

Oshin Dutta

PhD scholar

Indian Institute of Technology, Delhi.

Contact: oshin.dutta@ee.iitd.ac.in , oshindutta13@gmail.com

Website: <https://oshindutta.github.io/>



EDUCATION

Ph.D. in Efficient Deep Learning

Indian Institute of Technology Delhi

(2019 – present, Synopsis completed))

- **Area:** Model compression, Generative AI, Neural Architecture Search, Computer Vision, NLP
- **Thesis:** Optimizing Deep Learning Models for Resource Constrained Environments supervised by Sumeet Agarwal and Prathosh A.P.
- **Key Achievements:** Developed novel compression algorithms for popular LLMs and vision models; tested models on a wide range of classification and generative tasks. Published at high impact venues like ICML, WACV

Master of Technology in Electronics and Communication

Indian Institute of Technology Dhanbad

(2016 – 2018)

- **Area:** Machine Learning, Audio Signal Processing
- **CGPA:** 9.34/10
- **Thesis:** Tempo estimation and Octave Correction using vibrato suppression and Support Vector Machines

PROJECTS

- **Neural Network Model Compression**, IIT Delhi ([link](#)) (Sept 2019-2024)
 - Developed novel *VIB-LSTM* that achieves over 70× higher compression than previous state-of-art, 100× LSTM speedup for action recognition on Raspberry Pi
 - Developed *VTrans*, that is data-efficient, speedups up finetuning by 10× over previous state-of-art and compresses billion-parameter LLMs by over 50% with state-of-art accuracy.
 - Developed *TVA-prune* that enables better GPU utilization thus improving inference speedup up to 60% over previous SOTA with state-of-art accuracy of LLMs like LLaMA and Mistral
 - Developed *DCA-NAS* that enables hardware-aware search for optimal models on distributed GPUs and achieves up to 5× faster search on ImageNet
 - Collaborated with Samsung Research, and Cadence India, resulting in multiple high-impact publications at ICML, WACV
- **Rhythm estimation of various genres of music**, IIT Dhanbad ([link](#)) (June 2017 – May 2018)
 - Rhythm extraction in polyphonic music and tempo octave correction using Support Vector Machines. Published at a premier conference.
 - Analysis done with the help of MATLAB and Python
- **Fuel-Optimal Soft Lunar Landing Using Generalized Model Predictive Static Programming (GMPSP) algorithm**, IISc Bangalore (Feb 2015 - May 2015)
 - Coded and simulated a guidance algorithm on a TMS320C6748 DSP, optimizing memory usage and execution time to simulate precise, fuel-efficient lunar landings
- **Analysis of Hypersonic Shockwave Data for Missile Technology**, CMR Institute of Technology ([link](#))
 - Analysis and reduction of the shock waves registered during hypersonic speed of travel. (April to May 2014)

SKILLS

- **Programing Languages:** Python, C, Java, MATLAB
- **Frameworks and Libraries:** PyTorch, TensorFlow, OpenCV, PySpark
- **Scientific Paper Documentation:** LaTeX
- **Hardware:** Distributed Computing Systems, HPC, Edge Devices like Orin, Raspberry Pi, DSP
- **Efficient Techniques:** Data-efficient learning, HW-SW codesign, Model Compression, Quantization, PEFT, LoRA
- **Generative AI, Self-supervised Learning, Multimodal learning**

PUBLICATIONS

- O. Dutta, R. Gupta, and S. Agarwal. "Efficient LLM Pruning with Global Token-Dependency Awareness and Hardware-Adapted Inference." ([link](#)) Es-FoMo II@ **ICML** 2024
- O. Dutta, T. Kanvar, and S. Agarwal. "Search-Time Efficient Device Constraints-Aware Neural Architecture Search." ([link](#)) Springer, **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." ([link](#)) **WACV**, 2021

- *O. Dutta*, "Tempo Octave Correction Using Multiclass Support Vector Machine." ([link](#)) ICICCT, **IEEE**, 2018

Under Review:

- *O. Dutta*, R. Gupta, and S. Agarwal. "VTrans: Accelerating Transformer Compression with Variational Information Bottleneck based Pruning." ([link](#)) arXiv preprint

EXPERIENCE

- **Research Assistant, IIT Delhi,** (Oct 2019- 2024)
 - Delivered deployable AI solutions while collaborating on projects with Samsung Research and Cadence
 - Mentored and worked in a team with several undergrad and grad students and interns, leading to co-authored publications in high-impact venues
 - Presented research and attended prestigious conferences and workshops of ICML, ACML, PReMI, Google Research Week
- **Intern, Aerospace Dept., IISc Bangalore,** (Feb 2015- May 2015)
 - Converted the Generalized Model Predictive Static Programming (GMPS) control guidance algorithm for moon lander navigation into optimized C code.
 - Simulated the algorithm in MATLAB to analyse and validate the guidance path's accuracy
 - Evaluated throughput and computational efficiency on the TMS320C6748 DSP processor

SERVICES

- **Teaching Assistant** for courses- Cognitive and Intelligent Systems (2023), Introduction to Machine Learning (2022), Machine Intelligence and Learning (2021), Introduction to Electrical Engineering (2021)
- **Reviewer** for various conferences-WACV, Women in Machine Learning (WiML), AISTATS, ICML, IJCAI