Nandan Kumar Jha

PhD Candidate (Center for Cybersecurity, NYU)

Last updated: February, 2025

1032-3 10th Floor, 370 Jay Street
Brooklyn, New York, 11201
+1-(929)-513-1083
nj2049@nyu.edu
nankj.com

Research Interests

- Science of Large Language Models (LLMs) and Efficient Model Architectures
- Information-Theoretic Principles for Efficient LLM Architecture Design
- Cryptographically Secure Privacy-Preserving Machine Learning (PPML)

Education

2020–Present (Expected graduation: August 2025) **Ph.D, New York University**, Brooklyn, NY, USA,
Electrical and Computer Engineering Department .
- GPA: 3.78/4
- Supervisor: Prof. Brandon Reagen
- Thesis: Architectural and Algorithmic Innovations for Efficient Private Inference in LLMs and CNNs

2017–2020 **M.Tech, Indian Institute of Technology Hyderabad**, India,
Computer Science and Engineering Department.
- GPA: 9.27/10
- Supervisor: Dr. Sparsh Mittal
- Thesis: Hardware-Aware Co-Optimization of Deep Convolutional Neural Networks (Slides)

2009–2013 **B.Tech, National Institute of Technology Surat**, India,
Electronics and Communication Engineering Department.
- GPA: 8.20/10
- Supervisor: Dr. Upena Dalal
- Thesis: Simulation and Analysis of Joint Source and Channel Coding for Video Transmission

Publications

Peer-reviewed Conferences

2024 DeepReShape: Redesigning Neural Networks for Efficient Private Inference
Transactions on Machine Learning Research (TMLR)
Nandan Kumar Jha, Brandon Reagen
arXiv

2023 Characterizing and Optimizing End-to-End Systems for Private Inference
Architectural Support for Programming Languages and Operating Systems (ASPLOS)
Karthik Garimella, Zahra Ghodsi, **Nandan Kumar Jha**, Siddharth Garg, Brandon Reagen
arXiv

2021 DeepReDuce: ReLU Reduction for Fast Private Inference
International Conference on Machine Learning (ICML), ***Spotlight presentation***
Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen
arXiv, Press release, **[100+ citations]**

2021 Circa: Stochastic ReLUs for Private Deep Learning
Neural Information Processing Systems (NeurIPS)
Zahra Ghodsi, **Nandan Kumar Jha**, Brandon Reagen, Siddharth Garg
arXiv

2020 ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)
Rajat Saini*, **Nandan Kumar Jha***, Bedanta Das, Sparsh Mittal, C Krishna Mohan
arXiv (*Equal contributions.), **[100+ citations]**

2020 DRACO: Co-Optimizing Hardware Utilization and Performance of DNNs on Systolic Accelerator
IEEE Computer Society Annual Symposium on VLSI (ISVLSI)
Nandan Kumar Jha, Shreyas Ravishankar, Sparsh Mittal, Arvind Kaushik, Dipan Mandal, Mahesh Chandra
arXiv

2020 E2GC: Energy-efficient Group Convolution in Deep Neural Networks
International Conference on VLSI Design (VLSID)
Nandan Kumar Jha*, Rajat Saini*, Subhrajit Nag, Sparsh Mittal
arXiv (*Equal contributions.)

2019 Data-type Aware Arithmetic Intensity for Deep Neural Networks
IEEE International Conference on Computer Design (ICCD), (accepted as work in progress)
Nandan Kumar Jha, Sparsh Mittal, Sasikanth Avancha
Link

2019 The Ramifications of Making Deep Neural Networks Compact
International Conference on VLSI Design (VLSID)
Nandan Kumar Jha, Sparsh Mittal, Govardhan Mattela
arXiv

Peer-reviewed Journals

2020 Modeling Data Reuse in Deep Neural Networks by Taking Data-Types into Cognizance
IEEE Transactions on Computers (TC)
Nandan Kumar Jha, Sparsh Mittal
arXiv

2019 DeepPeep: Exploiting Design Ramifications to Decipher the Architecture of Compact DNNs
ACM Journal on Emerging Technologies in Computing Systems (JETC)
Nandan Kumar Jha, Sparsh Mittal, Binod Kumar, Govardhan Mattela
arXiv

Workshop Papers

2025 Entropy-Guided Attention for Private LLMs
The 6th Workshop on Privacy-Preserving Artificial Intelligence (PPAI@AAAI)
Nandan Kumar Jha, Brandon Reagen
arXiv, code

2024 ReLU's Revival: On the Entropic Overload in Normalization-Free Large Language Models
The 2nd Workshop on Attributing Model Behavior at Scale (ATTRIB@NeurIPS)
Nandan Kumar Jha, Brandon Reagen
arXiv, code

2021 Sisyphus: A Cautionary Tale of Using Low-Degree Polynomial Activations in Privacy-Preserving Deep Learning
Privacy Preserving Machine Learning Workshop (PPML@ACM CCS)
Karthik Garimella, **Nandan Kumar Jha**, Brandon Reagen
arXiv

Preprints

2024 AERO: Softmax-Only LLMs for Efficient Private Inference
Nandan Kumar Jha, Brandon Reagen
arXiv

2021 CryptoNite: Revealing the Pitfalls of End-to-End Private Inference at Scale
Karthik Garimella, **Nandan Kumar Jha**, Zahra Ghodsi, Siddharth Garg, Brandon Reagen
arXiv

2020 On the Demystification of Knowledge Distillation: A Residual Network Perspective
Nandan Kumar Jha*, Rajat Saini*, Subhrajit Nag, Sparsh Mittal
arXiv (*Equal contributions.)

Work Experience

2015–2017 **Seagate Technology HDD (India) Private Limited**, Bangalore, India.
- Designation: *Electrical Design Engineer*
- Job role: Design and verification of Solid State Drives (SSDs); Electrical characterization of DRAM and NAND; Signal integrity verification of NAND and DRAM datapath

2014–2015 **Indian Institute of Technology Bombay**, India.
- Designation: *Project Research Assistant*
- Job role: Unused licensed band in UHF used for wireless broadband in rural areas; LTE Wi-Fi dual connectivity using OFDM

Technical Skills

Proficient Python, PyTorch, Hugging Face Transformers, Scikit-learn, Git, Docker, Distributed ML, LaTeX

Used before Keras, TensorFlow, Caffe, OpenCV, Pandas, Verilog, VHDL, MATLAB, Synopsys EDA Tools

Awards

2021-2022 **Ernst Weber PhD Fellowship**
New York University

2019 **Certificate of Appreciation in Research**
Indian Institute of Technology Hyderabad

Reviewing

Conferences NeurIPS (2023 & 2024), ICLR (2024 & 2025), ICML (2024 & 2025), CVPR (2024 & 2025), ICCV 2025, AISTATS 2025, AAAI 2025

Journals TMLR 2025, TIFS 2025, JETC 2020

Outreaches

2024 **Guest Instructor, K12 Machine Learning Summer School**
New York University
- Taught the fundamentals of Large Language Models (LLMs) and guided a cohort of 30 students in using Hugging Face LLM libraries, focusing on practical implementations.

2023 **Lead Instructor and Mentor, K12 Machine Learning Summer School**
New York University
- Spearheaded three separate cohorts of K12 students, each through a two-week curriculum focusing on machine learning fundamentals, practical implementations, and hands-on projects.
- Facilitated interactive learning experiences, mentored students on their projects, and inspired a keen interest in Machine Learning domains.

2019 **Mentor, Artificial Intelligence and Emerging Technologies Summer School**
Indian Institute of Technology Hyderabad, India
- Mentored two student groups at AIET Summer School, IIT Hyderabad, steering capstone projects from conception to completion.
- Facilitated hands-on learning in machine learning, guiding projects on a food recommendation system and classification strategies for imbalanced datasets with probabilistic models.

Relevant Courses

AI-ML Track
- Introduction to Brain and Neuroscience
- Introduction to Deep Learning Systems
- Machine Learning for Cyber Security
- Foundations of Deep Learning
- Applied Machine Learning
- Video Content Analysis
- Visual Computing
- Deep Learning

SystemTrack
- Parallel and Customized Computer Architecture
- Hardware Architecture for Deep Learning
- Programming GPUs & Accelerators
- Advanced Computer Architecture
- Advanced Hardware Design
- Digital IC Design