Shuai Zhang

+1 408 858 6909 | @ cheungshuai@outlook.com | California, US

Current Position

Applied Scientist, Amazon Web Services AI — Santa Clara, US
Focus on foundation models related research, open-source projects, and products. — *Jan 2022 – Present*

- Products: Amazon TITAN, GLUE, Amazon Q, AutoGluon, D2L, Opensearch, etc.
- Research projects: CaMML, PipeRAG, POMP, CoMM, etc.

Academic Background

ETH Zurich — Zurich, Switzerland
Postdoc Researcherof department of Computer Science, mentored by Prof. Ce Zhang — *Feb 2020 – Jan 2022*

University of New South Wales — Sydney, Australia
PhD in Computer Science, mentored by Prof. Lina Yao — *2016 – 2019*

Nanjing University — Nanjing, China
Bachelor in Information System and Management — *2010 – 2014*

Honors and Awards

Area Chair Award, *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*	*2024*
Best of the Best Conference Paper Award, *Global Marketing Conference (GMC)*	*2023*
Outstanding Paper Award, *International Conference on Learning Representations (ICLR)*	*2021*
Best Paper Runner-up, *International Conference on Web Search and Data Mining (WSDM)*	*2020*
World's Top 2% Most Cited Scientist	*2021, 2022, 2023*
CSIRO Data61 PhD Scholarship	*2016-2019*

Publications

5.8k Google Scholar Citations, Equal contribution→∗, Corresponding author→♠.

1. Unraveling the Gradient Descent Dynamics of Transformers. Bingqing Song, Boran Han, **Shuai Zhang**, Jie Ding, Mingyi Hong. *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
2. CaMML: Context-Aware Multimodal Learner for Large Models (Area Chair Award). Yixin Chen∗, **Shuai Zhang**∗, Boran Han, Tong He, Bo Li. *The 62nd Annual Meeting of the Association for Computational Linguistics (ACL),* 2024
3. Discovering Bias in Latent Space: An Unsupervised Debiasing Approach. Dyah Adila, **Shuai Zhang**(♠), Boran Han, Bernie Wang. *Forty-first International Conference on Machine Learning (ICML)*, 2024.
4. Transferring Knowledge From Large Foundation Models to Small Downstream Models. Shikai Qiu, Boran Han, Danielle C. Maddix, **Shuai Zhang**, Bernie Wang, Andrew Gordon Wilson. *Forty-first International Conference on Machine Learning (ICML)*, 2024.
5. CoMM: Collaborative Multi-Agent, Multi-Reasoning-Path Prompting for Complex Problem Solving. Pei Chen, **Shuai Zhang**(♠), Boran Han. *The North American Chapter of the Association for Computational Linguistics (NAACL) Findings*, 2024.

6. Bridging Sources in Geospatial Sensing with Cross Sensor Pretraining. Boran Han, **Shuai Zhang**, Xingjian Shi, Markus Reichstein. *IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2024.
7. Data Augmentation for Object Detection via Controllable Diffusion Models. Haoyang Fang, Boran Han∗, **Shuai Zhang**∗, Su Zhou∗, Cuixiong Hu, Wen-Ming Ye. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024.
8. Prompt Pre-Training with Twenty-Thousand Classes for Open-Vocabulary Visual Recognition. Shuhuai Ren, Aston Zhang, Yi Zhu, **Shuai Zhang**, Shuai Zheng, Mu Li, Alex Smola, Xu Sun. *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
9. Data-Informed Geometric Space Selection. **Shuai Zhang**, Wenqi Jiang. *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
10. Rethinking Document-Level Relation Extraction: A Reality Check. Jing Li, Yequan Wang, **Shuai Zhang**, Min Zhang. *The 61st Annual Meeting of the Association for Computational Linguistics* (*ACL*) Findings, 2023
11. Co-design Hardware and Algorithms for Vector Search. Wenqi Jiang, Shigang Li, Yu Zhu, Johannes de Fine Licht, Zhenhao He, Runbin Shi, Cedric Renggli, **Shuai Zhang**, Theodoros Rekastsinas, Torsten Hoefler, Gustavo Alonso. *International Conference for High Performance Computing, Networking, Storage, and Analysis* (*SC*), 2023
12. Recommending the right Product to the right Customer with the right Message via Structural Neural Recommender (Best of the Best Conference Paper Award). **Shuai Zhang**, Lina Yao, Tianying Song, Jake An, Steven Lu. *Global Marketing Conference at Seoul*, 2023
13. xFraud: Explainable Fraud Transaction Detection. Susie Xi Rao, **Shuai Zhang**, Zhichao Han, Zitao Zhang, Wei Min, Zhiyao Chen, Yinan Shan, Yang Zhao, Ce Zhang. *The Proceedings of the VLDB Endowment* (*VLDB*), 2022.
14. Neural Methods for Logical Reasoning over Knowledge Graphs. Alfonso Amayuelas, **Shuai Zhang**(mentor), Susie Xi Rao, Ce Zhang. *The International Conference on Learning Representations* (*ICLR*), 2022. (co-author as a master thesis advisor).
15. Book Chapter: Recommender Systems. **Shuai Zhang**, Aston Zhang, Lina Yao. *Handbook of Machine Learning for Data Science*. Springer, 2022
16. Beyond Fully-Connected Layers with Quaternions: Parameterization of Hypercomplex Multiplications with 1/n Parameters (Outstanding Paper Award). Aston Zhang, Yi Tay, **Shuai Zhang**, Alvin Chan, Anh Tuan Luu, Siu Hui, Jie Fu. *The Ninth International Conference on Learning Representations* (*ICLR*), 2021
17. Knowledge Router: Learning Disentangled Representations for Knowledge Graphs. **Shuai Zhang**, Xi Rao, Yi Tay and Ce Zhang. *The Annual Conference of the North American Chapter of the Association for Computational Linguistics* (*NAACL-HLT*), 2021
18. Self-Instantiated Recurrent Units with Dynamic Soft Recursion. Aston Zhang, Yi Tay, Yikang Shen, Alvin Chan, **Shuai Zhang**. *Thirty-fifth Conference on Neural Information Processing Systems* (*NeurIPS*), 2021
19. Learning User Representations with Hypercuboids for Recommender Systems. **Shuai Zhang**, Huoyu Liu, Aston Zhang, Yue Hu, Ce Zhang, Yumeng Li, Tanchao Zhu, Shaojian He, Wenwu Ou. *The 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 2021
20. Book Chapter: Deep Learning for Recommender Systems. **Shuai Zhang**, Yi Tay, Lina Yao, Aixin Sun, Ce Zhang. *Recommender Systems Handbook, Third Edition*. Springer. 2021

21. DeGNN: Improving Graph Neural Networks with Graph Decomposition. Xupeng Miao∗, Nezihe Merve GÃijrel∗, Wentao Zhang, Zhichao Han, Bo Li, Wei Min, Susie Xi Rao, Hansheng Ren, Yinan Shan, Yingxia Shao, Yujie Wang, Fan Wu, Hui Xue, Yaming Yang, Zitao Zhang, Yang Zhao, **Shuai Zhang**, Yujing Wang, Bin Cui, Ce Zhang. *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (*KDD*), 2021

22. FleetRec: Large-Scale Recommendation Inference on Hybrid GPU-FPGA Clusters. Wenqi Jiang, Zhenhao He, **Shuai Zhang**, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, Gustavo Alonso. *The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (*KDD*), 2021

23. On Orthogonality Constraints for Transformers. Aston Zhang, Alvin Chan, Yi Tay, Jie Fu, Shuohang Wang, **Shuai Zhang**, Huajie Shao, Shuochao Yao and Roy Ka-Wei Lee. *The Annual Meeting of the Association for Computational Linguistics* (*ACL*) short paper, 2021

24. MicroRec: Accelerating Deep Recommendation Systems to Microseconds by Hardware and Data Structure Solutions. Wenqi Jiang, Zhenhao He, **Shuai Zhang**, Thomas B. PreuÃ§er, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, Gustavo Alonso. *The Fourth Conference on Machine Learning and Systems (MLSys)*, 2021

25. Ease.ML: A Lifecycle Management System for Machine Learning. Leonel Aguilar Melgar, David Dao, Shaoduo Gan, Nezihe Merve GÃijrel, Nora Hollenstein, Jiawei Jiang, Bojan KarlaÅą, Thomas Lemmin, Tian Li, Yang Li, Xi Rao, Johannes Rausch, Cedric Renggli, Luka Rimanic, Maurice Weber, **Shuai Zhang**, Zhikuan Zhao, Kevin Schawinski, Wentao Wu, Ce Zhang. *The Conference on Innovative Data Systems Research (CIDR)*, 2021

26. Suspicious Massive Registration Detection via Dynamic Heterogeneous Graph Neural Networks. Xi Rao, **Shuai Zhang**, Zhichao Han, Zitao Zhang, Wei Min, Mo Cheng, Yinan Shan, Yang Zhao, Ce Zhang. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*, 2021

27. Gradient Boosted Neural Decision Forest. Manqing Dong, Lina Yao, Xianzhi Wang, Boualem Benatallah, **Shuai Zhang**, Quan Z. Sheng. *IEEE Transactions on Services Computing*, 2021

28. Symmetric Metric Learning with Adaptive Margin for Recommendation. Mingming Li, **Shuai Zhang**, Fuqing Zhu, Liangjun Zang, Wanhui Qian, Jizhong Han, Songlin Hu. *The 34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020

29. HyperML: A Boosting Metric Learning Approach in Hyperbolic Space for Recommender Systems (Best Paper Award Runner-Up). Lucas Vinh Tran, Yi Tay, **Shuai Zhang**, Gao Cong, Xiaoli Li. *The ACM International Conference on Web Search and Data Mining (WSDM)*, 2020

30. Quaternion Knowledge Graph Embeddings. **Shuai Zhang**, Yi Tay, Lina Yao, Qi Liu. *The Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, 2019

31. Deep Learning based Recommender System: A Survey and New Perspectives. **Shuai Zhang**, Lina Yao, Aixin Sun, Yi Tay. *ACM Computing Surveys (CSUR)*, 2019

32. Quaternion Collaborative Filtering for Recommendation. **Shuai Zhang**, Lina Yao, Lucas Vinh Tran, Aston Zhang, Yi Tay. *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019

33. Lightweight and Efficient Neural Natural Language Processing with Quaternion Networks. Yi Tay, Aston Zhang, Anh Tuan Luu, Jinfeng Rao, **Shuai Zhang**, Shuohang Wang, Jie Fu, Siu Cheung Hui. *The 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019

34. Holographic Factorization Machines for Recommendation. **Shuai Zhang**∗, Yi Tay∗, Anh Tuan Luu, Siu Cheung Hui, Lina Yao, Vinh Tran. *The 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019

35. Next Item Recommendation with Self-Attentive Metric Learning. **Shuai Zhang**, Yi Tay, Lina Yao, Aixin Sun, Jake An. *AAAI Workshop on Recommender Systems Meet Natural Language Processing*, 2019

36. DeepRec: An Open-Source Toolkit for Deep Learning based Recommendation. **Shuai Zhang**, Yi Tay, Lina Yao, Bin Wu, Aixin Sun. *The 28th International Joint Conference on Artificial Intelligence (IJCAI Demo-track)*, 2019

37. Adaptive Cognitive Activity Recognition with Reinforced Attentive CNN. Xiang Zhang, Lina Yao, Xianzhi Wang, Wenjie Zhang, **Shuai Zhang**, Yunhao Liu. *The 19th IEEE International Conference on Data Mining (ICDM)*, 2019

38. Unraveling Metric Vector Spaces with Factorization for Recommendation. **Shuai Zhang**, Lina Yao, Bin Wu, Xiwei Xu, Xiang Zhang, Liming Zhu. *IEEE Transactions on Industrial Informatics*, 2019

39. Book Chapter: Deep Neural Networks meet Recommender Systems. **Shuai Zhang**, Lina Yao, Aixin Sun, Guibing Guo, Xiwei Xu, Liming Zhu. *Big Data Recommender Systems: Recent Trends and Advances*. Institution of Engineering and Technology, 2019

40. Thing-of-Interest Recommendation in the Internet of Things: Requirements, Challenges and Directions. Lina Yao, Xianzhi Wang, Quan Z. Sheng, Schahram Dustdar, **Shuai Zhang**. *IEEE Internet Computing*, 2019

41. NeuRec: On Nonlinear Transformation for Personalized Ranking. **Shuai Zhang**, Lina Yao, Aixin Sun, Sen Wang, Guodong Long, Manqing Dong. *The 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018

42. Internet of Things Meets Brain-Computer Interface: A Unified Deep Learning Framework for Enabling Human-Thing Cognitive Interactivity. Xiang Zhang, Lina Yao, **Shuai Zhang**, Salil S. Kanhere, Quan Z. Sheng, Yunhao Liu. *IEEE Internet of Things Journal*, 2018

43. Rating Prediction via Generative Convolutional Neural Networks based Regression. Xiaodong Ning, Lina Yao, Xianzhi Wang, Boualem Benatallah, Manqing Dong, **Shuai Zhang**. *Pattern Recognition Letters*, 2018

44. AutoSVD++: An Efficient Hybrid Collaborative Filtering Model via Contractive Auto-encoders. **Shuai Zhang**, Lina Yao, Xiwei Xu. *The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)* short paper, 2017

PREPRINTS

1. PipeRAG: Fast Retrieval-Augmented Generation via Algorithm-System Co-design. Wenqi Jiang, **Shuai Zhang**(♠), Boran Han, Jie Wang, Bernie Wang, Tim Kraska. 2024, https://arxiv.org/abs/2403.05676.

2. Lightweight In-context Tuning for Multimodal Unified Models. Yixin Chen, **Shuai Zhang**(♠), Boran Han, Jiaya Jia, 2023. https://arxiv.org/abs/2310.05109

REVIEWING AND OUTREACH

Invited Reviewer / Program Committee Member

Area Chair: ACL Rolling Review 2024, NeurIPS Workshop for Women in Machine Learning 2022

Organizer: RecSys 2024 Workshop - ROEGen

Senior PC: IJCAI 2021, CIKM 2021

Guest Editor: Frontiers of Big Data

Session Chair: CIKM 2021 (virtual)

OTHER EXPERIENCE

- Intern at Amazon Shanghai AI lab *2019.07-2020.01*
- Intern at Tencent *2018.10-2018.12*
- Software engineer at OPPO *2014.07-2016.06*

OPENSOURCE PROJECTS

- AutoGluon: https://auto.gluon.ai/ *6.8k GitHub stars*
- Dive into deep learning: https://d2l.ai/ *20k GitHub stars*
- DeepRec: cheungdaven/DeepRec *1k GitHub stars*