ZHENMEI SHI

Email: zhmeishi@gmail.com Tel: (+1) 608-698-2223 Homepage: http://zhmeishi.github.io/

PROFESSIONAL EXPERIENCE

Research Scientist *at* **Voyage AI**, Palo Alto, CA *01/2025-Now*

- Retrieval-Augmented Generation (RAG) Report to: **Tengyu Ma**

EDUCATION

University of Wisconsin-Madison, US
Ph.D. in Computer Sciences (Advised by **Yingyu Liang**) *09/2019-12/2024*
Master of Science in Computer Sciences *09/2019-12/2022*

Hong Kong University of Science and Technology, Hong Kong *09/2015-05/2019*
Bachelor of Science in Major GPA: 4.190 / 4.3
Computer Science and Pure Mathematics Advanced GPA: 3.974 / 4.3 (Top 2%)

ETH Zurich, Switzerland (Exchange) *02/2018-08/2018*

PHD THESIS

Understanding Training and Adaptation in Feature Learning: From Two-Layer Networks to Foundation Models 2024

PUBLICATIONS

* denotes equal contribution or alphabetical order.

1. Discovering the Gems in Early Layers: Accelerating Long-Context LLMs with 1000x Input Token Reduction
   Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, Shafiq Joty
   https://arxiv.org/abs/2409.17422 2024
2. Differential Privacy Mechanisms in Neural Tangent Kernel Regression
   Jiuxiang Gu*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2407.13621 WACV 2025
3. Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models
   Jiayu Wang, Yifei Ming, **Zhenmei Shi**, Vibhav Vineet, Xin Wang, Yixuan Li, Neel Joshi
   https://openreview.net/forum?id=cvaSru8LeO NeurIPS 2024
4. Multi-Layer Transformers Gradient Can be Approximated in Almost Linear Time
   Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
   https://openreview.net/forum?id=1LJIPZ4SvS Workshop NeurIPS 2024
5. Tensor Attention Training: Provably Efficient Learning of Higher-order Transformers
   Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
   https://openreview.net/forum?id=1vrVaR7FDC Workshop NeurIPS 2024
6. A Tighter Complexity Analysis of SparseGPT
   Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://openreview.net/forum?id=443oNjXhsT Workshop NeurIPS 2024

7. Differential Privacy of Cross-Attention with Provable Guarantee
   Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
   https://openreview.net/forum?id=GttuYQVARs Workshop NeurIPS 2024

8. Do Large Language Models Have Compositional Ability? An Investigation into Limitations and Scalability
   Zhuoyan Xu*, **Zhenmei Shi***, Yingyu Liang
   https://openreview.net/forum?id=iI1CzEhEMU COLM 2024

9. Why Larger Language Models Do In-context Learning Differently?
   Zhenmei Shi, Junyi Wei, Zhuoyan Xu, Yingyu Liang
   https://openreview.net/forum?id=WOa96EG26M ICML 2024

10. Towards Few-Shot Adaptation of Foundation Models via Multitask Finetuning
    Zhuoyan Xu, **Zhenmei Shi**, Junyi Wei, Fangzhou Mu, Yin Li, Yingyu Liang
    https://openreview.net/forum?id=1jbh2e0b2K ICLR 2024

11. Fourier Circuits in Neural Networks and Transformers: A Case Study of Modular Arithmetic with Multiple Inputs
    Chenyang Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Tianyi Zhou*
    https://openreview.net/forum?id=Gf7nZAfiaB Workshop ICLR 2024

12. Domain Generalization via Nuclear Norm Regularization
    Zhenmei Shi*, Yifei Ming*, Ying Fan*, Frederic Sala, Yingyu Liang
    https://openreview.net/forum?id=hJd66ZzXEZ **Oral** CPAL 2024

13. Provable Guarantees for Neural Networks via Gradient Feature Learning
    Zhenmei Shi*, Junyi Wei*, Yingyu Liang
    https://openreview.net/forum?id=5F04bU79eK NeurIPS 2023

14. A Graph-Theoretic Framework for Understanding Open-World Semi-Supervised Learning
    Yiyou Sun, **Zhenmei Shi**, Yixuan Li
    https://openreview.net/forum?id=ZITOHWeAy7 **Spotlight** NeurIPS 2023

15. When and How Does Known Class Help Discover Unknown Ones? Provable Understandings Through Spectral Analysis
    Yiyou Sun, **Zhenmei Shi**, Yingyu Liang, Yixuan Li
    https://openreview.net/forum?id=JHodnaW5WZ ICML 2023

16. The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning
    Zhenmei Shi*, Jiefeng Chen*, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, Somesh Jha
    https://openreview.net/forum?id=rvsbw2YthH_ (Accept Rate: 7.95%) **Spotlight** ICLR 2023

17. A Theoretical Analysis on Feature Learning in Neural Networks: Emergence from Inputs and Advantage over Fixed Features
    Zhenmei Shi*, Junyi Wei*, Yingyu Liang
    https://openreview.net/forum?id=wMpS-Z_AI_E ICLR 2022

18. Attentive Walk-Aggregating Graph Neural Networks
    Mehmet F. Demirel, Shengchao Liu, Siddhant Garg, **Zhenmei Shi**, Yingyu Liang
    https://openreview.net/forum?id=TWSTyYd2Rl TMLR 2022

19. Deep Online Fused Video Stabilization
    Zhenmei Shi, Fuhao Shi, Wei-Sheng Lai, Chia-Kai Liang, Yingyu Liang
    https://zhmeishi.github.io/dvs/ WACV 2022

20. Structured Feature Learning for End-to-End Continuous Sign Language Recognition

Zhaoyang Yang*, **Zhenmei Shi***, Xiaoyong Shen, Yu-Wing Tai
https://arxiv.org/abs/1908.01341 News, 2019

PREPRINT

* denotes equal contribution or alphabetical order.

1. Theoretical Constraints on the Expressive Power of RoPE-based Tensor Attention Transformers
   Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Mingda Wan*
   https://arxiv.org/abs/2412.18040 2024
2. Fast Gradient Computation for RoPE Attention in Almost Linear Time
   Yifang Chen*, Jiayan Huo*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2412.17316 2024
3. Curse of Attention: A Kernel-Based Perspective for Why Transformers Fail to Generalize on Time Series Forecasting and Beyond
   Yekun Ke*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Chiwun Yang*
   https://arxiv.org/abs/2412.06061 2024
4. The Computational Limits of State-Space Models and Mamba via the Lens of Circuit Complexity
   Yifang Chen*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2412.06148 2024
5. On the Expressive Power of Modern Hopfield Networks
   Xiaoyu Li*, Yuanpeng Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2412.05562 2024
6. Circuit Complexity Bounds for RoPE-based Transformer Architecture
   Bo Chen*, Xiaoyu Li*, Yingyu Liang*, Jiangxuan Long*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2411.07602 2024
7. Bypassing the Exponential Dependency: Looped Transformers Efficiently Learn In-context by Multi-step Gradient Descent
   Bo Chen*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2410.11268 2024
8. Beyond Linear Approximations: A Novel Pruning Approach for Attention Matrix
   Yingyu Liang*, Jiangxuan Long*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
   https://arxiv.org/abs/2410.11261 2024
9. Advancing the Understanding of Fixed Point Iterations in Deep Neural Networks: A Detailed Analytical Study
   Yekun Ke*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
   https://arxiv.org/abs/2410.11279 2024
10. Looped ReLU MLPs May Be All You Need as Practical Programmable Computers
    Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
    https://arxiv.org/abs/2410.09375 2024
11. Fine-grained Attention I/O Complexity: Comprehensive Analysis for Backward Passes
    Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
    https://arxiv.org/abs/2410.09397 2024
12. HSR-Enhanced Sparse Attention Acceleration
    Bo Chen*, Yingyu Liang*, Zhizhou Sha*, **Zhenmei Shi***, Zhao Song*
    https://arxiv.org/abs/2410.10165 2024

13. Fast John Ellipsoid Computation with Differential Privacy Optimization
Jiuxiang Gu*, Xiaoyu Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Junwei Yu*
https://arxiv.org/abs/2408.06395 2024

14. Toward Infinite-Long Prefix in Transformer
Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Chiwun Yang*
https://arxiv.org/abs/2406.14036 2024

15. Unraveling the Smoothness Properties of Diffusion Models: A Gaussian Mixture Perspective
Yingyu Liang*, **Zhenmei Shi***, Zhao Song*, Yufa Zhou*
https://arxiv.org/abs/2405.16418 2024

16. Conv-Basis: A New Paradigm for Efficient Attention Inference and Gradient Computation in Transformers
Yingyu Liang*, Heshan Liu*, **Zhenmei Shi***, Zhao Song*, Zhuoyan Xu*, Junze Yin*
https://arxiv.org/abs/2405.05219 2024

17. Exploring the Frontiers of Softmax: Provable Optimization, Applications in Diffusion Model, and Beyond
Jiuxiang Gu*, Chenyang Li*, Yingyu Liang*, **Zhenmei Shi***, Zhao Song*
https://arxiv.org/abs/2405.03251 2024

18. Dual Augmented Memory Network for Unsupervised Video Object Tracking
Zhenmei Shi*, Haoyang Fang*, Chi-Keung Tang, Yu-Wing Tai
https://arxiv.org/abs/1908.00777 https://zhmeishi.github.io/DAWN/, 2019

INTERNSHIP EXPERIENCE

Research Intern *at* **Google Cloud AI**, Sunnyvale, CA *10/2024-12/2024*

- Post-Training of LLMs.
  Supervised by: Sercan Arik

AI Research Scientist Intern *at* **Salesforce**, Palo Alto, CA *06/2024-09/2024*

- Long Context LLM Understanding, In-context Learning, and Inference Acceleration.
  Supervised by: Shafiq Joty

Research Scientist Intern *at* **Adobe**, Seattle, WA *12/2023-05/2024*

- Large Model (LLM, VLM, Diffusion) Emergent Ability and Learning Ability.
  Supervised by: Zhao Song

Software Engineering Intern *at* Google YouTube Ads ML, Mountain View, CA *06/2021-09/2021*

- Unsupervised Clustering in Recommendation System. Supervised by: Myra Nam

Software Engineering Intern *at* Google Pixel Camera, Mountain View, CA *05/2020-08/2020*

- Deep Video Stabilization. Supervised by: Fuhao Shi

Research Intern *at* Megvii (Face++) Foundation Model, Beijing *06/2019-08/2019*

- Neural Architecture Search. Supervised by: Xiangyu Zhang

Machine Learning Research Intern *at* Tencent YouTu, Shenzhen *12/2018-02/2019*

- Sign Language Recognition. Supervised by: Yu-Wing Tai News

Machine Learning Reseach Intern *at* Tencent YouTu, Shenzhen *12/2017-02/2018*

- Deep Colorization. Supervised by: Yu-Wing Tai News

Research Assistant *at* UW-Madison CS School *09/2019 - 12/2024*
Research Assistant *at* HKUST CS Department *02/2019-06/2019*
Research Assistant *at* Oak Ridge National Lab, US *05/2017-08/2017*
Part-Time Programmer *at* CUHK, HKBU, HKUST *06/2016-08/2016, 04/2017, 12/2018*

ACADEMIC SERVICES

Conference Reviewer *at* ICLR 2022-2025, AIStats 2025, NeurIPS 2022-2024, ICML 2022 and 2024-2025, WACV 2022 and 2025, ICCV 2021-2023, ECCV 2020-2022, CVPR 2021-2022 and 2025
Journal Reviewer *at* JVCI, IEEE Transactions on Information Theory

TEACHING EXPERIENCE

Teaching Assistant of CS220 (Data Programming I) *at* UW-Madison *Spring 2020*
Teaching Assistant of CS301 (Intro to Data Programming) *at* UW-Madison *Fall 2019*

EXTRA-CURRICULAR

Visiting Student *at* Microsoft Research Asia (MSRA) *08/2017*
Mentee *at* Hong Kong X-Tech *2017-2018*
Vice Chairman *at* Microsoft Student Club @ HKUST *2017-2018*
S.S. Chern Class Member *at* HKUST Mathematics Department *2016-2019*
Volunteer Teacher *at* Ociva, Maldives *12/2016-01/2017*
Software Team Member *at* HKUST Robotics Team *2015-2016*

AWARDS & SCHOLARSHIPS

NeurIPS 2023 Scholar Award *10/2023*
Student Research Grants Competition *from* UW-Madison *04/2023*
ICLR 2023 Financial Assistance *03/2023*
CS Departmental Scholarship *from* UW-Madison *09/2019*
Academic Achievement Medal *from* HKUST *11/2019*
S.S. Chern Class Achievement Scholarship *from* HKUST Math Department *07/2019*
Undergraduate Research Opportunity Program Award *from* HKUST *04/2019*
Best Student Team *from* HC2 (Switzerlands Biggest Programming Contest) *03/2018*
Champion of Micro Innovation Award *from* Tencent *02/2018*
1st Runner-up *from* HKUST x Radica Big Datathon *12/2017*
Technology Star *from* Beauty of Programming Competition held by MSRA *08/2017*
1st Prize of University's Scholarship for Undergraduate Student *from* HKUST *2016-2019*
Reaching Out Award *from* HKSAR Government Scholarship Fund *2017-2018*
Exchange Award, Lee Hysan Foundation Exchange Scholarship *from* HKUST *2017-2018*
The Cheng Foundation Scholarship for Mainland Students *from* HKUST *2016-2017*
Dean's List *from* HKUST *2015-2019*

TECHNICAL SKILLS

Python: JAX, TensorFlow, PyTorch, Numpy, sklearn, Pandas, BeautifulSoup, re, Spark
C++: Caffe, OpenCV, OpenMP, MPI
Java, MatLab, C, R, SQL, CUDA, LaTeX, Bash Script

INVITED TALKS

Provable Guarantees for Neural Networks via Gradient Feature Learning

- AI Time Idea Seminar, October 2023, Remote, `https://www.bilibili.com/video/BV1MG411y7gA/?vd_source=18fa90c33bc2626d02eca4a2c3df3601`

The Trade-off between Universality and Label Efficiency of Representations from Contrastive Learning

- AI Time Idea Seminar, June 2023, Remote, https://www.bilibili.com/video/BV1eo4y1T7Zb/?vd_source=18fa90c33bc2626d02eca4a2c3df3601
- MLOPT Idea Seminar, March 2023, Madison, USA, https://mlopt.ece.wisc.edu/idea-seminar/