Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics

Deep Convolutional Neural Networks on Cartoon Functions

P. Grohs and T. Wiatowski and H. Helmut Boelcskei

Research Report No. 2016-25
April 2016

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

Deep Convolutional Neural Networks on Cartoon Functions

Philipp Grohs*, Thomas Wiatowski†, and Helmut Bölcskei†
*Dept. Math., ETH Zurich, Switzerland, and Dept. Math., University of Vienna, Austria
†Dept. IT & EE, ETH Zurich, Switzerland,
*philipp.grohs@sam.math.ethz.ch, †{withomas, boelcskei}@nari.ee.ethz.ch

***Abstract*—Wiatowski and Bölcskei, 2015, proved that deformation stability and vertical translation invariance of deep convolutional neural network-based feature extractors are guaranteed by the network structure per se rather than the specific convolution kernels and non-linearities. While the translation invariance result applies to square-integrable functions, the deformation stability bound holds for band-limited functions only. Many signals of practical relevance (such as natural images) exhibit, however, sharp and curved discontinuities and are hence not band-limited. The main contribution of this paper is a deformation stability result that takes these structural properties into account. Specifically, we establish deformation stability bounds for the class of cartoon functions introduced by Donoho, 2001.**

I. Introduction

Feature extractors based on so-called deep convolutional neural networks have been applied with tremendous success in a wide range of practical signal classification tasks [1]–[3]. These networks are composed of multiple layers, each of which computes convolutional transforms, followed by the application of non-linearities and pooling operations.

The mathematical analysis of feature extractors generated by deep convolutional neural networks was initiated in a seminal paper by Mallat [4]. Specifically, Mallat analyzes so-called scattering networks, where signals are propagated through layers that compute semi-discrete wavelet transforms (i.e., convolutional transforms with pre-specified filters obtained from a mother wavelet through scaling operations), followed by modulus non-linearities. It was shown in [4] that the resulting wavelet-modulus feature extractor is horizontally translation-invariant [5] and deformation-stable, with the stability result applying to a function space that depends on the underlying mother wavelet.

Recently, Wiatowski and Bölcskei [5] extended Mallat's theory to incorporate convolutional transforms with filters that are (i) pre-specified and potentially structured such as Weyl-Heisenberg (Gabor) functions [6], wavelets [7], curvelets [8], shearlets [9], and ridgelets [10], (ii) pre-specified and unstructured such as random filters [11], and (iii) learned in a supervised [12] or unsupervised [13] fashion. Furthermore, the networks in [5] may employ general Lipschitz-continuous non-linearities (e.g., rectified linear units, shifted logistic sigmoids, hyperbolic tangents, and the modulus function) and pooling through sub-sampling. The essence of the results in [5] is that vertical translation invariance and deformation stability are induced by the network structure per se rather than the specific choice of filters and non-linearities. While the vertical translation invariance result in [5] is general in the sense of applying to the function space $L^2(\mathbb{R}^d)$, the deformation stability result in [5] pertains to square-integrable band-limited functions. Moreover, the corresponding deformation stability bound depends linearly on the bandwidth.

Many signals of practical relevance (such as natural images) can be modeled as square-integrable functions that are, however, not band-limited or have large bandwidth. Large bandwidths render the deformation stability bound in [5] void as a consequence of its linear dependence on bandwidth.

Contributions. The question considered in this paper is whether taking structural properties of natural images into account can lead to stronger deformation stability bounds. We show that the answer is in the affirmative by analyzing the class of cartoon functions introduced in [14]. Cartoon functions satisfy mild decay properties and are piecewise continuously differentiable apart from curved discontinuities along Lipschitz-continuous hypersurfaces. Moreover, they provide a good model for natural images such as those in the MNIST [15], Caltech-256 [16], and CIFAR-100 [17] datasets as well as for images of geometric objects of different shapes, sizes, and colors [18], [19]. The proof of our main result is based on the decoupling technique introduced in [5]. The essence of decoupling is that contractivity of the feature extractor combined with deformation stability of the signal class under consideration—under smoothness conditions on the deformation—establishes deformation stability for the feature extractor. Our main technical contribution here is to prove deformation stability for the class of cartoon functions. Moreover, we show that the decay rate of the resulting deformation stability bound is best possible. The results we obtain further underpin the observation made in [5] of deformation stability and vertical translation invariance being induced by the network structure per se.

Notation. We refer the reader to [5, Sec. 1] for the general notation employed in this paper. In addition, we will need the following notation. For $x \in \mathbb{R}^d$, we set $\langle x \rangle := (1+|x|^2)^{1/2}$. The Minkowski sum of sets $A, B \subseteq \mathbb{R}^d$ is $(A+B) := \{a+b \,|\, a \in A,\ b \in B\}$. A Lipschitz domain D is a set $D \subseteq \mathbb{R}^d$ whose boundary ∂D is "sufficiently regular" to be thought of as locally being the graph of a Lipschitz-continuous function, for a formal definition see [20, Def. 1.40]. The indicator function of a set $B \subseteq \mathbb{R}^d$ is defined as $\mathbb{1}_B(x) := 1$,

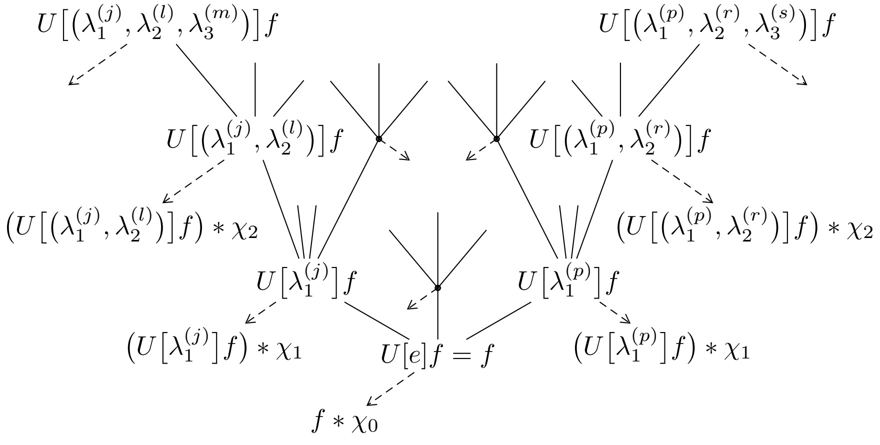
$U\big[\big(\lambda_1^{(j)}, \lambda_2^{(l)}, \lambda_3^{(m)}\big)\big]f$
$U\big[\big(\lambda_1^{(p)}, \lambda_2^{(r)}, \lambda_3^{(s)}\big)\big]f$
$U\big[\big(\lambda_1^{(j)}, \lambda_2^{(l)}\big)\big]f$
$U\big[\big(\lambda_1^{(p)}, \lambda_2^{(r)}\big)\big]f$
$\big(U\big[\big(\lambda_1^{(j)}, \lambda_2^{(l)}\big)\big]f\big) * \chi_2$
$\big(U\big[\big(\lambda_1^{(p)}, \lambda_2^{(r)}\big)\big]f\big) * \chi_2$
$U\big[\lambda_1^{(j)}\big]f$
$U\big[\lambda_1^{(p)}\big]f$
$\big(U\big[\lambda_1^{(j)}\big]f\big) * \chi_1$
$U[e]f = f$
$\big(U\big[\lambda_1^{(p)}\big]f\big) * \chi_1$
$f * \chi_0$

Fig. 1: Network architecture underlying the feature extractor (2). The index $\lambda_n^{(k)}$ corresponds to the k-th atom $g_{\lambda_n^{(k)}}$ of the collection Ψ_n associated with the n-th network layer. The function χ_n is the output-generating atom of the n-th layer.

for $x \in B$, and $\mathbb{1}_B(x) := 0$, for $x \in \mathbb{R}^d \backslash B$. For a measurable set $B \subseteq \mathbb{R}^d$, we let $\mathrm{vol}^d(B) := \int_{\mathbb{R}^d} \mathbb{1}_B(x) \mathrm{d}x = \int_B 1 \mathrm{d}x$.

II. Deep convolutional neural network-based feature extractors

We set the stage by briefly reviewing the deep convolutional feature extraction network presented in [5], the basis of which is a sequence of triplets $\Omega := \big((\Psi_n, M_n, R_n)\big)_{n \in \mathbb{N}}$ referred to as module-sequence. The triplet (Ψ_n, M_n, R_n)—associated with the n-th network layer—consists of (i) a collection $\Psi_n := \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ of so-called atoms $g_{\lambda_n} \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, indexed by a countable set Λ_n and satisfying the Bessel condition $\sum_{\lambda_n \in \Lambda_n} \|f * g_{\lambda_n}\|^2 \leq B_n \|f\|_2^2$, for all $f \in L^2(\mathbb{R}^d)$, for some $B_n > 0$, (ii) an operator $M_n : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ satisfying the Lipschitz property $\|M_n f - M_n h\|_2 \leq L_n \|f - h\|_2$, for all $f, h \in L^2(\mathbb{R}^d)$, and $M_n f = 0$ for $f = 0$, and (iii) a sub-sampling factor $R_n \geq 1$. Associated with (Ψ_n, M_n, R_n), we define the operator

$$U_n[\lambda_n]f := R_n^{d/2} \big(M_n(f * g_{\lambda_n})\big)(R_n \cdot), \tag{1}$$

and extend it to paths on index sets $q = (\lambda_1, \lambda_2, \ldots, \lambda_n) \in \Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_n := \Lambda_1^n$, $n \in \mathbb{N}$, according to

$$\begin{aligned} U[q]f =& U[(\lambda_1, \lambda_2, \ldots, \lambda_n)]f \\ :=& U_n[\lambda_n] \cdots U_2[\lambda_2] U_1[\lambda_1] f, \end{aligned}$$

where for the empty path $e := \emptyset$ we set $\Lambda_1^0 := \{e\}$ and $U[e]f := f$, for $f \in L^2(\mathbb{R}^d)$.

Remark 1. *The Bessel condition on the atoms g_{λ_n} is equivalent to $\sum_{\lambda_n \in \Lambda_n} |\widehat{g_{\lambda_n}}(\omega)|^2 \leq B_n$, for a.e. $\omega \in \mathbb{R}^d$ (see [5, Prop. 2]), and is hence easily satisfied even by learned filters [5, Remark 2]. An overview of collections $\Psi_n = \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ of structured atoms g_{λ_n} (such as, e.g., Weyl-Heisenberg (Gabor) functions, wavelets, curvelets, shearlets, and ridgelets) and non-linearities M_n widely used in the deep learning literature (e.g., hyperbolic tangent, shifted logistic sigmoid, rectified linear unit, and modulus function) is provided in [5, App. B-D].*

For every $n \in \mathbb{N}$, we designate one of the atoms $\Psi_n = \{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n}$ as the output-generating atom $\chi_{n-1} := g_{\lambda_n^*}$, $\lambda_n^* \in \Lambda_n$, of the $(n-1)$-th layer. The atoms $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n \backslash \{\lambda_n^*\}} \cup \{\chi_{n-1}\}$ are thus used across two consecutive layers in the sense of $\chi_{n-1} = g_{\lambda_n^*}$ generating the output in the $(n-1)$-th layer, and the remaining atoms $\{g_{\lambda_n}\}_{\lambda_n \in \Lambda_n \backslash \{\lambda_n^*\}}$ propagating signals to the n-th layer according to (1), see Fig. 1. From now on, with slight abuse of notation, we write Λ_n for $\Lambda_n \backslash \{\lambda_n^*\}$ as well.

The extracted features $\Phi_\Omega(f)$ of a signal $f \in L^2(\mathbb{R}^d)$ are defined as [5, Def. 3]

$$\Phi_\Omega(f) := \bigcup_{n=0}^{\infty} \{(U[q]f) * \chi_n\}_{q \in \Lambda_1^n}, \tag{2}$$

where $(U[q]f) * \chi_n$, $q \in \Lambda_1^n$, is a feature generated in the n-th layer of the network, see Fig. 1. It is shown in [5, Thm. 2] that for all $f \in L^2(\mathbb{R}^d)$ the feature extractor Φ_Ω is vertically translation-invariant in the sense of the layer depth n determining the extent to which the features $(U[q]f) * \chi_n$, $q \in \Lambda_1^n$, are translation-invariant. Furthermore, under the condition

$$\max_{n \in \mathbb{N}} \max\{B_n, B_n L_n^2\} \leq 1, \tag{3}$$

referred to as *weak admissibility condition* in [5, Def. 4] and satisfied by a wide variety of module sequences Ω (see [5, Sec. 3]), the following result is established in [5, Thm. 1]: The feature extractor Φ_Ω is stable on the space of R-band-limited functions $L_R^2(\mathbb{R}^d)$ w.r.t. deformations $(F_\tau f)(x) := f(x - \tau(x))$, i.e., there exists a universal constant $C > 0$ (that does not depend on Ω) such that for all $f \in L_R^2(\mathbb{R}^d)$ and all (possibly non-linear) $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq CR\|\tau\|_\infty \|f\|_2. \tag{4}$$

Here, the feature space norm is defined as $|||\Phi_\Omega(f)|||^2 := \sum_{n=0}^{\infty} \sum_{q \in \Lambda_1^n} \|(U[q]f) * \chi_n\|_2^2$.

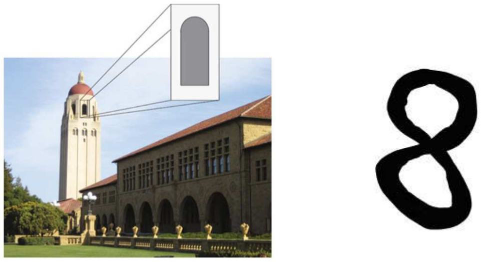

Fig. 2: Left: A natural image (image credit: [21]) is typically governed by areas of little variation, with the individual areas separated by edges that can be modeled as curved singularities. Right: An image of a handwritten digit.

For practical classification tasks, we can think of the deformation F_τ as follows. Let f be a representative of a certain signal class, e.g., f is an image of the handwritten digit "8" (see Fig. 2, right). Then, $\{F_\tau f \mid \|D\tau\|_\infty < \frac{1}{2d}\}$ is a collection of images of the handwritten digit "8", where each $F_\tau f$ may be generated, e.g., based on a different handwriting style. The bound $\|D\tau\|_\infty < \frac{1}{2d}$ on the Jacobian matrix of τ imposes a quantitative limit on the amount of deformation tolerated, rendering the bound (4) to implicitly depend on $D\tau$. The stability bound (4) now guarantees that the features corresponding to the images in the set $\{F_\tau f \mid \|D\tau\|_\infty < \frac{1}{2d}\}$ do not differ too much.

III. Cartoon functions

The bound in (4) applies to the space of square-integrable R-band-limited functions. Many signals of practical significance (e.g., natural images) are, however, not band-limited (due to the presence of sharp and possibly curved edges, see Fig. 2) or exhibit large bandwidths. In the latter case, the deformation stability bound (4) becomes void as it depends linearly on R.

The goal of this paper is to take structural properties of natural images into account by considering the class of cartoon functions introduced in [14]. These functions satisfy mild decay properties and are piecewise continuously differentiable apart from curved discontinuities along Lipschitz-continuous hypersurfaces. Cartoon functions provide a good model for natural images (see Fig. 2, left) such as those in the Caltech-256 [16] and CIFAR-100 [17] data sets, for images of handwritten digits [15] (see Fig. 2, right), and for images of geometric objects of different shapes, sizes, and colors [18], [19].

We proceed to the formal definition of cartoon functions.

Definition 1. *The function $f : \mathbb{R}^d \to \mathbb{C}$ is referred to as a cartoon function if it can be written as $f = f_1 + \mathbb{1}_B f_2$, where $B \subseteq \mathbb{R}^d$ is a compact Lipschitz domain with boundary of finite length, i.e., $vol^{d-1}(\partial B) < \infty$, and $f_i \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d, \mathbb{C})$, $i = 1, 2$, satisfies the decay condition*

$$|\nabla f_i(x)| \le C\langle x\rangle^{-d}, \quad i = 1, 2, \tag{5}$$

for some $C > 0$ (that does not depend on f_1, f_2). Furthermore, we denote by

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

the class of cartoon functions of maximal size $K > 0$.

We chose the term size to indicate the length $\mathrm{vol}^{d-1}(\partial B)$ of the boundary ∂B of the Lipschitz domain B. Furthermore, $\mathcal{C}^K_{\mathrm{CART}} \subseteq L^2(\mathbb{R}^d)$, for all $K > 0$; this simply follows from the triangle inequality according to $\|f_1 + \mathbb{1}_B f_2\|_2 \le \|f_1\|_2 + \|\mathbb{1}_B f_2\|_2 \le \|f_1\|_2 + \|f_2\|_2 < \infty$, where in the last step we used $f_1, f_2 \in L^2(\mathbb{R}^d)$. Finally, we note that our main results—presented in the next section—can easily be generalized to finite linear combinations of cartoon functions, but this is not done here for simplicity of exposition.

IV. Main results

We start by reviewing the decoupling technique introduced in [5] to prove deformation stability bounds for band-limited functions. The proof of the deformation stability bound (4) for band-limited functions in [5] is based on two key ingredients. The first one is a contractivity property of Φ_Ω (see [5, Prop. 4]), namely $|||\Phi_\Omega(f) - \Phi_\Omega(h)||| \le \|f - h\|_2$, for all $f, h \in L^2(\mathbb{R}^d)$. Contractivity guarantees that pairwise distances of input signals do not increase through feature extraction. The second ingredient is an upper bound on the deformation error $\|f - F_\tau f\|_2$ (see [5, Prop. 5]), specific to the signal class considered in [5], namely band-limited functions. Recognizing that the combination of these two ingredients yields a simple proof of deformation stability is interesting as it shows that whenever a signal class exhibits inherent stability w.r.t. deformations of the form $(F_\tau f)(x) = f(x - \tau(x))$, we automatically obtain deformation stability for the feature extractor Φ_Ω. The present paper employs this decoupling technique and establishes deformation stability for the class of cartoon functions by deriving an upper bound on the deformation error $\|f - F_\tau f\|_2$ for $f \in \mathcal{C}^K_{\mathrm{CART}}$.

Proposition 1. *For every $K > 0$ there exists a constant $C_K > 0$ such that for all $f \in \mathcal{C}^K_{\mathrm{CART}}$ and all (possibly non-linear) $\tau : \mathbb{R}^d \to \mathbb{R}^d$ with $\|\tau\|_\infty < \frac{1}{2}$, it holds that*

$$\|f - F_\tau f\|_2 \le C_K \|\tau\|_\infty^{1/2}. \tag{6}$$

Proof. see Appendix A. □

The Lipschitz exponent $\alpha = \frac{1}{2}$ on the right-hand side (RHS) of (6) determines the decay rate of the deformation error $\|f - F_\tau f\|_2$ as $\|\tau\|_\infty \to 0$. Clearly, larger $\alpha > 0$ results in the deformation error decaying faster as the deformation becomes smaller. The following simple example shows that the Lipschitz exponent $\alpha = \frac{1}{2}$ in (6) is best possible, i.e., it can not be larger. Consider $d = 1$ and $\tau_s(x) = s$, for a fixed s satisfying $0 < s < \frac{1}{2}$; the corresponding deformation F_{τ_s} amounts to a simple translation by s with $\|\tau_s\|_\infty = s < \frac{1}{2}$. Let $f = \mathbb{1}_{[-1,1]}$. Then $f \in \mathcal{C}^K_{\mathrm{CART}}$ for some $K > 0$ and $\|f - F_{\tau_s} f\|_2 = \sqrt{2s} = \sqrt{2}\|\tau\|_\infty^{1/2}$.

Remark 2. *It is interesting to note that in order to obtain bounds of the form $\|f - F_\tau f\|_2 \leq C\|\tau\|_\infty^\alpha$, for $f \in \mathcal{C} \subseteq L^2(\mathbb{R}^d)$, for some $C > 0$ (that does not depend on f, τ) and some $\alpha > 0$, we need to impose non-trivial constraints on the set $\mathcal{C} \subseteq L^2(\mathbb{R}^d)$. Indeed, consider, again, $d = 1$ and $\tau_s(x) = s$, for small $s > 0$. Let $f_s \in L^2(\mathbb{R}^d)$ be a function that has its energy $\|f_s\|_2 = 1$ concentrated in a small interval according to supp$(f_s) \subseteq [-s/2, s/2]$. Then, f_s and $F_{\tau_s} f_s$ have disjoint support sets and hence $\|f_s - F_{\tau_s} f_s\|_2 = \sqrt{2}$, which does not decay with $\|\tau\|_\infty^\alpha = s^\alpha$ for any $\alpha > 0$. More generally, the amount of deformation induced by a given function τ depends strongly on the signal (class) it is applied to. Concretely, the deformation F_τ with $\tau(x) = e^{-x^2}$, $x \in \mathbb{R}$, will lead to a small bump around the origin only when applied to a lowpass function, whereas the function f_s above will experience a significant deformation.*

We are now ready to state our main result.

Theorem 1. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a module-sequence satisfying the weak admissibility condition* (3)*. For every size $K > 0$, the feature extractor Φ_Ω is stable on the space of cartoon functions $\mathcal{C}_{\mathrm{CART}}^K$ w.r.t. deformations $(F_\tau f)(x) = f(x - \tau(x))$, i.e., 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 (possibly non-linear) $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|\tau\|_\infty < \frac{1}{2}$ and $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that*

$$|||\Phi_\Omega(F_\tau f) - \Phi_\Omega(f)||| \leq C_K \|\tau\|_\infty^{1/2}. \tag{7}$$

Proof. Applying the contractivity property $|||\Phi_\Omega(g) - \Phi_\Omega(h)||| \leq \|g - h\|_2$ with $g = F_\tau f$ and $h = f$, and using (6) yields (7) upon invoking the same arguments as in [5, Eq. 58] and [5, Lemma 2] to conclude that $f \in L^2(\mathbb{R}^d)$ implies $F_\tau f \in L^2(\mathbb{R}^d)$ thanks to $\|D\tau\|_\infty \leq \frac{1}{2d}$. □

The strength of the deformation stability result in Theorem 1 derives itself from the fact that the only condition we need to impose on the underlying module-sequence Ω is weak admissibility according to (3), which as argued in [5, Sec. 3], can easily be met by normalizing the elements in Ψ_n, for all $n \in \mathbb{N}$, appropriately. We emphasize that this normalization does not have an impact on the constant C_K in (7), which is shown in Appendix A to be independent of Ω. The dependence of C_K on K does, however, reflect the intuition that the deformation stability bound should depend on the signal class description complexity. For band-limited signals, this dependence is exhibited by the RHS in (4) being linear in the bandwidth R. Finally, we note that the vertical translation invariance result [5, Thm. 2] applies to all $f \in L^2(\mathbb{R}^d)$, and, thanks to $\mathcal{C}_{\mathrm{CART}}^K \subseteq L^2(\mathbb{R}^d)$, for all $K > 0$, carries over to cartoon functions.

Remark 3. *We note that thanks to the decoupling technique underlying our arguments, the deformation stability bounds* (4) *and* (7) *are very general in the sense of applying to every contractive (linear or non-linear) mapping Φ. Specifically, the identity mapping $\Phi(f) = f$ also leads to deformation stability on the class of cartoon functions (and the class of band-limited functions). This is interesting as it was recently demonstrated that employing the identity mapping as a so-called shortcut-connection in a subset of layers of a very deep convolutional neural network yields state-of-the-art classification performance on the ImageNet dataset [22]. Our deformation stability result is hence general in the sense of applying to a broad class of network architectures used in practice.*

For functions that do not exhibit discontinuities along Lipschitz-continuous hypersurfaces, but otherwise satisfy the decay condition (5), we can improve the decay rate of the deformation error from $\alpha = \frac{1}{2}$ to $\alpha = 1$.

Corollary 1. *Let $\Omega = \big((\Psi_n, M_n, R_n)\big)_{n\in\mathbb{N}}$ be a module-sequence satisfying the weak admissibility condition* (3)*. For every size $K > 0$, the feature extractor Φ_Ω is stable on the space $H_K := \{f \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d, \mathbb{C}) \mid |\nabla f(x)| \leq K\langle x\rangle^{-d}\}$ w.r.t. deformations $(F_\tau f)(x) = f(x - \tau(x))$, i.e., for every $K > 0$ there exists a constant $C_K > 0$ (that does not depend on Ω) such that for all $f \in H_K$, and all (possibly non-linear) $\tau \in C^1(\mathbb{R}^d, \mathbb{R}^d)$ with $\|\tau\|_\infty < \frac{1}{2}$ and $\|D\tau\|_\infty \leq \frac{1}{2d}$, it holds that*

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

Proof. The proof follows that of Theorem 1 apart from employing (12) instead of (6). □

Appendix A
Proof of Proposition 1

The proof of (6) is based on judiciously combining deformation stability bounds for the components f_1, f_2 in $(f_1 + \mathbb{1}_B f_2) \in \mathcal{C}_{\mathrm{CART}}^K$ and for the indicator function $\mathbb{1}_B$. The first bound, stated in Lemma 1 below, reads

$$\|f - F_\tau f\|_2 \leq CD\|\tau\|_\infty, \tag{8}$$

and applies to functions f satisfying the decay condition (11), with the constant $D > 0$ not depending on f and τ (see (14)). The bound in (8) needs the assumption $\|\tau\|_\infty < \frac{1}{2}$. The second bound, stated in Lemma 2 below, is

$$\|\mathbb{1}_B - F_\tau \mathbb{1}_B\|_2 \leq \big(2\,\mathrm{vol}^{d-1}(\partial B)\big)^{1/2} \|\tau\|_\infty^{1/2}. \tag{9}$$

We now show how (8) and (9) can be combined to establish (6). For $f = (f_1 + \mathbb{1}_B f_2) \in \mathcal{C}_{\mathrm{CART}}^K$, we have

$$\begin{aligned} &\|f - F_\tau f\|_2 \leq \|f_1 - F_\tau f_1\|_2 \\ &+ \|\mathbb{1}_B(f_2 - F_\tau f_2)\|_2 + \|(\mathbb{1}_B - F_\tau \mathbb{1}_B)(F_\tau f_2)\|_2 \qquad (10) \\ &\leq \|f_1 - F_\tau f_1\|_2 + \|f_2 - F_\tau f_2\|_2 + \|\mathbb{1}_B - F_\tau \mathbb{1}_B\|_2 \|F_\tau f_2\|_\infty, \end{aligned}$$

where in (10) we used $F_\tau(\mathbb{1}_B f_2)(x) = (\mathbb{1}_B f_2)(x - \tau(x)) = \mathbb{1}_B(x - \tau(x)) f_2((x - \tau(x))) = (F_\tau \mathbb{1}_B)(x)(F_\tau f_2)(x)$. With the upper bounds (8) and (9), invoking properties of the class of cartoon functions $\mathcal{C}_{\mathrm{CART}}^K$ (namely, (i) $\mathrm{vol}^{d-1}(\partial B) \leq K$, (ii) f_1,f_2 satisfy (11) and thus (8) with $C = K$, and (iii)

$\|F_\tau f_2\|_\infty = \sup_{x\in\mathbb{R}^d}|f_2(x-\tau(x))| \leq \sup_{y\in\mathbb{R}^d}|f_2(y)| = \|f_2\|_\infty \leq K$), this yields

$$\begin{aligned}\|f - F_\tau f\|_2 &\leq 2\,KD\,\|\tau\|_\infty + \sqrt{2}\,K^{3/2}\|\tau\|_\infty^{1/2}\\ &\leq \underbrace{2\max\{2KD,\sqrt{2}K^{3/2}\}}_{=:C_K}\|\tau\|_\infty^{1/2},\end{aligned}$$

which completes the proof of (6).

It remains to show (8) and (9).

Lemma 1. *Let $f \in L^2(\mathbb{R}^d)\cap C^1(\mathbb{R}^d,\mathbb{C})$ be such that*

$$|\nabla f(x)| \leq C\langle x\rangle^{-d}, \tag{11}$$

for some constant $C > 0$, and let $\|\tau\|_\infty < \frac{1}{2}$. Then,

$$\|f - F_\tau f\|_2 \leq CD\|\tau\|_\infty, \tag{12}$$

for a constant $D > 0$ that does not depend on f and τ.

Proof. We first upper-bound the integrand in $\|f-F_\tau f\|_2^2 = \int_{\mathbb{R}^d}|f(x)-f(x-\tau(x))|^2\mathrm{d}x$. Owing to the mean value theorem [23, Thm. 3.7.5], we have

$$\begin{aligned}|f(x)-f(x-\tau(x))| &\leq \|\tau\|_\infty \sup_{y\in B_{\|\tau\|_\infty}(x)}|\nabla f(y)|\\ &\leq \underbrace{C\|\tau\|_\infty \sup_{y\in B_{\|\tau\|_\infty}(x)}\langle y\rangle^{-d}}_{=:h(x)},\end{aligned}$$

where the last inequality follows by assumption. The idea is now to split the integral $\int_{\mathbb{R}^d}|h(x)|^2\mathrm{d}x$ into integrals over the sets $B_1(0)$ and $\mathbb{R}^d\backslash B_1(0)$. For $x\in B_1(0)$, the monotonicity of the function $x\mapsto\langle x\rangle^{-d}$ implies $h(x)\leq C\|\tau\|_\infty\langle 0\rangle^{-d} = C\|\tau\|_\infty$, and for $x\in\mathbb{R}^d\backslash B_1(0)$, we have $(1-\|\tau\|_\infty)\leq(1-\frac{\|\tau\|_\infty}{|x|})$, which together with the monotonicity of $x\mapsto\langle x\rangle^{-d}$ yields $h(x)\leq C\|\tau\|_\infty\langle(1-\frac{\|\tau\|_\infty}{|x|})x\rangle^{-d}\leq C\|\tau\|_\infty\langle(1-\|\tau\|_\infty)x\rangle^{-d}$. Putting things together, we hence get

$$\begin{aligned}\|f-F_\tau f\|_2^2 &\leq C^2\|\tau\|_\infty^2\Big(\mathrm{vol}^d\big(B_1(0)\big)\\ &+ 2^d\int_{\mathbb{R}^d}\langle u\rangle^{-2d}\mathrm{d}u\Big) &(13)\\ &\leq C^2\|\tau\|_\infty^2\underbrace{\Big(\mathrm{vol}^d\big(B_1(0)\big)+2^d\|\langle\cdot\rangle^{-d}\|_2^2\Big)}_{=:D^2}, &(14)\end{aligned}$$

where in (13) we used the change of variables $u = (1-\|\tau\|_\infty)x$, together with

$$\frac{\mathrm{d}u}{\mathrm{d}x} = (1-\|\tau\|_\infty)^d \geq 2^{-d}. \tag{15}$$

The inequality in (15) follows from $\|\tau\|_\infty < \frac{1}{2}$, which is by assumption. Since $\|\langle\cdot\rangle^{-d}\|_2 < \infty$, for $d\in\mathbb{N}$ (see, e.g., [24, Sec. 1]), and, obviously, $\mathrm{vol}^d\big(B_1(0)\big) < \infty$, it follows that $D^2 < \infty$, which completes the proof. □

We continue with a deformation stability result for indicator functions $\mathbb{1}_B$.

Lemma 2. *Let $B\subseteq\mathbb{R}^d$ be a compact Lipschitz domain with boundary of finite length, i.e., $vol^{d-1}(\partial B) < \infty$. Then,*

$$\|\mathbb{1}_B - F_\tau\mathbb{1}_B\|_2 \leq \big(2\,vol^{d-1}(\partial B)\big)^{1/2}\|\tau\|_\infty^{1/2}.$$

Proof. In order to upper-bound $\|\mathbb{1}_B - F_\tau\mathbb{1}_B\|_2^2 = \int_{\mathbb{R}^d}|\mathbb{1}_B(x)-\mathbb{1}_B(x-\tau(x))|^2\mathrm{d}x$, we first note that the integrand $h(x) := |\mathbb{1}_B(x)-\mathbb{1}_B(x-\tau(x))|^2$ satisfies $h(x) = 1$, for $x\in S$, where $S := \{x\in\mathbb{R}^d\,|\,x\in B \text{ and } x-\tau(x)\notin B\}\cup\{x\in\mathbb{R}^d\,|\,x\notin B \text{ and } x-\tau(x)\in B\}$, and $h(x) = 0$, for $x\in\mathbb{R}^d\backslash S$. Since $S\subseteq\big(\partial B + B_{\|\tau\|_\infty}(0)\big)$, where $(\partial B + B_{\|\tau\|_\infty}(0))$ is a tubular neighborhood of width $\|\tau\|_\infty$ around the boundary ∂B of B, we have $\|\mathbb{1}_B - F_\tau\mathbb{1}_B\|_2^2 = \int_{\mathbb{R}^d}|h(x)|^2\mathrm{d}x = \int_S 1\mathrm{d}x \leq \int_{\partial B + B_{\|\tau\|_\infty}(0)}1\mathrm{d}x \leq 2\,\mathrm{vol}^{d-1}(\partial B)\|\tau\|_\infty$, which completes the proof.

□

References

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

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

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[4] S. Mallat, "Group invariant scattering," *Comm. Pure Appl. Math.*, vol. 65, no. 10, pp. 1331–1398, 2012.

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

[6] K. Gröchening, *Foundations of time-frequency analysis.* Birkhäuser, 2001.

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

[8] E. J. Candès and D. L. Donoho, "Continuous curvelet transform: II. Discretization and frames," *Appl. Comput. Harmon. Anal.*, vol. 19, no. 2, pp. 198–222, 2005.

[9] G. Kutyniok and D. Labate, Eds., *Shearlets: Multiscale analysis for multivariate data.* Birkhäuser, 2012.

[10] E. J. Candès, "Ridgelets: Theory and applications," Ph.D. dissertation, Stanford University, 1998.

[11] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2146–2153.

[12] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional nets for generic object categorization," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 284–291.

[13] M. A. Ranzato, F. J. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[14] D. L. Donoho, "Sparse components of images and optimal atomic decompositions," *Constructive Approximation*, vol. 17, no. 3, pp. 353–382, 2001.

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

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

[17] A. Krizhevsky, "Learning multiple layers of features from tiny images," Master's thesis, University of Toronto, 2009.

[18] "The baby AI school dataset," http://www.iro.umontreal.ca/%7Elisa/twiki/bin/view.cgi/Public/BabyAISchool, 2007.

[19] "The rectangles dataset," http://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/RectanglesData, 2007.

[20] B. Dacorogna, *Introduction to the calculus of variations.* Imperial College Press, 2004.

[21] G. Kutyniok and D. Labate, "Introduction to shearlets," in *Shearlets: Multiscale analysis for multivariate data*, G. Kutyniok and D. Labate, Eds. Birkhäuser, 2012, pp. 1–38.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[23] M. Comenetz, *Calculus: The elements.* World Scientific, 2002.

[24] L. Grafakos, *Classical Fourier analysis*, 2nd ed. Springer, 2008.