# Yiyang Zhao

(+1) 617-306-0301 zyyaoiang@gmail.com -286947117 aoiang https://zhaoyiyang.me

## Summary

Yiyang Zhao has a diverse research background in machine learning, large language models, computer vision, and artificial intelligence. He has authored papers presented at top conferences and journals such as ICML, ICLR, AAAI, and JMLR, accumulating over 550 citations. His expertise in Neural Architecture Search has led to the design of deep learning models, pushing SoTA results across various AI tasks. He is currently conducting research on large language models.

## Education

**Worcester Polytechnic Institute, Worcester, MA** **2019.8 – 2024.6**
*Ph.D., Computer Science.*

**Northeastern University, Boston, MA** **2015.1 – 2016.12**
*M.S., Computer Engineering*

**Xidian University, Xi'an, China** **2010.8 – 2014.7**
*B.Eng., Electronic and Information Engineering*

## Technical Skills & Research

**Languages**: Python, Java, C++
**Research Areas**: Neural Architecture Search (**NAS**), Large Language Models (**LLMs**), Deep Learning, AutoML, Computer Vision, Optimization, Self-driving Database, High-Performance Computing
**Frameworks & Deployments**: Pytorch, NumPy, TensorFlow, Pandas, Scikit-learn, Vim, Git, Linux, Docker

## Work Experiences

**NVIDIA Corporation** **2022.5 – 2022.8**
*Deep Learning Algorithm Engineering Intern* *Santa Clara, CA*

- Contributed to the design of low-latency object detection models using Neural Architecture Search.
- Designed the GPU-YOLO series models, achieving SoTA benchmarks while maintaining reduced TensorRT inference latency.

**Tiktok - Infrastructure System Lab** **2021.5 – 2021.8**
*Research Intern* *Mountain View, CA*

- Applied Auto-Machine Learning to optimize database configurations.
- Achieved over a **60%** increase in throughput compared to configurations designed by DataBase administrators across various workloads(e.g., tpcc, twitter, small bank, and etc.)

## Projects

**Few-shot Neural Architecture Search** | *NAS, Optimization, Computer Vision. Github*

- Present Few-shot NAS concept that greatly improves NAS search performance.
- Achieved SoTA results in image classification, object detection, and AIGC tasks.
- Paper published in ICML 2021(**Long Oral**). Many extended works are applied in large language models, computer visions, and AIGC areas and published in top-tier AI conferences.

**Multi-objective Optimization** | *Optimization, Machine Learning, NAS, Computer Vision. Github*

- Build a new multi-objective black-box optimizer with high search efficiency.
- Achieve SoTA results across various tasks and areas. (e.g., Computer vision, NLP, Molecule discovery.)
- Paper published in ICLR 2022.

**Multi-Adapters Tuning for Memory injection on Large Language Models** | *LLMs, NAS, Optimization. Github*

- Constructed an efficient NAS agent for designing/turning adapters in LLMs.
- Customized LLMs by memory injection in NAS designed adapters while retaining LLM's intricate ability.
- Paper submitted to ACL 2024

**AlphaX** | *NAS, Monte Carto Tree Search, Computer vision. Github*

- Constructed an efficient NAS agent leveraging Monte Carlo Tree Search.
- The neural networks designed by AlphaX bolstered performance in downstream applications such as detection, style transfer, image captioning, and many others.
- Paper published in AAAI 2020.

## Publications

**Multi-Objective Neural Architecture Search by Learning Search Space Partitions** | *Link* **JMLR 2024**
- **Yiyang Zhao**, Linnan Wang, Tian Guo

Journal of Machine Learning Research

**Few-shot Neural Architecture Search** | *Link* **ICML 2021(Long Oral)**
- **Yiyang Zhao**, Linnan Wang, Yuandong Tian, Rodrigo Fonseca, Tian Guo

38th International Conference on Machine Learning

**Multi-objective Optimization by Learning Space Partitions** | *Link* **ICLR 2022**
- **Yiyang Zhao**, Linnan Wang, Kevin Yang, Tianjun Zhang, Tian Guo, Yuandong Tian

10th International Conference on Learning Representations

**Neural Architecture Search using Deep Neural Networks and Monte Carlo Tree Search** | *Link* **AAAI 2020**
- Linnan Wang*, **Yiyang Zhao***(equally contributed), Yuu Jinnai, Yuandong Tian, and Rodrigo Fonseca

34th AAAI Conference on Artificial Intelligence

**Efficient Communications in Training Large Scale Neural Networks** | *Link* **ACM-MM 2017**
- **Yiyang Zhao**, Linnan Wang, Jinmian Ye, Wei Wu, George Bosilca, Richard Vuduc, Wenqi Tang, Zenglin Xu

25th ACM international conference on Multimedia Thematic Workshop

**Dynamic GPU Memory Management for Training Nonlinear Deep Neural Networks** | *Link* **PPoPP 2018**
- Linnan Wang, Jinmian Ye, **Yiyang Zhao**, Wei Wu, Ang Li, Shuaiwen Leon Song, Zenglin Xu, Tim Kraska

23rd Principles and Practice of Parallel Programming

**Carbon-Efficient Neural Architecture Search** | *Link* **HotCarbon 2023**
- **Yiyang Zhao**, Tian Guo

2nd Workshop on Sustainable Computer Systems

**A New Method of Tipping Calibration for Ground-based Microwave Radiometer in Cloudy Atmosphere** | *Link* **TGRS 2014**
- Jiangman Li, Lixin Guo, Leke Lin, **Yiyang Zhao**, Xianhai Cheng

IEEE Transactions on Geoscience and Remote Sensing

**A dual-frequency method of eliminating liquid water radiation to remotely sense cloudy atmosphere by ground-based microwave radiometer** | *Link* **PIER 2013**
- Jiangman Li, Lixin Guo, Leke Lin, **Yiyang Zhao**, Zhenwei Zhao, Tingting Shu, Hengmin Han

Progress In Electromagnetics Research

## Academic Services

- Reviewer, International Conference on Machine Learning (ICML), 2024
- Reviewer, Neural Information Processing Systems (NeurIPS), 2022, 2023, 2024
- Reviewer, International Conference on Learning Representations (ICLR), 2024
- Reviewer, IEEE International Conference on Data Engineering (ICDE), 2023
- Reviewer, Elsevier Neural Networks, 2022
- Reviewer, IEEE Transactions on Evolutionary Computation, 2022