

Hao Yu

☎ +86 15955190101 | @ yuh@lamda.nju.edu.cn | 🌐 <http://www.lamda.nju.edu.cn/yuh>

EDUCATION

Nanjing University

Nanjing, China

Ph.D. in School of Artificial Intelligence;

Sep. 2019 – Dec. 2024 / Mar. 2025 (Expected)

LAMDA Group, State Key Laboratory for Novel Software Technology;

- **Supervisor:** [Prof. Jianxin Wu](#).
- **Dissertation:** Research on Approaches for Deep Model Compression, especially vision, language and recommendation models.

University of Science and Technology of China

Hefei, China

Bachelor in Computer Science, School of the Gifted Young;

Sep. 2015 – Jun. 2019

- **Dissertation:** Research on Accelerated Machine Learning Algorithms, especially K-Means Algorithm.

PUBLICATIONS

Reviving Undersampling for Long-Tailed Learning [[arXiv](#) | [code](#)]

- [Hao Yu](#), Yingxiao Du, Jianxin Wu.
- In arXiv preprint arXiv:2401.16811, 2024.

Unified Low-rank Compression Framework for Click-through Rate Prediction [[arXiv](#) | [code](#)]

- [Hao Yu](#), Minghao Fu, Jiandong Ding, Yusheng Zhou, Jianxin Wu.
- Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2024, CCF A Conference).

Compressing Transformers: Features Are Low-Rank, but Weights Are Not! [[Paper](#)]

- [Hao Yu](#), Jianxin Wu.
- Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023, CCF A Conference).
- This paper is awarded the Huawei Spark Award.

Training Vision Transformers with Only 2040 Images [[Paper](#) | [arXiv](#) | [code](#)]

- Yun-Hao Cao, [Hao Yu](#), Jianxin Wu.
- Proceedings of the 17th European Conference on Computer Vision (ECCV 2022, CCF B Conference).

A Unified Pruning Framework for Vision Transformers [[Paper](#) | [arXiv](#) | [code](#)]

- [Hao Yu](#), Jianxin Wu.
- SCIENCE CHINA Information Sciences (SCIS 2022, CCF A Journal).

Mixup without Hesitation [[Paper](#) | [arXiv](#) | [code](#)]

- [Hao Yu](#), Jianxin Wu.
- Proceedings of the 11th International Conference on Image and Graphics (ICIG 2021, Ei).

Fast K-Means Clustering with Anderson Acceleration [[arXiv](#)]

- Juyong Zhang, Yuxin Yao, Yue Peng, [Hao Yu](#), Bailin Deng.
- In arXiv preprint arXiv:1805.10638, 2018.

RESEARCH INTERESTS

My research interests include Machine Learning and Deep Learning. Currently, he focuses on:

- **Model Compression/Acceleration:** Quantization, pruning, low-rank approximation, knowledge distillation of vision, language and recommendation models.
- **Image Related Tasks:** Data Augmentation, Classification, Detection.

PROFESSIONAL EXPERIENCE

Alibaba, PAI Group

Research Intern;

Hangzhou, China

May 2023 – May 2024

- Worked with Mentor Shen Li and other colleagues on the compression of large language models;
- Propose a plug-and-play large language model quantization algorithm, which can be combined with various PTQ and QAT algorithms to improve the performance of low-bit quantization LLM. Submit to NeurIPS2024 and has been applied for a patent;
- Propose a fast compression algorithm for large language model KV Heads, which can effectively reduce the size of KV Caches. Submit to NeurIPS2024, and has been applied for a patent;

Minieye

Computer Vision Algorithm Intern;

Nanjing, China

Feb. 2019 – Aug. 2019

- Worked with Dr. Yizhang Xia and other colleagues on the automobile high beam intelligent control system;
- Responsible for the design and training of the tracker module;
- Responsible for managing the collection and labeling team of training data;
- The proposed high beam control system has been successfully applied to various ADAS L2 platforms;
- The work results have been applied for a patent, the application number is 201910933929.2;

CONTESTS & AWARDS

Contests

- Honorable mention in the competition of 2020 DIGIX Global AI Challenge. 2020.

Awards

- Huawei Spark Award. 2022.
- Excellent Graduate Cadre of Nanjing University. 2021.
- Excellent Youth League Member of Nanjing University. 2021.
- Xu Xin International Student Exchange Scholarship (Award 20000\$). 2019.
- Talent Program in Computer Science and Technology of USTC. 2016, 2017.

PROFESSIONAL SERVICE

Conference Reviewers

- CVPR2024, WWW2023, NeurIPS2023, AAAI2023, CVPR2022, ECCV2022, PAKDD2022, ECML2021, AAAI2021, IJCAI2021, ICPR2020.

Journal Reviewers

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).
- Knowledge and Information Systems (KIS).

Teaching Assistant

- Pattern Recognition, Spring, 2021.
- Data Mining for Complex Data Objects, Autumn, 2020.

SKILLS

Programming: Python, C/C++, Shell, LaTeX

Technologies: Git, PyTorch, TensorFlow, Caffe, NumPy, Matplotlib, Pandas

Operating Systems: Ubuntu, Windows