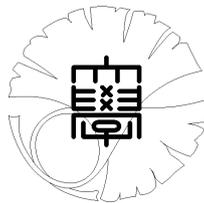


数理科学実践研究レター 2018–23 September 21, 2018

深層畳み込みニューラルネットワークによる
カートゥン関数の特徴量について

by

上坂正晃、田口和稔



UNIVERSITY OF TOKYO
GRADUATE SCHOOL OF MATHEMATICAL SCIENCES
KOMABA, TOKYO, JAPAN

深層畳み込みニューラルネットワークによるカートゥン関数の特徴量について

上坂正晃¹ (北海道大学電子科学研究所)

Masaaki Uesaka (Research Institute of Electron Sciences, Hokkaido University)

田口和稔² (東京大学大学院数理科学研究科)

Kazutoshi Taguchi (Graduate School of Mathematical Sciences, The University of Tokyo)

概要

Wiatowski et al. (2016) において、深層畳み込みニューラルネットワークによるカートゥン関数の特徴量は deformation $F_\tau(f)(\cdot) := f(\cdot + \tau(\cdot))$ に関して安定であることが証明された。ただし、 $\tau: \mathbb{R}^d \rightarrow \mathbb{R}^d$ はその大きさと勾配の大きさに関する制約が必要である。この論文では、 τ の仮定は不要であることを示す。

1 はじめに

この論文では、深層畳み込みニューラルネットワークによるカートゥン関数の特徴量の deformation-stability について考える。畳み込みニューラルネットワークとは、畳み込み層・プーリング層・活性化層からなるニューラルネットワークである。畳み込みニューラルネットワークは画像処理や画像認識に応用を持ち、特に、畳み込みニューラルネットワークによる特徴量抽出を用いた画像分類は大きな成功を収めている ([1], [2], [3])。深層ニューラルネットワークの研究の多くは数値実験によるものであるが、数学による研究もなされている。[4] において、いわゆる、散乱ネットワーク (scattering network) による特徴量抽出の数学解析が行われた。散乱ネットワークとは、各畳み込み層が半離散ウェーブレット変換から構成されている畳み込みニューラルネットワークである。[4] において、wavelet-modulus による特徴量はマザーウェーブレットにより生成される関数空間において、horizontally translation-invariant や deformation-stability を獲得することが数学的に証明された。続いて、[5] において、[4] の理論の拡張がなされた。畳み込み層のフィルタは、“事前指定かつ潜在的に構造化されたフィルタ、事前指定かつ構造化されていないフィルタ、教師あり又は教師なし学習されたフィルタ”を含む場合に一般化された。さらに、活性化層における活性化関数はリプシッツ連続関数に、deformation は $F_t(f) := f(x+t)$ から $F_\tau(f) := f(x+\tau(x))$ へとそれぞれ一般化された。ここで、 $t \in \mathbb{R}^d$ 、 $\tau \in \{\eta \in C^1(\mathbb{R}^d; \mathbb{R}^d) \mid \|\nabla \eta\|_\infty < 1/2d\}$ である。まとめると、畳み込みニューラルネットワークによる“ L^2 関数の特徴量の vertical translation-invariance”や“帯域制限された L^2 関数の特徴量の deformation-stability”はフィルターや活性化関数の選び方に依存せず、ネットワーク自身の構造により誘導されるものであることが証明された。ところが、帯域制限された L^2 関数は滑らかであるため、自然画像がもちうるエッジ等を十分に表現できない。そこで、[7] において、自然画像のエッジ等を十分に表現できるカートゥン関数の特徴量について研究がされ、deformation $F_\tau(f) := f(x+\tau(x))$ に関して安定であることを証明された。そこで、[7] において、自然画像のエッジ等を十分に表現できるカートゥン関数の特徴量について研究がされ、deformation $F_\tau(f) := f(x+\tau(x))$ に関して安定であることを証明された。ここで、 $\tau \in \{\eta \in C^1(\mathbb{R}^d; \mathbb{R}^d) \mid \|\nabla \eta\|_\infty < 1/2d, \|\tau\|_\infty < 1/2\}$ である。実際、カートゥン関数は、MNIST([9])、Caltech-256([10])、CIFAR-100([11]) などの自然画像のデータセットの良いモデルであることが知られている。[7] の結果は自然画像の特徴を捉えた関数の特徴量の deformation stability を証明するために、 τ に条件“ $\|\nabla \tau\|_\infty < 1/2d$ と $\|\tau\|_\infty < 1/2$ ”を課す必要がある。この論文では、この条件を外すことができることを示す。

謝辞 本研究にあたり興味深い課題を提供して下さった株式会社ニコン様に感謝いたします。

¹muesaka@es.hokudai.ac.jp

²ktaguchi@ms.u-tokyo.ac.jp

2 準備

本論文を通して、扱う記号について説明する。任意の $x \in \mathbb{R}^d$ に対して、 $\langle x \rangle := (1 + |x|^2)^{1/2}$ とおく。 \mathcal{H}^{d-1} によって、 $(d-1)$ 次元 Hausdorff 測度を表す。 \mathbb{R}^d の部分集合 B に対して、 $\mathbf{1}_B$ によって、 B の定義関数を表す、すなわち、

$$\mathbf{1}_B(x) := \begin{cases} 1, & x \in B, \\ 0, & x \notin B. \end{cases}$$

さらに、 ∂B によって、 \mathbb{R}^d における B の境界を表す。

2.1 カートゥン関数

定数 $K > 0$ に対して、 $f \in L^2(\mathbb{R}^d)$ は $f = f_1 + \mathbf{1}_B \cdot f_2$ と書けるとき、次数 K のカートゥン関数と呼ぶ。ここで、 B はリプシツ境界を持つコンパクト集合で

$$\mathcal{H}^{d-1}(\partial B) \leq K$$

を満たし、各 $f_j \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d)$ は減衰条件

$$\sup_{x \in \mathbb{R}^d} \langle x \rangle^d |\nabla f_j(x)| \leq K.$$

を満たす。ここで、 $K > 0$ は B, f_1, f_2 に依存しない定数である。さらに、 $K > 0$ に対して、 $C_{\text{CART}}^K(\mathbb{R}^d)$ によって、次数 K のカートゥン関数のクラスを表す、すなわち、

$$C_{\text{CART}}^K(\mathbb{R}^d) := \left\{ f := f_1 + \mathbf{1}_B \cdot f_2 \mid \begin{array}{l} f_j \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d), \quad \sup_{x \in \mathbb{R}^d} \langle x \rangle^d |\nabla f_j(x)| \leq K, \\ B \text{ はリプシツ境界をもつコンパクト集合, } \mathcal{H}^{d-1}(\partial B) \leq K \end{array} \right\}.$$

ここで、 $C_{\text{CART}}^K(\mathbb{R}^d)$ における B は f に依存してよいことを強調しておく。

2.2 畳み込みニューラルネットワークによる特徴量

[7] に従って、散乱ネットワークの定義を述べる。弱許容条件を満たす module-sequence $\Omega := \{(\Psi_n, M_n, R_n)\}_{n=1}^\infty$ が与えられたとする、すなわち、各 (Ψ_n, M_n, R_n) は次の条件を満たす：

- $\Psi_n := \{a_{\lambda_n}^n\}_{\lambda_n \in \Lambda_n} \subset L^2(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$, $\Lambda_n \subset \mathbb{Z}$ はバツセル条件を満たす、すなわち、ある $B_n \geq 0$ が存在して、任意の $f \in L^2(\mathbb{R}^d)$ に対して、

$$\sum_{\lambda_n \in \Lambda_n} \|f * a_{\lambda_n}^n\|_2 \leq B_n \|f\|_2.$$

- M_n は $L^2(\mathbb{R}^d)$ から $L^2(\mathbb{R}^d)$ の作用素で次の条件を満たす：
 - $f = 0$ ならば、 $M_n f = 0$.
 - ある $L_n \geq 0$ が存在して、任意の $f, g \in L^2(\mathbb{R}^d)$ に対して、 $\|M_n f - M_n g\|_{L^2} \leq L_n \|f - g\|_{L^2}$.
- (sub-sampling factor) $R_n \geq 1$.
- (弱許容条件) $\max\{B_n, B_n L_n^2\} \leq 1$.

このとき、 $q = (\lambda_1, \dots, \lambda_n) \in \Lambda_1 \times \dots \times \Lambda_n =: \Lambda_1^n$ に対して、

$$U[q]f := U[(\lambda_1, \dots, \lambda_n)]f := \left(\prod_{j=1}^n U_j[\lambda_j] \right) f$$

とおく。ただし、

$$U_n[\lambda_n]f(\cdot) := R_n^d(M_n(f * a_{\lambda_n}^n))(R_n \cdot)$$

である。また、各 $n \in \mathbb{N}$ に対して、 $\{a_{\lambda_n}^n\}_{\lambda_n \in \Lambda_n}$ から一つ取り出して、 $\chi_{n-1} := a_{\lambda_n}^n$ とおく。以後、記号を乱用して、 $\Lambda_n \setminus \{\lambda_n^*\}$ を Λ_n で表す。さらに、 $\text{Lay}_n(f) := \{U_n[q]f * \chi_{n-1}\}_{q \in \Lambda_n^1} \in L^2(\mathbb{R}^d)^{\Lambda_n^1}$ を Ω による f の第 n 層特徴量と呼び、 $\Phi_\Omega(f) := \{\text{Lay}_n(f)\}_{n=0}^\infty \in \prod_{n=0}^\infty L^2(\mathbb{R}^d)^{\Lambda_n^1}$ を Ω による f の特徴量と呼ぶ。

3 主結果

$\tau \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ に対して、 τ による deformation $F_\tau : L^2(\mathbb{R}^d) \rightarrow L^2(\mathbb{R}^d)$ を $F_\tau(f)(\cdot) := f(\cdot + \tau(\cdot))$ によって定める。

補題 1 任意の $K > 0$ に対して、次を満たす定数 $C_K \geq 0$ が存在する：任意の $f \in C_{\text{CART}}^K(\mathbb{R}^d)$ と任意の $\tau \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ に対して、

$$\|F_\tau(f) - f\|_2 \leq C_K(\|\tau\|_\infty + \|\tau\|_\infty^{1/2}) \quad (1)$$

が成り立つ。

証明 証明は補遺を参照せよ。

定理 2 弱許容条件を満たす *module-sequence* $\Omega := \{(\Psi_n, M_n, R_n)\}_{n=1}^\infty$ が与えられたと仮定する。このとき、任意の $K > 0$ に対して、特徴量写像 Φ_Ω はカートゥン関数のクラス $C_{\text{CART}}^K(\mathbb{R}^d)$ 上で deformation F_τ に関して安定である、すなわち、任意の $K > 0$ に対して、次を満たす (Ω に依存しない) 定数 $C_K \geq 0$ が存在する：任意の $\tau \in C^1(\mathbb{R}^d; \mathbb{R}^d)$ と任意の $f \in C_{\text{CART}}^K(\mathbb{R}^d)$ に対して、不等式

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

が成り立つ。ここで、 $g := \{\{g_q\}_{q \in \Lambda_n^1}\}_{n=0}^\infty \in \prod_{n=0}^\infty L^2(\mathbb{R}^d)^{\Lambda_n^1}$ に対して、 $\|g\|_* := \left(\sum_{n=0}^\infty \sum_{q \in \Lambda_n^1} \|g_q\|_2^2\right)^{1/2}$ である。

証明 Φ_Ω の縮小性 ([5]) より、 $\|\Phi_\Omega(F_\tau(f)) - \Phi_\Omega(f)\|_* \leq \|F_\tau(f) - f\|_2$ が成り立つ。さらに、(1) より、定理の不等式を得る。

3.1 まとめ

畳み込みニューラルネットワークによるカートゥン関数の特徴量の deformation stability について証明を行った。特に、我々の結果は、Theorem 1, [7] における変換 τ の仮定

- $\|\tau\|_\infty < \frac{1}{2}$,
- $\|\nabla \tau\|_\infty < \frac{1}{2d}$

を外すことに成功した。

4 補遺：補題 1 の証明

$f = f_1 + \mathbf{1}_B \cdot f_2 \in C_{\text{CART}}^K(\mathbb{R}^d)$ とかけた時、定理の証明は、 $f_j - F_\tau(f_j), j = 1, 2$ と $F_\tau(\mathbf{1}_B) - \mathbf{1}_B$ に関する二つの評価

$$\|F_\tau(f) - f\|_2 \leq KD\|\tau\|_\infty, \quad (2)$$

$$\|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2 \leq (2\mathcal{H}^{d-1}(\partial B))^{1/2} \|\tau\|_\infty \quad (3)$$

を組み合わせることで得られる。ここで、 $D > 0$ は f と τ に依存しない定数である。この二つの評価の証明は下で与えるので、二つの評価を認めて補題の証明をする。まず、

$$\begin{aligned} F_\tau(f) - f &= F_\tau(f_1 + \mathbf{1}_B \cdot f_2) - (f_1 + \mathbf{1}_B \cdot f_2) \\ &= (F_\tau(f_1) - f_1) + (F_\tau(\mathbf{1}_B) \cdot F_\tau(f_2) - \mathbf{1}_B \cdot F_\tau(f_2)) + (\mathbf{1}_B \cdot F_\tau(f_2) - \mathbf{1}_B \cdot f_2) \\ &= (F_\tau(f_1) - f_1) + (F_\tau(\mathbf{1}_B) - \mathbf{1}_B) \cdot F_\tau(f_2) + \mathbf{1}_B \cdot (F_\tau(f_2) - f_2) \end{aligned}$$

分解できることに注意する。ここで、 $F_\tau(\mathbf{1}_B \cdot f_2) = F_\tau(\mathbf{1}_B) \cdot F_\tau(f_2)$ を用いた。上の分解と三角不等式と Cauchy—Schwarz の不等式より、

$$\|F_\tau(f) - f\|_2 \leq \|F_\tau(f_1) - f_1\|_2 + \|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2 \|F_\tau(f_2)\|_\infty + \|\mathbf{1}_B \cdot (F_\tau(f_2) - f_2)\|_2$$

である。 $f_j \in C_{\text{CART}}^K(\mathbb{R}^d)$, $j = 1, 2$ と二つの評価 (2)、(3) を合わせると、

$$\begin{aligned} &\|F_\tau(f_1) - f_1\|_2 + \|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2 \|F_\tau(f_2)\|_\infty + \|\mathbf{1}_B \cdot (F_\tau(f_2) - f_2)\|_2 \\ &\leq 2KD \|\tau\|_\infty + \sqrt{2}K^{3/2} \|\tau\|_\infty^{1/2} \\ &\leq \max\{2DK, \sqrt{2}K^{3/2}\} (\|\tau\|_\infty + \|\tau\|_\infty^{1/2}) \end{aligned}$$

を得る。 $C_K := \max\{2DK, \sqrt{2}K^{3/2}\}$ として、補題の不等式 (1) を得る。

ここから、評価 (2),(3) を証明する。

補助定理 3 $f \in L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d)$ は減衰条件

$$\sup_{x \in \mathbb{R}^d} \langle x \rangle^d |\nabla f(x)| =: C < \infty$$

を満たすと仮定する。 $\tau: \mathbb{R}^d \rightarrow \mathbb{R}^d$ とする。このとき、

$$\|F_\tau(f) - f\|_2 \leq CD \|\tau\|_\infty$$

が成り立つ。ただし、 $D > 0$ は f に依存しない定数である。

証明 平均値の定理と f の減衰条件より、

$$\begin{aligned} |f(x + \tau(x)) - f(x)| &\leq |\tau(x)| \sup_{|x-y| \leq \|\tau\|_\infty} |\nabla f(y)| \\ &\leq |\tau(x)| \sup_{|x-y| \leq \|\tau\|_\infty} \langle y \rangle^{-d} \\ &\leq C |\tau(x)| h(x) \end{aligned}$$

である。ここで、

$$h(x) = \begin{cases} 1, & |x| < \|\tau\|_\infty \\ (1 + (|x| - \|\tau\|_\infty)^2)^{-d/2}, & |x| \geq \|\tau\|_\infty \end{cases}$$

である。この不等式とヘルダーの不等式を合わせると、

$$\|F_\tau(f) - f\|_2 \leq C \|\tau\|_\infty \left(\int_{\mathbb{R}^d} |h(x)|^2 dx \right)^{1/2} =: CD \|\tau\|_\infty.$$

補助定理 4 $B \subset \mathbb{R}^d$ をリプシッツ境界をもつコンパクト集合とする。このとき、

$$\|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2 \leq (2\mathcal{H}^{d-1}(\partial B))^{1/2} \|\tau\|_\infty$$

が成り立つ。

証明 $h(x) := \mathbf{1}_B(x + \tau(x)) - \mathbf{1}_B(x)$ とおくと、

$$h(x) = \begin{cases} 1, & x \in S, \\ 0, & x \notin S. \end{cases}$$

である。ただし、 $S := \{x \in B, x + \tau(x) \notin B\} \cup \{x \notin B, x + \tau(x) \in B\}$ である。よって、

$$\|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2^2 = \int_{\mathbb{R}^d} |h(x)|^2 dx \leq \int_S 1 dx.$$

さらに、 $S \subset \partial B + \{|x| \leq \|\tau\|_\infty\}$ なので、

$$\int_S 1 dx \leq \int_{\partial B + \{|x| \leq \|\tau\|_\infty\}} 1 dx \leq 2\mathcal{H}^{d-1}(\partial B)\|\tau\|_\infty.$$

よって、

$$\|F_\tau(\mathbf{1}_B) - \mathbf{1}_B\|_2 \leq \sqrt{2}\mathcal{H}^{d-1}(\partial B)^{1/2}\|\tau\|_\infty^{1/2}.$$

参考文献

- [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. Bolcski, A mathematical theory of deep convolutional neural networks for feature extraction, arXiv:1512.06293, 2015.
- [6] T. Wiatowski, M. Tschannen, A. Stanic, P. Grohs, and H. Bolcskei, Discrete deep feature extraction: A theory and new architectures, in Proc. of International Conference on Machine Learning (ICML), pp. 2149–2158, 2016.
- [7] P. Grohs, T. Wiatowski, and H. Bolcskei, Deep convolutional neural networks on cartoon functions, in Proc. of IEEE International Symposium on Information Theory (ISIT), pp. 1163–1167, 2016.
- [8] T. Wiatowski, P. Grohs, and H. Bolcskei, Topology reduction in deep convolutional feature extraction networks, Proc. of SPIE (Wavelets and Sparsity XVII), 2017, to appear.
- [9] Y. LeCun and C. Cortes, “The MNIST database of handwritten digits,” <http://yann.lecun.com/exdb/mnist/>, 1998.
- [10] G. Griffin, A. Holub, and P. Perona, “Caltech-256 object category dataset,” <http://authors.library.caltech.edu/7694/>, 2007.
- [11] A. Krizhevsky, “Learning multiple layers of features from tiny images,” Master’s thesis, University of Toronto, 2009.