Yang Li

(404)-512-1722 | Atlanta, GA | yangli.fm@gmail.com | github.com/leo-yangli | linkedin.com/in/mkdirleo

Ph.D. candidate in Computer Science at Georgia State University, specializing in **Parameter Efficient Fine-Tuning (PEFT), Model Compression, Large Language Models (LLMs)** and **AI for Sciences**. Proficient in leveraging deep learning and mathematical tools to address real-world challenges.

EDUCATION BACKGROUND

- **Georgia State University** — *August 2018 — July 2024*
  Ph.D of Computer Science, GPA:3.75/4.3
- **University of Science and Technology of China** — *September 2015 — June 2018*
  Master of Nuclear Science and Technology, GPA: 3.61/4.0
- **Anhui Normal University** — *September 2011 — July 2015*
  Bachelor of Computer Science and Technology, GPA: 3.93/5.0, Rank: **1**/97

SKILLS

- **Programming Languages:** Python, C++, Java, Swift, Matlab, Bash Script, Javascript, SQL
- **Framework & Tools:** Pytorch, Tensorflow, HuggingFace Transformers, HuggingFace PEFT, Linux, Latex

INTERN EXPERIENCE

Data Science Intern — **Remote**
Oracle — *May 2023 — August 2023*

– Improved data representations and model architecture to disaggregate energy usage from low-resolution (monthly) bills.
– Contributed code and unit tests to the deep learning library used internally at Oracle.

Research Intern — **Princeton, NJ**
NEC Laboratories America — *May 2022 — August 2022*

– Proposed an eigenvalue-based method in locating earthquakes in the ocean at span-level (60-90 kilometers) utilizing existing submarine fibers, achieving almost 100% accuracy. Published in the top conference *Optical Fiber Communication*[3].

SELECTED RESEARCH AND PROJECTS

Parameter Efficient Fine-Tuning (PEFT)
– Proposed VB-LoRA[1], an innovative parameter-efficient fine-tuning (PEFT) method for large language models, addressing the growing need for customizable models. Fine-tuning the Llama2-13B model with VB-LoRA demonstrated superior results using only **0.4%** of the parameters required by LoRA.

Network Compression
– Proposed two pruning algorithms, L0-ARM[6] and Dep-L0[4]. The former achieved competitive compression rates while maintaining accuracy, boosting inference speed by 5 times on edge devices; the latter improved compression performance on large models by explicitly considering dependencies between weights in different layers.

Human Intelligence Prediction and Brain Structure Exploration
– Designed an LSTM-based network to extract spatial temporal information from fMRI and predict human intelligence, achieving SOTA result. Applied l0-norm-based feature selection method to learn brain region activations when individuals perform different tasks[2].

SELECTED PUBLICATIONS

1. **Yang Li**, S. Han, S. Ji, **VB-LoRA: Extreme Parameter Efficient Fine-Tuning with Vector Banks,** 2024. [arxiv]
2. **Yang Li**,X. Ma, R. Sunderraman, S. Ji, S. Kundu, **Accounting for temporal variability in functional magnetic resonance imaging improves prediction of intelligence**, Human Brain Mapping, 2023. [wiley]
3. F. Yaman, **Yang Li**, S. Han, T. Inoue, E. Mateo, Y. Inada, **Polarization Sensing Using Polarization Rotation Matrix Eigenvalue Method**, Optical Fiber Communication Conference, 2023. [IEEE]
4. **Yang Li**, S. Ji, **Dep-L0: Improving L0-based Network Sparsification via Dependency Modeling**, European Conference on Machine Learning (ECML), 2021. [arxiv]
5. **Yang Li**, S. Ji, **Neural Plasticity Networks**, International Joint Conference on Neural Networks, 2021.[arxiv]
6. **Yang Li**, S. J. **L0-ARM: Network Sparsification via Stochastic Binary Optimization**. ECML, 2019. [arxiv]

AWARDS & ACHIEVEMENTS

Graduate Teaching Award of Georgia State University, *March 2022* | **Best Graduate Presentation** of CS Demo Day, *May 2019*
Outstanding Dissertation of Anhui Normal University, *June 2015* | **National Scholarship**, *November 2013*