Approximation to Smooth Functions by Low-Rank Swish Networks

Zimeng Li [1] Hongjun Li [2] Jingyuan Wang [1,3] Ke Tang [2]
{zimengli, jywang}@buaa.edu.cn {hongjunli, ketang}@tsinghua.edu.cn
[1]Beihang University [2]Tsinghua University [3]Engineering Research Center of Advanced Computer Application Technology

TL;DR: We offer a theoretical basis for low-rank compression from the perspective of universal approximation theory by proving any Hölder function can be approximated by a Swish network with low-rank weight matrices.

Motivation

Low-rank compression is a class of efficient and hardware-friendly neural network compression techniques that approximate weight matrices through matrix factorization. However, the universal effectiveness of low-rank compression in preserving model performance is not guaranteed theoretically.

In this paper, we partially explain this universal effectiveness by showing that for any $\varepsilon > 0$ and $f \in \mathcal{C}^{\beta,R}([0,1]^d)$ there exists a low-rank network nn such that $\sup_{x\in[0,1]^d} |nn(x) - f(x)| \leq \varepsilon$. This approximation result indicates that for a vast range of tasks there exists a good low-rank network solution, though it is not clear to us whether such a solution can be obtained by a specific low-rank compression algorithm.

Such an approximation rate can also be a key ingredient to derive consistency and convergence rate of low-rank network estimation.

Low-Rank Networks

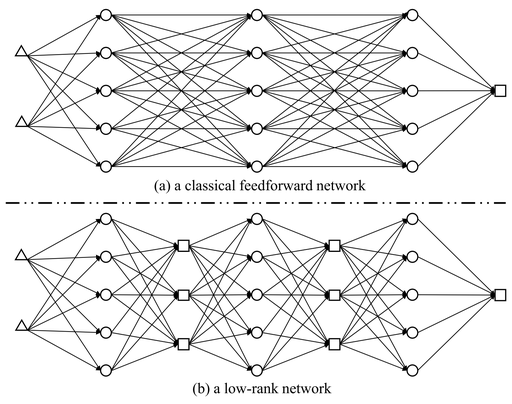

Figure 1. "△" stands for input neuron, "○" nonlinear neuron (i.e. neuron with activation function), and "□" linear neuron (i.e. neuron without activation function).

Hölder Functions

Let $d \in \mathbb{N}_+$, $\mathscr{X} \in \mathbb{R}^d$, and $R, \beta \in \mathbb{R}_+$. There exist $\kappa \in \mathbb{N}$ and $0 < \gamma \leq 1$ such that $\beta = \kappa + \gamma$. For a function $f : \mathscr{X} \to \mathbb{R}$, its Hölder norm is defined by

$$\|f\|_{\mathcal{C}^\beta} := \max\left\{ \sup_{|\boldsymbol{\alpha}|\leq\kappa} \|\partial^{\boldsymbol{\alpha}} f\|_\infty, \sup_{|\boldsymbol{\alpha}|=\kappa} \sup_{\boldsymbol{x}\neq\boldsymbol{y}} \frac{|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})|}{\|\boldsymbol{x}-\boldsymbol{y}\|_\infty^\gamma} \right\} \tag{1}$$

The Hölder space $\mathcal{C}^\beta([0,1]^d)$ is defined as the set

$$\left\{ f : \mathscr{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{C}^\beta} < \infty \right\} \tag{2}$$

equipped with Hölder norm $\|\cdot\|_{\mathcal{C}^\beta}$. And the Hölder ball $\mathcal{C}^{\beta,R}([0,1]^d)$ is defined by

$$\left\{ f : \mathscr{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{C}^\beta} \leq R \right\}. \tag{3}$$

Approximation Theorem

Let $\beta \in \mathbb{R}_+$, $\beta = \kappa + \gamma, \kappa \in \mathbb{N}, \gamma \in (0,1]$, and $R \in \mathbb{R}_+$. For all $f \in \mathcal{C}^{\beta,R}([0,1]^d)$, $M \in \mathbb{N}_+$, $\lambda \geq 2^{-\frac{1}{3}}$, and $\tau \geq 1$, there exists a low-rank Swish network $nn : [0,1]^d \to \mathbb{R}$ with depth

$$\max\left\{ \left\lceil \frac{\kappa}{2} \right\rceil, \lceil \log_2 d \rceil + 1 \right\} + 1, \tag{4}$$

width of nonlinear layers

$$2\binom{d+1}{d-1} + 4\binom{d+\kappa-2}{d-1} + 4\binom{d+\kappa-1}{d-1} + 6(M+1)^d, \tag{5}$$

width of linear hidden layers

$$\binom{d+1}{d-1} + \binom{d+\kappa-3}{d-1} + \binom{d+\kappa-2}{d-1} + 2(M+1)^d, \tag{6}$$

upper bound of absolute values of parameters

$$\max\left\{ (3M+2)\tau, 2\lambda^2 \max_{|\boldsymbol{\alpha}|\leq\kappa} \left\{ \sum_{\substack{\boldsymbol{\nu}\geq\boldsymbol{\alpha} \\ |\boldsymbol{\nu}|\leq\kappa}} \frac{R}{\boldsymbol{\nu}!} \prod_{i=1}^d \binom{\nu_i}{\alpha_i} \right\}, 2\lambda^2 \right\}, \tag{7}$$

and upper bound of number of nonzero parameters

$$c_1 + c_2(M+1)^d \tag{8}$$

such that

$$\begin{aligned} &|nn(\boldsymbol{x}) - f(\boldsymbol{x})| \\ &\leq c_3 \frac{(M+1)^d}{\lambda^2} + c_4 M^{-\beta} + c_5 (M+1)^d \tau e^{-\tau} \end{aligned} \tag{9}$$

for all $\boldsymbol{x} \in [0,1]^d$, where c_1, c_2, c_3, c_4, and c_5 are positive constants depending only on d, κ, and R.

Low-Rank Compression

When $\beta > 2$ (i.e. $\kappa \geq 2$), the width of linear hidden layers is always no more than one-third of the width of nonlinear layers, since

$$\begin{aligned} &3\left(\binom{d+1}{d-1} + \binom{d+\kappa-3}{d-1} + \binom{d+\kappa-2}{d-1} + 2(M+1)^d \right) \\ &\leq 2\binom{d+1}{d-1} + 4\binom{d+\kappa-2}{d-1} + 4\binom{d+\kappa-1}{d-1} + 6(M+1)^d \\ &\Leftarrow \binom{d+1}{d-1} \leq \binom{d+\kappa-2}{d-1} + \binom{d+\kappa-1}{d-1} \\ &\Leftarrow \binom{d+1}{d-1} \leq \binom{d}{d-1} + \binom{d+1}{d-1}. \end{aligned} \tag{10}$$

Proof Ideas

1. approximating any Hölder function f by a sum–product combination of Taylor polynomials $(P^\kappa_{\boldsymbol{m}})_{\boldsymbol{m}\in[M]^d}$ and approximate bump functions $(\phi^\tau_{\boldsymbol{m}})_{\boldsymbol{m}\in[M]^d}$, where $P^\kappa_{\boldsymbol{m}}(\boldsymbol{x}) := \sum_{|\boldsymbol{\alpha}|\leq\kappa} \frac{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{m}/M)}{\boldsymbol{\alpha}!} \left(\boldsymbol{x} - \frac{\boldsymbol{m}}{M}\right)^{\boldsymbol{\alpha}}$ and $\phi^\tau_{\boldsymbol{m}}(\boldsymbol{x}) := \prod_{i=1}^d \psi^\tau\left(3M\left(x_i - \frac{m_i}{M}\right)\right)$
2. approximating $(P^\kappa_{\boldsymbol{m}})_{\boldsymbol{m}\in[M]^d}$ by a low-rank Swish network $\mathcal{P}$
3. approximating $(\phi^\tau_{\boldsymbol{m}})_{\boldsymbol{m}\in[M]^d}$ by a low-rank Swish network $\mathcal{G}$
4. approximating $\sum_{\boldsymbol{m}\in[M]^d} P^\kappa_{\boldsymbol{m}} \phi^\tau_{\boldsymbol{m}}$ by the inner product of $\mathcal{P}$ and $\mathcal{G}$

Here, $M \in \mathbb{N}_+$, $[M] := \{0, 1, 2, \ldots, M\}$, ρ is the activation function and $\psi^\tau(x) := \frac{1}{\tau}(\rho(\tau(x+2)) - \rho(\tau(x+1)) - \rho(\tau(x-1)) + \rho(\tau(x-2)))$.

Curse of Dimensionality

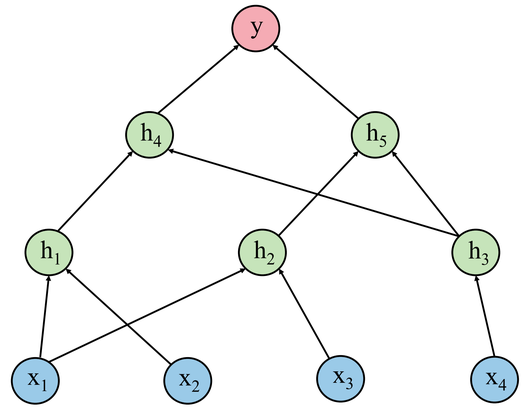

Figure 2. An illustration of a hierarchical composite function. x_1, x_2, x_3, and x_4 are input variables. $h_1 = h_1(x_1, x_2)$, $h_2 = h_2(x_1, x_3)$, $h_3 = h_3(x_4)$, $h_4 = h_4(h_1, h_3)$, $h_5 = h_5(h_2, h_3)$, and $y = y(h_4, h_5)$. Though the input dimension of the hierarchical composite function is 4, the input dimensions of its component functions do not exceed 2.

The curse of dimensionality refers to the phenomenon that as the input dimension d goes to infinity, the network size required to achieve a given approximation error grows fast or the approximation error grows fast when the network size is fixed.

Here we briefly introduce a class of high-dimensional functions, called hierarchical composite functions, which are universal in reality and can be approximated without being affected by the curse of dimensionality(Schmidt-Hieber, 2020; Kohler & Langer, 2021).

It is obvious that the network size required to approximate a hierarchical composite function is directly related to the input dimension of each component function and has no direct relation to the input dimension of the hierarchical composite function, because we can construct networks to approximate component functions respectively, then combine them into one network.

Experimental Validation

Table 1. Cross-validation results for classical feedforward networks and low-rank networks on various classification (top) and regression (bottom) datasets. L represents the depth (i.e. the number of nonlinear layers) of both networks and $\mathcal{H}$ represents the width of nonlinear layers of both networks.

DATASET	L	$\mathcal{H}$	ACC(%) classical	ACC(%) low-rank	t-statistic
Iris	4	20	95.3 ± 4.3	94.7 ± 5.0	0.36
Rice	2	35	92.7 ± 1.9	92.6 ± 2.0	1.00
BankMarketing	2	188	68.9 ± 15.3	71.1 ± 15.4	−2.01
Adult	2	540	85.8 ± 0.3	85.8 ± 0.3	−0.47
DATASET	L	$\mathcal{H}$	RMSE classical	RMSE low-rank	t-statistic
RealEstate	4	30	$.078 \pm .021$	$.077 \pm .020$	1.29
Abalone	3	50	$.077 \pm .022$	$.077 \pm .022$	−0.44
WineQuality	4	78	$.123 \pm .009$	$.123 \pm .009$	1.21
BikeSharing	4	60	$.100 \pm .036$	$.070 \pm .024$	3.90