Approximation Theory of Deep Convolutional Neural Networks

Ding-Xuan Zhou

School of Data Science, City University of Hong Kong

E-mail: mazhou@cityu.edu.hk

Supported in part by Research Grants Council of Hong Kong

Start

October 8, 2018

Outline of the talk

I. Fully connected neural networks

II. Universality of deep CNNs in depth

III. Approximation theory of deep CNNs

IV. Distributed approximation by deep CNNs

I. Fully connected neural networks

I.1. Deep learning

Deep learning has been very successful in many practical domains: speech recognition, computer vision, natural language processing, ...

LeCun-Bottou-Bengio-Haffner, Hinton-Osindero-Teh, Bengio, LeCun, Krizhevsky-Sutskever-Hinton, Mallat, ...

Mathematical theory of deep learning: at its infancy

I.2. Deep neural network architectures

Convolutional neural networks (CNNs)
Recursive neural networks
Deep belief networks: deep Boltzmann machines

A large literature on representational complexity: Goodfellow-Bengio-Courville, Montúfar-Pascanu-Cho-Bengio, Delalleau-Bengio, Poggio-Anselmi-Rosasco, ...

Approximation by **deep fully connected** neural networks: Eldan-Shamir, Telgarsky, Yarotsky, Kohler-Krzyżak, Bölcskei-Grohs-Kutyniok-Petersen, Petersen-Voigtlaender, Klusowski-Barron, McCane-Szymanski, Mhaskar-Poggio, Chui-Lin-Zhou, ...

Little is known on **approximation** or generalization ability of **structured** deep neural networks

I.3. Approximation theory of fully connected networks

Classical **shallow fully connected neural networks** to approximate functions or process data on $\mathbb{R}^d$:

$$f_N(x) = \sum_{i=1}^{N} c_i \sigma(\mathbf{w}_i \cdot x + b_i).$$

input vector $x := (x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^d$
activation function $\sigma : \mathbb{R} \to \mathbb{R}$
weights $\mathbf{w}_i \in \mathbb{R}^d$, biases $b = (b_i)_i$, coefficients $c = (c_i)_i \in \mathbb{R}^N$
Total number of **free parameters**: $\mathcal{N} = (d+2)N$
Matrix form with σ acting componentwise:

$$f_N(x) = c \cdot \sigma(T^w x + b), \tag{1}$$

where T^w is a full matrix without any imposed structures:

$$T^w = \begin{bmatrix} (\mathbf{w}_1)_1 & (\mathbf{w}_1)_2 & \cdots & (\mathbf{w}_1)_d \\ (\mathbf{w}_2)_1 & (\mathbf{w}_2)_2 & \cdots & (\mathbf{w}_2)_d \\ \vdots & \vdots & \vdots & \vdots \\ (\mathbf{w}_N)_1 & (\mathbf{w}_N)_2 & \cdots & (\mathbf{w}_N)_d \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

Universality of approximation: If σ is locally bounded, piecewise continuous, and not a polynomial, then for any compact $\Omega \subset \mathbb{R}^d$ and $f \in C(\Omega)$, there holds

$$\lim_{N\to\infty} \inf \left\{ \|f - f_N\|_{C(\Omega)} : \ \mathbf{w}_i \in \mathbb{R}^d, b \in \mathbb{R}^N, c \in \mathbb{R}^N \right\} = 0.$$

Approximation theory: Cybenko (1989), Hornik (1991), Barron (1993), Mhaskar (1994), Micchelli, Chui-Li-Mhaskar, Lin-Pinkus-Schocken, Maiorov, Petrushev, ...

Typical result (Mhaskar 1994): Assume a continuous activation functions σ satisfies with some $b \in \mathbb{R}$, $\sigma^{(k)}(b) \neq 0$ for any $k \in \mathbb{Z}_+$ and with some integer $\ell \neq 1$, $\lim_{u\to-\infty} \sigma(u)/|u|^\ell = 0$ and $\lim_{u\to\infty} \sigma(u)/u^\ell = 1$. Then for $f \in C^r([-1,1]^d)$,

$$\inf \left\{ \|f - f_N\|_{C(\Omega)} : \ \mathbf{w}_i, b, c \right\} = O(N^{-r/d}) = O\left((d+2)^{r/d} \mathcal{N}^{-r/d} \right).$$

ReLU with $\ell = 1$ was not included, and recent work of Klusowski-Barron covers this case.

I.4. Multi-layer fully connected neural networks

Multi-layer neural network with J hidden layers of neurons $\{h^{(j)} : \mathbb{R}^d \to \mathbb{R}^{d_j}\}_{j=1}^J$ with widths $\{d_j \in \mathbb{N}\}$ and $h^{(0)}(x) = x$

$$h^{(j)}(x) = \sigma\left(T^{(j)} h^{(j-1)}(x) + b^{(j)}\right). \qquad (2)$$

free parameters: **full matrix** $T^{(j)} \in \mathbb{R}^{d_j \times d_{j-1}}$, vector $b^{(j)} \in \mathbb{R}^{d_j}$

Approximation theory in 1990's

Recent work: approximation by fully-connected ReLU nets: Eldan-Shamir (2016), Telgarsky (2016), Yarotsky (2017), Klusowski-Barron (2016), Shaham-Cloningen-Coifman (2018), Chui-Lin-Zhou (2018), ...

approximating compositional functions: Mhaskar-Poggio (2016), Poggio-Mhaskar-Rosasco-Miranda-Liao (2016), ...

wavelet analysis and kernel methods for deep CNNs: Bruna-Mallat (2013), Mallat (2016), Steinwart-Thomann-Schmid (2016), ...

II. Universality of deep CNNs in depth

II.1. Structure of deep convolutional neural networks

Rectified linear unit (ReLU): $\sigma(u) = (u)_+ = \max\{u, 0\}$

sequence of convolutional filter masks $\mathbf{w} = \{w^{(j)} = (w_k^{(j)})_{k=-\infty}^{\infty}\}_j$

filter length s: Assume $w^{(j)}$ is supported in $\{0, 1, \ldots, s\}$

convolution: $w*x = \left(\sum_k w_k x_{i-k} = \sum_{k=0}^{s} w_{i-k} x_k\right)_i$ supported in $\{0, 1, \ldots, d+s\}$, so $(w*x)_{i=0}^{d+s} = T^w x$ with $T^w = (w_{i-k})_{0\le i\le d+s, 0\le k\le d}$:

$$T^w = \begin{bmatrix} w_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ w_1 & w_0 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ w_s & w_{s-1} & \cdots & w_0 & 0 & \cdots & 0 \\ 0 & w_s & \cdots & w_1 & w_0 & 0\cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_s & \cdots & w_1 & w_0 \\ 0 & \cdots & 0 & 0 & w_s & \cdots & w_1 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & w_s \end{bmatrix} \in \mathbb{R}^{(d+s)\times d}.$$

II.2. Deep CNNs

A **deep CNN of depth** J is a J-layer neural network with widths $\{d_j = d+js\}$ having a **convolutional sparse structure**:

$$h^{(j)}(x) = \sigma\left(T^{(j)}h^{(j-1)}(x) - b^{(j)}\right), \qquad j = 1, 2, \ldots, J, \tag{3}$$

where $T^{(j)} = T^{w^{(j)}}$ is a Toeplitz type $d_j \times d_{j-1}$ matrix generated by the filter mask $w = w^{(j)}$.

The generated **hypothesis space** of functions:

$$\mathcal{H}_J^{\mathbf{w},\mathbf{b}} = \left\{ \sum_{k=1}^{d+Js} c_k h_k^{(J)}(x) : c = (c_i)_{i=1}^{d+Js} \in \mathbb{R} \right\}. \tag{4}$$

Take $b^{(j)}$ of the form $b = [b_1 \ \ldots b_{s-1} \ b_s \ b_s \ \ldots \ b_s \ b_{d_j-s+1} \ \ldots b_{d_j}]^T$

Bias vector sequence $\mathbf{b} = \{b^{(j)}\}_{j=1}^J$

Total number of free parameters: $\mathcal{N} = 3s(J-1) + 2d + 2J$

II.3. Universality of deep CNNs

Theorem 1 *Let $2 \leq s \leq d$. For any compact subset Ω of $\mathbb{R}^d$ and any $f \in C(\Omega)$, there exist sequences $\mathbf{w}$ of filter masks, $\mathbf{b}$ of bias vectors, and $f_J^{\mathbf{w},\mathbf{b}} \in \mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ such that $\lim_{J\to\infty} \|f - f_J^{\mathbf{w},\mathbf{b}}\|_{C(\Omega)} = 0$.*

III. Approximation theory of deep CNNs

III.1. Same approximation as fully connected networks

Theorem 2 *Let $2 \le s \le d$, $N \in \mathbb{N}$, and $J \ge \lceil \frac{(N+1)d-1}{s-1} \rceil$. Then for any $f \in C(\Omega)$, there exist sequences $\mathbf{w}$ of filter masks, $\mathbf{b}$ of bias vectors, and $f_J^{\mathbf{w},\mathbf{b}} \in \mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ such that*

$$\begin{aligned} & \inf_{c\in\mathbb{R}} \left\| f - f_J^{\mathbf{w},\mathbf{b}} - c \right\|_{C(\Omega)} \\ \le & \inf_{\alpha_i\in\mathbb{R}^d,\ t_i,c_i\in\mathbb{R}} \left\{ \left\| f(x) - \alpha_0 \cdot x - t_0 - \sum_{i=1}^{N} c_i \sigma(\alpha_i \cdot x + t_i) \right\|_{C(\Omega)} \right\}. \end{aligned}$$

Numbers of parameters when $J = \lceil \frac{(N+1)d-1}{s-1} \rceil$:

deep CNN: $\mathcal{N} = 3s(J-1) + 2d + 2J \le 8(N+1)d$

shallow fully connected network: $\mathcal{N} = (d+2)N + d + 1 \ge (N+1)d$

III.2. Rate of approximation

Sobolev space $H^r(\mathbb{R}^d)$: F and all its partial derivatives up to order r are square integrable on $\mathbb{R}^d$.
Embedding Theorem: $H^r(\mathbb{R}^d) \subset C(\mathbb{R}^d)$ only when $r > \frac{d}{2}$.

Theorem 3 *If $\Omega \subseteq [-1,1]^d$ and f is the restriction to Ω of some function $F \in H^r(\mathbb{R}^d)$ on $\mathbb{R}^d$ with integer $r > 2 + \frac{d}{2}$ and $J \geq 2d/(s-1)$, then there exist $\mathbf{w}$, $\mathbf{b}$, and $f_J^{\mathbf{w},\mathbf{b}} \in \mathcal{H}_J^{\mathbf{w},\mathbf{b}}$ such that*

$$\|f - f_J^{\mathbf{w},\mathbf{b}}\|_{C(\Omega)} \leq \widetilde{c}\,\|F\|_{H^r}\sqrt{\log J}\,(1/J)^{\frac{1}{2}+\frac{1}{d}},$$

where $\widetilde{c}$ is an absolute constant and $\|F\|_{H^r}$ denotes the Sobolev space norm of F.

Corollary. *Take* $s = \lceil 1 + d^\tau/2 \rceil$ *and* $J = \lceil 4d^{1-\tau} \rceil L$ *with* $0 \leq \tau \leq 1$ *and* $L \in \mathbb{N}$. *Under the assumption of the theorem, we have*

$$\|f - f_J^{\mathbf{w},\mathbf{b}}\|_{C(\Omega)} \leq c \, \|F\|_{H^r} \sqrt{\frac{(1-\tau)\log d + \log L + \log 5}{4d^{1-\tau} L}},$$

while the widths of the deep CNN are bounded by $12Ld$ *and the total number of free parameters by*

$$5sJ + 2d - 2s + 1 \leq (65L + 2)d.$$

We can even take $L = 1$, $\tau = 1/2$, *filter length* $s = \lceil 1 + \sqrt{d}/2 \rceil$, *and depth* $\lceil 4\sqrt{d} \rceil$ *to get a bound for the relative error*

$$\frac{\|f - f_J^{\mathbf{w},\mathbf{b}}\|_{C(\Omega)}}{\|F\|_{H^r}} \leq c \sqrt{\frac{\log(5\sqrt{d})}{4\sqrt{d}}},$$

achieved by a deep CNN of at most $67d$ *free parameters, which decreases as the dimension* d *increases.*

IV. Distributed approximation by deep CNNs

Here we take equal widths: $d_j \equiv d$

The matrix $T^w := (w_{i-k})_{i,k=1,\ldots,d}$ is $d \times d$ given explicitly for $w = w^{(j)}$ by

$$T^w = \begin{bmatrix} w_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ w_1 & w_0 & 0 & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ w_s & w_{s-1} & \cdots & w_0 & 0 & \cdots & 0 \\ 0 & w_s & \cdots & w_1 & w_0 & 0 \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & w_s & \cdots & w_1 & w_0 \end{bmatrix} \in \mathbb{R}^{d \times d}.$$

constraints on the bias vectors: $b^{(j)}_{s+1} = b^{(j)}_{s+2} = \ldots = b^{(j)}_d$ for $j \leq J-1$ and $b^{(J)}_{s+1} = b^{(J)}_{s+2} = \ldots = b^{(J)}_{d-1}$

Theorem 4 *Let $2 \le s \le d$, $J \ge \lceil \frac{d-1}{s-1} \rceil$, and Ω be a compact subset of $\mathbb{R}^d$. Then for any $\xi \in \mathbb{R}^d$ and any $t \in \mathbb{R}$, there exist a sequence of filter masks $\mathbf{w}$ and a sequence of bias vectors $\mathbf{b}$ such that*

$$\left(h^{(J)}(x)\right)_d = (\xi \cdot x - t)_+, \qquad \forall x \in \Omega.$$

Denote $f_J^{\mathbf{w},\mathbf{b}}(x) := \left(h^{(J)}(x)\right)_d$

$\vec{\mathbf{w}} = \left\{\mathbf{w}^{[\ell]}\right\}_{\ell=0}^{m}$: array of filter mask sequences $\mathbf{w}^{[\ell]} = \{w^{(\ell,j)}\}_{j=1}^{J}$

$\vec{\mathbf{b}} = \left\{\mathbf{b}^{[\ell]}\right\}_{\ell=0}^{m}$: array of $\mathbf{b}^{[\ell]} = \{b^{(\ell,j)} \in \mathbb{R}^d\}_{j=1}^{J}$ with the same constraint

Hypothesis space for distributed approximation generated by $\vec{\mathbf{w}}, \vec{\mathbf{b}}$:

$$\mathcal{H}_{J,m}^{\vec{\mathbf{w}},\vec{\mathbf{b}}} = \left\{ \sum_{\ell=0}^{m} c_\ell f_J^{\mathbf{w}^{[\ell]},\mathbf{b}^{[\ell]}}(x) : \ c = (c_\ell)_{\ell=0}^{m} \in \mathbb{R}^{m+1} \right\}.$$

Theorem 5 *Let* $J \geq \lceil \frac{d-1}{s-1} \rceil$, $\Omega \subseteq [-1,1]^d$, *and* $m \in \mathbb{N}$. *If* f *is the restriction to* Ω *of some function* $F \in H^r(\mathbb{R}^d)$ *on* $\mathbb{R}^d$ *with integer* $r > 2 + \frac{d}{2}$, *then there exist* $\vec{\mathbf{w}}, \vec{\mathbf{b}}$ *such that*

$$\inf_{c \in \mathbb{R}^{m+1}} \left\{ \left\| f - \sum_{\ell=0}^{m} c_\ell f_J^{\mathbf{w}^{[\ell]}, \mathbf{b}^{[\ell]}}(x) \right\|_{C(\Omega)} \right\}$$
$$\leq \widetilde{c} \, \|f\|_{H^r(\Omega)} \left(\sqrt{\log m} + \sqrt{d} \right) (1/m)^{\frac{1}{2} + \frac{1}{d}},$$

where $\widetilde{c}$ *is an absolute constant.*

Thank you very much!