Shusen Wang

✉ Email: wssatzju@gmail.com
Homepage: wangshusen.github.io

About Me

I earned both my Bachelor's and Doctoral degrees in computer science from Zhejiang University, with the former being awarded in 2011 and the latter in 2016. Throughout my doctoral studies, I was honored with China's most prestigious fellowships and scholarships, namely the "Microsoft Research Asia Fellowship" and the "Baidu Scholarship". Subsequently, I continued my academic journey as a postdoctoral scholar at the Department of Statistics, University of California at Berkeley, from 2016 to 2018. I then served as a tenure-track assistant professor at the Department of Computer Science, Stevens Institute of Technology, from 2018 to 2021. In late 2021, I transitioned my career to the industry. I worked at Xiaohongshu (Shanghai) as a machine learning engineer and team leader, from 2021 to 2023. I worked on algorithmic trading for a short time, from 2023 to 2024. In 2024, I joint Meta as a staff engineer.

I possess a wide range of expertise encompassing machine learning, reinforcement learning, and numerical algorithms. I also have practical experience with industrial search engines, recommender systems, and NLP. In my leisure time, I published a book titled *Deep Reinforcement Learning (in Chinese)*, authored a book draft titled *Search Engines (in Chinese)*, and created an open course called *Industrial Recommender System (in Chinese)*.

Within the industry, my primary experience lies in search engines and NLP. From 2021 to 2022, I led the model team responsible for Xiaohongshu's search engine. During this time, my team successfully launched 14 experiments that positively impacted key performance indicators such as DAU, retention, and CTR, while also reducing CPU/GPU costs. From 2022 to 2023, I lead the NLP team, which supports various areas including search, recommendation, e-commerce, user growth, etc.

Work Experience

2024 – Now Meta (CA, USA), staff engineer
2021 – 2023 Xiaohongshu (Shanghai, China), machine learning engineer & manager
2018 – 2021 Stevens Institute of Technology (NJ, USA), Assistant Professor in Computer Science
2016 – 2018 UC Berkeley (CA, USA), Postdoc
2014 – 2015 Baidu Big Data Lab (Beijing, China), intern
2012 – 2012 Google (Beijing, China), intern
2011 – 2012 Microsoft Research Asia (Beijing, China), intern

Education

2011 – 2016 Zhejiang University (Hangzhou, China), Ph.D. in Computer Science
2007 – 2011 Zhejiang University (Hangzhou, China), B.Eng. in Computer Science

Industry Projects

2022 Improving CTR model for search ranking

- My team focused on enhancing the click-through rate (CTR) model for search ranking. We introduced several improvements to the neural network architecture and feature selection, resulting in significant increases in query CTR, document CTR, and overall user engagement.
- My team transitioned from CPU-based training and inference to GPU-based training and inference, consequently reducing costs and response times.

2022 Locality-sensitive search intents

- Over 1% of queries on Xiaohongshu were aimed at discovering nearby points of interest (POIs). However, our search engine seemed to be overlooking such queries and not catering to this specific user intent.
- In response to the issue, I initiated a project to address the needs of users searching for nearby POIs. I took the lead in designing the project pipeline, encompassing query understanding, retrieval, and ranking processes.

- As a result, we achieved a substantial increase in query CTR for searches with nearby intent. Due to its success, the project was recognized with the 2022 Q3 ExtraMile Prize, standing out as one of the top five projects within the company.

2022-2023 NLP techniques for search retrieval

- I conceptualized an innovative *inverse retrieval* method, which generates highly relevant queries offline and constructs an index to map queries to lists of related documents. My team implemented and launched this inverse retrieval strategy, leading to a notable 0.1% increase in both the app's daily active users (DAU) and user retention.
- My team used the inverse retrieval method to enhance the retrieval of newly published documents. When a new document is published, our nearline pipeline generates relevant queries for it and adds the corresponding $\langle \text{query}, \text{doc} \rangle$ pair to the index of query $\rightarrow$ List $\langle \text{doc} \rangle$. This project significantly improved the 24-hour new document impression ratio and led to a slight increase in both query and document click-through rates (CTR).

2022-2023 Offline search retrieval pipelines

- My team developed and implemented an innovative *cached retrieval* method. We maintain a table consisting of the top 5 million queries, and for each of these queries, we perform an offline analysis of the search log to extract highly relevant documents with impressive content quality, freshness, and substantial click numbers. The index for query $\rightarrow$ List $\langle \text{doc} \rangle$ is updated daily, serving as a retrieval channel. As a result of this project, we increased query click-through rates (CTRs).
- I proposed an offline search pipeline that targets top queries and initiates proactive, non-personalized retrieval and ranking during nighttime. This process establishes a key-value (KV) index for query $\rightarrow$ List $\langle \text{doc} \rangle$. In the online stage, the offline computation results replace the non-personalized retrieval channel. The project further reduced the GPU costs related to relevance by 21%, as well as improving click-through rates and other key performance indicators.

2022-2023 Pretrained BERT models.

- My team pretrained both 12-layer and 48-layer BERT models on a combination of public data and proprietary Xiaohongshu data.
- These pretrained models have been applied to various use cases, including search relevance and search queries, exhibiting significant enhancements in performance.

2023 GPT models.

- Utilizing the open-source LLAMA model as a foundation, we conduct continal pretrain and finetuning.
- We invented RefGPT that generates multi-turn dialogues, and we use the dialogues for finetuning GPT models.
- We developed LLM-driven applications such as conversational search engine.

Book

- Deep Reinforcement Learning (in Chinese).
  Shusen Wang, Yujun Li, and Zhihua Zhang.
  Posts & Telecom Press Co.,Ltd, 2022.
- Search Engines (in Chinese).
  Shusen Wang.
  Book draft available at [here], 2023.

Journal Papers

- Fast Randomized-MUSIC for Mm-Wave Massive MIMO Radars.
  Bin Li, **Shusen Wang**, Jun Zhang, Xianbin Cao, and Chenglin Zhao.
  IEEE Transactions on Vehicular Technology, 70(2):1952-1956, 2021.
- Fast Pseudo-spectrum Estimation for Automotive Massive MIMO Radar.
  Bin Li, **Shusen Wang**, Zhiyong Feng, Jun Zhang, Xianbin Cao, and Chenglin Zhao.
  IEEE Internet of Things Journal, 2021.

- Randomized Approximate Channel Estimator in Massive-MIMO Communication.
  Bin Li, **Shusen Wang**, Xianbin Cao, Jun Zhang, and Chenglin Zhao.
  IEEE Communications Letters, 24(10):2314 - 2318, 2020.
- A Bootstrap Method for Error Estimation in Randomized Matrix Multiplication.
  Miles E. Lopes, **Shusen Wang**, Michael W. Mahoney.
  Journal of Machine Learning Research (**JMLR**), 20(39):1-40, 2019.
- Scalable Kernel K-Means Clustering with Nystrom Approximation: Relative-Error Bounds.
  Shusen Wang, Alex Gittens, and Michael W. Mahoney.
  Journal of Machine Learning Research (**JMLR**), 20(12):1-49, 2019.
- Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging.
  Shusen Wang, Alex Gittens, and Michael W. Mahoney.
  Journal of Machine Learning Research (**JMLR**), 18:1-50, 2018.
- Efficient Data-Driven Geologic Feature Characterization from Pre-stack Seismic Measurements using Randomized Machine-Learning Algorithm.
  Youzuo Lin, **Shusen Wang**, Jayaraman Thiagarajan, George Guthrie, and David Coblentz.
  Geophysical Journal International, ggy385, 2018.
- Alchemist: An Apache Spark <=> MPI Interface.
  Alex Gittens, Kai Rothauge, Michael W. Mahoney, **Shusen Wang**, Lisa Gerhardt, Prabhat, Jey Kottalam, Michael Ringenburg, and Kristyn Maschhoff.
  Concurrency and Computation Practice and Experience, Special Issue on the Cray User Group, 2018.
- Towards More Efficient SPSD Matrix Approximation and CUR Matrix Decomposition.
  Shusen Wang, Zhihua Zhang, and Tong Zhang.
  Journal of Machine Learning Research (**JMLR**), 17(210):1-49, 2016.
- SPSD Matrix Approximation vis Column Selection: Theories, Algorithms, and Extensions.
  Shusen Wang, Luo Luo, and Zhihua Zhang.
  Journal of Machine Learning Research (**JMLR**), 17(49):1-49, 2016.
- Improving CUR Matrix Decomposition and the Nystrom Approximation via Adaptive Sampling.
  Shusen Wang and Zhihua Zhang.
  Journal of Machine Learning Research (**JMLR**), 14: 2729-2769, 2013.
- EP-GIG Priors and Applications in Bayesian Sparse Learning.
  Zhihua Zhang, **Shusen Wang**, Dehua Liu, and Michael I. Jordan.
  Journal of Machine Learning Research (**JMLR**), 13: 2031-2061, 2012.

Conference Papers

- RefGPT: Dialogue Generation of GPT, by GPT, and for GPT.
  Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, **Shusen Wang**, Hai Zhao.
  In *EMNLP Findings* , 2023.
- Federated Reinforcement Learning with Environment Heterogeneity.
  Hao Jin, Yang Peng, Wenhao Yang, **Shusen Wang**, and Zhihua Zhang.
  In *Artificial Intelligence and Statistics* (**AISTATS**), 2022.
- Learning by Interpreting.
  Xuting Tang, Abdul Rafae Khan, **Shusen Wang**, and Jia Xu.
  In *International Joint Conference on Artificial Intelligence* (**IJCAI**), 2022.
- Matrix Sketching for Secure Collaborative Machine Learning.
  Mengjiao Zhang and **Shusen Wang**.
  In *International Conference on Machine Learning* (**ICML**), 2021.
- Communication-Efficient Distributed SVD via Local Power Iterations.
  Xiang Li, **Shusen Wang**, Kun Chen, and Zhihua Zhang.
  In *International Conference on Machine Learning* (**ICML**), 2021.

- On the Convergence of FedAvg on Non-IID Data.
  Xiang Li, Kaixuan Huang, Wenhao Yang, **Shusen Wang**, and Zhihua Zhang.
  In *International Conference on Learning Representations* (**ICLR**), 2020.
- Do Subsampled Newton Methods Work for High-Dimensional Data?
  Xiang Li, **Shusen Wang**, and Zhihua Zhang.
  In *AAAI Conference on Artificial Intelligence* (**AAAI**), 2020.
- Cola-GNN: Cross-Location Attention based Graph Neural Networks for Long-term ILI Prediction.
  Songgaojun Deng, **Shusen Wang**, Huzefa Rangwala, Lijing Wang, and Yue Ning.
  In *Conference on Information and Knowledge Management* (**CIKM**), 2020.
- Sharper Generalization Bound for the Divide-and-Conquer Ridge Regression.
  Shusen Wang.
  In *AAAI Conference on Artificial Intelligence* (**AAAI**), 2019.
- GIANT: Globally Improved Approximate Newton Method for Distributed Optimization.
  Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W. Mahoney.
  In *Advances in Neural Information Processing Systems* (**NIPS**), 2018.
- Error Estimation for Randomized Least-Squares Algorithms via the Bootstrap.
  Miles E. Lopes, **Shusen Wang**, and Michael W. Mahoney.
  In *International Conference on Machine Learning* (**ICML**), 2018.
- Accelerating Large-Scale Data Analysis by Offloading to High-Performance Computing Libraries using Alchemist.
  Alex Gittens, Kai Rothauge, **Shusen Wang**, Michael W. Mahoney, Lisa Gerhardt, Prabhat, Jey Kottalam, Michael Ringenburg, and Kristyn Maschhoff.
  In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD**), 2018.
- OverSketch: Approximate Matrix Multiplication for the Cloud.
  Vipul Gupta, **Shusen Wang**, Thomas Courtade, and Kannan Ramchandran.
  In *IEEE International Conference on Big Data*, 2018.
- Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging.
  Shusen Wang, Alex Gittens, and Michael W. Mahoney.
  In *International Conference on Machine Learning* (**ICML**), 2017.
- Towards Real-Time Geologic Feature Detection from Seismic Measurements using a Randomized Machine-Learning Algorithm.
  Youzuo Lin, **Shusen Wang**, Jayaraman Thiagarajan, George Guthrie, and David Coblentz.
  In *Proceeding of Society of Exploration Geophysics* (**SEG**), 2017.
- Open Domain Short Text Conceptualization: A Generative + Descriptive Modeling Approach.
  Yangqiu Song, **Shusen Wang**, and Haixun Wang.
  In *International Joint Conference on Artificial Intelligence* (**IJCAI**), 2015.
- Improving the Modified Nystrom Method Using Spectral Shifting.
  Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang.
  In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (**KDD**), 2014.
- Efficient Algorithms and Error Analysis for the Modified Nystrom Method.
  Shusen Wang and Zhihua Zhang.
  In *International Conference on Artificial Intelligence and Statistics*, (**AISTATS**), 2014.
- Making Fisher Discriminant Analysis Scalable.
  Bojun Tu, Zhihua Zhang, **Shusen Wang**, and Hui Qian.
  In *International Conference on Machine Learning* (**ICML**), 2014.
- Exact Subspace Clustering in Linear Time.
  Shusen Wang, Bojun Tu, Congfu Xu, and Zhihua Zhang.
  In *the 28th AAAI Conference on Artificial Intelligence* (**AAAI**), 2014.

- Using The Matrix Ridge Approximation to Speedup Determinantal Point Processes Sampling Algorithms.
  Shusen Wang, Chao Zhang, Hui Qian, and Zhihua Zhang.
  In *the 28th AAAI Conference on Artificial Intelligence* (**AAAI**), 2014.
- Transfer Understanding from Head Queries to Tail Queries.
  Yangqiu Song, Haixun Wang, Weizhu Chen, and **Shusen Wang**.
  In *ACM International Conference on Information and Knowledge Management* (**CIKM**), 2014.
- Nonconvex Relaxation Approaches to Robust Matrix Recovery.
  Shusen Wang, Dehua Liu, and Zhihua Zhang.
  In *International Joint Conference on Artificial Intelligence* (**IJCAI**), 2013.
- A Scalable CUR Matrix Decomposition Algorithm: Lower Time Complexity and Tighter Bound.
  Shusen Wang and Zhihua Zhang.
  In *Advances in Neural Information Processing Systems* (**NIPS**), 2012.
- Colorization by Matrix Completion.
  Shusen Wang and Zhihua Zhang.
  In *AAAI Conference on Artificial Intelligence* (**AAAI**), 2012.
- Efficient Subspace Segmentation via Quadratic Programming.
  Shusen Wang, Xiaotong Yuan, Tiansheng Yao, Shuicheng Yan, and Jialie Shen.
  In *AAAI Conference on Artificial Intelligence* (**AAAI**), 2011.

Teaching

2021 Fall	CS600: Advanced Algorithms
2021 Spring	CS583: Deep Learning (remote)
2020 Fall	CS600: Advanced Algorithms (remote), with students' rating of 3.90/4.0
2020 Spring	CS583: Deep Learning, with students' rating of 3.89/4.0
2019 Fall	CS583: Deep Learning, with students' rating of 3.83/4.0
2021 Spring	CS583: Deep Learning, with students' rating of 3.71/4.0
Open Course	YouTube Chinese Channel: `https://www.youtube.com/c/ShusenWang`
	YouTube English Channel: `https://www.youtube.com/c/ShusenWangEng`
	Bilibili Chinese Channel: `https://space.bilibili.com/1369507485`

Honors & Awards

2014	Baidu Scholarship, awarded to 8 Chinese students in the world, US$30,000
2013	Microsoft Research Asia Fellow, awarded to 10 students in Asia Pacific, US$10,000
2012	Scholarship Award for Excellent Doctoral Student Granted by Ministry of Education, US$5,000
2012 – 2014	National Scholarship for Graduate Students, 3 times, each time US$5,000

Academic Service

Journal Reviewer

- Journal of Machine Learning Research, 2015 – 2021
- SIAM Journal on Scientific Computing, 2017
- ACM Transactions on Mathematical Software, 2017
- Journal of Econometrics, 2017
- SIAM Journal on Matrix Analysis and Applications, 2017, 2019
- International Journal of Data Science and Analytics, 2018
- IEEE Transactions on Signal Processing, 2018
- IEEE Transactions on Information Theory, 2019
- IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 2020

Conference Committee Member

- NIPS 2014, 2015, 2017, 2018, 2020
- ICML 2017, 2018, 2019
- IJCAI 2015, 2017, 2018, 2019, 2020
- AAAI 2017, 2018, 2020
- AISTATS 2019, 2020
- UAI 2019, 2020
- Supercomputing 2019
- KDD 2020
- ICLR 2021

Conference Senior Committee Member

- AAAI 2021
- IJCAI 2021