Eager Pruning: Algorithm and Architecture Support for Fast Training of Deep Neural Networks

Corresponding author

Jiaqi Zhang, Xiangru Chen, Mingcong Song, Tao Li, University of Florida, USA

Keywords

neural network training, neural network pruning, software-hardware co-design

Summary

*Challenge*

1. Several works demonstrate redundancy in DNNs.
2. Ranking of the significance of the weights changes slightly during training.

*Contribution*

1. Propose Eager Pruning, which speeds up DNN training by moving pruning before original training. (Algorithm)
2. Co-design a novel architecture to transform the reduced training computation into performance improvement. (Architecture)

*Innovation Points*

1. Due to the irregular and changeable sparsity in EP causes resource underutilization. They break all the fixed connections between the processing elements (PEs) so that each PE can be assigned independently.
2. The DRACT can dynamically collect and accumulate the partial results with the design of path gate (PG). PG can realize the dynamic combination of PEs operations.

*Result*

Eager Pruning system gains an average of 1.91 speedup over state-of-the-art hardware accelerator and 6.31 energy-efficiency over Nvidia GPUs.