Parsa Kavehzadeh

parsareal@gmail.com Github LinkedIn Google Scholar parsareal.github.io

EXPERIENCE

Huawei Technologies Canada Co. Markham, Ontario
Associate Researcher *May 2023-Present*

- **Introduced Sorted LLaMA**, a many-in-one architecture with **8 nested sub-models**, enabling text generation up to **70% smaller and faster** with efficient single-round training.
- **Developed a confidence-based early exiting mechanism** for Sorted LLaMA, achieving a **55% reduction in inference time** compared to autoregressive generation. Published the Sorted LLaMA paper at **EACL 2024**.
- **Reduced the performance gap** between standard fine-tuning and Sorted networks by up to **33%** using **smart optimized sub-model loss weighting** during training.
- **Accelerated LLM inference** by training a **nested draft model** with adaptive token drafting, achieving up to **2x speedup on 70B models** without performance loss.
- Integrated cutting-edge research ideas like **Adaptive Attention Tree** into nested draft models, achieving a further **2.6x speedup** on 70B models.
- **Led inference acceleration research**, managing a sub-team of four and collaborating with headquarters bi-weekly, achieving **2.55x speedup** on Huawei's 38B Pangu models on NPU hardware.
- **Expert in Python, Hugging Face, and PyTorch**, implementing Nested Training and Sorted Speculative Sampling, and leveraging **DeepSpeed** and **FSDP** for distributed LLM training on diverse datasets.
- **Skilled in tensor and pipeline parallelism**, training 38B models on **128 Ascend NPUs across 16 nodes** using a **200,000-sample multilingual dataset**, with experience in Huawei's Pangu LLMs and the Megatron codebase.
- **Recognized as MVP** in the first year as a researcher at Huawei Noah's Ark Lab.

Intelligent Visualization Lab - York University Toronto, Ontario
Graduate Student Researcher *Sep 2021-Aug 2023*

- Researched **natural language interactions with visualizations**, emphasizing **chart comprehension and reasoning** through **multimodal NLP and computer vision** methods.
- Authored a comprehensive survey on **chart question answering**, categorizing key subdomains such as input, output, and modeling aspects while identifying research opportunities in each category.
- Developed a novel end-to-end **chart pretraining** approach capable of addressing multiple chart understanding tasks, such as chart question answering and summarization.
- Curated a pre-training corpus with over **7 million synthetic and real chart images**, paired with natural language queries and responses, to support diverse chart comprehension tasks.
- Addressed the lack of large-scale chart-summary datasets using **knowledge distillation**, fine-tuning **Flan-T5** on **4,500 summaries**, enabling generation for **450,000 charts** without costly API calls.
- Pretrained and finetuned a vision-language model on 8 NVIDIA A100 GPUs, achieving **66% exact match accuracy** on ChartQA and state-of-the-art performance on chart tasks. Published this work as **UniChart** at **EMNLP 2023**.

Manulife Toronto, Ontario
Data Science Internship *May 2022-Aug 2022*

- Deployed **BERTopic** on **Azure Machine Learning** and built an ETL pipeline to structure unstructured data into an **Azure SQL** warehouse, enhancing topic modeling and search with **SQL full-text indexing**.
- Automated ML inference and optimized workflows with **Azure Scheduler**, integrating a **Kubernetes (AKS)** interface to streamline database queries and boost operational efficiency.

RECENT PUBLICATIONS

- *A.Masry, **P.Kavehzadeh**, XL.Do, E.Hoque, S.Joty*, "UniChart: A Universal Vision-language Pretrained Model for Chart Comprehension and Reasoning", **EMNLP 2023**.
- ***P.Kavehzadeh**, M.Pourreza, M.Valipour, T.Zhu, H.Bai, A.Ghodsi, B.Chen, M.Rezagholidzadeh*, "S2D: Sorted Speculative Decoding For More Efficient Deployment of Nested Large Language Models", **ENLSP Workshop at NeurIPS 2024**.
- ***P.Kavehzadeh**, M.Valipour, M.Tahaei, A.Ghodsi, B.Chen, M.Rezagholidzadeh*, "Sorted LLaMA: Unlocking the Potential of Intermediate Layers of Large Language Models for Dynamic Inference", **EACL 2024**.

EDUCATION

York University
*Master of Science in **Computer Science**, GPA: 8.8/9* *Sep 2021-Aug 2023*

Amirkabir University of Technology
*Bachelor of Science in **Computer Engineering**, GPA: 3.86/4* *Sep 2016-July 2022*