**Title:** Hardware-aware Design, Search and Optimization of Deep Neural Networks

**Abstract:** Convolutional Neural Networks (CNNs) are commonly used in computer vision, virtual reality, and autonomous driving applications. Pruning and Quantization are two effective Deep Neural Network (DNN) compression methods for efficient inference on various hardware platforms. Pruning is a method to reduce the size of a DNN model by removing the redundant parameters. Quantization is a technique to lower the precision (bit-width) of CNN parameters. The pruned and low precision models result in smaller and faster inference models on hardware platforms with almost the same accuracy as the unoptimized network.

Mixed Precision Quantization is a promising solution to overcome the difficulties of single-precision assignment by allotting different precision to different layers in the neural network. Neural Architecture Search (NAS) methods search network architectures that are more accurate and hardware-efficient compared to the handcrafted/manually designed models. In parallel to efficient neural network design, advances in accelerator hardware design also occurred for efficient processing of CNN forward propagation. Hardware-Aware NAS (HW-NAS) aims to search for a neural network architecture that maximizes the accuracy and performance metrics (Ex: latency, MACs) at the same time for a given dataset.

We developed “Hardware Dimension aware Pruning” method, which prunes neural networks with respect to the size and shape of the hardware dimension for efficient inference. We also developed “Mixed Sparse and Precision” method to search for optimized models on a fixed predefined search space. We illustrate the effectiveness of our methods targeting Nvidia’s Tensor Core-enabled GPUs and Google’s Tensor Processing Unit by searching efficient networks to achieve better accuracy-latency trade-off models compared to the previous neural network models. Our research is particularly useful for researchers willing to port their applications on hardware accelerators.

**Bio:** Krishna Teja Chitty-Venkata received the Bachelor of Engineering degree in Electronics and Communication Engineering from Osmania University, India in 2017. He is currently a Ph.D. Candidate in Computer Engineering at Iowa State University. He primarily works at the intersection of Deep Neural Network (DNN) algorithm design and hardware acceleration. His research lies in optimizing DNNs for efficient inference on special-purpose and general-purpose hardware platforms. He interned at Intel, AMD and Argonne National Laboratory during his graduate studies at Iowa State.