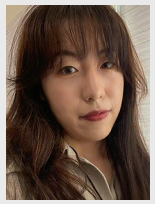
TensorTrim: Dynamic Tensor-Train Decomposition for Efficient Neural Network Compression



In the field of model compression, choosing an appropriate rank for tensor decomposition is pivotal for balancing model compression rate and efficiency. However, this selection, whether done manually or through optimization-based automatic methods, often increases computational complexity. Manual rank selection lacks efficiency and scalability, often requiring extensive trial-and-error, while optimization-based automatic methods significantly increase the computational burden. To address this, we introduce a novel, automatic, and budget-aware rank selection method for efficient model compression, which employs Layer-Wise Imprinting Quantitation (LWIQ). LWIQ quantifies each layer's significance within a neural network by integrating a proxy classifier. This classifier assesses the layer's impact on overall model performance, allowing for a more informed adjustment of tensor rank. Furthermore, our approach includes a scaling factor to cater to varying computational budget constraints. This budget awareness eliminates the need for repetitive rank recalculations for different budget scenarios. Experimental results on the CIFAR-10 dataset show that our LWIQ improved by 63.2% in rank search efficiency, and the accuracy only dropped by 0.86% with 3.2x less model size on the ResNet-56 model as compared to the state-of-the-art proxy-based automatic tensor rank selection method.

Shiyi Luo, *Mingshuo Liu, Shangping Ren, and Yu Bai*

This research is supported by the SDSU Presidential Graduate Research Fellowship, the University of CA, Irvine, and the Computational Science Research Center (CSRC) at San Diego State University