Oshin Dutta

PhD scholar
Indian Institute of Technology, Delhi.
oshin.dutta@ee.iitd.ac.in

SUMMARY

Passionate and driven researcher specializing in model compression and optimization for power, space-constrained devices. Proven expertise in designing, implementing, and testing compressed AI models on both classification and generative tasks. Adept at applied research and development, with a track record of publishing high-impact research and delivering practical solutions for industrial applications.

EDUCATION

Ph.D. in Compressing AI Models
Indian Institute of Technology Delhi (2019 – present, Synopsis completed))

- Area: Model compression, Generative AI, Computer Vision, NLP
- Thesis: Optimizing Deep Learning Models for Resource Constrained Environments supervised by Sumeet Agarwal and Prathosh A.P.
- Relevant Coursework: Machine Learning, Deep Learning and Generative AI, Compressed Sensing
- Key Achievements: Developed novel compression algorithms for popular AI models like LLMs; tested models on a wide range of classification and generative tasks.

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

PUBLICATIONS

- **Dutta, O**., Gupta, R., & Agarwal, S. (2024). Efficient LLM Pruning with Global Token-Dependency Awareness and Hardware-Adapted Inference. In Workshop on Efficient Systems for Foundation Models II@ **ICML** 2024. (link)
- **Dutta, O.,** Kanvar, T., & Agarwal, S. (2023, December). Search-Time Efficient Device Constraints-Aware Neural Architecture Search. In International Conference on Pattern Recognition and Machine Intelligence (**PReMI**) (pp. 38-48). Cham: Springer Nature Switzerland. (link)
- **Dutta, O**., Srivastava, A., AP, P., Agarwal, S., & Gupta, J. (2020). A Variational Information Bottleneck Based Method to Compress Sequential Networks for Human Action Recognition. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (**WACV**). (link)
- **Dutta, O**. (2018, April). Tempo Octave Correction Using Multiclass Support Vector Machine. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1333-1337). **IEEE**. (link)

Under Review:

- **Dutta, O**., Gupta, R., & Agarwal, S. (2024). VTrans: Accelerating Transformer Compression with Variational Information Bottleneck based Pruning. arXiv preprint arXiv:2406.05276. (link)

SKILLS

Programing Languages: Python, C, Java, MATLAB
Frameworks and Libraries: PyTorch, TensorFlow, OpenCV
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

PROJECTS

- **Neural Network Model Compression,** IIT Delhi (link) (Sept 2019-present)
  - Developing algorithms to obtain memory and computation efficient models for deployment to edge applications. Pushing the boundaries of Neural Architecture Search for resource constrained devices.
  -Framework used- PyTorch
- **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)
  -Coding and simulating an advanced guidance algorithm on a TMS320C6748 digital signal processor to achieve soft lunar landing
  -Software used: MATLAB, CC studio
- **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)
  -Software used: MATLAB

EXPERIENCE

- **Research Assistant**, **Indian Institute of Technology, Delhi**, Oct 2019- 2024
  -Compression of popular deep neural networks like transformer-based LLMs, Diffusion Models
  -Attended International Conferences and workshops of ICML, ACML, PReMI, Google Research Week
- **Intern, Aerospace Dept., IISc Bangalore**, Feb 2015- May 2015
  -Simulation of an optimal control guidance algorithm on DSP processor TMS320C6748 and MATLAB.

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),Signal Processing (2014)
- **Reviewer** for various conferences-WACV, Women in Machine Learning (WiML), AISTATS, ICML, IJCAI

-