Generic, Sparse Tensor Core for Neural Networks

Xiaolong Wu
Purdue University
West Lafayette, IN, USA
wu1565@purdue.edu

Yang Yi
Virginia Tech
Blacksburg, VA, USA
cindy_yangyi@vt.edu

Dave Tian
Purdue University
West Lafayette, IN, USA
daveti@purdue.edu

Jiajia Li
Pacific Northwest National Laboratory
Richland, WA, USA
Jiajia.Li@pnnl.gov

***Abstract*—Sparse neural networks attract broad attention to model compression, fast execution, and power reduction. The state-of-the-art designs of sparse tensor cores for NVIDIA GPUs target on structured and static sparsity, and do not support generic or dynamic sparsity well. We design a sparse tensor core architecture to support generic sparsity pruning with a novel hybrid and blocked sparse matrix storage format, HB-ELL, which saves computation and storage while keeping the most significant elements, as well as supporting dynamic sparsity for data flow in neural networks. We achieve better performance in our preliminary results than the state of the art on an NVIDIA GPU simulator.**

BIOGRAPHY

Jiajia Li is a staff scientist at High Performance Computing group of Pacific Northwest National Laboratory (PNNL), Richland, WA. Her research emphasizes on high performance computing with a focus on the interaction among applications, numerical methods, data structures, algorithms, automatic performance tuning, and computer architectures. She is eager to pursue high performance sparse (multi-)linear algebra, solvers, and tensor decompositions for large-scale data analytics and domain applications on diverse computer architectures. She has received her Ph.D. degree in Computational Science & Engineering at Georgia Institute of Technology. She has received Rising Stars in Computational and Data Sciences, Best Student Paper Award, and IBM PhD Fellowship. In the past, she has received a Ph.D. degree from Institute of Computing Technology at Chinese Academy of Sciences.