Pengcheng Xu

+1(217)550-1337 | px6@illinois.edu | https://github.com/explcre | https://linkedin.com/in/pengcheng-xu-ryan
Personal Website: https://explcre.github.io/ | ChatGPT version of me on this website: https://explcre.github.io/mychat/

EDUCATION

University of Illinois Urbana-Champaign Aug 2022 - Dec 2023

Master's Degree in Computer Engineering, GPA: 3.8/4.0. Research Scientist at AI@UIUC.

Core Courses: Distributed Systems, Communication Network(A+), Computer Vision(A), Advanced Distributed Systems, Transfer Learning(A), Engineering Entrepreneurship(A+), Machine Learning(Ongoing), Applied Parallel Programming(Ongoing).

University of Michigan - Shanghai Jiao Tong University Joint Institute Sep 2018 - Aug 2022

Bachelor's Degree in Electrical and Computer Engineering. Minor in Data Science. **Outstanding Graduate (School Level).**

Core Courses: Data structures and Algorithms, Intro to Computer Organization (A, Teaching Assistant), Computational Methods for Statistics and Data Science (A+), Intro to Operating Systems, Intro to Data Science (A), Undergraduate Research (A+).

PUBLICATIONS

(Under Review) **Pengcheng Xu*,** Jinpu Cai*, Yulin Gao, Ziqi Rong, Hongyi Xin. "MIRACLE: Multi-task learning based Interpretable Regulation of Autoimmune diseases through Common Latent Epigenetics",2023. https://arxiv.org/abs/2306.13866

(Under Review) Tao Feng*, **Pengcheng Xu***, Tianfan Fu, Jimeng Sun. "Molecular de-novo design through Transformer-based Reinforcement Learning",2023. http://arxiv.org/abs/2310.05365

(Under Review) Ziqi Rong, Jinpu Cai, Jiahao Qiu, **Pengcheng Xu**, Lana Garmire, Qiuyu Lian, Hongyi Xin. "Balancing Information Preservation and Computational Efficiency: L2 Normalization and Geodesic Distance in Manifold Learning",2023.

(To be submitted) **Pengcheng Xu**, Kaiyang Chen, Yuanrui Zhang, Indranil Gupta. "Pipe-Déjàvu: Hardware-aware Latency Predictable, Differentiable Search for Faster Config and Convergence of Distributed ML Pipeline Parallelism",2023.

Zhikai Yang, **Pengcheng Xu**, Dekun Yang, Yufeng Chen, Yancong Ma. "Vascular Intervention Training System Based on Electromagnetic Tracking Technology", ICVRV, 2020. https://ieeexplore.ieee.org/document/9479727

RESEARCH EXPERIENCE

Molecular de-novo design through Decision Transformer and Oracle-feedback reinforcement learning May 2023-Present

Advisor: Tianfan Fu(Incoming Assistant Professor at Rensselaer Polytechnic Institute), Jimeng Sun(Professor at CS, UIUC)

- ➢ Implemented a decision transformer architecture to improve the AUC for over fifteen molecular optimization tasks for 5% each on average.
- ➢ Applied Oracle-feedback reinforcement learning on the downstream tasks to reach higher performance than pretrained model.
- ➢ Carried out ablation study and investigation into loss curve and conditional probability over the next token as a function of previously chosen ones according to the model.

Hardware-aware Latency Predictable, Differentiable Search for Faster Config and Convergence of Distributed ML Pipeline Parallelism Advisor: Indranil Gupta, Professor of CS UIUC | Advanced Distributed Systems| Researcher | Feb 2023 – May,2023

- ➢ Implemented a predictive model that considers communication cost, model computational cost, and hardware information to predict latency and resources of parallel configurations, saving time on pre-profiling before searching the parallel configuration.
- ➢ Proposed a differentiable parallel configuration search space inspired by DARTS, can potentially reach optimal configuration faster than the original dynamic programming.
- ➢ Employed parallel random initialization using sampling algorithms like Bayesian Optimization for faster train loss convergence.

Multi-task learning based interpretable gene-level methylation estimations |Research Assistant Sep 2021 - Present

Advisor: Hongyi Xin, Associate Professor of UM-SJTU Joint Institute, Shanghai Jiao Tong University

- ➢ Explored adaptable and interpretable neural network to find common genotype given 480k dimension sites, hundreds of sample.
- ➢ Designed an explainable site-gene-pathway ontology constraint to NN to discover new biomarkers by checking weights.
- ➢ Implemented a Variational Auto-Encoder to support gene-level embedding shared among datasets to obtain multi-task learning.
- ➢ Optimized a pretrain-finetune training scheme to increase accuracy by over 10%, wrote the paper under review in 2023.

Augmented reality simulation of cardiovascular interventional surgery | Research Assistant Mar 2020 - May 2021

Advisor: Lixu Gu, Professor of Biomedical Engineering, Shanghai Jiao Tong University

- ➢ Developed the framework of an augmented reality surgery training assistant system for medical student and surgery.
- ➢ Predicted the operation trajectory using LSTM and used KD-Tree to calculate the distance for operation safety warning.

- Displayed vascular model in AR with OpenGL and designed the UI interface to support translation.
- Used the aruco library in OpenCV to coordinate positioning of the QR code.
- Published *Vascular Intervention Training System Based on Electromagnetic Tracking Technology* on ICVRV as second author.

PROFESSIONAL EXPERIENCE

Amazon Web Services | VMware Cloud on AWS (Brio) | Software Engineer Intern | Seattle, Washington May 2023 – Aug 2023

- Designed UI to automate the workflow to update Rate Card for Brio resource console using React, TypeScript and JavaScript.
- Integrated UI with backend API using Java, API Gateway and Amazon Lambda. Supported validation and creating and updating rate cards. Preview visualizes modifications and identifies pricing errors. (5k+ lines of code in total)
- Designed anomaly detection algorithms for disbursement generator to identify usage spikes or subscription anomalies using Java, SQL, Amazon Athena, S3 Buckets, DynamoDB Table. Utilized CloudWatch to create tickets and alarms to engineers.

4Paradigm Co., Ltd | OpenMLDB | GitLink Code Camp 2022 | Open-Source Developer | Shanghai July 2022 – Oct 2022

- Developed an automated feature engineering pipeline, including feature generation and selection, based on AutoX and OpenMLDB sql with Python (600+ lines of code). See pull request: https://github.com/4paradigm/OpenMLDB/pull/2381.
- Saving data scientists in dozens of companies' weeks of time by automatic feature selection pipeline.
- The data is transformed by OpenMLDB sql to get time series and statistics features, and then we select the most important K features based on some algorithms like Adversarial Validation, GRN feature selection or Reinforcement Learning.

Intel Co., Ltd | DL Model Optimization Department | Deep Learning Software Engineer Intern | Shanghai Nov 2021 - June 2022

- Implemented new features to Intel® Neural Compressor, Cross Layer Equalization, a data free quantization to rescale different layers' weight range to reduce over 5% in the drop in accuracy after quantization like FP32 to INT8, and edited documents.
- Studied the open-source software like Triton Inference Server (around 70k lines of code) and AI Model Efficiency Tool and gave presentations to around 100 colleagues in whole department to introduce the design and technique detail of them.
- Cooperated to design and implement class of AI inference server software by C++, which enables the team to deploy trained AI models from multiple frameworks and deploy more models on GPU or CPU based infrastructure to simplify AI inferencing.

SHUKUN (Beijing) Technology Co., Ltd | R & D Department | Algorithm Intern | Shanghai Dec 2020 - Apr 2021

- Implemented multi-node and multi-GPU training with **horovod** framework, NVIDIA clara train sdk, OpenMPI, and NCCL2, 4GPU (2 nodes each 2GPU) can reach 2.5x speed up compared to 1GPU (Discovered that communication is bottleneck).
- Compared the efficiencies of models like 3D-UNet when multi-node training with different GPU configurations in Python and wrote the training process documents, saving algorithm scientists days in each training job(~0.65n times faster with n nodes).
- Used Java to add multi-machine and multi-GPU training functions to the company's back-end web page for algorithm scientists.

COURSE PROJECTS

CS 425 Distributed Systems: Distributed ML Inference System (C++ and Python) (100/100) Aug 2022 – Dec 2022

- Implemented scheduling algorithm to obtain fair-time inference, making each ML job query rate within 20% difference.
- Design a distributed file system to maintain datasets with SWIM-like membership protocol (nearly 1e-2 false positive rate).
- Supported fault-tolerance and can recover when N members and coordinator leave or fail. Leader re-election is within 10s.

CS 543 Computer Vision (A): (Python) Aug 2022 – Dec 2022

- Channel Alignment: Designed multi-scale algorithm to align 3 channels of images, evaluating with NCC and Fourier Transform.
- Laplacian Blob Detection: Implemented a scale-space blob detection with a Laplacian scale space using Laplacian of Gaussian filter and performed non-maximum suppression in scale space.
- Stitching: Stitching pairs of images using OpenCV, SIFT descriptors and RANSAC to estimate homography mapping of images.
- 3D Shape from shading: Estimate the albedo and surface normal and compute the surface height map by integration to reconstruct 3D surface from image of shading.

CS 438 Communication Network (A+): TCP Protocol (C++) Aug 2022 – Dec 2022

- Implemented transport protocol with properties equivalent to TCP based on unreliable UDP, which can tolerate packet drops, allow other concurrent connections fair chance (0.5x to 2x of TCP), doesn't give up entire bandwidth to other connections.

Sonar detection water pipe state system based on machine vision and neural network compression Sep 2021 – Dec 2021

ECE4710 Advanced Embedded Systems (MDE, Thesis). | Instructor: Zou An, Assistant Professor of UM-SJTU Joint Institute

- Processed the source data of sonar as a matrix and use Python to generate polar images in real time.
- Used OpenCV to perform subtraction and corrosion operations to complete hole or bulge detection.
- Designed a compressed and binarized YOLO architecture to realize defect area detection with accuracy 99%.

- Used HLS to deploy to FPGA, matrix parallel operation, real-time embedded system is implemented.

Honors & Awards

2022 Shanghai Jiao Tong University Outstanding Graduate (School Level)	Aug 2022
2021 Microsoft Imagine Cup Global Competition - Third Prize in China	Jan 2021
2020 Mathematical Contest in Modeling - Meritorious Winner (Top 6%)	Apr 2020
2020 “Jidong Cup” CCVR China Virtual Reality Competition, Product Creative Group - Second Prize	Nov 2020
2018-2019 and 2020-2021 Academic Year Undergraduate Excellence Scholarship	Nov 2019/2021

Skills

- **Programming:** C/C++(Proficient),Python (Proficient), MATLAB, R, Verilog, Java, RISC-V Assembly, SQL, Shell, TypeScript, CUDA Programming.
- **Frameworks & Libraries:** PyTorch, TensorFlow, horovod, Keras, Sk-learn, Pandas, NumPy, OpenCV, Matplotlib, Selenium.
- **Developer Tools:** Docker, Git, LaTeX, PSpice, Linux.