Pooya Moradi

Email: mo.pooya@gmail.com Mobile: +1-604-728-1235

Experience

- **Microsoft** — Canada
  Applied Scientist - Bing Chat — *Nov 2022 - Present*
  - Working on **optimizing inference compute/memory** through techniques like sparse attention, speculative decoding, optimized CoT reasoning, and more.
  - **Decreased latency of product about 30%** by optimizing various avenues leading to **5% higher product usage**. Got acknowledged by senior leadership of Microsoft.
  - Augmented the LLM behind Bing chat with **advertisement generation capabilities**.
  - Iterated over various LLM prompts to make **Bing chat** accessible to millions of users in the wait list.
- **Microsoft** — Canada
  Applied Scientist — *Aug 2020 - Nov 2022*
  - Designed, trained, distilled and deployed various robust deep neural models for automatic Wikipedia creation from millions of enterprise text documents which **lowered product defect by 50%.**
  - Was in charge of **leading the project** that involved **working with 3 cross-org teams**, including converting high-level ambiguous business requirement into narrowed-down technical milestones.
  - Helped to optimize one of our pipelines to reduce processing time from multiple days to 10 hours and memory requirement from 1TB to 200GB by discovering and improving performance and memory bottlenecks.
- **Microsoft** — USA
  Applied Scientist Intern — *Mar 2020 - May 2020*
  - Large-scale image super-resolution: **Kick-started a project that is now part of the Edge browser beta**. Did literature review on image super-resolution problem, trained various Transformers-based model for it and proposed various techniques for model improvement.
- **Simon Fraser University** — Canada
  Researcher — *Sep 2018 - Sep 2020*
  - Worked on analyzing and improving interpretability of attention mechanism in deep sequence-to-sequence models. Published two papers into **top tier conferences** (e.g. EMNLP).
- **Resid** — Iran
  Software Engineer — *Sep 2017 - Sep 2018*
  - Was responsible for **leading the design and implementation of architecture** and back-end APIs for the first P2P mobile payment solution in Iran.
  - Migrated a live code-base and database schema to a completely new one with near-zero downtime
- **Divar** — Iran
  Software Engineer (Search & AI) — *Mar 2016 - Sep 2017*
  - Implemented various machine learning microservices for semantic information extraction from search queries and ads. **Increased search recall by 7%** .
  - Trained and deployed a vision model into production for automatic identification of vehicle makes and models from the ad's image. **Reduced average-time-to-ad-submit metric by 10%.**
  - Analyzed **500M users' action logs** via Spark to analyze users' behaviors and proposed product improvement ideas accordingly.
  - Worked on the largest classified ads platform in Iran: **Led the effort of extracting search stack** from a giant monolithic codebase and converting it into various dockerized microservices deployed on kubernetes.
  - Designed a fault-tolerant and scalable microservice architecture on top of rabbitmq and elasticsearch, which later **served 10M queries per day.**
- **Taskulu** — Iran
  Software Engineer — *Aug 2013 - Aug 2014*
  - Contributed to design and development of the backend of a project management platform startup.
  - Improved response time of the web app by 20% by efficient use of cache hierarchy ranging from database-level caching to application-level Redis-based caching

Publications

- **P. Moradi**, N. Kambhatla, and A. Sarkar. *Measuring and improving faithfulness of attention in neural machine translation.* EACL 2021
- **P. Moradi**, N. Kambhatla, and A. Sarkar. *Interrogating the Explanatory Power of Attention in Neural Machine Translation.* EMNLP 2019 Workshops.
- H. Zamani, **P. Moradi**, and A. Shakeri. *Adaptive User Engagement Evaluation via Multi-task Learing.* SIGIR 2015.
- H. Zamani, A. Shakeri, **P. Moradi**. *Regression and Learning to Rank Aggregation for User Engagement Evaluation.* ACM RecSys Workshops 2014

Education

- **Simon Fraser University** — Burnaby, Canada
  Master of Science - Computer Science — *Sep 2018 - Sep 2020*
  Thesis: *Improving interpretability of attention mechanism in sequence-to-sequence models.*
- **University of Tehran** — Tehran, Iran
  Bachelor of Science - Software Engineering — *Sep 2013 - Sep 2017*