Yonghui Xiao http://yohu.me

yohuxiao@gmail.com 404-772-7097
1 Hacker Way, Attn yhx, Menlo Park, CA 94025

EXPERIENCE

Tech Lead, Staff Software Engineer on AI Agent Dec 2024 - present
GenAI for monetization, Meta, Menlo Park, CA, USA

- Led the business AI agent creation from 0 to 1 based on LLM, RAG and toolset with APIs.
- Launched the AI agent for real-world customers (demo is available).

Tech Lead, Senior Software Engineer on Gemini ASR May 2019 - Dec 2024
DeepMind & Google, Mountain view, CA, USA (joined DeepMind in Oct 2024)

- **Best Gemini multimodal LLM for ASR**. I led several novel algorithms for multimodal LLM to tackle the fundamental problem of LLM, the hallucination under multimodality. My algorithms improved the multimodal Gemini 1.5 at both pre-training and fine-tuning stages and achieved the best Gemini ASR performance under extensive evaluations.
- **Personalized Gemini model**. I designed LoRA adapters to personalized Gemini models.
- **Best ASR model under federated learning (FL)**. I led the federated training of Conformer based ASR model which was deployed on millions of real-world smartphones. This was the largest Conformer model trained with FL in the real world.
- **Real-time query classification for Assistant**. I created both LLM based and pQRNN based query understanding models to classify the real-time Google Assistant queries. My models were deployed in the Google Assistant production with verified positive prod metrics.

Software Engineer May 2017 - May 2019
Google, Sunnyvale, CA, USA

- Gmail security. I fixed a massive Gmail security loophole that could affect billions of Gmail users. It took more than 1 year to re-program 100k+ Google prod servers and collaborate with more than 1000 Google teams to finish the massive infrastructure update.

iOS App Development May 2013 - June 2016
Based on my research of location privacy, I developed an iPhone app LocLok (search LocLok in Apple's app store; supported by NSF I-Corps award)

Download on the App Store

EDUCATION

Emory University, USA, Ph.D, Computer Science August 2011 - April 2017
Tsinghua University, China, Master, Computer Science September 2008 - July 2011
Xi'an Jiaotong University, China September 2000 – July 2005
Undergraduate student, 3 bachelor degrees (Computer science, mechanics and management)

PUBLICATIONS

Patent

- [Google patents] 3 in-progress Multimodal LLM patents that improves Gemini model quality
- [Google patent] Decentralized Learning of Large Machine Learning Models
- [Google patent] On-the-Fly Parameter Compression and Decompression to Facilitate Forward and/or Back Propagation at Clients during Federated Learning
- [PhD work] Methods and Systems for Determining Protected Location Information Based on Temporal Correlations

Conference and Journal Papers

- **Yonghui Xiao**, Yuxin Ding et al. Federated Learning of Large ASR Models in the Real World. https://arxiv.org/abs/2408.10443, 2024
- Xuan Kan, **Yonghui Xiao** et al. Parameter-Efficient Transfer Learning under Federated Learning for Automatic Speech Recognition. https://arxiv.org/abs/2408.11873, 2024
- Renkun Ni, **Yonghui Xiao** et al. FedAQT: Accurate Quantized Training with Federated Learning. ICASSP 2024
- Yuxin Ding, **Yonghui Xiao** et al. Improved Federated Learning for Handling Long-tail Words. Defensive publication.
- Tien-Ju Yang, **Yonghui Xiao** et al. Online model compression for federated learning with large models. ICASSP 2023.
- Rongmei Lin, Yonghui Xiao et al. Federated pruning: Improving neural network efficiency with federated learning. INTERSPEECH 2023.
- Dhruv Guliani, **Yonghui Xiao** et al. Enabling on-device training of speech recognition models with federated dropout. ICASSP 2022.
- Qiuchen Zhang, Jing Ma, **Yonghui Xiao**, Jian Lou, and Li Xiong. Broadening Differential Privacy for Deep Learning Against Model Inversion Attacks. IEEE BigData 2020.
- Yang Cao, **Yonghui Xiao**, Shun Takagi, Li Xiong, Masatoshi Yoshikawa, Yilin Shen, Jinfei Liu, Hongxia Jin, and Xiaofeng Xu. Customizable and Rigorous Location Privacy through Policy Graph, 25th European Symposium on Research in Computer Security (ESORICS), 2020
- Yang Cao, Shun Takagi, **Yonghui Xiao**, Li Xiong, Masatoshi Yoshikawa. PANDA: Policy-aware Location Privacy for Epidemic Surveillance. 46rd International Conference on Very Large Database (VLDB) demo 2020
- Yang Cao, **Yonghui Xiao**, Li Xiong, Liquan Bai and Masatoshi Yoshikawa. Protecting Spatiotemporal Event Privacy in Continuous Location-Based Services. IEEE Transactions on Data and Knowledge Engineering (TKDE) 2019
- Yang Cao, **Yonghui Xiao**, Li Xiong, Liquan Bai, Masatoshi Yoshikawa. PriSTE: Protecting Spatiotemporal Event Privacy in Continuous Location-Based Services. 45rd International Conference on Very Large Database (VLDB) demo 2019.
- Yang Cao, **Yonghui Xiao**, Li Xiong, Liquan Bai. PriSTE: From Location Privacy to Spatiotemporal Event Privacy. International Conference on Data Engineering (ICDE) 2019
- Yang Cao, Li Xiong, Masatoshi Yoshikawa, **Yonghui Xiao**, Si Zhang. ConTPL: Controlling Temporal Privacy Leakage in Differentially Private Continuous Data Release. VLDB, 2018
- Yang Cao, Masatoshi Yoshikawa, **Yonghui Xiao**, Li Xiong. Quantifying Differential Privacy in Continuous Data Release under Temporal Correlations. IEEE Transactions on Data and Knowledge Engineering, 2018
- **Yonghui Xiao**, Li Xiong, Si Zhang, Yang Cao. LocLok: Location Cloaking with Differential Privacy via Hidden Markov Model. International Conference on Very Large Database (VLDB), 2017
- Yang Cao, Masatoshi Yoshikawa, **Yonghui Xiao**, Li Xiong. Quantifying Differential Privacy under Temporal Correlations. IEEE International Conference on Data Engineering (ICDE), 2017
- Xiaofeng Xu, Li Xiong, Vaidy Sunderam, **Yonghui Xiao**. A Markov Chain Based Pruning Method for Predictive Range Queries. ACM SIGSPATIAL, 2016
- **Yonghui Xiao**, Li Xiong. Protecting Locations with Differential Privacy under Temporal Correlations. ACM Conference on Computer and Communications Security (CCS), 2015
- **Yonghui Xiao**, Li Xiong, Liyue Fan, Slawomir Goryczka, Haoran Li. DPCube: Differentially Private Histogram Release through Multidimensional Partitioning. Transactions on Data Privacy, 2014
- **Yonghui Xiao**, James Gardner, Li Xiong. DPCube: Releasing Differentially Private Data Cubes for Health Information. IEEE International Conference on Data Engineering (ICDE), 2012
- James Gardner, Li Xiong, **Yonghui Xiao**, Jingjing Gao, Andrew Post, Xiaoqian Jiang, Lucila Ohno-Machado. SHARE: System Design and Case Studies for Statistical Health Information Release. Journal of the American Medical Informatics Association (JAMIA), 2012
- **Yonghui Xiao**, Li Xiong, Chun Yuan. Differentially Private Data Release through Multidimensional Partitioning. 7th VLDB Workshop on Secure Date Management, 2010