RAVI RAJU

11009 Alterra Pwky
Austin, TX 78758, USA

262-957-6388
ravi.raju0594@gmail.com

PUBLICATIONS

R. S. Raju, K. Daruwalla, M. Lipasti, *Accelerating Deep Learning with Dynamic Data Pruning*, November, 2021. [pdf]

Ravi Raju, Dibakar Gope, Urmish Thakker, and Jesse Beu. 2020. Understanding the Impact of Dynamic Channel Pruning on Conditionally Parameterized Convolutions. In Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things (AIChallengeIoT '20). Association for Computing Machinery, New York, NY, USA, 27–33. DOI:https://doi.org/10.1145/3417313.3429381 [pdf]

R. S. Raju and M. Lipasti, "BlurNet: Defense by Filtering the Feature Maps," 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2020, pp. 38-46, doi: 10.1109/DSN-W50199.2020.00016. [pdf]

EXPERIENCE

Senior Software Engineer 11/29/2022-Present
SambaNova Systems

- Creating high quality data via deduplication, domain filtering, and templating to train billon parameter LLMs, as well as evaluating these models through community benchmarks and human evaluation.
- Applying techniques such as pruning/quantization (up to 50% sparsity on 176B model) in order to reduce memory bandwidth requirement to improve inference latency

UW-Madison Research Assistant 8/21/2018 – 08/16/2022
UW-Madison ECE Department

- Accelerated training deep learning models with novel **dynamic data pruning** using active learning/RL methods which enables aggressively pruning input data to obtain 2x speedup with no loss in performance. When more data is pruned, our method's performance degradation is minimal compared to prior works using static methods.
- Developed defense algorithms like **BlurNet** for which reduce the impact of adversarial attacks like Robust Physical Perturbations on neural networks, dropping attack success rate from 90% to 17.5%.

Machine Learning Research Intern 5/17/2021 – 8/20/2021
ARM ML Research

- Applied Saliency-based patches to accelerate differentiable neural architecture search (NAS) for 4x speedup on search phase for transformer-based (DieT) architectures.
- Met the target operation count resulting in similar representation of models from NAS on ImageNet and similar accuracy target.

Machine Learning Research Intern 5/19/2020 – 8/14/2020
ARM ML Research

- Analyzed conditional execution methods like dynamic channel pruning and conditionally parameterized convolution for efficient inference.
- Wrote paper on **Understanding the Impact of Dynamic Channel Pruning on Conditionally Parameterized Convolutions** and showed 7.2% savings in computational costs at iso-accuracy and 1.01% improvement in accuracy at iso-computational costs over the state-of-art Dynamic Channel Pruning technique.

UW-Madison Teaching Assistant 8/21/2017 – 8/21/2018
UW-Madison ECE Department

- Provided instruction on Digital System Fundamentals/Intro to. CompE in active learning setting.
- Administered lab sessions with Quartus II software for logic circuit design.
- Held office hours and mentored students during office hours on course material.

EDUCATION

UW Madison 05/2018 – 08/2022
PhD in Computer Engineering
GPA: 3.67/4.0
Graduation Date: 08/2022

UW Madison 09/2016 – 05/2018
Masters in Computer Engineering
GPA: 3.67/4.0

Milwaukee School of Engineering 09/2012 – 05/2016
Bachelor’s in Electrical Engineering
GPA: 3.7/4.0

Courses
ECE-532 Theory and Applications of Pattern Recognition
ECE-752 Advanced Computer Architecture I
ECE-759 High Performance Computing
ECE-524 Introduction to Optimization
ECE-757 Advanced Computer Architecture2
ECE-533 Image Processing

SKILLS

Computer Skills/Programming
Programming in C, Unix, OpenMP, CUDA, Matlab, Python, TensorFlow, Pytorch