ABSTRACT

Tensorized Neural Networks for Faster Training and Better Co-design

Authors:

Chunxing Yin, CSE, Georgia Tech
Richard Vuduc, CSE, Georgia Tech

In this work we studied the potential of tensorized neural networks and its implication on co-design. Studies have shown that we will soon reach the computational limits if progress continues along current lines. We explored to what extent the large neural networks can be trained in a reduced form using the techniques of low-rank tensor train decomposition. We focused on 2 types of networks, convolutional neural networks and deep learning recommendation models. Both networks can scale to tens of GB and can't be trained on a single GPU. Our core idea is to replace the largest modules, such as large convolution kernels and embedding tables with their Tensor Train decomposition, which is a sequence of small tensor products. In recommenders, we show that the model can be compressed by 100x with no loss of accuracy and only 10% overhead on training time. In Wide Resnet, our approach was able to compress the network by 5 times and accelerate the training by 1.3 times with marginal loss of accuracy.