Ben Athiwaratkun

in /benathi
benathi.github.io

Education

PhD in Statistics with Special Masters in Computer Science, *Cornell University, 2014-2019*, PhD thesis on density representation for words and hierarchical data.

Bachelor of Arts in Mathematics and Economics, *Williams College, 2008-2012*, Honors in Mathematics with undergraduate thesis on ergodic theory, Benedict First Prize in Mathematics, Magna Cum Laude, Phi Beta Kappa and Sigma Xi.

Patents

Bifurcated Attention for Memory-IO Efficient Inference. A method to increase throughput and reduce latency for single context large batch sampling. Submitted March 2023.

Token Alignment via Constrained Generation for Subword Completion. Helps LLMs complete text that ends with a partial token. Submitted November 2022.

Programmatic Conversion of Code Evaluation Data to Different Programming Languages Enables scaling of evaluation data to multiple programming languages for measuring code generation abilities. Submitted November 2022.

Work Experience

August 2022 - present **Senior ML Scientist**, *AWS CodeWhisperer/AWS AI Labs*, Amazon Web Services.

- Led the research to investigate the effectiveness of **sparse attention** as a way to reduce context encoding latency.
- Collaborated with teammates on the research for **compression-optimized tokenizer** to increase model performance via higher data compression and reduce latency.
- Collaborated with teammates on investigating **position embeddings** in order to extend language model context length.
- Led the research on **multi-group attention** (generalized multi-query attention) including model training, evaluation, and inference latency benchmarking, with wide adoption beyond the team.
- Proposed an approach to augment language model's with **code insertion** abilities and drove the project to production.
- Proposed a product direction to offer **customized models** for AWS CodeWhisperer, and brainstormed the initial research direction.

August 2019 - August 2022 **Applied Scientist**, *AWS CodeWhisperer/AWS AI Labs*, Amazon Web Services.

- Developed a method to reduce memory IO during inference for large batch sampling called **bifurcated attention**. This research led to a patent submission, deployment for real-world usage, and a paper.
- Developed a **multi-lingual evaluation framework** in order to gauge language models' code generation abilities on various programming languages. This research led to an ICLR paper (spotlight), a patent submission, and is used widely within the team for various benchmarking purposes.
- Collaborated with teammates on **model finetuning** to increase language model's code generation capabilities, including data augmentation by bootstrapping with integration test.

- Invented the method of **token alignment** which helps language models *complete* text or code when the prompt ends with subword. Wrote production code and worked with engineering partners to deploy the technique in AWS CodeWhisperer, submitted a patent, and a paper under review.
- Collaborated with team for the development of **internal language models for code generation**, including data processing, tokenizer design, pre-training, inference logic, and evaluation. Gained **full stack** knowledge on language models for code and participated in discussion to help brainstorm in various workstreams.
- Conducted research in the area of few-shot learning using **language models for structured prediction**, being one of the early advocates (before GPT-3) in the team to use language models' generalization abilities for few-shot learning.

Summer 2017 **Research Intern**, *AWS*.
Explored the use of subword structure to multimodal word distributions for modeling polysemies jointly with subword information. Supervisor: Anima Anandkumar.

Summer 2016 **Research Intern**, *Microsoft Research*.
Built neural language models (RNN, GRU, LSTM and character-level CNN) with unsupervised training to learn the language of computer APIs. The trained representation proves useful for downstream tasks on malware detection. Supervisor: Jack Stokes.

Summer 2011 **Intern**, *RelSci*.
Performed data processing for relationship extraction at RelSci, a platform that provides research driven data for relationship management.

Summer 2010 **Undergraduate Physics Research Assistant**, *Williams College*.
Developed a Fortran algorithm based on mT2 method to group clusters of by-products from proton-proton collision generated by Monte Carlo simulation of Large Hadron Collider in order to identify multiple mass eigenstates of top squark (predicted by supersymmetry). Succeeded in distinguishing two mass eigenstates, if exists, with resolution of 100 MeV. Supervisor: David Tucker Smith.

Winter 2009 **Undergraduate Physics Research Assistant**, *Williams College*.
Assisted a faculty member on maximizing the precision of computer-generated phase-regime holograms that controls light patterns for high precision trapping of small objects such as computer chips. Accomplished in determining optimal holograms for grid-like laser patterns. Supervisor: Ward Lopes

Select Publications

Greener yet Powerful: Taming Large Code Generation Models with Quantization. Xiaokai Wei, Sujan K. Gonugondla, Wasi Uddin Ahmad, Shiqi Wang, Baishakhi Ray, Haifeng Qian, Xiaopeng Li, Varun Kumar, Zijian Wang, Yuchen Tian, Qing Sun, Ben Athiwaratkun, Mingyue Shang, Murali Krishna Ramanathan, Parminder Bhatia, and Bing Xiang. In *Preprint*, 2023.

FusionToken: Enhancing Compression and Efficiency in Language Model Tokenization. Robert Kwiatkowski, Zijian Wang, Varun Kumar Robert Giaquinto, Xiaofei Ma, Bing Xiang, and Ben Athiwaratkun. In *Preprint*, 2023.

Token Alignment via Character Matching for Subword Completion. Ben Athiwaratkun, Shiqi Wang, Mingyue Shang, Yuchen Tian, Zijian Wang, Sujan Kumar Gonugondla, Sanjay Krishna Gouda, Rob Kwiatowski, Ramesh Nallapati, and Bing Xiang. In *Preprint*, 2023.

Multi-lingual Evaluation of Code Generation Models. Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng Li, Yuchen Tian, Ming Tan, Wasi Uddin Ahmad, Shiqi Wang, Qing Sun, Mingyue Shang, Sujan Kumar Gonugondla, Hantian Ding, Varun Kumar, Nathan Fulton, Arash Farahani, Siddhartha Jain, Robert Giaquinto, Haifeng Qian, Murali Krishna Ramanathan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatia, Sudipta Sengupta, Dan Roth, and Bing Xiang. In *ICLR (Spotlight)*, 2023.

Structured Prediction as Translation between Augmented Natural Languages. Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. In *ICLR (Spotlight)*, 2021.

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average. Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. In *ICLR*, 2019.

Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. Xilun Chen, Yu Sun, Ben Athiwaratkun, Kilian Q. Weinberger, and Claire Cardie. 2018.

Hierarchical Density Order Embeddings. Ben Athiwaratkun and Andrew Gordon Wilson. In *ICLR*, 2018.

Probabilistic FastText for Multi-Sense Word Embeddings. Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. ACL (Oral), 2018.

Multimodal Word Distributions. Ben Athiwaratkun and Andrew Gordon Wilson. ACL, 2017.

Malware classification with LSTM and GRU language models and a character-level CNN. Ben Athiwaratkun and Jack W. Stokes. In *ICASSP*, 2017.

Program Committess

Served as a reviewer for ICLR, ICML, Neurips, AISTATS, ACL, EMNLP. Best reviewer award Neurips 2019.

Blog and Open Source Contribution

I blog about technical details in AI such as illustrated tensor parallelism, multi-query attention, benchmarking GPT-3.5/GPT-4's code generation and physics problem solving abilities, to name a few.

Contributed open source implementation such as generalized multi-query in megatron-deepspeed, multi-lingual execution framework for code generation abilities and MBXP datasets, etc.

Skills

Programming Languages

Python, R, C++, Mathematica, Matlab, JavaScript, HTML, SQL

Deep Learning Frameworks

Pytorch, Megatron-LM, DeepSpeed, DeepSpeed Inference, Tensorflow

Competitions

Top-500 in Putnam Mathematical Competition, 2012.

Silver Medal, International Physics Olympiad, 2007.

Gold Medal, Thailand Physics Olympiad, 2006.