Sandeep Polisetty

25 High street, Unit 3, Greenfield MA-01301.
☎ 413-801-3855 ✉ spolisetty@umass.edu Linkedin Google Scholar

Education

University of Massachusetts Amherst Aug 2020 – Present
Ph.D., Computer Science *Amherst, MA*

University of Massachusetts Amherst Aug 2016 – May 2020
M.S, Computer Science *Amherst, MA*

Indian Institute of Technology Sep 2008 – May 2013
B.Tech, M.Tech, Naval Architecture *Kharagpur, India*

Goals

I aspire to build frameworks to support efficient and scalable machine learning utilizing both algorithm and hardware properties. During my PhD, I have worked accross the machine learning stack, from writing CUDA kernels, to higher level framewormks and compilers upto machine learning applications.

Selected Publications

GSplit: Scaling graph neural network training on large graphs via split-parallelism under submission
S Polisetty, J Liu, K Falus, YR Fung, SH Lim, H Guan, M Serafini

Accelerating graph sampling for graph machine learning using GPUs Eurosys
A Jangda, S Polisetty, A Guha, M Serafini *2021*

Graphmini: Accelerating graph pattern matching using auxiliary graphs PACT
J Liu, S Polisetty, H Guan, M Serafini *2023*

Work Experience

Intel May 2020 – August 2020
AI Frameworks Engineer Intern *Santa Clara, CA*

- Worked on extending pytorch to support intel hardware
- Benchmarked state of the art models and optimizations on Intel Hardware

Reservoir Labs (acquired by Qualcomm) May 2019 – August 2019
Research Engineer Intern *Remote*

- Extended TVM intermediate representation Relay to support dynamic operators (eg. embedding bag) and generate c-code for downstream polyhedral compiler.
- Wrote transformation passes such as shape propagation and affine loop identification to increase the scope for polyhedral optimizations which are restricted to affine loops

Research Experience

Scaling graph machine learning
advised by Prof. Marco Serafini and Prof. Hui Guan January 2021

- Designed a novel distributed sampling and training paradigm to eliminate redundancy across multiple mini batches in GNN training across multiple GPUs
- Designed abstractions to handle all inter-GPU communication, duplicates and indexes needed for sampling and training loops with minimal code change.
- Wrote efficient CUDA kernels inside DGL codebase to shuffle intermediate activation data between GPUs and construct indexes efficiently.Our GSplit approach allows for deeper GNN models and outperforms the state of the art benchmarks such as Quiver and DGL by 30%.

GPU sampling for GNN
advised by Prof. Marco Serafini November 2020

- Integrated a novel in-GPU transit node-centric sampler called Nextdoor across various state-of-the-art graph training pipelines using torch extension. (Published at Eurosys - 2021)

Large Scale Semantic Scholar Paper Recommendation
at John Hopkins Speech and Language Workshop June 2023 - Aug 2023

- Used temporal partitioning of the graph and UVA to train on the large graph containing over 500 GB of data on commodity 16GB V100.
- Decreased training time further by using edge selection to limit training edges. The resulting GAT model delivered competitive accuracy to other techniques such as node2vec and matrix factorization, decreasing training time to under a day from one week.

Technical Skills

Languages: C/ C++ , CUDA, python, java
Frameworks: tensorflow, pytorch, DGL, TVM, pybind

Services

- Reviewer SC-2024
- Artifact Evaluater OSDI-2024, ATC-2024
- Artifact Evaluater OSDI-2020