NANDAN KUMAR JHA

8018 6th ave FL2, Brooklyn, New York, 11209
+1 (929)-513-1083 ◇ nj2049@nyu.edu ◇ www.nankj.com

RESEARCH INTEREST

Privacy-preserving machine learning; Applied homomorphic encryption; Multi-party computation; Hardware accelerator design for privacy-enhancing technology; Deep Learning; DNN Security

EDUCATION

New York University Sept 2020 - present
Doctor of Philosophy (Electrical and Computer Engineering)
Advisor: Prof. Brandon Reagen
Research: *Privacy-preserving computation; System design for privacy-enhancing technology; Applied homomorphic encryption, Secret sharing, Garbled circuit.*

Indian Institute of Technology Hyderabad (IIT-H) Aug 2017 - June 2020
Master of Technology - RA (Computer Science and Engineering) GPA: 9.27/10
Advisor: Dr. Sparsh Mittal
Research: *Computational intensity aware dataflow design in DNN accelerators (inference); Energy-efficient group convolution to balance the computational complexity with the degree of data reuse in DNNs; Devising metric to incorporate the dynamics of data reuse across a wide range of DNNs; Designing secure and robust DNN models; Multi-scale and multi-frequency feature learning in DNNs.*

National Institute of Technology Surat (SVNIT) July 2009 - May 2013
Bachelor of Technology (Electronics and Communication Engineering) GPA: 8.20/10
Undergraduate Thesis Advisor: Dr. (Mrs) Upena. D. Dalal
Undergraduate Research: *Simulation and analysis of joint source and channel coding for video transmission (H.264); Implementation of decoder heavy model using Distributed Video Coding (DVC).*

PUBLICATIONS

Peer-reviewed conferences

1. Zahra Ghodsi, Nandan Kumar Jha, Brandon Reagen, Siddharth Garg "Circa: Stochastic ReLUs for Private Deep Learning", 2021 Thirty-fifth Conference on Neural Information Processing Systems (**NeurIPS**) (*acceptance rate* 26%)

2. Karthik Garimella, Nandan Kumar Jha, Brandon Reagen "Sisyphus: A Cautionary Tale of Using Low-Degree Polynomial Activations in Privacy-Preserving Deep Learning", 2021 ACM CCS 4th Workshop on Privacy-preserving Machine Learning (**PPML**)

3. Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen "DeepReDuce: ReLU Reduction for Fast Private Inference", 2021 Thirty-eighth International Conference on Machine Learning (**ICML**) (*acceptance rate* 21.5%)

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

5. Nandan Kumar Jha*, Rajat Saini*, Subhrajit Nag, Sparsh Mittal, "E2GC: Energy-efficient Group Convolution in Deep Neural Networks", 2020 33rd International Conference on VLSI Design and 2020 19th International Conference on Embedded Systems (**VLSID**) (*acceptance rate* 12.8%)

6. Rajat Saini*, Nandan Kumar Jha*, Bedanta Das, Sparsh Mittal, C Krishna Mohan, "ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks", 2020 IEEE Winter Conference on Applications of Computer Vision (**WACV**)
7. Nandan Kumar Jha, Sparsh Mittal, Govardhan Mattela, "The Ramifications of Making Deep Neural Networks Compact", 2019 32nd International Conference on VLSI Design and 2019 18th International Conference on Embedded Systems (**VLSID**)
8. Nandan Kumar Jha, Sparsh Mittal, Sasikanth Avancha, "Data-type Aware Arithmetic Intensity for Deep Neural Networks", 2019 37th IEEE International Conference on Computer Design (**ICCD**) (*accepted as work in progress*)

Peer-reviewed journals

1. Nandan Kumar Jha, Sparsh Mittal, "Modeling Data Reuse in Deep Neural Networks by Taking Data-types into Cognizance", 2020, Special Issue on Machine-Learning Architectures and Accelerators, IEEE Transactions on Computers (**TC**) (IF 2.71)
2. Nandan Kumar Jha, Sparsh Mittal, Binod Kumar, Govardhan Mattela, "DeepPeep: Exploiting Design Ramifications to Decipher the Architecture of Compact DNNs", 2020, ACM Journal on Emerging Technologies in Computing Systems (**JETC**) (IF 1.65)

Arxiv preprint

1. Karthik Garimella, Nandan Kumar Jha, Zahra Ghodsi, Siddharth Garg, Brandon Reagen "CryptoNite: Revealing the Pitfalls of End-to-End Private Inference at Scaling", ArXiv 2021
2. Nandan Kumar Jha*, Rajat Saini*, Sparsh Mittal, "On the Demystification of Knowledge Distillation: A Residual Network Perspective", 2020

SELECTED RESEARCH PROJECTS

1. Energy-efficient DNN acceleration for self-driving cars (SRC funded)
M.Tech Thesis (IIT-Hyderabad); Advisor: Dr. Sparsh Mittal *May 2019 - present*

To optimize the energy efficiency of Mask R-CNN, group convolution with constant group size, which balances the computational complexity and arithmetic intensity of each layer in ResNet-101, has been employed; Further, the sparsity introduced by group convolution in ResNet-101 has been exploited to improve the predictive performance of Mask R-CNN; At present, to optimize the PE utilization and energy efficiency on Eyeriss (simulator), data-reuse aware dataflow design is in progress.

2. Lightweight subspace attention mechanism for compact DNNs
Research Project; Advisor: Dr. C Krishna Mohan *Jan 2019 - Sept 2019*

A lightweight and novel (subspace) attention mechanism has been devised to gather and distribute features in different feature subspace; Learning separate attention map for each feature map subspace enabled: (1) multi-scale and multi-frequency feature representation which is more desirable for fine-grained image classification tasks, (2) parameter-efficient gathering and distribution of features, and also reduces channel as well as spatial redundancy in DNNs; ImageNet-1K, Caltech birds, Stanford dogs, and Food-101 datasets have been used for experiments.

3. Deciphering the architecture of compact DNNs through side-channel attacks on GPU
Research project; Advisor: Dr. Sparsh Mittal *Jan 2019 - July 2019*

Devised two-stage attack methodology, termed "DeepPeep", to predict the architectural building blocks in compact DNNs; used the percentage of cuBLAS kernels `Gemmv2T`, `Gemmv2N`, and `Gemmk1` to predict the fine-grained architectural components such as residual connections, dense connections, branching, asymmetrical filter decomposition, etc; Proposed and implemented secure MobileNet-V1.

4. Real-time action recognition in videos using motion vectors

Visual Computing course project, Advisor: Dr. C Krishna Mohan *Jan 2018 - Apr 2018*

A variant of popular two-stream ConvNets for action recognition on UCF-101 and HMDB51 datasets has been deployed; To enable real-time extraction of motion information from videos, motion vectors, which encodes motion information between consecutive frames in video, has been used; motion vectors extracted the features with 656 FPS while traditional method (optical flow) with 18FPS.

PROFESSIONAL EXPERIENCE

Seagate Technology HDD (India) Private Limited Sept. 2015 - July 2017

Designation: Electrical Design Engineer

Job role: *Creation, development, and execution of Solid State Drives (SSDs); Performance measurement and optimization of DRAM; Electrical characterization of DRAM and NAND; Signal integrity verification of NAND and DRAM datapath; Design power delivery circuit for M.2 (SSDs).*

Indian Institute of Technology Bombay Nov 2014 - May 2015

Designation: Project Research Assistant

Research: *TV white Space: Unused licensed band in UHF used for wireless broadband in rural areas; LTE Wi-Fi dual connectivity using OFDM.*

AWARDS

Awarded **Certificate of Appreciation in Research** from IIT Hyderabad. March 2019

RELEVANT COURSES

AI Track

Deep Learning
Visual Computing
Video Content Analysis
Applied Machine Learning
Machine Learning for Cyber Security
Introduction to Brain & Neuroscience

System Track

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

SKILLS

Programming Skills	C, C++, Python, OpenCV, CUDA, Verilog
Tools & frameworks	PyTorch, Caffe, Eyeriss simulator, Synopsys EDA Tools, MATLAB