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 Compressing AI Models
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, PReMI

Master of Technology, Electronics and Communication
Indian Institute of Technology Dhanbad (2016 – 2018)

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

PROJECTS

- **Neural Network Model Compression,** IIT Delhi (link) (Sept 2019-2024)
  - Achieved 100x LSTM speedup for action recognition on Raspberry Pi via novel *VIB-LSTM* algorithm design.
  - Developed *VTrans*, a novel algorithm that compresses billion-parameter models like LLaMA by over 50% while maintaining competitive accuracy
  - Achieved over 10x faster search for resource-constrained devices using our novel *DCA-NAS* approach
  - Collaborated with PolicyBazaar, Samsung Research, and Cadence India, resulting in multiple high-impact publications at ICML, WACV
- **Rhythm estimation of various genres of music,** IIT (ISM) Dhanbad (link) (June 2017 – May 2018)
  - Rhythm extraction in polyphonic music and tempo octave correction using Support Vector Machines
  - Dominant Technologies: MATLAB, 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 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, Edge Devices like Orin, Raspberry Pi, Digital Signal Processors
- **Data-efficient Learning, Generative AI, Self-supervised Learning, Explainable AI**

PUBLICATIONS

- *O. Dutta*, R. Gupta, and S. Agarwal. "Efficient LLM Pruning with Global Token-Dependency Awareness and Hardware-Adapted Inference." In Workshop on Efficient Systems for Foundation Models II@ **ICML** 2024. (link)
- *O. Dutta*, T. Kanvar, and S. Agarwal. "Search-Time Efficient Device Constraints-Aware Neural Architecture Search." In **PReMI**, 2023, Cham: Springer Nature Switzerland. (link)
- *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." In **WACV**, 2021. (link)
- *O. Dutta*, "Tempo Octave Correction Using Multiclass Support Vector Machine." In ICICCT, **IEEE**, 2018. (link)

Under Review:

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

EXPERIENCE

- **Research Assistant**, **IIT Delhi**, (Oct 2019- 2024)
  - Delivered deployable AI solutions while collaborating on projects with PolicyBazar, 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 aattended 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 (GMPSP) control guidance algorithm for moon lander navigation into optimized C code.
  - Simulated the algorithm in MATLAB to analyze 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

-