Embedded Systems Week

Accelerating Large-Scale Graph Neural Network Training on Crossbar Diet

Chukwufumnanya Ogbogu†, Aqeeb Iqbal Arka†, Biresh Kumar Joardar*, Janardhan Rao Doppa†, Hai (Helen) Li*, Krishnendu Chakrabarty*, Partha Pratim Pande† .
Washington State University†, Duke University*

Introduction

- Training machine learning (ML) models at the edge (training on-chip or on embedded systems) can address many pressing challenges, including data privacy/security.
- Resistive random-access memory (ReRAM) based processing-in-memory (PIM) architectures can be used to address this problem.
- We propose a crossbar-aware pruning technique called **DietGNN** (GNN pruning on a crossbar diet) to address the storage, computation, and communication challenges of ReRAM-based GNN accelerators.
- DietGNN-enabled ReRAM-based PIM architecture achieves low energy- and storage-efficient GNN computation

Background and Overview

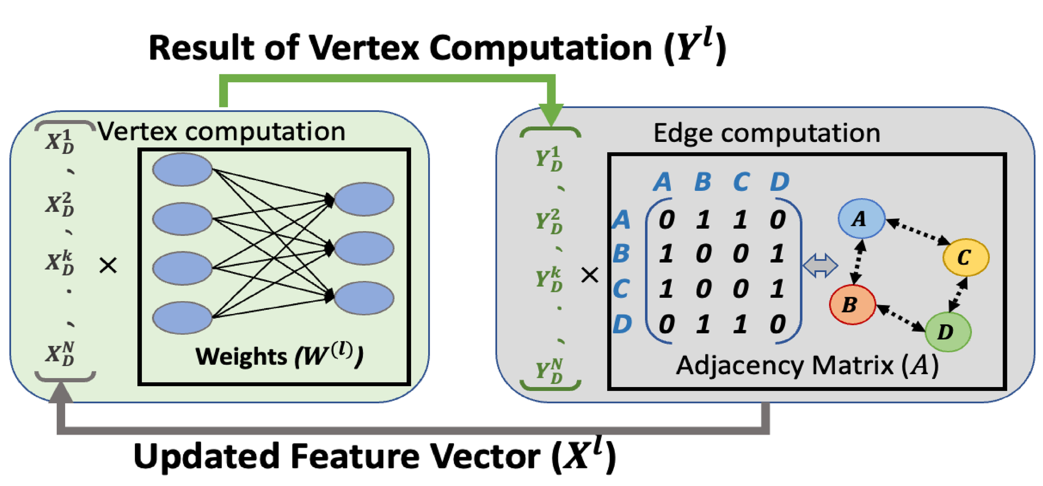


Fig. 1: Two phases of the GNN computation kernel.

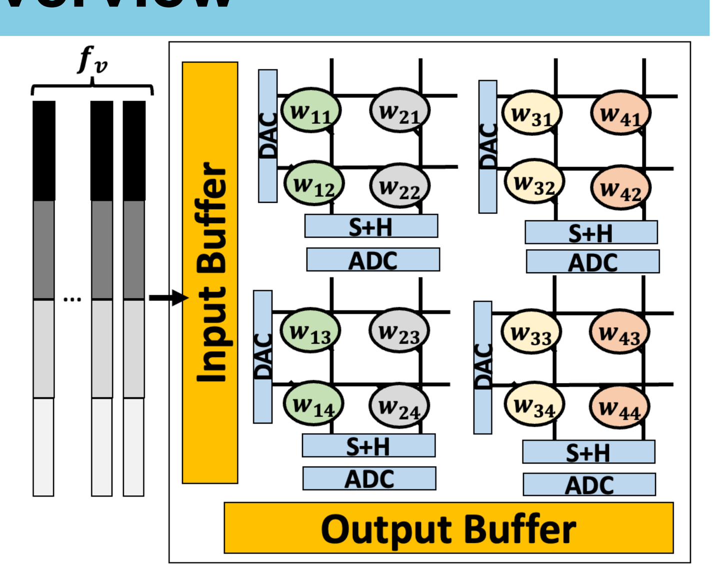


Fig. 2: Mapping the weights of a GNN layer to ReRAM crossbars.

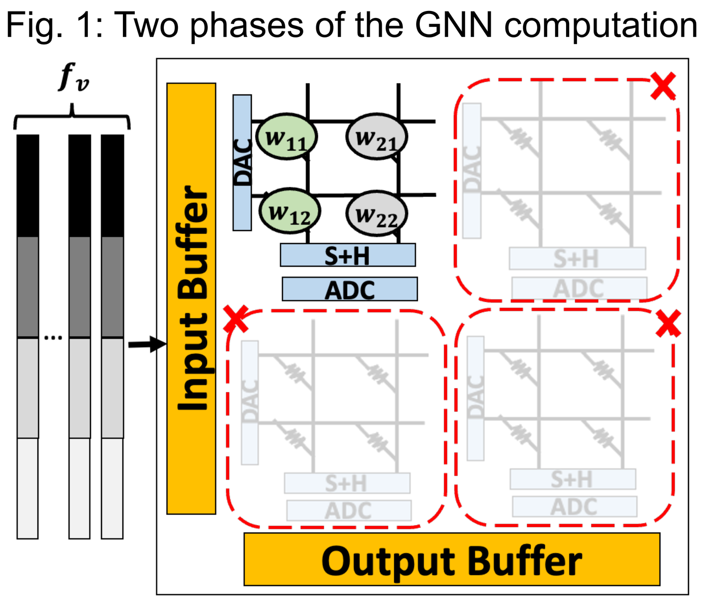


Fig. 3: Mapping the DietGNN Pruned weights to ReRAM crossbars

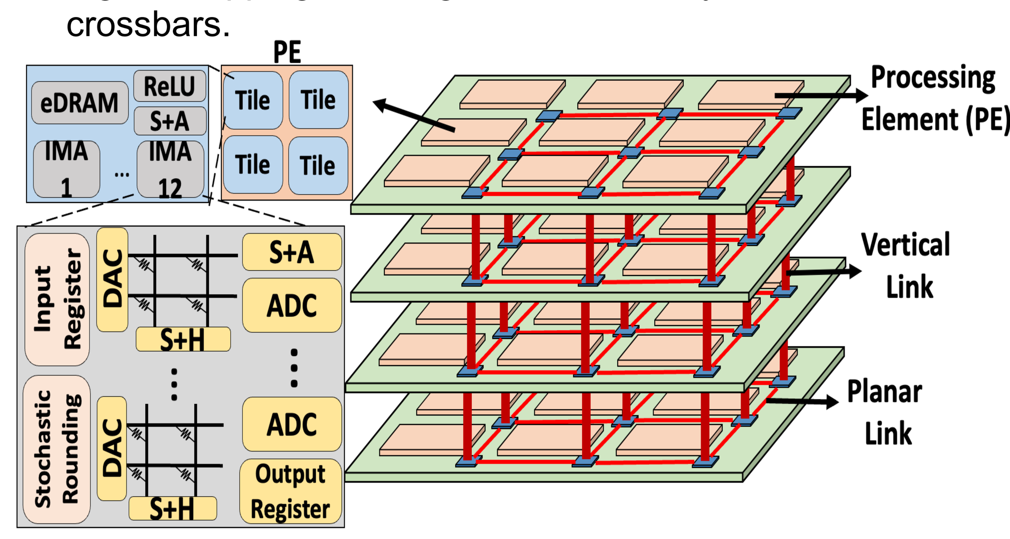


Fig. 4: Illustration of the 3D ReRAM-based PIM Architecture

Methodology

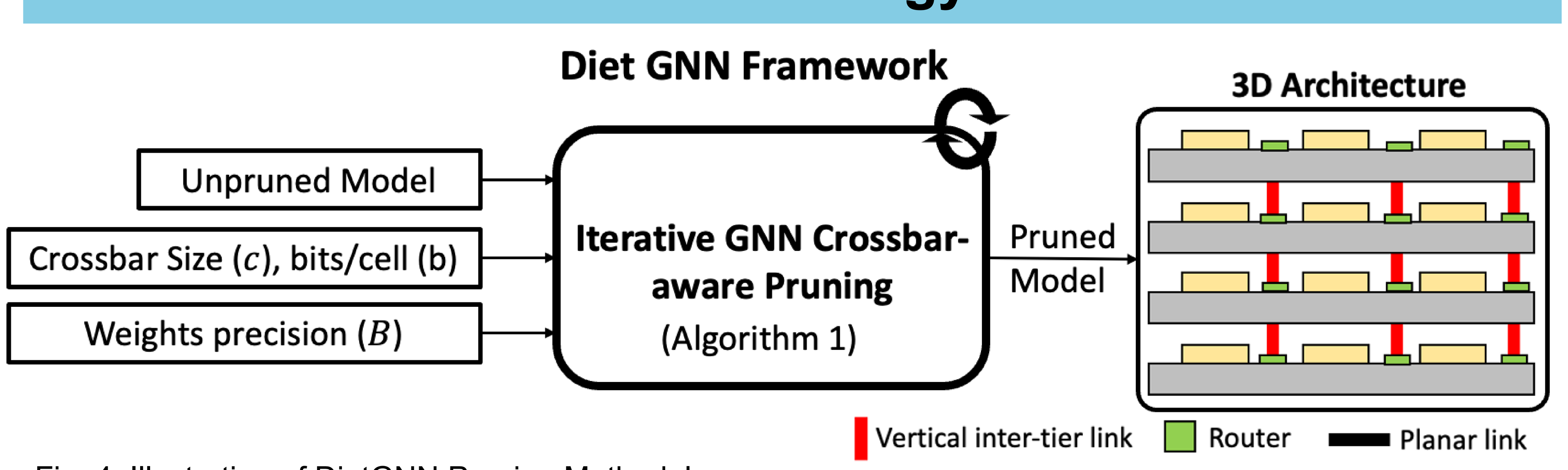


Fig. 4: Illustration of DietGNN Pruning Methodology

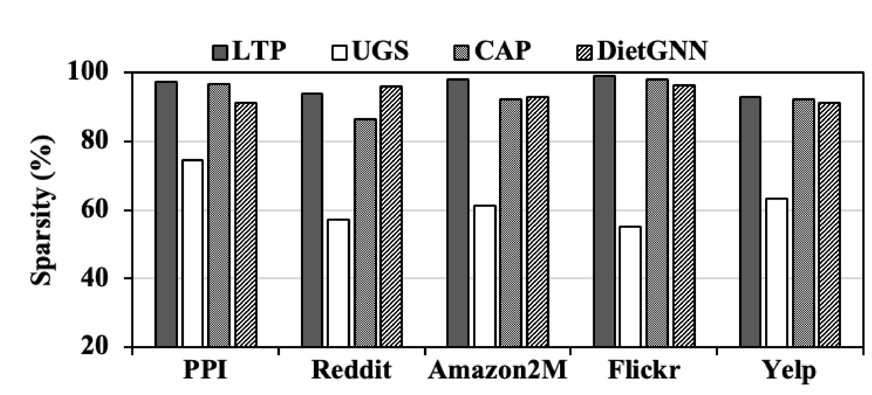
Algorithm 1. Pruning with DietGNN

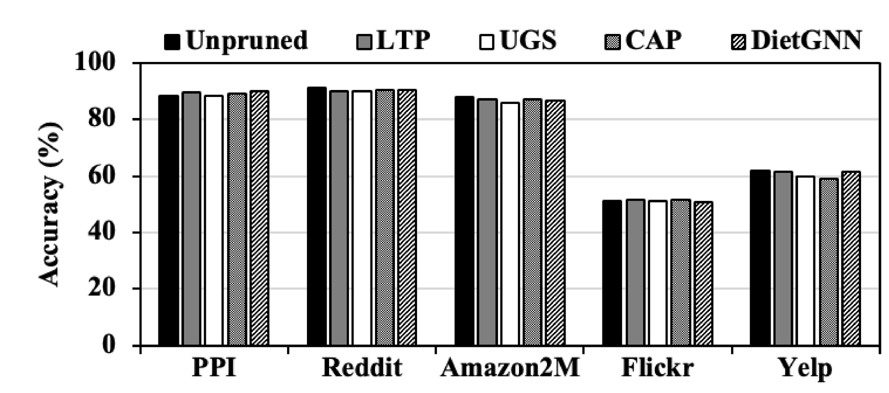
Input: GNN model, crossbar structure, prune percentage p
Output: Pruned GNN model or winning ticket
Algorithm:

1: **Initialize**: $W^l \leftarrow W_{initial}$;
2: **Partition** W^l into blocks (B^l) of size $c \times \left(c * \frac{b}{B}\right)$
3: **While** $itr < n$:
4: **Train** for E epochs
5: **Prune** $p\%$ of B^l based on average magnitude
6: **Reinitialize** remaining **weights** with $W_{initial}$
7: **Return** *Pruned Model (Hardware-friendly winning ticket)*

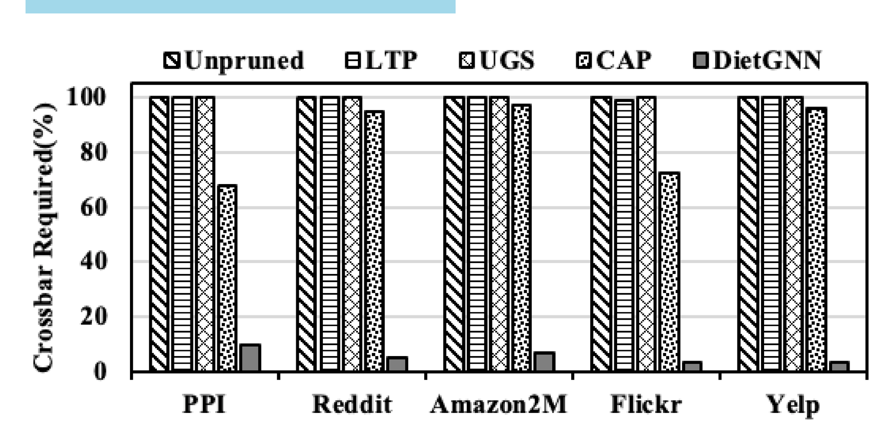
Results

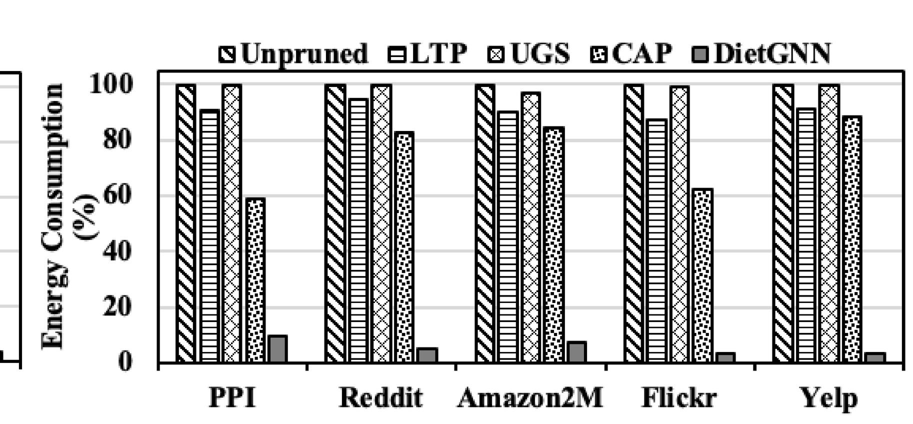
Accuracy & Sparsity



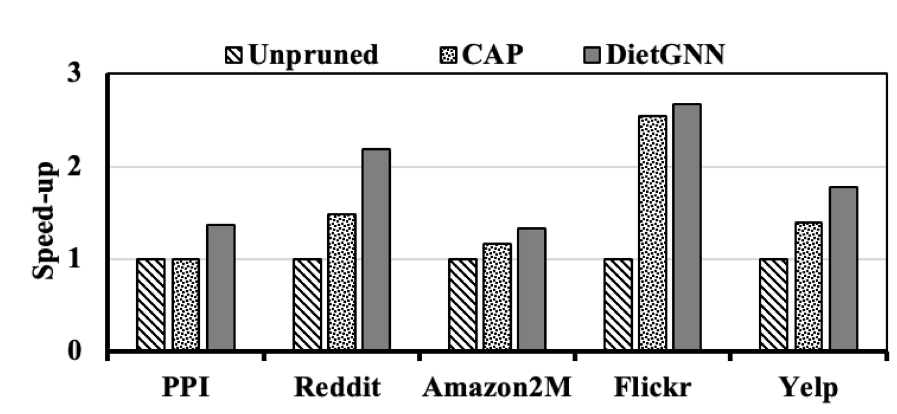


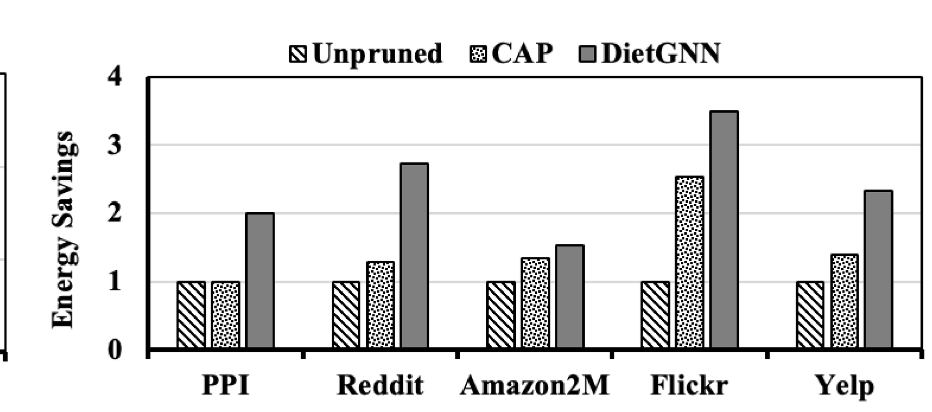
Area & Energy



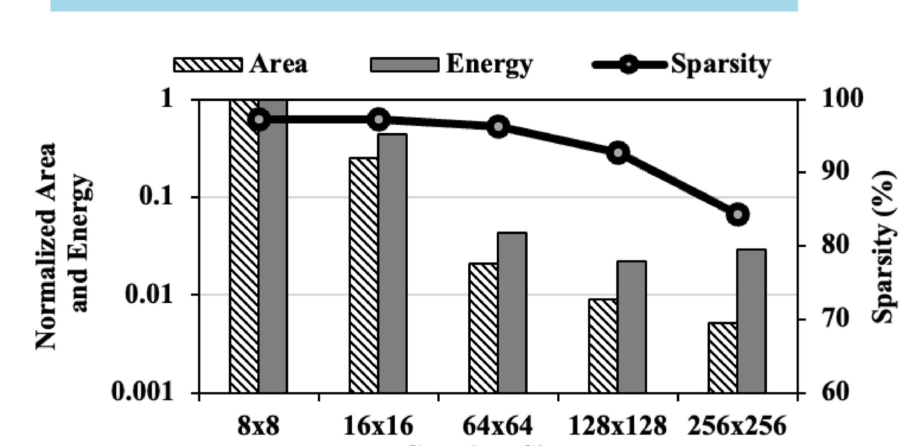


Overall System Speed-up & Energy Savings

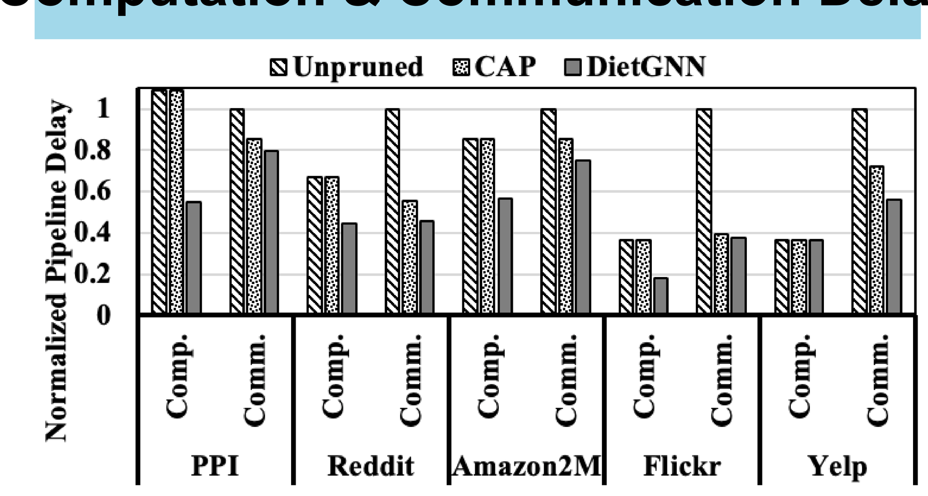




Sparsity, Area, Energy tradeoffs



Computation & Communication Delay



Conclusion

We have presented a crossbar-aware pruning technique called DietGNN, which can be trained from scratch, achieves high sparsity, and enables significant reduction in energy consumption and area overhead. DietGNN achieves ~2.7 × speedup while being 3.5 × more energy efficient when compared to its unpruned version on an ReRAM-based manycore platform.

Acknowledgements

We thank our collaborators at Duke University for their insightful feedback and suggestions throughout the project development. The paper got accepted at CASES 2022 at the ESWEEK Conference.

WASHINGTON STATE UNIVERSITY