ZHUMAKHAN NAZIR

+77754428074 ⋄ zhumakhan.nazir@nu.edu.kz ⋄ github.com/zhumakhan

ACADEMIC BACKGROUND

M.Sc in Computer Science, Nazarbayev University (NU) *May 2024*
B.Sc in Computer Science, Nazarbayev University (NU) *May 2021*

- Interpretable ML enhanced Performance Analysis of cuBLAS, cuDNN and TensorRT [ACM SAC'23]
- A CNN Inference micro-benchmark for Performance Analysis and Optimization on GPU [IEE SMC'22]
- A Machine Learning Model Selection considering Tradeoffs between Accuracy and Interpretability [ASTES journal'22]
- Dean's list award for Spring semester (2022)
- 1 st place in V Yassawi Republic Olympiad in physics (2016)
- 3 rd place in regional olympiad physics (2015)

WORK EXPERIENCE/PROJECTS

ML Research Engineer intern at Ivy (unify.ai) Dec'22-current *London, UK (remote)*

- Implementing various machine learning/linear algebra functions. Developing, testing and optimizing graph compiler of ivy framework. Leading TensorRT backend of Ivy's framework.

Machine Learning intern at Verigram.ai, Dec'22-Feb'23 *Almaty, Kazakhstan*

- Applied model interpretability methods using torch's captum library to explain anomalies in the model's detection. Tested effect of brightness, noise level, background objects and crop methods to the accuracy. Cropping method improved accuracy by 10%. Quantized face recognition models to increase speed in CPU using Intel's OpenVino by 15-100% with drop in accuracy around 5%, could be improved further by quantizing over full dataset.

Software Engineer Intern at Meta (Facebook), Jul-Oct 2022 *London, UK*

- Improved the accuracy and stability of location finder algorithm using bluetooth signals by center of mass method. Added new functionalities to admin panel of office devices.

Teaching assistant for GPU programming class at NU, Jan-May 2022 *Astana, Kazakhstan*

- Graded assignments, hosted QA sessions, troubleshooted programming issues.

Research assistant at Embedded Systems Lab at NU, Sep'21-Jun'22 *Astana, Kazakhstan*

- Researched and attempted to optimize matrix multiplication and convolution operations for Nvidia GPU's using CUDA. Compared perforamance of GPU libraries like cuBLAS, cuDNN and TensorRT, and identified important profiling metrics. Published paper on ACM Symposium On Applied Computing (SAC) conference.

TECHNICAL STRENGTHS

- Linear Algebra, Probability, Graph Theory, Numerical Methods, Convex Optimization
- C/C++, Python, SQL, redis, Tibco EMS, kafka, prometheus, grafana, websockets, django, fastapi
- docker, git, CI/CD, jenkins, OpenMP, POSIX thread, CUDA programming, TensorRT, pytorch