Souvik Kundu

Address: 3165 S. Sepulveda Blvd, Apt 202, Los Angeles, CA 90034, USA
Personal-web-home

Email : souvikku@usc.edu
Mobile : +1-213-431-9768

EDUCATION

- **University of Southern California** — Los Angeles, CA
  Doctor of Philosophy (Ph.D.) in Electrical Engineering; GPA: 3.98 /4.00 — *August 2017 – May 2022*
  Thesis Title: Algorithms and Frameworks for Deep Neural Network Models Addressing Energy-Efficiency, Robustness, and Privacy. **Advisors**: Dr. Peter A. Beerel and Dr. Massoud Pedram.

- **Indian Institute of Technology** — Kharagpur, India
  Master of Technology in Microelectronics and VLSI design; GPA: 9.44 /10.0 (Departmental Rank: 1) — *August 2013 – May 2015*

EXPERIENCE

- **Intel AI Labs** — San Diego/Remote, USA
  Research Scientist — *June 2022 - present*
  - **AI algorithm hardware co-design**: Work on various algorithm-hardware co-design aspect of machine learning for complex computer vision and NLP tasks for their efficient and robust deployment at the cloud/edge.
  - **Intel Labs** — San Diego, USA
    Deep Learning Research Intern — *June 2021 - Nov 2021*
    * **Neural Architecture Search**: **Application to resource constrained hardware**
  - **Intel Labs** — Santa Clara, USA
    Deep Learning Research Intern — *June 2020 - Dec 2020*
    * **Self-attention for vision**:
      1. As a part of the Intel AI Labs team I am closely involved in various distillation and computer-vision related projects and having the opportunity to work on SOTA CV models and enhance my skills on Pytorch, efficient model training.
      2. Worked towards development of various computer vision models for joint image upsampling tasks.
  - **University of Southern California** — Los Angeles, CA
    Ph.D. Graduate Student (Advisors: Prof. Peter A. Beerel and Prof. Massoud Pedram.) — *August 2017 - present*
    * **Machine Learning**: **Pre-defined sparsity based network model search**
      1. Currently involved in real time machine learning network model design for energy and storage constrained application to make the model deploy able fully or partially in embedded systems or edge devices for real-time training and inference. We have also **developed** a more **efficient** form of **compressed sparse representation** scheme to represent our notion of sparsity. This representation leverages the advantage of lower transfer of data from high cost DRAM to processing elements (PEs) where the multiply-accumulation operation takes place. Thus it ensures to minimize the cost of data transfer in our proposed sparse representation for domain specific ASIC based application.
    * **Machine Learning: Training framework in beyond CMOS technology**:
      1. Pre-defined sparse neuromorphic ex-situ training framework for Memristive accelerators for MLPs.
    * **Machine Learning: Novel network pruning driven by optimization**:
      1. Currently working on an energy efficient model pruning framework design driven by pruning and quantization. Future scope involves, architecture searches for efficient training and inference.
    * **Machine Learning: Analysis of model privacy and robustness**:
      1. Developed a novel training scheme to yield robust yet compressed DNN models via a single training iteration.
      2. Analyze improve the model and data-privacy in various training framework including distillation.
    * **Algorithm + architecture: Energy-efficient Event-driven Inference**:
      1. Developed novel neuro-inspired learning algorithms with reduced latency which can yield orders of magnitude improvement in energy-efficiency for image classification tasks compared to traditional deep neural networks.
      2. Working on a novel architecture that can process a set of input spikes using an output-stationary dataflow model, along with a hybrid on-chip memory to accelerate the processing of Spiking Neural Networks (SNNs).
    * **Machine Learning: In-pixel computing for efficient object detection and video tracking**:
      1. Has recently received a 1 M dollar grant from DARPA for exploration of algorithm-hardware co-design of in-pixel computation based accelerator for object detection and tracking.
  - **Texas Instruments** — Bangalore, India
    Digital Design Engineer — *June 2016 - July 2017*
    * **System Verilog, UVM and MATLAB**: As a part of high speed converter group I was responsible for designing an automatic gain controller (AGC) block in Verilog. This block was responsible for controlling the gain of the low noise amplifier (LNA) which captures the input analog signal, to avoid signal amplitude saturation for the filter blocks which is after LNA in the datapath.
  - **Synopsys** — Bangalore, India
    R and D Engineer II — *June 2015 - May 2016*
    * **VCS and Assertion**: Developed skills in VCS and System Verilog assertion.

SERVICES

- **Summer Mentoring** : Mentored and guided students from IIT Gandhinagar through Viterbi IUSSTF Summer intern program in 2018, 2020 (remote).
- **Reviewing and membership services**:
  Journals: More than 20 journals including IEEE Transactions on Circuits and Systems I and II, Computers, Neural Networks and Learning Systems, CAD, MICRO, MDPI. **Conferences:** More than 80 including ISCAS 2020, DAC 2020, BMVC 2020, 21, EMNLP 2020 [outstanding reviewer], WACV 2021, 22, NeurIPS 2021, ICLR 2022, ACL 2022, ICASSP 2022.

SELECTED 1^{st} AUTHOR PUBLICATIONS [GOOGLE SCHOLAR: SOUVIKKUNDU]

- [**DATE 2022**] **S. Kundu**, S. Wang, Q. Sun, P. A. Beerel, M. Pedram, "BMPQ: Bit-Gradient Sensitivity Driven Mixed-Precision Quantization of DNNs from Scratch".
- [**ACM TECS**] **S. Kundu**, Y. Fu, B. Ye, P. A. Beerel, M. Pedram, "Towards Adversary Aware Non-Iterative Model Pruning Through Dynamic Network Rewiring of DNNs", 2022.
- [**NeurIPS 2021**] **S. Kundu**, Q. Sun, Y. Fu, M. Pedram, P. A. Beerel, "Analysing the Confidentiality of Undistillable Teachers in Knowledge Distillation". (AR: 26%, HI: 245) [paper]
- [**ICCV 2021**] **S. Kundu**, M. Pedram, P. A. Beerel, "HIRE-SNN: Harnessing the Inherent Robustness of Energy-Efficient Deep Spiking Neural Networks by Training with Crafted Input Noise". (AR: 25.9%, HI: 184)[paper]
- [**ICASSP 2021**] **S. Kundu**, S. Sundaresan, "AttentionLite: Towards Efficient Self-Attention Models for Vision". (HI: 96) [paper]
- [**WACV 2021**] **S. Kundu**, G. Datta, M. Pedram, P. A. Beerel, "Spike-Thrift: Towards Energy-Efficient Deep Spiking Neural Networks by Limiting Spiking Activity via Attention-Guided Compression".(HI: 62) [paper]
- [**ASP-DAC 2021**] **S. Kundu**, M. Nazemi, P. A. Beerel, M. Pedram, "DNR: A Tunable Robust Pruning Framework Through Dynamic Network Rewiring of DNNs". [paper]
- [**IEEE Trans. on Comp.**] **S. Kundu**, M. Nazemi, M. Pedram, K. M. Chugg, P. A. Beerel, "Pre-defined Sparsity for Low-Complexity Convolutional Neural Networks", July 2020. (AR: $< 20\%$) [paper]
- [**Allerton Conf. 2019**] **S. Kundu***, S. Prakash*, H. Akrami, P. A. Beerel, K. M. Chugg, "pSConv: A Pre-defined Sparse Kernel Based Convolution for Deep CNNs". [paper]

  (*= equal contribution, AR= acceptance rate, HI= h5 index)

Featured Research and Patents

- [**Feature**] Achieving Greater Parameter and Computational Efficiency, by Intel Labs, USA, 2021.
- [**Feature**] USC ECE at NeurIPS and ICCV: Q&A With Souvik Kundu, by ECE, University of Southern California, USA, 2021.
- [**US Patent App.**] **S. Kundu**, M. Pedram, P. A. Beerel, "Efficient training of Low-latency and Robust Spiking Neural Networks", filed in July, 2021.
- [**US Patent App.**] S. Sundaresan, **S. Kundu**, "Methods of Pruning and Distillation for Compute-Efficient Structured and Unstructured Pruned Large-Scale DNN Models", filed in August, 2021.
- [**US Patent App.**] D. Cummings, **S. Kundu**, et al. "Communication Efficient Search for NAS Model Swap", filed in August, 2021.

PROJECTS

- **Deep Learning Mini Prject**: A bag of feature based convolution with feature of interpretability (won **best Research project** in EE599 course consisting of nearly 100 students). (April 2019 - May 2019)
- **VLSI Design Mini Project**: Design of a General Purpose 5-stage Pipelined Microprocessor with Software and Hardware Components in Cadence. (Nov 2017 - Dec 2017)
- **DFT Mini Project**: Implementation of ATPG and Fault Simulator for combinatorial circuits. (Nov 2017 - Dec 2017)

ACADEMIC ACHIEVEMENTS

- **USC Order De Arete Award**: For being best overall graduate student from USC.
- **USC ECE Ph.D. Achievement Award**: Received this award from the Ming Hsieh Dept. of Electrical and Computer Engineering, and also the nomination for Ph.D. achievement award from the graduate school.
- **USC MHI Scholar Finalist, 2020-21**: Was one of the 11 finalists out of all the Ph.D. students in the department of Electrical and Computer Engineering.
- **Best poster award**: Won the best poster out of 136 posters in the USC MHI research festival event, held on 11-08-2019. The poster topic was: Toward low complexity CNN models for both training and inference.
- **USC Annenberg Fellowship**: Top few incoming Ph.D. students at USC.

SKILLS & ACTIVITIES

- **Programming Languages**: Python (Experienced with API, viz. pyTorch, Keras on Tensorflow), C, C++; worked on online cloud platform, viz. AWS;
  HDL related: Verilog, System-verilog, UVM, DFT, Digital Design.
  Language: English, Hindi, Bengali.

Last updated: June 8, 2022.