Benyamin Jami

NLP Research Associate at Huawei

226-899-0692 — benyamin.jami76@gmail.com — Benyamin Jami — benyaminjami — Benyamin Jami — Toronto, On, Canada

Summary

- 4+ years of experience in Machine Learning, AI Research, and Model Optimization Expertise in Generative Models, Transformer Models, Speculative Decoding, Hybrid Models with Mamba, and Knowledge Distillation.
- Proven track record in training and fine-tuning large-scale language models (up to 40B parameters) using 3D Parallelization and frameworks like Nanotron, Megatron, and DeepSpeed.
- Strong background in debugging complex systems, optimizing workflows, and implementing research-driven solutions.

Experience

Huawei Technologies Canada — May 2024 – Present

Research Associate — *Toronto, Canada*

- **Led research initiatives** to enhance transformer model efficiency for training and inference, focusing on cutting-edge methodologies, including **linear attention, hybrid models with Mamba, speculative decoding**, and **nested model architectures**.
- **Implemented and trained speculative techniques** like **Eagle and Medusa** on Huawei's **Pangu (40B parameter language model)** using Huawei's **NPU**s, and conducted comprehensive **benchmarking with VLLM** to evaluate performance gains.
- **Pretrained and fine-tuned large-scale language models (1B–40B parameters)** in multi-node environments, leveraging **3D parallelization** with Nanotron, Megatron, and DeepSpeed to optimize computational performance.
- Successfully transformed a fully transformer-based model into a **50% Mamba-layer hybrid model using knowledge distillation**, achieving equivalent performance while reducing computational complexity.
- **Mentored and guided two interns**, fostering their technical growth and ensuring the successful completion of key research projects.

University of Waterloo — Sep 2021 – Dec 2023

Research Assistant — *Waterloo, Canada*

- Implemented a transformer-based model to **translate antigens into antibody sequences**, leveraging a **semi-supervised framework and back translation** for training on **2M unsupervised samples**.
- **Enhanced model performance by 30%**, outperforming state-of-the-art sequential methods in antibody sequence prediction.
- Integrated a **graph neural network** with a **protein language model**, employing a **non-autoregressive training approach**, and achieved a **60% improvement in antibody affinity optimization**.
- Achieved a **7% improvement over state-of-the-art models in antibody sequence recovery**, demonstrating the efficacy of the integrated model.

General Motors Canada — Sep. 2022 – Apr. 2023

Data Scientist (Co-op) — *Toronto, Canada*

- **Designed and implemented a cutting-edge dashboard solution** for the Warranty Support Center (WSC), integrating data from disparate sources, optimizing reporting processes, and delivering a **15% reduction in manual errors** and a **20% faster claims processing time**.
- **Coordinated with teams across Canada and the U.S.** to align strategic objectives, streamline processes, and achieve a **50% reduction in project delays**.
- Leveraged **Greenplum, Oracle, and Hue databases** for data collection, organization, and analysis.

Digikala — Jul. 2020 – Aug. 2021

Machine Learning Engineer — *Tehran Iran*

- **Developed different features for the search engine**, such as related search using **likelihood optimization**, boosting **search queries per session by 8%**.
- **Enhanced search ranking** by developing a **Named Entity Recognition module** for the Learning to Rank model, using the **Hidden Markov Model**, increasing **conversion rate by 7%**.
- Leveraged **PySpark to process and analyze data for 10M users**, enabling the creation of a robust **Item Factorization algorithm** that revolutionized main page carousel recommendations for **personalized user experiences**.

Education

University of Waterloo — Sep 2021 – Dec 2023

M.Math. of Computer Science – Supervisors: Dr. Ali Ghodsi and Dr. Mohammad Kohandel — *Waterloo, Canada*

Thesis: Advancing Antibody Design: Integrating Protein Language Models for Enhanced Computational Strategies. (link)

Sharif University of Technology — Sep 2021 – Dec 2023

B.Sc. of Computer Engineering - Supervisor: Dr. Mahdi Jafari Siavoshani — Tehran, Iran

Thesis: Deep learning approach with Variational Autoencoder architecture in image transfer over noisy channels