OSHIN DUTTA

https://oshindutta.github.io/ (+91) 7003132699 ◇ oshind.phd@gmail.com

EDUCATION

Indian Institute of Technology (IIT) Delhi,India 2019 - 2025(expected)
Ph.D. in Efficient AI, CGPA: 8.0/10.0

Indian Institute of Technology (IIT), Dhanbad, India 2016 - 2018
M.Tech. in Electronics and Communication, CGPA: 9.34/10

Visvesvaraya Technological University (VTU), India 2011 - 2015
B.E. in Electronics and Communication, Overall: 80.3% Award: Academic Excellence

EXPERIENCE

Ph.D. Scholar, IIT Delhi July 2019 - Present

Accelerating LLMs (With Dr. Sumeet Agarwal and **Cadence India**)

- Developed a pruning algorithm that is data-efficient, speedups up pruning and finetuning LLMs by 10× over previous state-of-art. LLMs like LLaMA and Mistral were compressed by over 50% and inference speedup improved up to 60% with state-of-art performance.

Compressing Action Recognition Models (With Dr. Prathosh A.P. and **Samsung Research**)

- Developed compression algorithm that achieves over 70× higher compression than previous state-of-art for Human Action Recognition on large scale datasets. Model deployed achieved about 100× speedup on Raspberry Pi over full-sized LSTM models.

Student Researcher, IIT Dhanbad June 2017- June 2018

Rhythm extraction in polyphonic music and tempo octave correction using ML techniques. Published at a premier IEEE conference.

Intern, Aerospace Dept., IISc Bangalore Feb 2015 - May 2015

Coded and simulated a guidance algorithm for precise fuel-efficient lunar landings- Evaluated throughput and computational efficiency on the TMS320C6748 DSP processor

SKILLS

Programing Frameworks: Python, C, Java, MATLAB, Pytorch, TensorFlow, OpenCV
AI Models handled: CNNs, RNNs, GANs, LLMs, ViTs, Neural Architecture Search
Efficient AI Techniques: Data-efficient learning, HW-SW codesign, Quantization, PEFT, LoRA
Hardware: Distributed Computing Systems, HPC, NVIDIA A100, V100, Orin, Raspberry Pi, DSP
Generative AI, Model Optimization, Scalable Deployment, NLP, Computer Vision

POSITIONS OF RESPONSIBILITY

Research Associate 2020 - 2024

Mentored and worked in a team with 10 undergraduate, graduate students and collaborators and further leading coauthored publications in high-impact venues such as ICML, WACV. Also lead knowledge transfer with industry collaborators.

Teaching Assistant 2021 - 2023

Taught and assisted in several courses such as Cognitive and Intelligent Systems (2023), Introduction to Machine Learning (2022), Machine Intelligence and Learning (2021) and Introduction to Electrical Engineering (2021)

PUBLICATIONS

https://scholar.google.co.in/citations?user=SOzYDkEAAAAJ

- **O. Dutta**, R. Gupta, and S. Agarwal.,"Efficient LLM Pruning with Global Token-Dependency Awareness and Hardware Adapted Inference", Es-FoMo II@ International Conference on Machine Learning **(ICML)** 2024
- **O. Dutta**, T. Kanvar, and S. Agarwal., "Search-Time Efficient Device Constraints-Aware Neural Architecture Search", International Conference on Pattern Recognition and Machine Intelligence **(PReMI)**, 2023
- **O. Dutta**, A. Srivastava, P. AP, S. Agarwal, and J. Gupta., "A Variational Information Bottleneck Based Method to Compress Sequential Networks for Human Action Recognition", IEEE/CVF Winter Conference on Applications of Computer Vision **(WACV)**, 2021
- **O. Dutta**, R. Gupta, and S. Agarwal., "VTrans: Accelerating Transformer Compression with Variational Information Bottleneck based Pruning", PrePrint
- **O. Dutta**, "Tempo Octave Correction Using Multiclass Support Vector Machine", International Conference on Inventive Communication and Computational Technologies (ICICCT),**IEEE**, 2018