A Review of An Energy-Efficient Deep-Neural-Network Learning Processor With Stochastic Coarse–Fine Level Weight Pruning and Adaptive Input/Output/Weight Zero Skipping

1) Aryan Langhanoja
2) Jay Mangukiya
Department of Information and Communication Technology
Marwadi University
Rajkot , India

Abstract—

The paper introduces PNPU, an energy-efficient deep-neural-network (DNN) learning processor designed to reduce the complexity of over-parameterized networks during training. The authors highlight the need for DNN training at the edge for high accuracy and domain-specific adaptation. They also discuss the importance of weight pruning to remove unnecessary parameters and improve energy efficiency. However, existing methods for weight pruning and zero skipping have limitations in backpropagation and weight gradient update. To address these challenges, the authors propose PNPU with three key features: stochastic coarse-fine level pruning, adaptive input/output/weight zero skipping, and a weight pruning unit with weight sparsity balancer.

Keywords— *Deep learning, deep neural network, pruning, training, DNN accelerator, energy efficiency.*

Introduction—

The abstract provides a concise overview of the paper's content. It mentions the limitations of existing DNN learning processors and introduces PNPU as a solution. The three key features of PNPU are briefly described, along with the achieved energy efficiency. The abstract concludes by emphasizing the importance of DNN training at the edge and the need for energy-efficient processors.

Body—

The body of the paper provides a detailed explanation of the proposed PNPU architecture and its key features. It starts by discussing the challenges of designing the processor, including the computational overhead of coarse level pruning, the degradation of zero skipping efficiency during DNN learning, and the issue of weight sparsity imbalance.

The authors then introduce stochastic coarse-fine level pruning (SCFLP) as a solution to reduce the computational overhead of coarse level pruning. They explain how SCFLP combines fine level pruning (FLP) and coarse level pruning (CLP) to generate high weight sparsity while maintaining accuracy. The structured pruning method is also introduced to increase the ratio of consecutive zeros, improving coarse level sparsity.

Next, the authors describe the adaptive input/output/weight zero skipping (AIOWS) feature of PNPU. They explain how hierarchical zero skipping (HZS) is implemented to efficiently handle dual zero skipping. The details of the HZS controller and the fine skipping controllers (FSC) for input/weight and output/weight zero skipping are provided.

The paper then discusses the weight pruning unit of PNPU, which includes the random channel allocator (RCA) for stochastic grouping of output channels and the weight sparsity balancer to address workload imbalance caused by weight sparsity. The authors explain how weight data sharing is used to eliminate additional weight memory access.

Finally, the authors present the measurement results of PNPU, including chip performance and energy efficiency. They compare PNPU with previous DNN learning hardware and highlight its superior performance under the same activation and weight sparsity conditions.

Conclusion—

In conclusion, the paper introduces PNPU as an energy-efficient DNN learning processor that addresses the limitations of existing processors. The proposed architecture incorporates stochastic coarse-fine level pruning, adaptive input/output/weight zero skipping, and a weight pruning unit with weight sparsity balancer. The measurement results demonstrate the high energy efficiency and performance of PNPU compared to previous DNN learning hardware. The authors emphasize the importance of DNN training at the edge and the need for energy-efficient processors like PNPU.