David Anugraha

Email: david.anugraha@gmail.com
Website: davidanugraha.github.io

RESEARCH INTERESTS

My current research interests primarily lie in the areas of Language Models, Multilingual NLP, and Low-Resource NLP. Additionally, I am also interested in data-driven decision-making in other research areas, including Parallel and Distributed Databases and Drug Discovery using Machine Learning.

EDUCATION

B.Sc (Hons), University of Toronto, Toronto, Canada — June 2024
Computer Science Specialist, Statistics Major, Chemistry Minor — GPA: 3.97/4.0

AWARDS

Dean's List Scholarship (University of Toronto) — 2020 - 2024
University of Toronto Excellence Award (Statistical Sciences) *(declined)* — 2023
Later Life Learning Scholarship (University of Toronto) — 2020 - 2022

PUBLICATIONS

David Anugraha, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, En-Shiun Annie Lee. 2024
ProxyLM: Predicting Language Model Performance on Multilingual Tasks via Proxy Models
arXiv preprint arXiv:2406.0933

Eric Khiu, Hasti Toossi, **David Anugraha**, Jinyu Liu, Jiaxu Li, Juan Armando Parra Flores, Leandro Arcos Roman, A. Seza Dogruöz, En-Shiun Annie Lee. 2024
Predicting Machine Translation Performance on Low-Resource Languages: The Role of Domain Similarity
Findings of the Association for Computational Linguistics: EACL 2024

TALKS

Toronto Machine Learning Summit 2024
ProxyLM: Predicting Language Model Performance on Multilingual Tasks via Proxy Models

EXPERIENCE

Associate Research Engineer, Markham, Canada — June 2024 - Present
Distributed Data Storage and Management Lab at Huawei Canada

- Researching efficient distributed sorting and windowing algorithms for mix of batch and streaming execution.
- Explored data-driven cost query estimation models to optimize query execution.

Undergraduate Research Assistant, Toronto, Canada — August 2023 - Present
Advised by Annie En-Shiun Lee

- Led and managed the development of ProxyLM, a novel, scalable, and efficient method to predict language models performance on multilingual tasks using proxy models.
- Investigated domain similarity in machine translation performance for low-resource languages. This study was accepted as Findings at EACL 2024.

Undergraduate Research Volunteer, Toronto, Canada — August 2023 - Present
ParaMathics (Maryam Mehri Dehnavi's Group)

- Fine-tuned large language models (LLMs) and language models (LMs), including LLaMA-2 and BERT, using various compression techniques, and conducted data analysis on their performance against multiple benchmark evaluations.
- Developed sparse kernels in CUDA to implement sparsity in the weights of language models for more efficient pre-training and inference.

Assistant Research Engineer, Markham, Canada May 2022 – August 2023
Distributed Data Storage and Management Lab at Huawei Canada

- Contributed to the MindPandas project by developing 16 map, reduce, and window operators in both lazy batch and streaming mode, resulting in a 5x increase in performance compared to Pandas and receiving an outstanding team award.
- Conducted research on efficient shuffling algorithms for a potential patent in Huawei's next AI Analytics Engine.
- Maintained and handled 23 issues and requirements from headquarters, researching and implementing possible performance improvements on the MindData codebase.

PROJECTS

MindSpore (Open-source deep learning training/inference framework)

- Designed and implemented support for compressed TFRecord dataset in MindData pipeline, benefiting MindSpore users migrating from TensorFlow for better performance in MindSpore.
- Added documentation to 448 test files and reorganized 284 source files using multiple code check tools to prevent future errors in the CI/CD pipeline.

Drug Synergy Prediction

- Designed a preprocessing pipeline for feature extraction from drug and cell data to predict synergy scores for cancer treatment, using DrugComb as the benchmark.
- Implemented a graph-based deep neural network using Torch in Python, improving prediction accuracy by 2x compared to state-of-the-art benchmarks.

Solubility Prediction

- Developed a machine learning algorithm to estimate the solubility of compounds in water, a critical task in pharmaceutical chemistry on expediting drug discovery processes.
- Implemented deep neural networks, RandomForest, and XGBoost using TensorFlow in Python, achieving RMSE of 0.81, surpassing results from related papers such as SolTransNet in 2021 with RMSE of 1.141 and Graph Convolutional Neural Network in 2023 with RMSE of 0.86.

Personalized Education Algorithm

- Developed an algorithm focused on improving educational strategies and personalized learning that estimates the students' ability level.
- Designed machine learning models using KNN, Rasch model, and neural network in Python, achieving an accuracy rate of 72%.

SERVICES

Peer Mentor May 2020 – September 2021
Innis Mentorship (University of Toronto)

- Mentored and supported a diverse group of international first-year students during their transition to the University of Toronto.
- Provided guidance and assistance by offering accurate information on majors, program prerequisites, and other essential resources.
- Facilitated regular check-ins and maintained open lines of communication, resulting in positive feedback and improved student satisfaction.

International Committee Council September 2019 – September 2020
Innis Residence Council (University of Toronto)

- Successfully orchestrated 7 engaging international-themed events at Innis Residence, fostering cultural understanding and fostering a sense of belonging among international students.

- Presented event proposals to the council, leading to their approval and ensuring seamless coordination and execution of each event.
- Demonstrated strong event planning and organizational abilities while creating a welcoming and comfortable atmosphere that enhanced the overall student experience at the residence.

Math and Chemistry Tutor January 2020 – June 2020
The Saturday Program

- Achieved an average grade improvement of 10% among students by providing targeted math and chemistry tutoring and personalized support.
- Developed tailored lesson plans and study materials to address individual learning needs.

ADDITIONAL SKILLS

Programming Languages: Python, C, C++, Java, CUDA, SQL, Bash, Assembly.
Libraries and Frameworks: Torch, TensorFlow, pandas, NumPy, Spark.
Applications: Docker, Kubernetes, PostgreSQL, Vim, Git, SLURM, LaTeX.
Operating Systems: Unix, Linux, Mac OSX, Windows.