Ben Athiwaratkun

(413) 884-3087
ben.athiwaratkun
@gmail.com

EDUCATION

Cornell University, PhD in Statistics

September 2014 - May 2019

PhD in Statistics with concentration in artificial intelligence. PhD thesis on density representation for words and hierarchical data.

Advisor: Andrew Gordon Wilson

Williams College, BA in Mathematics (Honors) and Economics

August 2008 - June 2012

Honors in Mathematics with undergraduate thesis on ergodic theory.

SELECT RESEARCH PROJECTS

Multilingual Evaluation of Code Generation Models - ICLR 2023 (spotlight)

Ben Athiwaratkun, et al. (+25 authors)

• Invented a method to scale the execution-based evaluation of code generation abilities to many programming languages, by transpiling the original Python datasets (HumanEval, MBPP) to 10+ languages.
• Demonstrated emergent abilities such as out-of-domain generalization to other programming languages and zero-shot code translation abilities.

Structured prediction as translation between augmented natural languages — ICLR 2021 (spotlight)

Giovanni Paolini, **Ben Athiwaratkun**, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, Stefano Soatto

Demonstrated that the augmented natural language framework (details below) can elegantly solve various structured prediction tasks under a generative setting (20+ tasks) within a single shared model. This paper shows the power of generative models as a general task solver, and provides clear contrast to other baseline models which design specific methods to perform each task.

Augmented Natural Language for Generative Sequence Labeling — EMNLP 2020 (oral)

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, Bing Xiang

Demonstrated an effective approach to model sequence labeling with a generative model by using a proposed augmented natural language as a proposed output format. Our model learns from limited data efficiently and sets a new state-of-the-art record for few-shot slot labeling.

EMBEDDED LINKS

Google Scholar
benathi.github.io

SKILLS

Distributed training / 3D parallelism

PyTorch, Megatron-LM, DeepSpeed Zero, DeepSpeed Inference

PATENTS

Constrained generation for subword handling (submitted, 2022)

Programmatic conversion of code evaluation data (submitted, 2022)

No-duplicate broadcasting for memory IO reduction (in preparation)

There Are Many Consistent Explanations of Unlabeled Data: Why You Should Average — ICLR 2019 (poster)

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, Andrew Gordon Wilson

Demonstrated that semi-supervised learning can lead to multiple consistent explanations of the unlabeled data and found that weight averaging can significantly bridge the gap, achieving SOTA in all resource levels for CIFAR-10 and CIFAR-100.

Adversarial Deep Averaging Networks for Cross-Lingual Domain Adaptation — TACL 2018 (poster)

Xilun Chen, Yu Sun, **Ben Athiwaratkun**, Claire Cardie, Kilian Weinberger

Demonstrated that a simple deep averaging network with an adversarial training can help transfer knowledge learned from labeled data in the resource-rich setting to low-resource languages using only unlabeled data.

Probabilistic FastText for Multi-Sense Word Embeddings — ACL 2018 (oral)

Ben Athiwaratkun, Andrew Gordon Wilson, Anima Anandkumar

Incorporated subword information to a mixture of Gaussian representation, improving the semantic representation of out-of-vocabulary words as well as foreign languages (via latin roots).

Hierarchical Density Order Embeddings — ICLR 2018 (poster)

Ben Athiwaratkun, Andrew Gordon Wilson

Adopted Gaussian representation to capture data hierarchies, with demonstrated benefits on tasks such as hypernym prediction and lexical entailment.

Multimodal Word Distributions — ACL 2017 (poster)

Ben Athiwaratkun, Andrew Gordon Wilson

Proposed a Gaussian mixture representation for words, through which multiple semantics emerge from unsupervised training based on distributional hypothesis of word similarity.

WORK AND INTERNSHIP EXPERIENCE

AWS AI Labs. New York, NY — *ML Scientist*

August 2019 - PRESENT

• Working on language models for code generation on various aspects including training (familiar with 3d parallelism, Deepspeed), efficient inference (Deepspeed inference with custom module for multi-group attention), evaluation at scale (data parallel evaluation with large-scale sampling).
• Proposed a method that incorporates the right context to perform code insertion for language models.
• Developed a lossless tokenizer based on a wrapper over SentencePiece.
• Innovated impactful ideas that are used in real-world products along with patent submissions.
• Collaborated cross teams for research projects such as generative modeling for structured prediction tasks.

AWARDS

Magna Cum Laude, Phi Beta Kappa and Sigma Xi

Top-500 in Putnam Competition, 2012

Benedict First Prize in Mathematics, 2010

International Physics Olympiad 2007, Silver Medal

AWS. Palo Alto, CA — *Research Intern*

May 2017 - August 2017

Explored the idea of balancing dictionary-level and word-level representations for word embeddings using group sparsity regularization. Pivoted the idea and applied the subword structure to multimodal word distributions. Supervisor: Anima Anandkumar.

Microsoft Research. Redmond, WA — *Research Intern*

May 2016 - August 2016

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.