Contribution ID: **21** Type: **Talk**

An Early Exploration into the Interplay between Quantization and Pruning of Neural Networks

Tuesday, 1 December 2020 15:10 (8 minutes)

Machine Learning (ML) is already being used as a powerful tool in High Energy Physics, but the typically high computational cost associated with running ML workloads is often a bottleneck in data processing pipelines. Even on high performance hardware such as Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs) the speed and size of these models are often heavily constrained by available hardware resources. Various model optimization techniques, such as pruning and quantization, have been used in an attempt to alleviate the high performance costs, but often not together, or without fully understanding the interplay of the two techniques. We attempt to explore this interplay between quantization and pruning in order to better understand how they interact. Targeting FPGAs and ASICs, we attempt to determine how to yield the best performance with both quantization and pruning. In this presentation, we explore these techniques by optimizing the HLS4ML 3 Hidden Layer Jet Substructure tagging model, finding that we can successfully optimize the model down to 3-5% of its original size while retaining comparable performance to the original network. Finally, we discuss some next steps into understanding how the different optimization techniques affect the model internally, beyond standard performance metrics.

Primary authors: Mr HAWKS, Benjamin (Fermi National Accelerator Laboratory); TRAN, Nhan Viet (Fermi National Accelerator Lab. (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US))

Presenter: Mr HAWKS, Benjamin (Fermi National Accelerator Laboratory)