}$$

and extend it to paths on index sets

$$q = (\lambda_1, \lambda_2, \ldots, \lambda_d) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_d := \Lambda_1^d,$$

for $1 \leq d \leq D$, according to

$$\begin{aligned} U[q]f &= U[(\lambda_1, \lambda_2, \ldots, \lambda_d)]f \\ &:= U_d[\lambda_d] \cdots U_2[\lambda_2]U_1[\lambda_1]f. \end{aligned} \tag{5}$$

For the empty path $e := \emptyset$ we set $\Lambda_1^0 := \{e\}$ and let $U[e]f := f$, for all $f \in H_{N_1}$.

The network output in the d-th layer is given by $(U[q]f) * \chi_d$, $q \in \Lambda_1^d$, where $\chi_d \in H_{N_{d+1}}$ is referred to as output-generating atom. Specifically, we let χ_d be (i) the delta

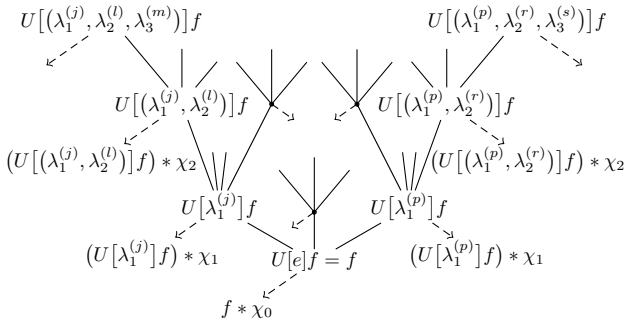

Figure 1. Network architecture underlying the feature extractor (6). The index $\lambda_d^{(k)}$ corresponds to the k-th atom $g_{\lambda_d^{(k)}}$ of the convolutional set Ψ_d associated with the d-th network layer. The function χ_d is the output-generating atom of the d-th layer. The root of the network corresponds to $d=0$.

function $\delta[n]$, $n \in I_{N_{d+1}}$, if we want the output to equal the unfiltered features $U[q]f$, $q \in \Lambda_1^d$, propagated down to layer d, or (ii) any other signal of length N_{d+1}, or (iii) $\chi_d = 0$ if we do not want layer d to contribute to the feature vector. From now on we formally add χ_d to the set $\Psi_{d+1} = \{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}}$, noting that $\{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}} \cup \{\chi_d\}$ forms a convolutional set Ψ'_{d+1} with Bessel bound $B'_{d+1} \leq B_{d+1} + \max_{k\in I_{N_{d+1}}} |\widehat{\chi}_d[k]|^2$. We emphasize that the atoms of the augmented set $\{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}} \cup \{\chi_d\}$ are employed across two consecutive layers in the sense of χ_d generating the output in the d-th layer according to $(U[q]f) * \chi_d$, $q \in \Lambda_1^d$, and the remaining atoms $\{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}}$ propagating the signals $U[q]f$, $q \in \Lambda_1^d$, from the d-th layer down to the $(d+1)$-st layer according to (4), see Fig. 1. With slight abuse of notation, we shall henceforth write Ψ_d for Ψ'_d and B_d for B'_d as well.

We are now ready to define the feature extractor Φ_Ω based on the module-sequence Ω.

Definition 3. *Let $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{1\leq d\leq D}$ be a module-sequence. The feature extractor Φ_Ω based on Ω maps $f \in H_{N_1}$ to its features*

$$\Phi_\Omega(f) := \bigcup_{d=0}^{D-1} \Phi_\Omega^d(f), \tag{6}$$

where $\Phi_\Omega^d(f) := \{(U[q]f) * \chi_d\}_{q\in\Lambda_1^d}$ is the collection of features generated in the d-th network layer (see Fig. 1).

The dimension of the feature vector $\Phi_\Omega(f)$ is given by $\varepsilon_0 N_1 + \sum_{d=1}^{D-1} \varepsilon_d N_{d+1}\big(\prod_{k=1}^{d} \mathrm{card}(\Lambda_k)\big)$, where $\varepsilon_d = 1$, if an output is generated (either filtered or unfiltered) in the d-th network layer, and $\varepsilon_d = 0$, else. As $N_{d+1} = \frac{N_d}{S_d} = \cdots = \frac{N_1}{S_1\cdots S_d}$, for $d \geq 1$, the dimension of the overall feature vector is determined by the pooling factors S_k and, of course, the layers that contribute to the feature vector.

Remark 1. *It was argued in (Bruna & Mallat, 2013; Andén & Mallat, 2014; Oyallon & Mallat, 2014) that the features $\Phi_\Omega^1(f)$ when generated by wavelet filters, modulus non-linearities, without intra-layer pooling, and by employing output-generating atoms with low-pass characteristics, describe mel frequency cepstral coefficients (Davis & Mermelstein, 1980) in 1-D, and SIFT-descriptors (Lowe, 2004; Tola et al., 2010) in 2-D.*

4. Sampled cartoon functions

While our main results hold for general signals f, we can provide a refined analysis for the class of sampled cartoon functions. This allows to understand how certain structural properties of the input signal, such as the presence of sharp edges, are reflected in the feature vector. Cartoon functions—as introduced in continuous time in (Donoho, 2001)—are piecewise "smooth" apart from curved discontinuities along Lipschitz-continuous hypersurfaces. They hence provide a good model for natural images (see Fig. 2, left) such as those in the Caltech-256 (Griffin et al., 2007) and the CIFAR-100 (Krizhevsky, 2009) datasets, for images of handwritten digits (LeCun & Cortes, 1998) (see Fig. 2, middle), and for images of geometric objects of different shapes, sizes, and colors as in the Baby AI School dataset[2].

Bounds on deformation sensitivity for cartoon functions in continuous-time DCNNs were recently reported in (Grohs et al., 2016). Here, we analyze deformation sensitivity for sampled cartoon functions passed through discrete DCNNs.

Definition 4. *The function $c : \mathbb{R} \to \mathbb{C}$ is referred to as a cartoon function if it can be written as $c = c_1 + \mathbb{1}_{[a,b]}c_2$, where $[a,b] \subseteq [0,1]$ is a closed interval, and $c_i : \mathbb{R} \to \mathbb{C}$, $i = 1,2$, satisfies the Lipschitz property*

$$|c_i(x) - c_i(y)| \leq C|x-y|, \quad \forall x, y \in \mathbb{R},$$

for some $C > 0$. Furthermore, we denote by

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

the class of cartoon functions of variation $K > 0$, and by

$$\mathcal{C}_{\mathrm{CART}}^{N,K} := \Big\{ f[n] = c(n/N),\ n \in \{0,1,\ldots,N-1\} \ \Big|\ \\ c = (c_1 + \mathbb{1}_{[a,b]}c_2) \in \mathcal{C}_{\mathrm{CART}}^K \text{ with} \\ a, b \notin \Big\{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\Big\}\Big\}$$

the class of sampled cartoon functions of length N and variation $K > 0$.

We note that excluding the boundary points a, b of the interval $[a,b]$ from being sampling points n/N in the definition

[2] http://www.iro.umontreal.ca/%7Elisa/twiki/bin/view.cgi/Public/BabyAISchool

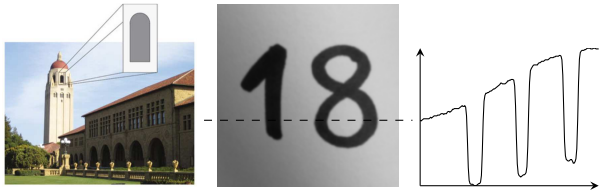

Figure 2. Left: A natural image (image credit: (Kutyniok & Labate, 2012b)) is typically governed by areas of little variation, with the individual areas separated by edges that can be modeled as curved singularities. Middle: Image of a handwritten digit. Right: Pixel values corresponding to the dashed row in the middle image.

of $\mathcal{C}_{\text{CART}}^{N,K}$ is of conceptual importance (see Remark D.1 in Appendix D in the Supplement). Moreover, our results can easily be generalized to classes $\mathcal{C}_{\text{CART}}^{N,K}$ consisting of functions $f[n] = c(n/N)$ with c containing multiple "1-D edges" (i.e., multiple discontinuity points) according to $c = c_1 + \sum_{l=1}^{L} \mathbb{1}_{[a_l,b_l]} c_2$ with $\cap_{l=1}^{L}[a_l, b_l] = \emptyset$. We also note that $\mathcal{C}_{\text{CART}}^{K}$ reduces to the class of Lipschitz-continuous functions upon setting $c_2 = 0$.

A sampled cartoon function in 2-D models, e.g., an image acquired by a digital camera (see Fig. 2, middle); in 1-D, $f \in \mathcal{C}_{\text{CART}}^{N,K}$ can be thought of as the pixels in a row or column of this image (see Fig. 2 right, which shows a cartoon function with 6 discontinuity points).

5. Analytical results

We analyze global and local feature vector properties with globality pertaining to characteristics brought out by the union of features across all network layers, and locality identifying attributes made explicit in individual layers.

5.1. Global properties

Theorem 1. *Let $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{1\le d\le D}$ be a module-sequence. Assume that the Bessel bounds $B_d > 0$, the Lipschitz constants $L_d > 0$ of the non-linearities ρ_d, and the Lipschitz constants $R_d > 0$ of the pooling operators P_d satisfy*

$$\max_{1\le d\le D} \max\{B_d, B_d R_d^2 L_d^2\} \le 1. \tag{7}$$

i) The feature extractor Φ_Ω is Lipschitz-continuous with Lipschitz constant $L_\Omega = 1$, i.e.,

$$|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \le \|f - h\|_2, \tag{8}$$

for all $f, h \in H_{N_1}$, where the feature space norm is defined as

$$|||\Phi_\Omega(f)|||^2 := \sum_{d=0}^{D-1} \sum_{q\in\Lambda_1^d} \|(U[q]f) * \chi_d\|_2^2. \tag{9}$$

ii) If, in addition to (7), for all $d \in \{1, \dots, D-1\}$ the non-linearities ρ_d and the pooling operators P_d satisfy $\rho_d(0) = 0$ and $P_d(0) = 0$ (as all non-linearities and pooling operators in Sections 2.2.1 and 2.3.1, apart from the logistic sigmoid non-linearity, do), then

$$|||\Phi_\Omega(f)||| \le \|f\|_2, \quad \forall f \in H_{N_1}. \tag{10}$$

iii) For every variation $K > 0$ and deformation F_τ of the form

$$(F_\tau f)[n] := c(n/N_1 - \tau(n/N_1)), \quad n \in I_{N_1}, \tag{11}$$

where $\tau : \mathbb{R} \to [-1, 1]$, the deformation sensitivity is bounded according to

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \le 4K N_1^{1/2} \|\tau\|_\infty^{1/2}, \tag{12}$$

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

Proof. See Appendix C in the Supplement. □

The Lipschitz continuity (8) guarantees that pairwise distances of input signals do not increase through feature extraction. As an immediate implication of the Lipschitz continuity we get robustness of the feature extractor w.r.t. additive bounded noise $\eta \in H_{N_1}$ in the sense of

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

for all $f \in H_{N_1}$.

Remark 2. *As detailed in the proof of Theorem 1, the Lipschitz continuity (8) combined with the deformation sensitivity bound (see Proposition D.1 in Appendix D in the Supplement) for the signal class under consideration, namely sampled cartoon functions, establishes the deformation sensitivity bound (12) for the feature extractor. This insight has important practical ramifications as it shows that whenever we have deformation sensitivity bounds for a signal class, we automatically obtain deformation sensitivity guarantees for the corresponding feature extractor.*

From (12) we can deduce a statement on the sensitivity of Φ_Ω w.r.t. translations on $\mathbb{R}$. To this end, we first note that setting $\tau_t(x) = t$, $x \in \mathbb{R}$, for $t \in [-1, 1]$, (11) becomes

$$(F_{\tau_t} f)[n] = c(n/N_1 - t), \quad n \in I_{N_1}.$$

Particularizing (12) accordingly, we obtain

$$|||\Phi_\Omega(F_{\tau_t} f) - \Phi_\Omega(f)||| \le 4K N_1^{1/2} |t|^{1/2}, \tag{13}$$

which shows that small translations $|t|$ of the underlying analog signal $c(x)$, $x \in \mathbb{R}$, lead to small changes in the feature vector obtained by passing the resulting sampled signal through a discrete DCNN. We shall say that (13) is a translation sensitivity bound. Analyzing the impact of deformations and translations over $\mathbb{R}$ on the discrete feature vector generated by the sampled analog signal closely models real-world phenomena (e.g., the jittered acquisition of an analog signal with a digital camera, where different values of N_1 in (11) correspond to different camera resolutions).

We note that, while iii) in Theorem 1 is specific to cartoon functions, i) and ii) apply to all signals in H_{N_1}.

The strength of the results in Theorem 1 derives itself from the fact that condition (7) on the underlying module-sequence Ω is easily met in practice. To see this, we first note that B_d is determined by the convolutional set Ψ_d, L_d by the non-linearity ρ_d, and R_d by the pooling operator P_d. Condition (7) is met if

$$B_d \leq \min\{1, R_d^{-2} L_d^{-2}\}, \quad \forall\, d \in \{1, 2, \ldots, D\}, \tag{14}$$

which, if not satisfied by default, can be enforced simply by normalizing the elements in Ψ_d. Specifically, for $C_d := \max\{B_d, R_d^2 L_d^2\}$ the set $\widetilde{\Psi}_d := \{C_d^{-1/2} g_{\lambda_d}\}_{\lambda_d \in \Lambda_d}$ has Bessel bound $\widetilde{B_d} = \frac{B_d}{C_d}$ and hence satisfies (14). While this normalization does not have an impact on the results in Theorem 1, there exists, however, a tradeoff between energy preservation and deformation (respectively translation) sensitivity in Φ_Ω^d as detailed in the next section.

5.2. Local properties

Theorem 2. *Let $\Omega = \big((\Psi_d, \rho_d, P_d)\big)_{1 \leq d \leq D}$ be a module-sequence with corresponding Bessel bounds $B_d > 0$, Lipschitz constants $L_d > 0$ of the non-linearities ρ_d, Lipschitz constants $R_d > 0$ of the pooling operators P_d, and output-generating atoms χ_d. Let further $L_\Omega^0 := \|\chi_0\|_1$ and*[3]

$$L_\Omega^d := \|\chi_d\|_1 \Big(\prod_{k=1}^{d} B_k L_k^2 R_k^2 \Big)^{1/2}, \qquad d \geq 1. \tag{15}$$

i) The features generated in the d-th network layer are Lipschitz-continuous with Lipschitz constant L_Ω^d, i.e.,

$$|||\Phi_\Omega^d(f) - \Phi_\Omega^d(h)||| \leq L_\Omega^d \|f - h\|_2, \tag{16}$$

for all $f, h \in H_{N_1}$, where $|||\Phi_\Omega^d(f)|||^2 := \sum_{q \in \Lambda_1^d} \|(U[q]f) * \chi_d\|_2^2$.

ii) If the non-linearities ρ_k and the pooling operators P_k satisfy $\rho_k(0) = 0$ and $P_k(0) = 0$, respectively, for all $k \in \{1, \ldots, d\}$, then

$$|||\Phi_\Omega^d(f)||| \leq L_\Omega^d \|f\|_2, \qquad \forall f \in H_{N_1}. \tag{17}$$

iii) For all $K > 0$ and all $\tau : \mathbb{R} \to [-1, 1]$, the features generated in the d-th network layer satisfy

$$|||\Phi_\Omega^d(F_\tau f) - \Phi_\Omega^d(f)||| \leq 4 L_\Omega^d K N^{1/2} \|\tau\|_\infty^{1/2}, \tag{18}$$

for all $f \in \mathcal{C}_{\mathrm{CART}}^{N_1, K}$, where $F_\tau f$ is defined in (11).

iv) If the module-sequence employs sub-sampling, average pooling, or max-pooling with corresponding pooling factors $S_d \in \mathbb{N}$, then

$$\Phi_\Omega^d(T_m f) = T_{\frac{m}{S_1 \ldots S_d}} \Phi_\Omega^d(f), \tag{19}$$

for all $f \in H_{N_1}$ and all $m \in \mathbb{Z}$ with $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$. Here, $T_m \Phi_\Omega^d(f)$ refers to element-wise application of T_m, i.e., $T_m \Phi_\Omega^d(f) := \{T_m h \mid \forall h \in \Phi_\Omega^d(f)\}$.

Proof. See Appendix E in the Supplement. □

One may be tempted to infer the global results (8), (10), and (12) in Theorem 1 from the corresponding local results in Theorem 2, e.g., the energy bound in (10) from (17) according to $|||\Phi_\Omega(f)||| = \Big(\sum_{d=0}^{D-1} |||\Phi_\Omega^d(f)|||^2 \Big)^{1/2} \leq \sqrt{D} \|f\|_2$, where we employed $L_\Omega^d \leq 1$ owing to (7). This would, however, lead to the "global" Lipschitz constant $L_\Omega = 1$ in (8), (10), and (12) to be replaced by $L_\Omega = \sqrt{D}$ and thereby render the corresponding results much weaker.

Again, we emphasize that, while iii) in Theorem 2 is specific to cartoon functions, i), ii), and iv) apply to all signals in H_{N_1}.

For a fixed network layer d, the "local" Lipschitz constant L_Ω^d determines the noise sensitivity of the features $\Phi_\Omega^d(f)$ according to

$$|||\Phi_\Omega^d(f + \eta) - \Phi_\Omega^d(f)||| \leq L_\Omega^d \|\eta\|_2, \tag{20}$$

where (20) follows from (16). Moreover, L_Ω^d via (18) also quantifies the impact of deformations (or translations when $\tau_t(x) = t$, $x \in \mathbb{R}$, for $t \in [-1, 1]$) on the feature vector. In practice, it may be desirable to have the features Φ_Ω^d become more robust to additive noise and less deformation-sensitive (respectively, translation-sensitive) as we progress deeper into the network. Formally, this vertical sensitivity reduction can be induced by ensuring that $L_\Omega^{d+1} < L_\Omega^d$. Thanks to $L_\Omega^d = \frac{\|\chi_d\|_1 B_d^{1/2} L_d R_d}{\|\chi_{d-1}\|_1} L_\Omega^{d-1}$, this can be accomplished by choosing the module-sequence such that $\|\chi_d\|_1 B_d^{1/2} L_d R_d < \|\chi_{d-1}\|_1$. Note, however, that owing to (17) this will also reduce the signal energy contained in the features $\Phi_\Omega^d(f)$. We therefore have a tradeoff between deformation (respectively translation) sensitivity and energy preservation. Having control over this tradeoff through the choice of the module-sequence Ω may come in handy in practice.

For average pooling with uniform weights $\alpha_k^d = \frac{1}{S_d}$, $k = 0, \ldots, S_d - 1$ (noting that the corresponding Lipschitz constant is $R_d = S_d^{-1/2}$, see Section 2.3.1), we get $L_\Omega^d = \|\chi_d\|_1 \Big(\prod_{k=1}^{d} \frac{B_k L_k^2}{S_k} \Big)^{1/2}$, which illustrates that pooling can have an impact on the sensitivity and energy properties of Φ_Ω^d.

We finally turn to interpreting the translation covariance result (19). Owing to the condition $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$, we get translation covariance only on the rough grid induced by the product of the pooling factors. In the absence of pooling,

[3] We note that $\|\chi_d\|_1$ in (15) can be upper-bounded (and hence substituted) by B_{d+1}, see Remark E.1 in Appendix E in the Supplement.

i.e., $S_k = 1$, for $k \in \{1, \dots, d\}$, we obtain translation covariance w.r.t. the fine grid the input signal $f \in H_{N_1}$ lives on.

Remark 3. *We note that ScatNets (Bruna & Mallat, 2013) are translation-covariant on the rough grid induced by the factor 2^J corresponding to the coarsest wavelet scale. Our result in (19) is hence in the spirit of (Bruna & Mallat, 2013) with the difference that the grid in our case is induced by the pooling factors S_k.*

6. Experiments[4]

We consider the problem of handwritten digit classification and evaluate the performance of the feature extractor Φ_Ω in combination with a support vector machine (SVM). The results we obtain are competitive with the state-of-the-art in the literature. The second line of experiments we perform assesses the importance of the features extracted by Φ_Ω in facial landmark detection and in handwritten digit classification, using random forests (RF) for regression and classification, respectively. Our results are based on a DCNN with different non-linearities and pooling operators, and with tensorized (i.e., separable) wavelets as filters, sensitive to 3 directions (horizontal, vertical, and diagonal). Furthermore, we generate outputs in all layers through low-pass filtering. Circular convolutions with the 1-D filters underlying the tensorized wavelets are efficiently implemented using the *algorithme à trous* (Holschneider et al., 1989).

To reduce the dimension of the feature vector, we compute features along frequency decreasing paths only (Bruna & Mallat, 2013), i.e., for every node $U[q]f$, $q \in \Lambda_1^{d-1}$, we retain only those child nodes $U_d[\lambda_d]U[q]f = P_d\big(\rho_d((U[q]f) * g_{\lambda_d})\big)$ that correspond to wavelets g_{λ_d} with scales larger than the maximum scale of the wavelets used to get $U[q]f$. We refer to (Bruna & Mallat, 2013) for a detailed justification of this approach for scattering networks.

6.1. Handwritten digit classification

We use the MNIST dataset of handwritten digits (LeCun & Cortes, 1998) which comprises 60,000 training and 10,000 test images of size 28×28. We set $D = 3$, and compare different network configurations, each defined by a single module (i.e., we use the same filters, non-linearity, and pooling operator in all layers). Specifically, we consider Haar wavelets and reverse biorthogonal 2.2 (RBIO2.2) wavelets (Mallat, 2009), both with $J = 3$ scales, the non-linearities described in Section 2.2.1, and the pooling operators described in Section 2.3.1 (with $S_1 = 1$ and $S_2 = 2$). We use a SVM with radial basis function (RBF) kernel for classification. To reduce the dimension of the feature vectors from 18,424 (or 50,176, for the configurations without pooling) down to 1000, we employ the supervised orthogonal least squares feature selection procedure described in (Oyallon & Mallat, 2014). The penalty parameter of the SVM and the localization parameter of the RBF kernel are selected via 10-fold cross-validation for each combination of wavelet filter, non-linearity, and pooling operator.

[4]Code available at http://www.nari.ee.ethz.ch/commth/research/

	Haar				RBIO2.2			
	abs	ReLU	tanh	LogSig	abs	ReLU	tanh	LogSig
n.p.	0.55	0.57	1.41	1.49	0.50	0.54	1.01	1.18
sub.	0.60	0.58	1.25	1.45	0.59	0.62	1.04	1.13
max.	0.61	0.60	0.68	0.76	0.55	0.56	0.71	0.75
avg.	0.57	0.58	1.26	1.44	0.51	0.60	1.04	1.18

Table 1. Classification error in percent for handwritten digit classification using different configurations of wavelet filters, non-linearities, and pooling operators (sub.: sub-sampling; max.: max-pooling; avg.: average-pooling; n.p.: no pooling).

Table 1 shows the resulting classification errors on the test set (obtained for the SVM trained on the full training set). Configurations employing RBIO2.2 wavelets tend to yield a marginally lower classification error than those using Haar wavelets. For the tanh and LogSig non-linearities, max-pooling leads to a considerably lower classification error than other pooling operators. The configurations involving the modulus and ReLU non-linearities achieve classification accuracy competitive with the state-of-the-art (Bruna & Mallat, 2013) (class. err.: 0.43%), which is based on directional non-separable wavelets with 6 directions without intra-layer pooling. This is interesting as the separable wavelet filters employed here can be implemented more efficiently.

6.2. Feature importance evaluation

In this experiment, we investigate the "importance" of the features generated by Φ_Ω corresponding to different layers, wavelet scales, and directions in two different learning tasks, namely, facial landmark detection and handwritten digit classification. The primary goal of this experiment is to illustrate the practical relevance of the notion of local properties of Φ_Ω as established in Section 5.2. For facial landmark detection we employ a RF regressor and for handwritten digit classification a RF classifier (Breiman, 2001). In both cases, we fix the number of trees to 30 and select the tree depth using out-of-bag error estimates (noting that increasing the number of trees does not significantly increase the accuracy). The impurity measure used for learning the node tests is the mean square error for facial landmark detection and the Gini impurity for handwritten digit classification. In both cases, feature importance is assessed using the Gini importance (Breiman et al., 1984), averaged over all trees. The Gini importance $I(\theta, T)$ of feature θ in the (trained) tree T is defined as

$I(\theta, T) = \sum_{\ell \in T:\, \varphi(\ell)=\theta} \frac{n_\ell}{n_{\text{tot}}} (\hat{\imath}_\ell - \frac{n_{\ell_L}}{n_\ell} \hat{\imath}_{\ell_L} - \frac{n_{\ell_R}}{n_\ell} \hat{\imath}_{\ell_R})$, where $\varphi(\ell)$ denotes the feature determined in the training phase for the test at node ℓ, n_ℓ is the number of training samples passed through node ℓ, $n_{\text{tot}} = \sum_{\ell \in T} n_\ell$, $\hat{\imath}_\ell$ is the impurity at node ℓ, and ℓ_L and ℓ_R denote the left and right child node, respectively, of node ℓ. For the feature extractor Φ_Ω we set $D = 4$, employ Haar wavelets with $J = 3$ scales and the modulus non-linearity in every network layer, no pooling in the first layer and average pooling with uniform weights $1/S_d^2$, $S_d = 2$, in layers $d = 2, 3$.

Facial landmark detection. We use the Caltech 10,000 Web Faces data base (Angelova et al., 2005). Each of the 7092 images in the data base depicts one or more faces in different contexts (e.g., portrait images, groups of people). The data base contains annotations of the positions of eyes, nose, and mouth for at least one face per image. The learning task is to estimate the positions of these facial landmarks. The annotations serve as ground truth for training and testing. We preprocess the data set as follows. The patches containing the faces are extracted from the images using the Viola-Jones face detector (Viola & Jones, 2004). After discarding false positives, the patches are converted to grayscale and resampled to size 120×120 (using linear interpolation), before feeding them to the feature extractor Φ_Ω. This procedure yields a dataset containing a total of 8776 face images. We select 80% of the images uniformly at random to form a training set and use the remaining images for testing. We train a separate RF for each facial landmark. Following (Dantone et al., 2012) we report the localization error, i.e., the ℓ_2-distance between the estimated and the ground truth landmark positions, on the test set as a fraction of the (true) inter-ocular distance. The errors obtained are: left eye: 0.062; right eye: 0.064; nose; 0.080, mouth: 0.095. As an aside, we note that these values are comparable with the ones reported in (Dantone et al., 2012) for a conditional RF using patch comparison features (evaluated on a different dataset and a larger set of facial landmarks).

Handwritten digit classification. For this experiment, we again rely on the MNIST dataset. The training set is obtained by sampling uniformly at random $1,000$ images per digit from the MNIST training dataset and we use the complete MNIST test set. We train two RFs, one based on unmodified images, and the other one based on images subject to a random uniform displacement of at most 4 pixels in (positive and negative) x and y direction to study the impact of offsets on feature importance. The resulting RFs achieve a classification error of 4.2% and 9.6%, respectively.

Discussion. Figure 3 shows the cumulative feature importance (per triplet of layer index, wavelet scale, and direction, averaged over all trees in the respective RF) in handwritten digit classification and in facial landmark detection. Table 2 shows the corresponding cumulative feature importance for each layer.

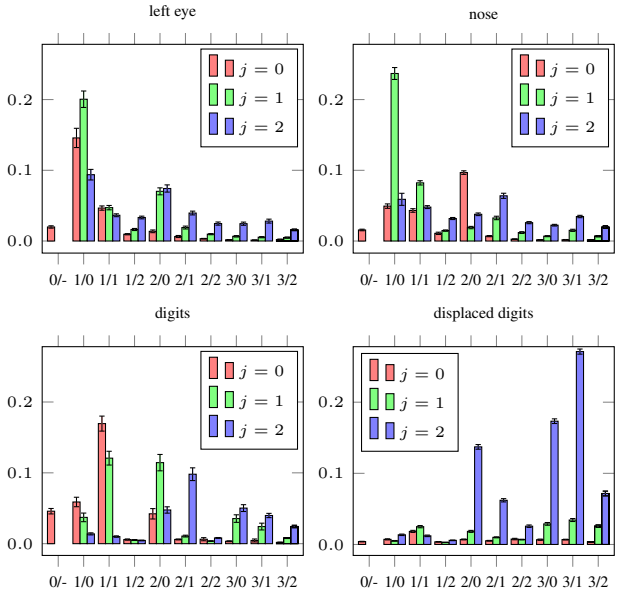

Figure 3. Average cumulative feature importance and standard error for facial landmark detection and handwritten digit classification. The labels on the horizontal axis indicate layer index d/wavelet direction (0: horizontal, 1: vertical, 2: diagonal).

	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

Table 2. Cumulative feature importance per layer. Columns 1–4: facial landmark detection. Columns 5 and 6: handwritten digit classification.

For facial landmark detection, the features in layer 1 clearly have the highest importance, and the feature importance decreases with increasing layer index d. For handwritten digit classification using the unshifted MNIST images, the cumulative importance of the features in the second/third layer relative to those in the first layer is considerably higher than in facial landmark detection (see Table 2). For the translated MNIST images, the importance of the features in the second/third layer is significantly higher than those in the 0-th and in the first layer. An explanation for this observation could be as follows: In a classification task, small sensitivity to translations is beneficial. Now, according to our theory (see Section 5.2) translation sensitivity, indeed, decreases with increasing layer index for average pooling as used here. For localization of landmarks, on the other hand, the RF needs features that are covariant on the fine grid of the input image thus favoring features in the layers closer to the root.

Acknowledgments

The authors would like to thank C. Geiger for preliminary work on the experiments in Section 6.2 and M. Lerjen for help with computational issues.

References

Andén, J. and Mallat, S. Deep scattering spectrum. *IEEE Trans. Sig. Process.*, 62(16):4114–4128, 2014.

Angelova, A., Abu-Mostafa, Y., and Perona, P. Pruning training sets for learning of object categories. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 494–501, 2005.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

Bölcskei, H. and Hlawatsch, F. Discrete Zak transforms, polyphase transforms, and applications. *IEEE Trans. Sig. Process.*, 45(4):851–866, 1997.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC Press, 1984.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35 (8):1872–1886, 2013.

Candès, E. J., Demanet, L., Donoho, D., and Ying, L. Fast discrete curvelet transforms. *Multiscale Modeling and Simulation*, 5(3):861–899, 2006.

Dantone, M., Gall, J., Fanelli, G., and Van Gool, L. Real-time facial feature detection using conditional regression forests. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 2578–2585, 2012.

Daubechies, I. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, and Signal Process.*, 28(4):357–366, 1980.

Donoho, D. Sparse components of images and optimal atomic decompositions. *Constructive Approximation*, 17 (3):353–382, 2001.

Folland, G. B. *A course in abstract harmonic analysis*, volume 29. CRC Press, 2015.

Golub, G. H. and Van Loan, C. F. *Matrix computations*. Johns Hopkins University Press, 2013.

Griffin, G., Holub, A., and Perona, P. Caltech-256 object category dataset. http://authors.library.caltech.edu/7694/, 2007.

Grohs, P., Wiatowski, T., and Bölcskei, H. Deep convolutional neural networks on cartoon functions. In *Proc. of IEEE Int. Symp. on Inform. Theory (ISIT),* to appear. 2016.

Holschneider, M., Kronland-Martinet, R., Morlet, J., and Tchamitchian, P. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pp. 286–297. Springer, 1989.

Huang, F. J. and LeCun, Y. Large-scale learning with SVM and convolutional nets for generic object categorization. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 284–291, 2006.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., and LeCun, Y. What is the best multi-stage architecture for object recognition? In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 2146–2153, 2009.

Krizhevsky, A. Learning multiple layers of features from tiny images. MS thesis, University of Toronto, 2009.

Kutyniok, G. and Labate, D. (eds.). *Shearlets: Multiscale analysis for multivariate data*. Birkhäuser, 2012a.

Kutyniok, G. and Labate, D. Introduction to shearlets. In *Shearlets: Multiscale analysis for multivariate data*, pp. 1–38. Birkhäuser, 2012b.

LeCun, Y. and Cortes, C. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proc. of the IEEE*, pp. 2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521:436–444, 2015.

Lowe, D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

Mallat, S. *A wavelet tour of signal processing: The sparse way*. Academic Press, 3rd edition, 2009.

Mallat, S. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, 2012.

Mutch, J. and Lowe, D. G. Multiclass object recognition with sparse, localized features. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 11–18, 2006.

Oyallon, E. and Mallat, S. Deep roto-translation scattering for object classification. *arXiv:1412.8659*, 2014.

Pinto, N., Cox, D. D., and DiCarlo, J. J. Why is real-world visual object recognition hard. *PLoS Computational Biology*, 4(1):151–156, 2008.

Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. Efficient learning of sparse representations with an energy-based model. In *Proc. of Int. Conf. on Neural Information Processing Systems (NIPS)*, pp. 1137–1144, 2006.

Ranzato, M. A., Huang, F. J., Boureau, Y. L., and LeCun, Y. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 1–8, 2007.

Serre, T., Wolf, L., and Poggio, T. Object recognition with features inspired by visual cortex. In *Proc. of IEEE Conf. Comp. Vision Pattern Recog. (CVPR)*, pp. 994–1000, 2005.

Tola, E., Lepetit, V., and Fua, P. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5(815–830), 2010.

Viola, P. and Jones, M. J. Robust real-time face detection. *International Journal of Computer Vision*, 57(2): 137–154, 2004.

Wiatowski, T. and Bölcskei, H. A mathematical theory of deep convolutional neural networks for feature extraction. *arXiv:1512.06293*, 2015.

A. Appendix: Additional numerical results

A.1. Handwritten digit classification

For the handwritten digit classification experiment described in Section 6.1, Table 3 shows the classification error for Daubechies wavelets with 2 vanishing moments (DB2).

	DB2			
	abs	ReLU	tanh	LogSig
n.p.	0.54	0.51	1.29	1.40
sub.	0.60	0.58	1.16	1.34
max.	0.57	0.57	0.75	0.67
avg.	0.52	0.61	1.16	1.27

Table 3. Classification errors in percent for handwritten digit classification using DB2 wavelet filters, different non-linearities, and different pooling operators (sub.: sub-sampling; max.: max-pooling; avg.: average-pooling; n.p.: no pooling).

A.2. Feature importance evaluation

For the feature importance experiment described in Section 6.2, Figure 4 shows the cumulative feature importance (per triplet of layer index, wavelet scale, and direction, averaged over all trees in the respective RF) in facial landmark detection (right eye and mouth).

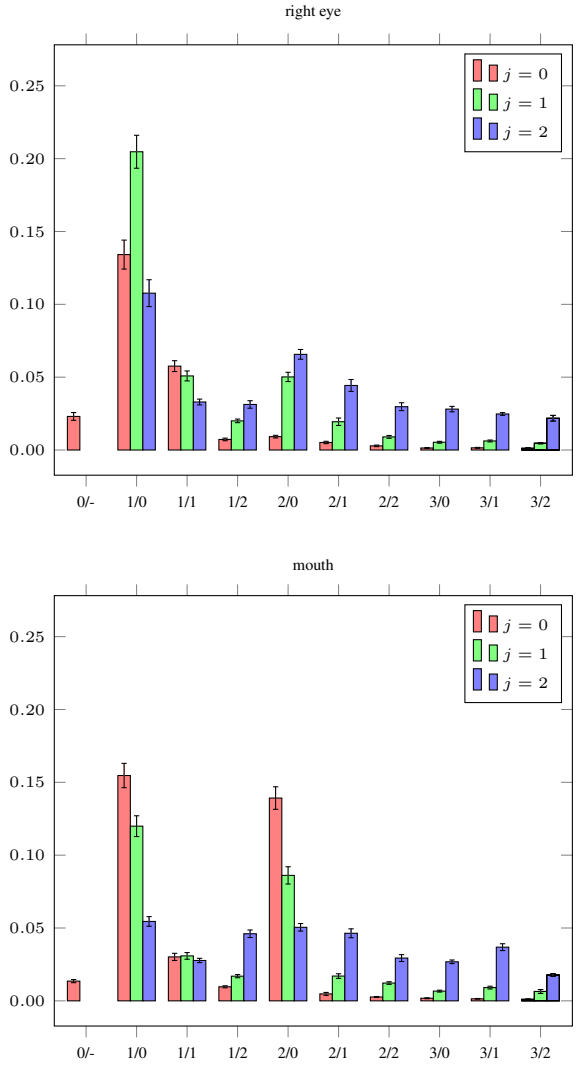

Figure 4. Average cumulative feature importance and standard error for facial landmark detection. The labels on the horizontal axis indicate layer index *d*/wavelet direction (0: horizontal, 1: vertical, 2: diagonal).

B. Appendix: Lipschitz continuity of pooling operators

We verify the Lipschitz property

$$\|P(f) - P(h)\|_2 \leq R\|f - h\|_2, \quad \forall f, h \in H_N,$$

for the pooling operators in Section 2.3.1.

Sub-sampling: Pooling by sub-sampling is defined as

$$P : H_N \to H_{N/S}, \quad P(f)[n] = f[Sn], \quad n \in I_{N/S},$$

where $N/S \in \mathbb{N}$. Lipschitz continuity with $R = 1$ follows from

$$\|P(f) - P(h)\|_2^2 = \sum_{n \in I_{N/S}} |f[Sn] - h[Sn]|^2$$
$$\leq \sum_{n \in I_N} |f[n] - h[n]|^2 = \|f - h\|_2^2, \quad \forall f, h \in H_N.$$

Averaging: Pooling by averaging is defined as

$$P : H_N \to H_{N/S}, \quad P(f)[n] = \sum_{k=Sn}^{Sn+S-1} \alpha_{k-Sn} f[k],$$

for $n \in I_{N/S}$, where $N/S \in \mathbb{N}$. We start by setting $\alpha' :=$

$\max_{k\in\{0,\dots,S-1\}}|\alpha_k|$. Then,

$$\begin{aligned}
&\|P(f)-P(h)\|_2^2\\
&=\sum_{n\in I_{N/S}}\Big|\sum_{k=Sn}^{Sn+S-1}\alpha_{k-Sn}(f[k]-h[k])\Big|^2\\
&\le\sum_{n\in I_{N/S}}\Big|\sum_{k=Sn}^{Sn+S-1}\alpha'|f[k]-h[k]|\Big|^2\\
&\le\alpha'^2S\sum_{n\in I_{N/S}}\sum_{k=Sn}^{Sn+S-1}\Big|f[k]-h[k]\Big|^2 \qquad \text{(B.1)}\\
&=\alpha'^2S\sum_{n\in I_N}\Big|f[k]-h[k]\Big|^2=\alpha'^2S\|f-h\|_2^2,
\end{aligned}$$

where we used $\sum_{k\in I_S}|f[k]-h[k]|\le S^{1/2}\|f-h\|_2$, $f,h\in H_S$, to get (B.1), see, e.g., (Golub & Van Loan, 2013).

Maximization: Pooling by maximization is defined as

$$P:H_N\to H_{N/S},\quad P(f)[n]=\max_{k\in\{Sn,\dots,Sn+S-1\}}|f[k]|,$$

for $n\in I_{N/S}$, where $N/S\in\mathbb{N}$. We have

$$\begin{aligned}
&\|P(f)-P(h)\|_2^2\\
&=\sum_{n\in I_{N/S}}\Big|\max_{k\in\{Sn,\dots,Sn+S-1\}}|f[k]|\\
&-\max_{k\in\{Sn,\dots,Sn+S-1\}}|h[k]|\Big|^2\\
&\le\sum_{n\in I_{N/S}}\max_{k\in\{Sn,\dots,Sn+S-1\}}\Big|f[k]-h[k]\Big|^2 \qquad \text{(B.2)}\\
&\le\sum_{n\in I_{N/S}}\sum_{k=0}^{S-1}|f[Sn+k]-h[Sn+k]|^2 \qquad \text{(B.3)}\\
&=\|f-h\|_2^2,
\end{aligned}$$

where we employed the reverse triangle inequality $\big|\|f\|_\infty-\|h\|_\infty\big|\le\|f-h\|_\infty$, $f,h\in H_S$, to get (B.2), and in (B.3) we used $\|f\|_\infty\le\|f\|_2$, $f\in H_S$, see, e.g., (Golub & Van Loan, 2013).

C. Appendix: Proof of Theorem 1

We start by proving i). The key idea of the proof is—similarly to the proof of Proposition 4 in (Wiatowski & Bölcskei, 2015)—to employ telescoping series arguments. For ease of notation, we let $f_q:=U[q]f$ and $h_q:=U[q]h$, for $f,h\in H_{N_1}$, $q\in\Lambda_1^d$. With (9) we have

$$|||\Phi_\Omega(f)-\Phi_\Omega(h)|||^2=\sum_{d=0}^{D-1}\underbrace{\sum_{q\in\Lambda_1^d}||(f_q-h_q)*\chi_d||_2^2}_{=:a_d}.$$

The key step is then to show that a_d can be upper-bounded according to

$$a_d\le b_d-b_{d+1},\qquad d=0,\dots,D-1, \qquad \text{(C.1)}$$

with $b_d:=\sum_{q\in\Lambda_1^d}\|f_q-h_q\|_2^2$, for $d=0,\dots,D$, and to note that

$$\begin{aligned}
\sum_{d=0}^{D-1}a_d&\le\sum_{d=0}^{D-1}(b_d-b_{d+1})=b_0-\underbrace{b_D}_{\ge0}\le b_0\\
&=\sum_{q\in\Lambda_1^0}\|f_q-h_q\|_2^2=\|f-h\|_2^2,
\end{aligned}$$

which then yields (8). Writing out (C.1), it follows that we need to establish

$$\begin{aligned}
&\sum_{q\in\Lambda_1^d}\|(f_q-h_q)*\chi_d\|_2^2\le\sum_{q\in\Lambda_1^d}||f_q-h_q\|_2^2\\
&-\sum_{q\in\Lambda_1^{d+1}}\|f_q-h_q\|_2^2,\qquad d=0,\dots,D-1. \qquad \text{(C.2)}
\end{aligned}$$

We start by examining the second sum on the right-hand side (RHS) in (C.2). Every path

$$\tilde q\in\Lambda_1^{d+1}=\underbrace{\Lambda_1\times\cdots\times\Lambda_d}_{=\Lambda_1^d}\times\Lambda_{d+1}$$

of length $d+1$ can be decomposed into a path $q\in\Lambda_1^d$ of length d and an index $\lambda_{d+1}\in\Lambda_{d+1}$ according to $\tilde q=(q,\lambda_{d+1})$. Thanks to (5) we have $U[\tilde q]=U[(q,\lambda_{d+1})]=U_{d+1}[\lambda_{d+1}]U[q]$, which yields

$$\begin{aligned}
&\sum_{\tilde q\in\Lambda_1^{d+1}}\|f_{\tilde q}-h_{\tilde q}\|_2^2=\sum_{q\in\Lambda_1^d}\sum_{\lambda_{d+1}\in\Lambda_{d+1}}\|U_{d+1}[\lambda_{d+1}]f_q\\
&-U_{d+1}[\lambda_{d+1}]h_q\|_2^2. \qquad \text{(C.3)}
\end{aligned}$$

Substituting (C.3) into (C.2) and rearranging terms, we obtain

$$\begin{aligned}
&\sum_{q\in\Lambda_1^d}\Big(\|(f_q-h_q)*\chi_d\|_2^2 \qquad \text{(C.4)}\\
&+\sum_{\lambda_{d+1}\in\Lambda_{d+1}}\|U_{d+1}[\lambda_{d+1}]f_q-U_{d+1}[\lambda_{d+1}]h_q\|_2^2\Big) \qquad \text{(C.5)}\\
&\le\sum_{q\in\Lambda_1^d}||f_q-h_q\|_2^2,\qquad d=0,\dots,D-1. \qquad \text{(C.6)}
\end{aligned}$$

We next note that the sum over the index set Λ_{d+1} inside the brackets in (C.4)-(C.5) satisfies

$$\begin{aligned}
&\sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|U_{d+1}[\lambda_{d+1}]f_q - U_{d+1}[\lambda_{d+1}]h_q\|_2^2 \\
&= \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|P_{d+1}\big(\rho_{d+1}(f_q * g_{\lambda_{d+1}})\big) \\
&\qquad - P_{d+1}\big(\rho_{d+1}(h_q * g_{\lambda_{d+1}})\big)\|_2^2 \\
&\leq R_{d+1}^2 \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|\rho_{d+1}(f_q * g_{\lambda_{d+1}}) \qquad \text{(C.7)} \\
&\qquad - \rho_{d+1}(h_q * g_{\lambda_{d+1}})\|_2^2 \qquad \text{(C.8)} \\
&\leq R_{d+1}^2 L_{d+1}^2 \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|(f_q - h_q) * g_{\lambda_{d+1}}\|_2^2, \qquad \text{(C.9)}
\end{aligned}$$

where we employed the Lipschitz continuity of P_{d+1} in (C.7)-(C.8) and the Lipschitz continuity of ρ_{d+1} in (C.9). Substituting the sum over the index set Λ_{d+1} inside the brackets in (C.4)-(C.5) by the upper bound (C.9) yields

$$\begin{aligned}
&\sum_{q\in\Lambda_1^d} \Big(\|(f_q - h_q) * \chi_d\|_2^2 \\
&+ \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|U_{d+1}[\lambda_{d+1}]f_q - U_{d+1}[\lambda_{d+1}]h_q\|_2^2\Big) \\
&\leq \sum_{q\in\Lambda_1^d} \max\{1, R_{d+1}^2 L_{d+1}^2\}\Big(\|(f_q - h_q) * \chi_d\|_2^2 \qquad \text{(C.10)} \\
&+ \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|(f_q - h_q) * g_{\lambda_{d+1}}\|_2^2\Big), \qquad \text{(C.11)}
\end{aligned}$$

for $d = 0, \ldots, D-1$. As $\{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}} \cup \{\chi_d\}$ are atoms of the convolutional set Ψ_{d+1}, and $f_q, h_q \in H_{N_{d+1}}$, we have

$$\begin{aligned}
&\|(f_q - h_q) * \chi_d\|_2^2 + \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|(f_q - h_q) * g_{\lambda_{d+1}}\|_2^2 \\
&\leq B_{d+1}\|f_q - h_q\|_2^2,
\end{aligned}$$

which, when used in (C.10)-(C.11) yields

$$\begin{aligned}
&\sum_{q\in\Lambda_1^d} \Big(\|(f_q - h_q) * \chi_d\|_2^2 \\
&\quad + \sum_{\lambda_{d+1}\in\Lambda_{d+1}} \|U_{d+1}[\lambda_{d+1}]f_q - U_{d+1}[\lambda_{d+1}]h_q\|_2^2\Big) \\
&\leq \sum_{q\in\Lambda_1^d} \max\{B_{d+1}, B_{d+1}R_{d+1}^2 L_{d+1}^2\}\|f_q - h_q\|_2^2,
\end{aligned} \qquad \text{(C.12)}$$

for $d = 0, \ldots, D-1$. Finally, invoking (7) in (C.12) we get (C.4)-(C.6) and hence (C.1). This completes the proof of i).

We continue with ii). The key step in establishing (10) is to show that for $\rho_d(0) = 0$ and $P_d(0) = 0$, for $d \in \{1, \ldots, D-1\}$, the feature extractor Φ_Ω satisfies $\Phi_\Omega(0) = 0$, and to employ (8) with $h = 0$ which yields

$$|||\Phi(f)||| \leq \|f\|,$$

for $f \in H_{N_1}$. It remains to prove that $\Phi_\Omega(h) = 0$ for $h = 0$. For $h = 0$, the operator U_d, $d \in \{1, 2, \ldots, D\}$, defined in (4) satisfies

$$(U_d[\lambda_d]h) = P_d\big(\underbrace{\rho_d\big(\underbrace{\underbrace{h * g_{\lambda_d}}_{=0}}_{=0}\big)}_{=0}\big),$$

for $\lambda_d \in \Lambda_d$, by assumption. With the definition of $U[q]$ in (5) this then yields $(U[q]h) = 0$ for $h = 0$ and all $q \in \Lambda_1^d$. $\Phi_\Omega(0) = 0$ finally follows from

$$\Phi_\Omega(h) = \bigcup_{d=0}^{D-1} \big\{\underbrace{(U[q]h) * \chi_d}_{=0}\big\}_{q\in\Lambda_1^d} = 0. \qquad \text{(C.13)}$$

We proceed to iii). The proof of the deformation sensitivity bound (12) is based on two key ingredients. The first one is the Lipschitz continuity result stated in (8). The second ingredient, stated in Proposition D.1 in Appendix D, is an upper bound on the deformation error $\|f - F_\tau f\|_2$ given by

$$\|f - F_\tau f\|_2 \leq 4KN_1^{1/2}\|\tau\|_\infty^{1/2}, \qquad \text{(C.14)}$$

where $f \in C_{\text{CART}}^{N_1,K}$. We now show how (8) and (C.14) can be combined to establish (12). To this end, we first apply (8) with $h := (F_\tau f)$ to get

$$|||\Phi_\Omega(f) - \Phi_\Omega(F_\tau f)||| \leq \|f - F_\tau f\|_2, \qquad \text{(C.15)}$$

for $f \in C_{\text{CART}}^{N_1,K} \subseteq H_{N_1}$, $N_1 \in \mathbb{N}$, and $K > 0$, and then replace the RHS of (C.15) by the RHS of (C.14). This completes the proof of iii).

D. Appendix: Proposition D.1

Proposition D.1. *For every $N \in \mathbb{N}$, every $K > 0$, and every $\tau : \mathbb{R} \to [-1, 1]$, we have*

$$\|f - F_\tau f\|_2 \leq 4KN^{1/2}\|\tau\|_\infty^{1/2}, \qquad \text{(D.1)}$$

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

Remark D.1. *As already mentioned at the end of Section 4, excluding the interval boundary points a, b in the definition of sampled cartoon functions $\mathcal{C}_{\text{CART}}^{N,K}$ (see Definition 4) is necessary for technical reasons. Specifically, without imposing this exclusion, we can not expect to get deformation sensitivity results of the form (D.1). This can be seen as follows. Let us assume that we seek a bound of the*

form $\|f - F_\tau f\|_2 \leq C_{N,K}\|\tau\|_\infty^\alpha$, *for some* $C_{N,K} > 0$ *and some* $\alpha > 0$, *that applies to all* $f[n] = c(n/N)$, $n \in I_N$, *with* $c \in \mathcal{C}_{\mathrm{CART}}^K$. *Take* $\tau(x) = 1/N$, *in which case the deformation* $(F_\tau f)[n] = c(n/N - 1/N)$ *amounts to a simple translation by* $1/N$ *and* $\|\tau\|_\infty = 1/N \leq 1$. *Let* $c(x) = \mathbb{1}_{[0,2/N]}(x)$. *Then* $c \in \mathcal{C}_{\mathrm{CART}}^K$ *for* $K = 1$ *and* $\|f - F_\tau f\|_2 = \sqrt{2}$, *which obviously does not decay with* $\|\tau\|_\infty^\alpha = N^{-\alpha}$ *for some* $\alpha > 0$. *We note that this phenomenon occurs only in the discrete case.*

Proof. The proof of (D.1) is based on judiciously combining deformation sensitivity bounds for the sampled components $c_1(n/N), c_2(n/N)$, $n \in I_N$, in $(c_1 + \mathbb{1}_{[a,b]}c_2) \in \mathcal{C}_{\mathrm{CART}}^K$, and the sampled indicator function $\mathbb{1}_{[a,b]}(n/N)$, $n \in I_N$. The first bound, stated in Lemma D.1 below, reads

$$\|f - F_\tau f\|_2 \leq CN^{1/2}\|\tau\|_\infty, \tag{D.2}$$

and applies to discrete-time signals $f[n] = f(n/N)$, $n \in I_N$, with $f : \mathbb{R} \to \mathbb{C}$ satisfying the Lipschitz property with Lipschitz constant C. The second bound we need, stated in Lemma D.2 below, is given by

$$\|\mathbb{1}_{[a,b]}^N - F_\tau \mathbb{1}_{[a,b]}^N\|_2 \leq 2N^{1/2}\|\tau\|_\infty^{1/2}, \tag{D.3}$$

and applies to sampled indicator functions $\mathbb{1}_{[a,b]}^N[n] := \mathbb{1}_{[a,b]}(n/N)$, $n \in I_N$, with $a, b \notin \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. We now show how (D.2) and (D.3) can be combined to establish (D.1). For a sampled cartoon function $f \in \mathcal{C}_{\mathrm{CART}}^{N,K}$, i.e.,

$$\begin{aligned} f[n] &= c_1(n/N) + \mathbb{1}_{[a,b]}(n/N)c_2(n/N) \\ &=: f_1[n] + \mathbb{1}_{[a,b]}^N[n]f_2[n], \quad n \in I_N, \end{aligned}$$

we have

$$\begin{aligned} \|f - F_\tau f\|_2 &\leq \|f_1 - F_\tau f_1\|_2 + \|\mathbb{1}_{[a,b]}^N(f_2 - F_\tau f_2)\|_2 \\ &+ \|(\mathbb{1}_{[a,b]}^N - F_\tau \mathbb{1}_{[a,b]}^N)(F_\tau f_2)\|_2 \qquad \text{(D.4)} \\ &\leq \|f_1 - F_\tau f_1\|_2 + \|f_2 - F_\tau f_2\|_2 \\ &+ \|\mathbb{1}_{[a,b]}^N - F_\tau \mathbb{1}_{[a,b]}^N\|_2 \|F_\tau f_2\|_\infty, \end{aligned}$$

where in (D.4) we used

$$\begin{aligned} \big(F_\tau(\mathbb{1}_{[a,b]}^N f_2)\big)[n] &= (\mathbb{1}_{[a,b]}c_2)(n/N - \tau(n/N)) \\ &= \mathbb{1}_{[a,b]}(n/N - \tau(n/N))c_2((n/N - \tau(n/N))) \\ &= (F_\tau \mathbb{1}_{[a,b]}^N)[n](F_\tau f_2)[n]. \end{aligned}$$

With the upper bounds (D.2) and (D.3), invoking properties of $\mathcal{C}_{\mathrm{CART}}^{N,K}$ (namely, (i) c_1, c_2 satisfy the Lipschitz property with Lipschitz constant $C = K$ and hence $f_1[n] = c_1(n/N), f_2[n] = c_2(n/N)$, $n \in I_N$, satisfy (D.2) with $C = K$, and (ii) $\|F_\tau f_2\|_\infty = \sup_{n \in I_N} |(F_\tau f_2)[n]| = \sup_{n \in I_N} |c_2(n/N - \tau(n/N))| \leq \sup_{x \in \mathbb{R}} |c_2(x)| = \|c_2\|_\infty \leq K$), this yields

$$\begin{aligned} \|f - F_\tau f\|_2 &\leq 2KN^{1/2}\|\tau\|_\infty + 2KN^{1/2}\|\tau\|_\infty^{1/2} \\ &\leq 4KN^{1/2}\|\tau\|_\infty^{1/2}, \end{aligned}$$

where in the last step we used $\|\tau\|_\infty \leq \|\tau\|_\infty^{1/2}$, which is thanks to the assumption $\|\tau\|_\infty \leq 1$. This completes the proof of (D.1). □

It remains to establish (D.2) and (D.3).

Lemma D.1. *Let* $c : \mathbb{R} \to \mathbb{C}$ *be Lipschitz-continuous with Lipschitz constant* C. *Let further* $f[n] := c(n/N)$, $n \in I_N$. *Then,*

$$\|f - F_\tau f\|_2 \leq CN^{1/2}\|\tau\|_\infty.$$

Proof. Invoking the Lipschitz property of c according to

$$\begin{aligned} \|f - F_\tau f\|_2^2 &= \sum_{n \in I_N} |f[n] - (F_\tau f)[n]|^2 \\ &= \sum_{n \in I_N} |c(n/N) - c(n/N - \tau(n/N))|^2 \\ &\leq C^2 \sum_{n \in I_N} |\tau(n/N)|^2 \leq C^2 N \|\tau\|_\infty^2 \end{aligned}$$

completes the proof. □

We continue with a deformation sensitivity result for sampled indicator functions $\mathbb{1}_{[a,b]}(x)$.

Lemma D.2. *Let* $[a, b] \subseteq [0, 1]$ *and set* $\mathbb{1}_{[a,b]}^N[n] := \mathbb{1}_{[a,b]}(n/N)$, $n \in I_N$, *with* $a, b \notin \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$. *Then, we have*

$$\|\mathbb{1}_{[a,b]}^N - F_\tau \mathbb{1}_{[a,b]}^N\|_2 \leq 2N^{1/2}\|\tau\|_\infty^{1/2}.$$

Proof. In order to upper-bound

$$\begin{aligned} \|\mathbb{1}_{[a,b]}^N - F_\tau \mathbb{1}_{[a,b]}^N\|_2^2 &= \sum_{n \in I_N} |\mathbb{1}_{[a,b]}^N[n] - (F_\tau \mathbb{1}_{[a,b]}^N)[n]|^2 \\ &= \sum_{n \in I_N} |\mathbb{1}_{[a,b]}(n/N) - \mathbb{1}_{[a,b]}(n/N - \tau(n/N))|^2, \end{aligned}$$

we first note that the summand $h(n) := |\mathbb{1}_{[a,b]}(n/N) - \mathbb{1}_{[a,b]}(n/N - \tau(n/N))|^2$ satisfies $h(n) = 1$, for $n \in S$, where

$$\begin{aligned} S := &\left\{ n \in I_N \,\middle|\, \frac{n}{N} \in [a, b] \text{ and } \frac{n}{N} - \tau\Big(\frac{n}{N}\Big) \notin [a, b] \right\} \\ &\cup \left\{ n \in I_N \,\middle|\, \frac{n}{N} \notin [a, b] \text{ and } \frac{n}{N} - \tau\Big(\frac{n}{N}\Big) \in [a, b] \right\}, \end{aligned}$$

and $h(n) = 0$, for $n \in I_N \backslash S$. Thanks to $a, b \notin \{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\}$, we have $S \subseteq \Sigma$, where

$$\begin{aligned} \Sigma := &\left\{ n \in \mathbb{Z} \,\middle|\, \left| \frac{n}{N} - a \right| < \|\tau\|_\infty \right\} \\ &\cup \left\{ n \in \mathbb{Z} \,\middle|\, \left| \frac{n}{N} - b \right| < \|\tau\|_\infty \right\}. \end{aligned}$$

The cardinality of the set Σ can be upper-bounded by $2\,\frac{2\|\tau\|_\infty}{1/N}$, which then yields

$$\|\mathbb{1}^N_{[a,b]} - F_\tau \mathbb{1}^N_{[a,b]}\|_2^2 = \sum_{n\in I_N} |h(n)|^2 = \sum_{n\in S} 1 \leq \sum_{n\in\Sigma} 1 \leq 4N\|\tau\|_\infty. \tag{D.5}$$

This completes the proof.

Remark D.2. *For general $a, b \in [0,1]$, i.e., when we drop the assumption $a, b \notin \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$, it follows that $S \subseteq \Sigma'$, where*

$$\Sigma' := \Big\{n \in \mathbb{Z} \,\Big|\, \Big|\frac{n}{N} - a\Big| \leq \|\tau\|_\infty\Big\} \cup \Big\{n \in \mathbb{Z} \,\Big|\, \Big|\frac{n}{N} - b\Big| \leq \|\tau\|_\infty\Big\}.$$

Noting that the cardinality of Σ' can be upper-bounded by $2\left(\frac{2\|\tau\|_\infty}{1/N} + 1\right) = 4N\|\tau\|_\infty + 2$, this then yields (similarly to (D.5))

$$\|\mathbb{1}^N_{[a,b]} - F_\tau \mathbb{1}^N_{[a,b]}\|_2^2 \leq \sum_{n\in\Sigma} 1 \leq 4N\|\tau\|_\infty + 2,$$

which shows that the deformation error—for general $a, b \in [0,1]$—does not decay with $\|\tau\|_\infty^\alpha$ for some $\alpha > 0$ (see also the example in Remark D.1).

□

E. Appendix: Theorem 2

We start by establishing i). For ease of notation, again, we let $f_q := U[q]f$ and $h_q := U[q]h$, for $f, h \in H_{N_1}$, $q \in \Lambda_1^d$. We have

$$|||\Phi^d_\Omega(f) - \Phi^d_\Omega(h)|||^2 = \sum_{q\in\Lambda_1^d} ||(f_q - h_q) * \chi_d||_2^2 \tag{E.1}$$

$$\leq \|\chi_d\|_1^2 \underbrace{\sum_{q\in\Lambda_1^d} ||(f_q - h_q)||_2^2}_{=:a_d}, \tag{E.2}$$

where (E.2) follows by Young's inequality (Folland, 2015).

Remark E.1. *We emphasize that (E.1) can also be upper-bounded by $B_{d+1}\sum_{q\in\Lambda_1^d} ||(f_q - h_q)||_2^2$, which follows from the fact that $\{g_{\lambda_{d+1}}\}_{\lambda_{d+1}\in\Lambda_{d+1}} \cup \{\chi_d\}$ are atoms of the convolutional set Ψ_{d+1} with Bessel bound B_{d+1}. Hence, one can substitute $\|\chi_d\|_1$ in (15) by B_{d+1}.*

The key step is then to show that a_d can be upper-bounded according to

$$a_k \leq (B_k L_k^2 R_k^2) a_{k-1}, \qquad k = 1, \dots, d, \tag{E.3}$$

and to note that

$$a_d \leq (B_d L_d^2 R_d^2) a_{d-1} \leq \cdots \leq \Big(\prod_{k=1}^d B_k L_k^2 R_k^2\Big) a_0 = \Big(\prod_{k=1}^d B_k L_k^2 R_k^2\Big) \sum_{q\in\Lambda_1^0} \|f_q - h_q\|_2^2 = \Big(\prod_{k=1}^d B_k L_k^2 R_k^2\Big) \|f - h\|_2^2,$$

which yields (16). We now establish (E.3). Every path

$$\tilde{q} \in \Lambda_1^k = \underbrace{\Lambda_1 \times \cdots \times \Lambda_{k-1}}_{=\Lambda_1^{k-1}} \times \Lambda_k$$

of length k can be decomposed into a path $q \in \Lambda_1^{k-1}$ of length $k-1$ and an index $\lambda_k \in \Lambda_k$ according to $\tilde{q} = (q, \lambda_k)$. Thanks to (5) we have $U[\tilde{q}] = U[(q, \lambda_k)] = U_k[\lambda_k]U[q]$, which yields

$$\sum_{\tilde{q}\in\Lambda_1^k} \|f_{\tilde{q}} - h_{\tilde{q}}\|_2^2 = \sum_{q\in\Lambda_1^{k-1}} \sum_{\lambda_k\in\Lambda_k} \|U_k[\lambda_k]f_q - U_k[\lambda_k]h_q\|_2^2. \tag{E.4}$$

We next note that the term inside the sums on the RHS in (E.4) satisfies

$$\begin{aligned} &\|U_k[\lambda_k]f_q - U_k[\lambda_k]h_q\|_2^2 \\ &= \|P_k\big(\rho_k(f_q * g_{\lambda_k})\big) - P_k\big(\rho_k(h_q * g_{\lambda_k})\big)\|_2^2 \\ &\leq L_k^2 R_k^2 \|(f_q - h_q) * g_{\lambda_k}\|_2^2, \end{aligned} \tag{E.5}$$

where we used the Lipschitz continuity of P_k and ρ_k with Lipschitz constants $R_k > 0$ and $L_k > 0$, respectively. As $\{g_{\lambda_k}\}_{\lambda_k\in\Lambda_k} \cup \{\chi_{k-1}\}$ are the atoms of the convolutional set Ψ_k, and $f_q, h_q \in H_{N_k}$ by (5), we have

$$\sum_{\lambda_k\in\Lambda_k} \|(f_q - h_q) * g_{\lambda_k}\|_2^2 \leq B_k \|f_q - h_q\|_2^2,$$

which, when used in (E.5) together with (E.4), yields

$$\sum_{\tilde{q}\in\Lambda_1^k} \|f_{\tilde{q}} - h_{\tilde{q}}\|_2^2 \leq B_k L_k^2 R_k^2 \sum_{q\in\Lambda_1^{k-1}} \|f_q - h_q\|_2^2,$$

and hence establishes (E.3), thereby completing the proof of i).

We now turn to ii). The proof of (17) follows—as in the proof of ii) in Theorem 1 in Appendix C—from (16) together with $\Phi^d_\Omega(h) = \{(U[q]h) * \chi_d\}_{q\in\Lambda_1^d} = 0$ for $h = 0$, see (C.13).

We continue with iii). The proof of the deformation sensitivity bound (18) is based on two key ingredients. The

first one is the Lipschitz continuity result in (16). The second ingredient is, again, the deformation sensitivity bound (D.1) stated in Proposition D.1 in Appendix D. Combining (16) and (D.1)—as in the proof of iii) in Theorem 1 in Appendix C—then establishes (18) and completes the proof of iii).

We proceed to iv). For ease of notation, again, we let $f_q := U[q]f$, for $f \in H_{N_1}$, $q \in \Lambda_1^d$. Thanks to (5), we have $f_q \in H_{N_{d+1}}$, for $q \in \Lambda_1^d$. The key step in establishing (19) is to show that the operator U_k, $k \in \{1, 2, \ldots, d\}$, defined in (4) satisfies the relation

$$(U_k[\lambda_k]T_m f) = T_{m/S_k}(U_k[\lambda_k]f), \tag{E.6}$$

for $f \in H_{N_k}$, $m \in \mathbb{Z}$ with $\frac{m}{S_k} \in \mathbb{Z}$, and $\lambda_k \in \Lambda_k$. With the definition of $U[q]$ in (5) this then yields

$$(U[q]T_m f) = T_{m/(S_1 \cdots S_d)}(U[q]f), \tag{E.7}$$

for $f \in H_{N_1}$, $m \in \mathbb{Z}$ with $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$, and $q \in \Lambda_1^d$. The identity (19) is then a direct consequence of (E.7) and the translation-covariance of the circular convolution operator (which holds thanks to $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$):

$$\begin{aligned}\Phi_\Omega^d(T_m f) &= \big\{\big(U[q]T_m f\big) * \chi_d\big\}_{q\in\Lambda_1^d} \\ &= \big\{\big(T_{m/(S_1\cdots S_d)}U[q]f\big) * \chi_d\big\}_{q\in\Lambda_1^d} \\ &= \big\{T_{m/(S_1\cdots S_d)}\big((U[q]f) * \chi_d\big)\big\}_{q\in\Lambda_1^d} \\ &= T_{m/(S_1\cdots S_d)}\Phi_\Omega^d(f),\end{aligned}$$

for $f \in H_{N_1}$ and $m \in \mathbb{Z}$ with $\frac{m}{S_1 \ldots S_d} \in \mathbb{Z}$. It remains to establish (E.6):

$$\begin{aligned}(U_k[\lambda_k]T_m f) &= \Big(P_k\big(\rho_k((T_m f) * g_{\lambda_k})\big)\Big) \\ &= \Big(P_k\big(\rho_k(T_m(f * g_{\lambda_k}))\big)\Big) \qquad \text{(E.8)} \\ &= \Big(P_k\big(T_m(\rho_k(f * g_{\lambda_k}))\big)\Big), \qquad \text{(E.9)}\end{aligned}$$

where in (E.8) we used the translation covariance of the circular convolution operator (which holds thanks to $m \in \mathbb{Z}$), and in (E.9) we used the fact that point-wise non-linearities commute with the translation operator thanks to

$$\begin{aligned}(\rho_k T_m f)[n] &= \rho_k((T_m f)[n]) \\ &= \rho_k(f[n-m]) = (T_m \rho_k f)[n],\end{aligned}$$

for $f \in H_{N_k}$, $n \in I_{N_k}$, and $m \in \mathbb{Z}$. Next, we note that the pooling operators P_k in Section 2.3.1 (namely, sub-sampling, average pooling, and max-pooling) can all be written as $(P_k f)[n] = (P_k' f)[S_k n]$, for some P_k' that commutes with the translation operator, namely, for (i) sub-sampling $(P_k' f)[n] = f[n]$, with $(P_k' T_m f)[n] = (T_m f)[n] = f[n-m] = (T_m P_k' f)[n]$, (ii) average pooling $(P_k' f)[n] = \sum_{l=n}^{n+S_k-1} \alpha_{l-n} f[l]$ with

$$\begin{aligned}(P_k' T_m f)[n] &= \sum_{l=n}^{n+S_k-1} \alpha_{l-n} f[l-m] \\ &= \sum_{l'=(n-m)}^{(n-m)+S_k-1} \alpha_{l-(n-m)} f[l'] \\ &= (T_m P_k' f)[n],\end{aligned}$$

and for (iii) max-pooling $(P_k' f)[n] = \max_{l\in\{n,\ldots,n+S_k-1\}} |f[l]|$ with

$$\begin{aligned}(P_k' T_m f)[n] &= \max_{l\in\{n,\ldots,n+S_k-1\}} |f[l-m]| \\ &= \max_{(l-m)\in\{n-m,\ldots,(n-m)+S_k-1\}} |f[l-m]| \\ &= \max_{l'\in\{(n-m),\ldots,(n-m)+S_k-1\}} |f[l']| \\ &= (T_m P_k' f)[n],\end{aligned}$$

in all three cases for $f \in H_{N_k}$, $n \in I_{N_k}$, and $m \in \mathbb{Z}$. This then yields

$$\begin{aligned}(P_k T_m f)[n] &= (P_k' T_m f)[S_k n] = (T_m P_k' f)[S_k n] \\ &= P_k'(f)[S_k n - m] \\ &= P_k'(f)[S_k(n - S_k^{-1} m)] \\ &= P_k(f)[n - S_k^{-1} m] \\ &= (T_{m/S_k} P_k f)[n], \qquad \text{(E.10)}\end{aligned}$$

for $f \in H_{N_k}$ and $n \in I_{N_{k+1}}$. Here, we used $m/S_k \in \mathbb{Z}$, which is by assumption. Substituting (E.10) into (E.9) finally yields

$$(U_k[\lambda_k]T_m f) = T_{m/S_k} U_k[\lambda_k] f,$$

for $f \in H_{N_k}$, $m \in \mathbb{Z}$ with $\frac{m}{S_k} \in \mathbb{Z}$, and $\lambda_k \in \Lambda_k$. This completes the proof of (E.6) and hence establishes (19).