1050 Lenox Park Blvd NE, APT 7408
Atlanta, GA 30319, USA
☎ *+1 (832) 571 3421*
✉ *ankush@gatech.edu*
✉ *ankushmandal19@gmail.com*
Linkedin Profile
Homepage

Ankush Mandal

Research Interests

Parallel Computing, Parallel Randomized Algorithms for Big-Data, Compiler Optimizations, Performance Optimization of Approximate Algorithms on Modern Architectures (e.g. Multi-core, Many-core, SIMD, GPU processors), High Performance Libraries for Machine Learning Kernels

Education

2017–2020 **Doctor of Philosophy**, *Georgia Institute of Technology*, Atlanta, GA, USA, (August 2020).
Advisor: Vivek Sarkar, Habanero Extreme Scale Software Research Laboratory, Georgia Institute of Technology
Co-advisor: Anshumali Shrivastava, RUSHLab, Rice University
Major: Computer Science
PhD thesis
Title: Enabling Parallelism and Optimizations in Data Mining Algorithms for Power-law Data
Committee: Vivek Sarkar (Chair), Anshumali Shrivastava, Hyesoon Kim, Santosh Pande, Richard Vuduc

2014–2017 **Master of Science**, *Rice University*, Houston, TX, USA.
Major: Computer Science.
Master's thesis
Title: Optimizing Convolutions in State-of-the-art Convolutional Neural Networks on Intel Xeon Phi
Committee: Vivek Sarkar (Chair), Rajkishore Barik, Keith D. Cooper, Anshumali Shrivastava

2008–2012 **Bachelor of Engineering**, *Jadavpur University*, Kolkata, India.
Major: Electronics and Telecommunication Engineering.

Work Experience

Aug, 2017 - Present **Research Assistant**, *Georgia Institute of Technology*, Atlanta, GA, USA.
Research projects:

- **Optimizing Word2Vec (word embedding method) on latest x86 CPUs with wide SIMD units**
  - Developed insights into performance issues in Stochastic Gradient Descent (SGD) inside Word2Vec
  - Our solution involved both compiler optimizations (static multi-version code generation with novel vector register blocking scheme) and algorithmic modifications (reduced computation and improved data locality).
  - Achieved 9.5× speedup on SGD and 2.5× speedup on training time over state-of-the-art methods with AVX-512 ISA on Intel® Xeon® Platinum 8280 CPU (Cascade Lake architecture)
- **Improving concurrency in approximate frequency estimation methods on modern GPUs**
  - Proposed new nested sketching strategy suitable for GPU-scale parallelism
  - Our approach exploits power-law behavior in data to reduce contention in atomic updates, gave detailed theoretical analysis
  - Attained throughput improvement of 32× over competing GPU-based method on nVidia® Tesla® V100 GPU and 272× over state-of-the-art sequential CPU-based method on Intel® Xeon® Platinum 8180 CPU (Skylake architecture)
- **Approximate K most frequent elements finding on massively parallel distributed + shared memory systems**
  - Tackled important data mining problem of finding TopK frequent items in distributed data streams
  - Combined sketch-based and counter-based approaches in unique way to aid parallelization while retaining fast update time
  - Implemented using MPI for multi-node parallelism and OpenMP for multi-core parallelism
  - Demonstrated 2.5× speedup over competing methods on clusters of Intel® Westmere and IBM Power®7 CPUs

Technology C, C++, OpenMP, MPI, CUDA

May, 2018 - July, 2018 **Intern for R&D of Energy and Performance Analysis**, *Intel*, Austin, TX, USA.

Mentor: David Kuck

- Performed energy and performance analysis of convolutions in popular Convolutional Neural Networks on x86 CPUs (Broadwell and Skylake architectures)
- Came up with performance-energy trade-off variation for direct convolution kernel when applying different compiler optimizations

Technology C, OpenMP

Aug, 2014 - Aug, 2017 **Research Assistant**, *Rice University*, Houston, TX, USA.

Research projects:

- Focused on improving performance of parallel machine learning algorithms
- Studied locality-sensitive hashing and heavy hitter detection on different architectures
- Worked on parallelizing forward-backward algorithm over profile-Hidden Markov Model in bioinformatics application.

Technology C, C++, OpenMP, CUDA

Jan, 2017 - May, 2017 **Graduate Intern**, *Intel Labs*, Santa Clara, CA, USA.

Mentor: Rajkishore Barik

- Optimized direct convolution kernel for convolutions in popular Convolutional Neural Networks on x86 CPUs targeting High-Performance Computing, specifically Intel Xeon Phi Knights Landing CPU.
- Contributed to open source LIBXSMM library
- Achieved ninja performance via JIT-based runtime code specialization and compiler optimizations
- Showed orders of magnitude performance improvement compared to popular matrix-multiplication (GEMM) based approach employing Intel® MKL

Technology C, OpenMP

June, 2016 - Aug, 2016 **Intern**, *AMD*, Austin, TX, USA.

Mentor: Mayank Daga

- Worked on in-house auto-tuning GEMM framework and Caffe (popular Deep Learning framework)
- Focused on analyzing and improving performance of auto-tuning GEMM framework for Caffe related problems on GPU architecture.
- Showed 5× performance improvement over ViennaCL for forward pass on convolution layers of AlexNet.

Technology C++, OpenCL

Aug, 2015 - May, 2016 **Teaching Assistant**, *Rice University*, Houston, TX, USA.

Courses:

- "Introduction to Computer Systems" COMP 321 - focuses on underlying aspects of computer systems
- "Parallel Computing" COMP 422 - introduction to foundations of parallel computing including the principles of parallel algorithm design, programming models for shared- and distributed-memory systems, parallel computer architectures

Technology C, C++, Cilk, OpenMP, Pthread, MPI, CUDA

Publications (selected)

Ankush Mandal, Anshumali Shrivastava, and Vivek Sarkar. NinjaVec: Learning Word Embeddings with Word2Vec at Lightning Speed. (In submission).

Ankush Mandal, Anshumali Shrivastava, and Vivek Sarkar. Matryoshka: a Nested Sketching Strategy for GPU-scale Parallelism and Skewed Data. (In submission).

Ankush Mandal, He Jiang, Anshumali Shrivastava, and Vivek Sarkar. Topkapi: Parallel and Fast Sketches for Finding Top-K Frequent Elements. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10898–10908, 2018.

Ankush Mandal, Rajkishore Barik, and Vivek Sarkar. Using Dynamic Compilation to Achieve Ninja Performance for CNN Training on Many-Core Processors. In *European Conference on Parallel Processing (Euro-Par)*, pages 265–278. Springer, 2018.

Swagatam Das, Ankush Mandal, and Rohan Mukherjee. An Adaptive Differential Evolution Algorithm for Global Optimization in Dynamic Environments. *IEEE Transactions on Cybernetics*, 44(6):966–978, 2013.

Personal Information

Visa Status: F1