Songlin Yang

PhD student, Massachusetts Institute of Technology

Address: 711-D, 51 Vassar St, Cambridge, MA 02139

Phone: 857-756-0739

E-mail: yangsl66@mit.edu

Homepage: sustcsonglin.github.io

Last updated: Tuesday 22nd October, 2024

Education

Massachusetts Institute of Technology, Cambridge, MA, USA — 2023-now
Ph.D. in Computer Science
Advisor: Yoon Kim

ShanghaiTech University, Shanghai, China — 2020-2023
M.S. in Computer Science
Advisor: Kewei Tu

Southern University of Science and Technology, Shenzhen, China — 2016-2020
B.S. in Computer Science

Research Internships

NVIDIA Research, Santa Clara, CA, USA — May. 2024-August. 2024
Host: Ali Hatamizadeh, Jan Kautz

Teaching

Teaching Assistant at ShanghaiTech University
Natural Language Processing   Instructor: Kewei Tu — Spring 2022

Teaching Assistant at ShanghaiTech University
Artificial Intelligence   Instructor: Kewei Tu — Fall 2021

Teaching Assistant at Southern University of Science and Technology
Artificial Intelligence   Instructor: Ke Tang — Fall 2019

Referred Conference Publications

[1] **Songlin Yang**, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. *Parallelizing Linear Transformers with the Delta Rule over Sequence Length*. In: *Thirty-eighth Annual Conference on Neural Information Processing Systems* (*NeurIPS*). 2024.

[2] Yu Zhang*, **Songlin Yang***, Ruijie Zhu, Yue Zhang, Leyang Cui, Yiqiao Wang, Bolun Wang, Freda Shi, Bailin Wang, Wei Bi, Peng Zhou, and Guohong Fu. *Gated Slot Attention for Efficient Linear-Time Sequence Modeling*. In: *Thirty-eighth Annual Conference on Neural Information Processing Systems* (*NeurIPS*). 2024.

[3] Zhen Qin*, **Songlin Yang***, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. *HGRN2: Gated Linear RNNs with State Expansion*. In: *First Conference on Language Modeling* (*COLM*). 2024.

[4] **Songlin Yang***, Bailin Wang*, Yikang Shen, Rameswar Panda, and Yoon Kim. *Gated Linear Attention Transformers with Hardware-Efficient Training*. In: *Forty-first International Conference on Machine Learning* (*ICML*). 2024.

[5] Zhen Qin*, **Songlin Yang***, and Yiran Zhong. *Hierarchically Gated Recurrent Neural Network for Sequence Modeling*. In: *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*. 2023.

[6] Wei Liu*, **Songlin Yang***, Yoon Kim, and Kewei Tu. *Simple Hardware-Efficient PCFGs with Independent Left and Right Productions*. In: *Findings of the Association for Computational Linguistics (EMNLP Findings)*. 2023.

[7] Pengyu Ji, **Songlin Yang**, and Kewei Tu. *Improving Span Representation by Efficient Span-Level Attention*. In: *Findings of the Association for Computational Linguistics (EMNLP Findings)*. 2023.

[8] Zhaohui Yan, **Songlin Yang**, Wei Liu, and Kewei Tu. *Joint Entity and Relation Extraction with Span Pruning and Hypergraph Neural Networks*. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023.

[9] **Songlin Yang**, Roger Levy, and Yoon Kim. *Unsupervised Discontinuous Constituency Parsing with Mildly Context-Sensitive Grammars*. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 2023.

[10] **Songlin Yang** and Kewei Tu. *Don't Parse, Choose Spans! Continuous and Discontinuous Constituency Parsing via Autoregressive Span Selection*. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 2023.

[11] Wei Liu, **Songlin Yang**, and Kewei Tu. *Structured Mean-Field Variational Inference for Higher-Order Span-Based Semantic Role Labeling*. In: *Findings of the Association for Computational Linguistics (ACL Findings)*. 2023.

[12] **Songlin Yang** and Kewei Tu. *Headed-Span-Based Projective Dependency Parsing*. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2022.

[13] **Songlin Yang** and Kewei Tu. *Bottom-Up Constituency Parsing and Nested Named Entity Recognition with Pointer Networks*. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2022.

[14] Chao Lou, **Songlin Yang**, and Kewei Tu. *Nested Named Entity Recognition as Latent Lexicalized Constituency Parsing*. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2022.

[15] **Songlin Yang***, Wei Liu*, and Kewei Tu. *Dynamic Programming in Rank Space: Scaling Structured Inference with Low-Rank HMMs and PCFGs*. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. 2022.

[16] **Songlin Yang** and Kewei Tu. *Combining (Second-Order) Graph-Based and Headed-Span-Based Projective Dependency Parsing*. In: *Findings of the Association for Computational Linguistics (ACL Findings)*. 2022.

[17] **Songlin Yang** and Kewei Tu. *Semantic Dependency Parsing with Edge GNNs*. In: *Findings of the Association for Computational Linguistics (EMNLP Findings)*. 2022.

[18] **Songlin Yang**, Yanpeng Zhao, and Kewei Tu. *PCFGs Can Do Better: Inducing Probabilistic Context-Free Grammars with Many Symbols*. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. 2021.

[19] **Songlin Yang**, Yanpeng Zhao, and Kewei Tu. *Neural Bi-Lexicalized PCFG Induction*. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*. 2021.

[20] **Songlin Yang**, Yong Jiang, Wenjuan Han, and Kewei Tu. *Second-Order Unsupervised Neural Dependency Parsing*. In: *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. 2020.

Invited Talks

[21] **Songlin Yang**. *Linear Transformers for Efficient Sequence Modeling*. Invited talk, Stanford University. Aug. 2024.

[22] **Songlin Yang**. *Gated Linear Recurrence for Efficient Sequence Modeling*. Invited talk, LinkedIn Corporation. June 2024.

[23] **Songlin Yang**. *Gated Linear Recurrence for Efficient Sequence Modeling*. Invited talk, Cornell Tech. Apr. 2024.

Open-Sourced Project

1. **Flash Linear Attention**
   A Triton-Based Library for Hardware-Efficient Implementations of Linear Attention Mechanism.
   `https://github.com/sustcsonglin/flash-linear-attention/`