Tensorizing Neural Networks

Team 28

Skoltech

December 20, 2018

Outline

Outline

Introduction

- Deep Neural Networks demonstrate state-of-the-art performance in several domains
- But it is difficult to train too many parameters of the DNN
- It is also difficult to store too many parameters of the DNN
- Our goal is to efficiently compress the neural nets using tensor decompositions

Outline

Problem statement

- Fully-connected layers of DNN consume up to 90% of memory
- Hard to store them on the portable devices with with small amount of memory
- Compression of dense weight matrices might be a good idea

How to reduce size of weight matrix W in FC-layers?

Problem statement

- Fully-connected layers of DNN consume up to 90% of memory
- Hard to store them on the portable devices with with small amount of memory
- Compression of dense weight matrices might be a good idea

How to reduce size of weight matrix W in FC-layers?

Usually weight matrix W has low rank, it can be efficiently represented using much less parameters

Outline

Tensor Train factorization

$$A \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_d} \text{ is d-dimensional tensor}$$

$$A(l_1, l_2, \ldots, l_d) \overset{\text{TTF}}{=} G_1(l_1)G_2(l_2)\ldots G_d(l_d),$$

where $G_k \in \mathbb{R}^{p_k \times r_{k-1} \times r_k}$, $l_k \in [1, p_k] \forall k \in [1, d]$ and $r_0 = r_d = 1$

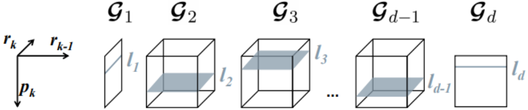
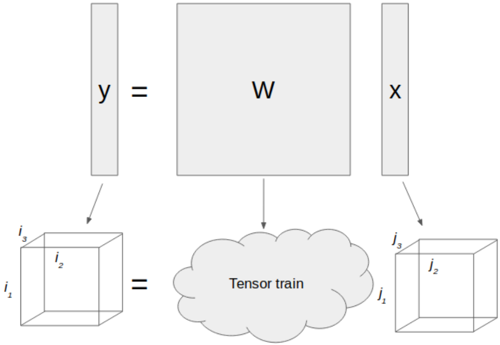


Figure 1: Tensor-Train Factorization Model: To reconstruct one entry in the target tensor, one performs a sequence of vector-matrix-vector multiplications, yielding a scalar.

Fully connected layer



Representation of W in the TT-format

$$\mathbf{y}(i_1, ..., i_d) = \sum_{j_1,...,j_d} \mathbf{G}_1[i_1, j_1]...\mathbf{G}_d[i_d, j_d]\mathbf{x}(j_1, ..., j_d)$$

is called Tensor-Train Layer (Novikov et al. 2015)

Tensor train layer

In theory

- Choose shape of input tensor
- Choose shape of output tensor
- Calculate $W \overset{\text{TTF}}{=} G_1(l_1)G_2(l_2)...G_d(l_d)$, ranks r_k are defined by rank of W

In practice

- Choose shape of input tensor
- Choose shape of output tensor
- Choose $r_1...r_{d-1} \rightarrow$
- Train neural network. TT factorization is low-rank approximation of W

New hyper parameters: shape of input, shape of output, TT ranks

Tensor train layer

Advantages of TT-format for FC-layers representation

- Memory saving
- Linear algebra operations without reconstruction of W
- Training without reconstruction of W

Outline

Experiments

- MNIST

max TT core rank	compression rate	accuracy
1	6.6e-5	0.971
2	2.0e-4	0.975
3	3.9e-4	0.980
4	6.7e-4	0.982

- CIFAR10

max TT core rank	max accuracy
1	0.51
2	0.55
3	0.56
4	0.56

Experiments

Experiments

Experiments

Outline

Conclusion

- High redundancy in the current neural network parameterization could be reduced using TT-decomposition of the FC-layers
- In some cases TT-layer representation allows us to train the fully-connected layers compressed by up to $100000\times$ compared with the explicit parameterization without significant error increase.
- Compact TT-representation allows to train really huge FC-layers
- Exist more advanced extensions of TT-format for CNNs and RNNs

Thank you for your attention!