YANG LI

Google Scholar ◇ Github ◇ LinkedIn
Phone: (+1) 404-512-1722 ◇ Email: yangli.fm@gmail.com

PERSONAL SUMMARY

- Ph.D. in Computer Science. Former intern at Oracle and NEC Labs.
- Extensive hands-on experience with LLMs such as BERT, GPT, Llama, Mistral, Gemma, and more. Worked on a variety of NLP tasks including natural language understanding (NLU), natural language generation (NLG), instruction tuning, and mathematical reasoning.
- Specializations include parameter-efficient fine-tuning, neural network compression, and data mining and modeling. Authored 7 research papers published in prestigious conferences and journals such as ECML, IEEE IJCNN, OFC, and Human Brain Mapping. Authored 1 US Patent.
- Strong coding skills with proficiency in Python, C++, Java, SQL, PyTorch, and TensorFlow. Contributed code under the HuggingFace PEFT framework.

EDUCATION

Georgia State University *Aug. 2018 — Jul. 2024*
Ph.D of Computer Science
GPA:3.75/4.3

University of Science and Technology of China (USTC) *Sept. 2015 — Jun. 2018*
Master of Nuclear Science and Technology
GPA: 3.61/4.0

Anhui Normal University *Sept. 2011 — Jul. 2015*
Bachelor of Computer Science and Technology
GPA: 3.93/5.0, Rank: **1**/97

PROJECTS

Research on LLMs Finetuning (Parameter Efficient Finetuning) [1]
Released in May 2024. Received 3 citations within the first 2 months. [Paper], [Code]

· Proposed VB-LoRA which aims at reducing the trainable/storage parameters required for fine-tuning large language models (LLMs). Introduced a novel divide-and-share strategy, where parameters from different layers and modules are decomposed and shared in a global *vector bank*.
· Performed instruction tuning on Llama2-13B, Mistral-7B, and Gemma-7B: VB-LoRA uses approximately 0.4% of the parameters compared to LoRA, while achieving better performance.

Neural Network Pruning [2]–[4]

· Proposed L0-ARM and Dep-L0, learning binary gates with L_0-based regularization in an end-to-end manner. Pruned 43.9% of the parameters and reduced 38.1% FLOPs of ResNet-50 on the ImageNet task while maintaining comparable performance.
· Proposed NPNs, a unified framework for network pruning and expansion.

Human Intelligence Prediction and Brain Structure Exploration [5]
Collaborated with Columbia University and MD Anderson Cancer Center. [Paper]

· Designed an LSTM-based network to extract spatiotemporal information from fMRI data and predict human intelligence, achieving state-of-the-art results.
· Learned the importance scores of brain regions in an end-to-end manner. Our findings aligned with results from neuroscience research.

Network Device Classification and Anomaly Detection
A project with VMWare

- Designed and curated a dataset of network traffic by analyzing TCP and UDP packet headers; performed feature engineering to select the top 10 most informative features for device classification tasks.
- Designed and implemented LSTM and CNN models for device classification (iOS, Android, Linux, Windows, MacOS, webcams, and routers), achieving an accuracy rate of approximately 95%.
- Identified unique traffic patterns for each device type and developed language models to detect anomalous behavior in real-time.

INTERN EXPERIENCE

Data Science Intern @ Oracle May 2023 - Aug 2023

- Designed and implemented an LSTM model to disaggregate energy usage from monthly electric bills.
- Proposed a new feature representation to improve model performance by 10%.
- Contributed code and unit tests to the deep learning library used internally at Oracle.

Research Intern @ NEC Laboratories America May 2022 - Aug 2022

- Proposed an eigenvalue-based algorithm for locating ocean earthquakes at span-level (60-90 kilometers) using existing submarine fibers, achieving nearly 100% accuracy.
- Published the results in the top conference Optical Fiber Communication [6] and filed a patent [7].

PUBLICATIONS

[1] Y. Li, S. Han, and S. Ji, "Vb-lora: Extreme parameter efficient fine-tuning with vector banks," *arXiv preprint arXiv:2405.15179*, 2024.

[2] Y. Li and S. Ji, "Dep-l0: Improving l 0-based network sparsification via dependency modeling," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2021, pp. 167–183.

[3] Y. Li and S. Ji, "L0-ARM: Network sparsification via stochastic binary optimization," in *The European Conference on Machine Learning (ECML)*, 2019.

[4] Y. Li and S. Ji, "Neural plasticity networks," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–9.

[5] Y. Li, X. Ma, R. Sunderraman, S. Ji, and S. Kundu, "Accounting for temporal variability in functional magnetic resonance imaging improves prediction of intelligence," *Human Brain Mapping*, vol. 44, no. 13, pp. 4772–4791, 2023.

[6] F. Yaman, Y. Li, S. Han, T. Inoue, E. Mateo, and Y. Inada, "Polarization sensing using polarization rotation matrix eigenvalue method," in *Optical Fiber Communication Conference*, Optica Publishing Group, 2023, W1J–7.

[7] F. Yaman, H. Shaobo, E. F. M. RODRIGUEZ, Y. Li, Y. Inada, and T. Inoue, *Fiber sensing by monitoring polarization function of light on supervisory path of cables*, US Patent App. 18/369,041, Mar. 2024.

ACHIEVEMENTS

Graduate Teaching Award, awarded by Georgia State University *March 2022*
Best Graduate Presentation, awarded by Georgia State University *May 2019*
Outstanding Dissertation, awarded by Anhui Normal University *June 2015*
National Scholarship, awarded by Ministry of Education, PRC *November 2013*

SKILLS

Programming Languages	Python, C++, Java, Swift, Matlab, Bash Script, Javascript, SQL
Machine Learning Tools	Pytorch, Tensorflow, HuggingFace Transformers / PEFT