Expressive Power of Tensor-Train Networks With Equal TT-cores

Student: Emil Alkin [1,2]
MSc program: Data Science

Research Advisor: Ivan Oseledets[1], Dr. Sc., Professor

[1]Skolkovo Institute of Science and Technology
[2]Moscow Institute of Physics and Technology

07 June 2024

Background

Motivation

One of the relevant directions in the study of the theoretical foundations of deep neural networks is the study of the expressive power of various deep network architectures. New methods in studying the *depth power* are demonstrated in papers [1] and [2]. They established the relationship between convolutional and recurrent neural network architectures with different tensor decompositions and applied for the first time *tensor analysis* to the study of the expressive power of networks.

[1] Nadav Cohen et al. On the expressive power of deep learning: A tensor analysis. (2016)
[2] Valentin Khrulkov et al. Expressive power of recurrent neural networks (2017)

Background

Let us define our hypotheses space by the following formula for the score function h_y for a class label y:

$$h_y(x_1, x_2, \ldots, x_d) = \sum_{i_1,\ldots,i_d=1}^{m} \mathcal{W}_y^{i_1 \ldots i_d} \cdot f_{\theta_{i_1}}(x_1) \cdot \ldots \cdot f_{\theta_{i_d}}(x_d), \qquad (1)$$

where $\mathcal{W}_y$ is a *coefficient tensor* of order d and dimension m in each mode.

Connection between Tensor Train tensor decomposition and RNNs

Khrulkov, V. et al. showed that a recurrent-type neural network with d layers, each implementing a bilinear function, lies in the hypotheses space defined by formula 1, with the parameters of each of the d layers being exactly equal to the TT-cores $G_1, \ldots, G_d$ of the corresponding Tensor Train decomposition of the coefficient tensor $\mathcal{W}_y$.

Background

Connection between Tensor Train tensor decomposition and RNNs

Moreover, the TT-ranks $r_1, r_2, \ldots, r_{d-1}$ of the corresponding Tensor Train decomposition of $\mathcal{W}_y$ are equal to the dimensions of the hidden states after, respectively, the first layer, the second layer and so on up to the $(d-1)$-th layer.

General problem

One of the main results of paper [2] is the theorem about the lower bound on the CP-rank for a Tensor Train network with random parameters. A closer analogue of the traditional recurrent neural network is the Tensor Train network with equal inner layers, i.e. Tensor Train networks such that $G_2 = G_3 = \ldots = G_{d-1}$. For such networks the lower estimate on CP-rank is not proved.

[2] Valentin Khrulkov et al. Expressive power of recurrent neural networks (2017)

Aim and Objectives

The main task is to obtain new results on the expressive power of Tensor Train networks with equal TT-cores. In particular,

- To prove Hypothesis 1 from the paper [2];
- To obtain an empirical estimate of the CP-rank of "random" Tensor Train networks with equal TT-cores;
- To show that the Tensor Train networks with equal TT-cores are complex enough to obtain high quality on real world datasets such.

Main result

The following theorem states that RNNs with equal layers are exponentially more expressive than shallow networks.

Theorem

Suppose that $d = 2k$ is even. Define the following set

$$B := \left\{ \mathcal{X} \in \mathcal{M}^{eq}_{\mathsf{n},\mathsf{r}} : \mathsf{rank}_{CP}\, \mathcal{X} < q^{\frac{d}{2}} \right\},$$

where $q = \min\{n, r-1\}$.
Then

$$\mu(B) = 0,$$

where μ is the standard Lebesgue measure on $\mathcal{M}^{eq}_{\mathsf{n},\mathsf{r}}$.

Sketch of the proof

1. Consider the following subsets of $\mathcal{M}^{eq}_{n,r}$:

$$B := \left\{ \mathcal{X} \in \mathcal{M}^{eq}_{n,r} : \operatorname{rank}_{CP} \mathcal{X} < q^{\frac{d}{2}} \right\},$$

$$B^{(s,t)} := \left\{ \mathcal{X} \in \mathcal{M}^{eq}_{n,r} : \operatorname{rank} \mathcal{X}^{(s,t)} < q^{\frac{d}{2}} \right\},$$

   where $q = \min\{n, r-1\}, s = \{1, 3, \ldots, d-1\}$, $t = \{2, 4, \ldots, d\}$.
2. Any matricization $\mathcal{X}^{(s,t)}$ of a tensor $\mathcal{X}^{i_1 i_2 \ldots i_d}$ of CP-rank r has the ordinary matrix rank at most r. Thus, we have $B \subset B^{(s,t)}$.;
3. Since $B^{(s,t)}$ is an algebraic subset of the irreducible algebraic variety $\mathcal{M}^{eq}_{n,r}$, then $B^{(s,t)}$ either equal to $\mathcal{M}^{eq}_{n,r}$ or has measure 0.

To finish the proof we need find at least one $\mathcal{X}$ such that $\operatorname{rank} \mathcal{X}^{(s,t)} \geqslant q^{\frac{d}{2}}$.

Sketch of the proof

Let us define the following tensors:

$$G_1^{i_1 \alpha_1} = [i_1 = \alpha_1], \quad G_1 \in \mathbb{R}^{1 \times n \times r}$$

$$G_k^{\alpha_{k-1} i_k \alpha_k} = \begin{cases} [\alpha_{k-1} = i_k] \cdot [\alpha_k = q+1], & \text{if } \alpha_{k-1} \leqslant q \\ [i_k = \alpha_k], & \text{if } \alpha_{k-1} = q+1 \\ 0, & \text{if } \alpha_{k-1} > q+1 \end{cases}, \quad G_k \in \mathbb{R}^{r \times n \times r},$$

$$G_d^{\alpha_{d-1} i_d} = [\alpha_{d-1} = i_d], \quad G_d \in \mathbb{R}^{r \times n \times 1},$$

where $[\,\cdot\,]$ is the Iverson bracket notation.

Sketch of the proof

Lemma

Consider $(i_1, i_2, \ldots, i_d) \in [q]^d$ *and* $(\alpha_1, \alpha_2, \ldots, \alpha_{d-1}) \in [r]^{d-1}$ *such that* $G_1^{i_1\alpha_1} \cdot \ldots \cdot G_d^{\alpha_{d-1} i_d} \neq 0$. *Then*

$$\alpha_k = \begin{cases} i_k, & \text{if } k \text{ is odd} \\ q+1, & \text{if } k \text{ is even} \end{cases} \qquad \text{for any } k \in [d-1].$$

The following identity holds true for any values of indices such that $i_k = 1, \ldots, q$, $k = 1, \ldots, d$.

$$\mathcal{X}^{(i_1,i_3,\ldots,i_{d-1}),(i_2,i_4,\ldots,i_d)} = \sum_{\alpha_1,\ldots,\alpha_{d-1}} G_1^{i_1\alpha_1} \ldots G_d^{\alpha_{d-1} i_d} = (\text{by Lemma}) =$$

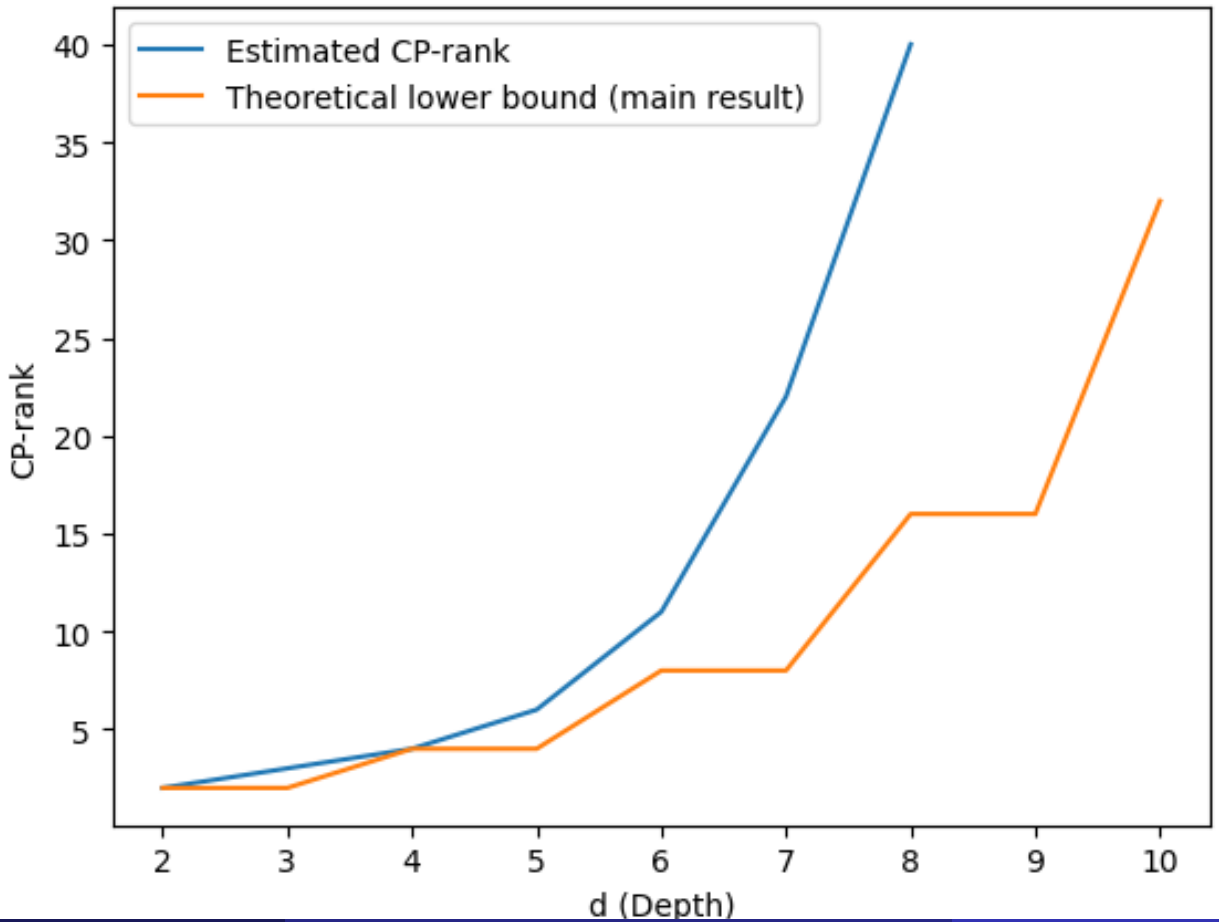
$$= G_1^{i_1 i_1} G_2^{i_1 i_2 (q+1)} G_3^{(q+1) i_3 i_3} \ldots G_d^{i_{d-1} i_d} =$$

$$= [i_1 = i_2] \cdot [i_3 = i_4] \cdot \ldots \cdot [i_{d-1} = i_d] = I^{(i_1,i_3,\ldots,i_{d-1}),(i_2,i_4,\ldots,i_d)},$$

where I is the identity matrix of size $q^{\frac{d}{2}} \times q^{\frac{d}{2}}$.

Numerical experiments

Using MATLAB package Tensorlab for tensor computations, we estimated CP-rank of a random tensor from $\mathcal{M}_{2,3}^{eq}$. The following graph compares the theoretical lower CP-rank estimate (orange line) and the numerical CP-rank estimate of the random tensor from $\mathcal{M}_{2,3}^{eq}$ (blue line).



Numerical experiments

To train the TT-Networks, we implemented them in PyTorch and used Adam optimizer with batch size 64 and learning rate 5e-4. For the experiment, we use computer vision datasets MNIST which is a collection of 70000 handwritten digits. We use the same training scheme for MNIST as in the paper [2]. For MNIST, both TT-Networks and TT-Networks with equal cores show reasonable performance (see the table).

Table: The results of training TT-Networks on MNIST dataset during 20 epochs.

	TT-net with different cores	TT-net with equal cores
Train accuracy	96.36	90.78
Test accuracy	95.32	90.66

Conclusions

- Steps (1), (2), (3) of the proof of Hypothesis 1 are analogous to the proof of the main theorem of the paper [2];
- The novelty of the proof is the last part;
- We have proved a theorem on the lower bound of rank of almost all tensors of fixed size and bounded Tensor Train rank having equal inner TT-cores;
- We were convinced that traditional recurrent neural networks are exponentially more expressive than shallow neural networks. Thus, we close the question of the exponential expressiveness of traditional recurrent neural networks.

[2] Valentin Khrulkov et al. Expressive power of recurrent neural networks (2017)

Thank you for your attention!