Expressive Power of Recurrent Neural Networks, II

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

Research Advisor: Ivan Oseledets[2]

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

18 May 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]. In them, the authors consider the following hypotheses space

$$h_y(x_1, \ldots, x_N) = \sum_{d_1, \ldots, d_N = 1}^{M} A^y_{d_1, \ldots, d_N} \prod_{i=1}^{N} f_{\theta_{d_i}}(x_i),$$

where $A^y_{d_1, \ldots, d_N}$ are trainable parameters, $f_{\theta_{d_i}}(x_i)$ are feature vectors and $h_y(x_1, \ldots, x_N)$ is a score function for label y.

[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

General problem

The authors of the paper [2] showed the connection between RNNs and Tensor Train decomposition of a tensor, and also proved the theorem on the “lower bound” of the CP-rank of a tensor corresponding to an RNN in which each layer has its own set of weights. A similar result (Hypothesis 1) for RNNs with the number of weights independent of the depth of the network has not been proven, but formulated as a hypothesis.

[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 various deep network architectures. In particular,

- To prove Hypothesis 1 from the paper [2];
- To show that the architectures discussed in the paper [2] are complex enough to obtain high quality on real world datasets such.

Connection between Tensor Train tensor decomposition and RNNs

A tensor $\mathcal{X}$ is said to be represented in the Tensor Train (TT) format if each element of $\mathcal{X}$ can be computed as follows

$$\mathcal{X}^{i_1 i_2 \dots i_d} = \sum_{\alpha_1=1}^{r_1} \sum_{\alpha_2=1}^{r_2} \dots \sum_{\alpha_{d-1}=1}^{r_{d-1}} G_1^{i_1 \alpha_1} G_2^{\alpha_1 i_2 \alpha_2} \dots G_d^{\alpha_{d-1} i_d}$$

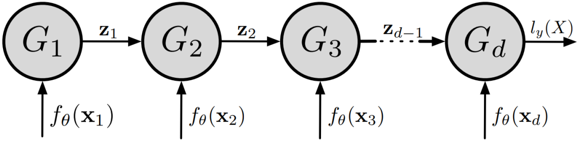


Figure: Recurrent-type neural architecture that corresponds to the Tensor Train decomposition. Gray circles are bilinear maps.

Terms and definitions

Let us denote $\mathrm{n} = (n_1, n_2 \ldots n_d)$. Set of all tensors $\mathcal{X}$ with mode sizes n representable in TT-format with

$$\operatorname{rank}_{TT} \mathcal{X} \leqslant \mathrm{r},$$

for some vector of positive integers r (inequality is understood entry-wise) forms an irreducible algebraic variety, which we denote by $\mathcal{M}_{\mathrm{n},\mathrm{r}}$.
By $\mathcal{M}^{eq}_{\mathrm{n},\mathrm{r}}$ we denote such tensors in $\mathcal{M}_{\mathrm{n},\mathrm{r}}$ whose TT-cores are equal (except the first core and the last core).

Results

Hypothesis 1 from the paper [2], formulated as a theorem:

Theorem

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

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

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

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

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

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

Sketch of the proof

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

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

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

   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 $\mathcal{M}^{eq}_{\mathrm{n,r}}$, then $B^{(s,t)}$ either equal to $\mathcal{M}^{eq}_{\mathrm{n,r}}$ 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 $(i_1, i_2, \dots, i_d) \in [q]^d$ *and* $(\alpha_1, \alpha_2, \dots, \alpha_{d-1}) \in [r]^{d-1}$ *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 $i_k = 1, \dots, q$, $k = 1, \dots, d$.

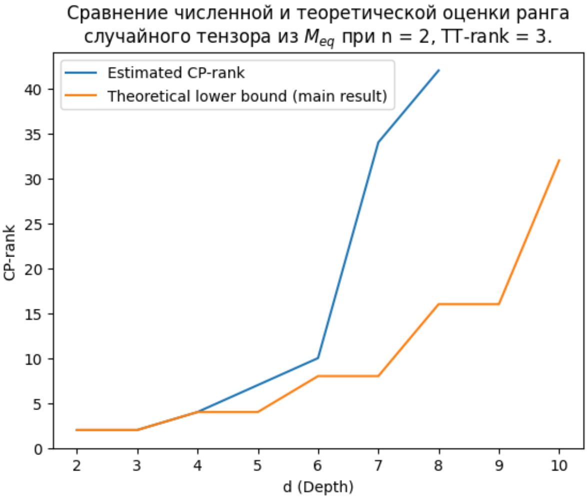
$$\mathcal{X}^{(i_1, i_3, \dots, i_{d-1}), (i_2, i_4, \dots, i_d)} = \sum_{\alpha_1, \dots, \alpha_{d-1}} G_1^{i_1 \alpha_1} \dots G_d^{\alpha_{d-1} i_d} = (\text{by Lemma}) =$$

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

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

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

Numerical experiments



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	95.9	75.3
Test accuracy	95.4	76.2

Discussion of results

- 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;
- The result obtained can be interpreted as follows: the exponential growth in the expressive complexity of recurrent-type models is true even for recurrent nets with equal layers;
- The result obtained is quite theoretical and is not directly applicable to practical tasks. However, to substantiate the usefulness of the result, it is important to show that a TT-network with a high rank can give high results when solving a computer vision problem.

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

Conclusions

- 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 internal TT-cores;
- The experimental results showed us that the TT-networks with equal cores are expressive enough to obtain good accuracy on real-world datasets.

Thank you for your attention!