Lingfan Yu

Email lingfan.yu@nyu.edu
Homepage https://cs.nyu.edu/~lingfan/

Address 60 Fifth Ave Room 450
New York, NY, U.S.

Education

2016.9 - Present PhD Candidate in Computer Science, New York University, U.S.
Advised by Professor Jinyang Li, GPA: 4.0

2012.9 - 2016.6 B.S. in Computer Science and Technology, Nanjing University, China
GPA: 1st year: 4.62 / 5, rank 1 / 151
2nd year: 4.57 / 5, rank 4 / 144
3rd year: 4.68 / 5, rank 1 / 144

2015.9 - 2016.4 University Exchange Program, University of Waterloo, Canada

Publications

- **Stateful Large Language Model Serving with Pensieve**
  Lingfan Yu, Jinyang Li
  Under Review
- **Scalable Graph Neural Networks for Heterogeneous Graphs**
  Lingfan Yu, Jiajun Shen, Jinyang Li, Adam Lerer
  ArXiv preprint arXiv:2011.09679
- **Deep Graph Library: Towards Efficient and Scalable Deep Learning on Graphs**
  Minjie Wang, **Lingfan Yu**, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander Smola and Zheng Zhang
  ICLR Workshop on Representation Learning on Graphs and Manifolds, 2019
- **Low Latency RNN Inference with Cellular Batching**
  Pin Gao, **Lingfan Yu** (equal contributions), Yongwei Wu, Jinyang Li
  The 13th European Conference on Computer Systems (Eurosys 2018)
- **The Efficient Server Audit Problem, Deduplicated Re-execution, and the Web**
  Cheng Tan, **Lingfan Yu**, Joshua B. Leners, Michael Walfish
  The 26th ACM Symposium on Operating Systems Principles (SOSP 2017, Best Paper Award)

Research Projects

- **Pensieve**
  - A serving system for Transformer-based models when used in conversation scenarios
  - Memorizes intermediate results of conversation history to avoid redundant computation
  - Develops GPU kernels to support page-based memory management to reduce memory fragmentation

- **NARS**
  - An approach to extend graph feature smoothing techniques to heterogeneous graphs
  - Trains a classifier on neighbor-averaged features of randomly samples relation subgraphs
  - Achieves a new state-of-the-art performance on several node classification datasets
- **Deep Graph Library**
  - An open source project to build system for learning graph-structured data
  - Provides intuitive and expressive interfaces and comprehensive system optimizations to make learning on graphs easy and efficient
  - Compatible with popular tensor frameworks (PyTorch, MXNet, etc.)
- **BatchMaker**
  - A serving system for Recurrent Neural Network
  - Proposed Cellular Batching to achieve both low latency and high throughput
- **OROCHI**
  - An audit system to check if service provider faithfully executes program
  - Collects trace and logs hints during online execution, and applies techniques like SIMD-on-demand to efficiently verify by re-execution

Work Experience

- **Research Internship at Facebook AI Research, Facebook, Inc** **2020.5 - 2020.9**
  - Experimented the effectiveness of GNN and graph feature smoothing approaches
  - Developed NARS, an approach to extend feature smoothing to heterogeneous graphs
  - Achieved a new state-of-the-art performance on several heterogeneous benchmarks
- **Applied Scientist Internship at Amazon Web Services, Inc.** **2019.6 - 2019.8**
  - Analyzed and identified performance issue of DGL's sparse kernel framework
  - Investigated state-of-the-art solutions and designed algorithm to improve the framework
  - Matched speed of NVIDIA CuSPARSE while being more generic and customizable
- **Quantum Cube Corporation, Waterloo, Canada (part-time)** **2015.9 - 2015.12**
  - Company focus is commodity futures trading and developing research system for future trading analysis and automation
  - Designed and implemented backend system to provide support for the project

Teaching Experience

Spring 2019	Recitation Leader & Grader of Computer System Organization (CSCI-UA.0201)
Fall 2018	Recitation Leader & Grader of Computer System Organization (CSCI-UA.0201)
Fall 2017	Teaching Assistant of Distributed System (CSCI-GA.3033-002)

Selected Course Projects

- **Paxos-based Key/Value Store** **2016.9 - 2016.12**
  - Course project for Distributed System
  - Implemented a key/value store service on top of Paxos consensus protocol
  - Supported features like sharding, fault-tolerance and recovery
- **Simplified C Compiler** **2015.2 - 2015.6**
  - Course project for Compiler's Principle
  - Implemented building blocks of compilers: lexical analysis, syntax analysis, syntax-directed translation, intermediate code generation, and machine code generation
  - Implemented a compiler of simplified C language, which supports most C syntax except pointer
  - Applied code optimization algorithms to improve efficiency of results
- **Network Protocol Implementation** **2015.2 - 2015.6**
  - Course project for Computer Network
  - Implemented simplified version of TCP/IP protocol using C language to support data transmission between application layers
  - Implemented an efficient BitTorrent client for resource sharing
- **Optimization model for flat folding tables** **2014.9.12 - 9.15**
  - Problem from China's Undergraduate Mathematical Contest in Modeling (CUMCM)
  - Designed an optimization model to minimize cost of producing flat folding tables
  - Simulated dynamic folding process of folding tables
  - Won national first prize in contest
- **MIPS CPU Design** **2014.2 - 2014.6**
  - Course project for Computer System Organization
  - Implemented mono-cycle MIPS CPU and multi-cycle MIPS CPU on FPGA
  - Designed and implemented 32-bit MIPS pipelining CPU on FPGA
  - Implemented features like forwarding, 2-bit dynamic prediction, and exception handling to improve efficiency and robustness
  - The 32-bit pipelining CPU reached clock rate of 0.7 GHz
- **Nanos-based Operating System** **2014.2 - 2014.6**
  - Course project for Operating Systems
  - Implemented simplified operating system kernel that supports features like creation and switching of threads
  - Supported locking mechanism and message transferring among threads using message queue

Honors & Awards & Scholarships

- EuroSys 2018 Travel Grant, Apr 2018

- NYU Dean of the Graduate School of Arts and Science Travel Grant, Feb 2018
- SOSP 2017 Best Paper Award, Oct 2017
- SOSP 2017 Travel Grant, Oct 2017
- NYU Dean of the Graduate School of Arts and Science Travel Grant, Oct 2017
- Henry M. MacCracken Fellowship, New York University, Sept 2016
- Scholarship of No.14 Electronic Technology Institute of China for academic excellence, Nov 2015
- Liu Jimin Scholarship for Exchange Program at the University of Waterloo, Jun 2015
- Scholarship of Nanjing Fujitsu Software Technology Co., Ltd for academic excellence, Dec 2014
- Outstanding Student of Nanjing University, Dec 2014
- National First Prize of China Undergraduate Mathematical Contest in Modeling, Dec 2014
- Outstanding Student of Computer Science Department, Dec 2013
- National Scholarship of China for academic excellence, Nov 2013

Voluntary Experience

- **Volunteers Association of Computer Science Department** **2013.9 - 2014.6**
    - Served as Minister of Project Department of the Volunteers Association
    - Organized and participated in many voluntary events held by Volunteers Association
- **Benefaction 100 Love Package** **2012.9 - 2012.11**
    - Charity event held by China Foundation for Poverty Alleviation (CFPA)
    - Aimed at raising awareness of caring about children living in poor conditions
    - Served as volunteer for over 40 hours