Yu Chin Fabian Lim

Google Scholar (click) . Patents (click) . Github (click) ✉
fabian.lim@gmail.com . ☎ +65-9783-8760 . Singapore citizen .

Tech professional with 20+ years experience in industry and academic research. Wide background in Generative AI, Blockchain, and Hardware Acceleration.

Skills: large language models, generative AI, open source software, optimization, devops, project managmenet, system architecture & design

Professional Experience

Senior Research Scientist, IBM Research

Aug 2022 – Present, Singapore

- Worked on AI modeling fine-tuning stack `fms-hf-tuning` for training of generative AI models.
- Tech lead for `fms-acceleration`, a companion repository to the fms stack that enables quantization, distributed training (e.g., packing), triton kernels, etc.
- Contributed model tuning improvements and fixes to Huggingface opensource.
- Contributed to IBM `instructlab` for model training, enabling advanced features such as elastic checkpointing.

Skills: Python (Programming Language) · Open-Source Software · Artificial Intelligence (AI) · Generative AI

Research Scientist, IBM Research

Jul 2020 – Aug 2022, Singapore

- Worked on internal AIOps projects involving log analysis tools, deployed prototype to aid service engineers to investigate issues that customers face, for variety of products such as IBM WebSphere.

Skills: Scholarly Research · Artificial Intelligence (AI) · Applied Probability · Natural Language Processing (NLP) · Software Deployment

Research Manager, IBM Research

Aug 2018 – Feb 2020, Singapore

- Led a team on asset tokenization platform on Hyperledger Fabric, to deliver a minimum viable research prototype.
- People management for a team of scientist and engineers, and also play a key role for the leadership and growth of the lab.

Skills: R&D Management · Software Project Management

Research Staff Member / Scientist

Mar 2016 - Aug 2018, Singapore

- Technical lead for private sharing protocols for permission-ed blockchains for the Shared KYC project with Deutsche, HSBC, MUFC banks and Cargill.
- Founding member of the IBM Center for Blockchain Innovation in Singapore, setup in late 2016
- Worked with IBM Researchers from IBM Research, Melbourne

Skills: Software Project Management · Blockchain · Cryptography

Senior Staff Engineer, SK Hynix Memory Solutions

Nov 2014 – Mar 2016, San Jose, CA

- Analyzed and developed models for NAND flash storage systems.
- Developed simulator platform for prototyping and testing on NAND data and NAND models.
- Developed error recovery algorithms for NAND flash storage systems.

Skills: Error Correction Codes, Algorithm Architecture, Hardware-Software Verification

Staff Engineer, LSI Corporation (an Avago Technologies Company)

June 2013 – Nov 2014, San Jose, CA

- Involved in algorithm development and architecture design for hard drive data controllers.
- Expertise in modern iterative-type error-correction codes aimed at pushing recording densities. Expertise in low-density parity check codes.

Skills: Error Correction Codes, Algorithm Architecture, Hardware-Software Verification

Awards & Recognition

- National Science Foundation Grant EECS-1128226, amount $359,987,
  "Energy-efficient compressed sensing: A joint algorithmic/implementation approach using deterministic sensing",
  Massachusetts Institute of Technology. Aug 2011- Aug 2014.
- Lockheed Martin Orincon Scholarship, University of Hawaii, Manoa, Aug 2006- Jul 2007.
- Research Scholarship, National University of Singapore, Aug 2004-2005

Recent Projects

Foundation Model Stack (FMS) Acceleration (2024)

- Core Developer & Maintainer of a series of ML acceleration plugins for IBM Foundation Model Stack used in WatsonX, for introducing training speedups (e.g., quantization, kernels, packing).

FSDP and DeepSpeed (blog) (2024)

- Main investigator and author of Huggingface blog and concept guide on the equivalence of two ML distributed training frameworks MS DeepSpeed and PyTorch Fully-Sharded Data Parallel (FSDP).

Mixture-of-Experts Distributed Training (2024)

- Core Investigator into incorporating expert parallel techniques for speeding up MoE Training (full-FT or finetuning).

Optional Transport With Order Constraints (2022)

- Companion repository for our 2022 ICML Paper Order Constraints in Optimal Transport.

Education

Postdoctoral Associate

- Massachusetts Institute of Technology, Cambridge, MA, 2013

Doctor of Philosophy in Electrical Engineering

- University of Hawaii, Manoa, 2010
- Thesis Title: Ordered Statistics Decoding for Intersymbol Interference Channels, 2010

Master of Engineering in Electrical Engineering

- National University Of Singapore, Singapore, 2006
- Thesis Title: Optimal Precompensation in High Density Magnetic Recording, 2006

Bachelors of Engineering in Electrical Engineering

- National University Of Singapore, Singapore, 2003

Selected Publications and Patents

F. Lim, "FSDP to DeepSpeed and Back Again", *Pytorch Conference 2024*, San Franscisco, CA, Aug 2024.

A. Kundu, **F. Lim**, A. Chew, L. Wynter, P. Chong, R. D. Lee, "Efficiently Distilling LLMs for Edge Applications," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp 52-62, Mexico City, Mexico, Jun 2024

S. Samanta, P. Mohapatra, **F. Lim**, M. Madugula, X. Liu and S. Lalithsena, "LogInsights - Understanding and Extracting Information from Logs for Fast Fault Classification by Weak Supervision," *2023 IEEE International Conference on Software Services Engineering (SSE)*, Chicago, IL, USA, 2023, pp. 20-26

F. Lim, L. Wynter, S. H. Lim, "Order Constraints in Optimal Transport", *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162:13313-13333, 2022.

K. Bhaskaran, P. Ilfrich, D. Liffman, C. Vecchiola, P. Jayachandran, A. Kumar, **F. Lim**, K. Nandakumar, Z. Qin, V. Ramakrisna, E. G. S. Teo, C. H. Suen, "Double-Blind Consent-Driven Data Sharing on Blockchain," in *IEEE Proc. Internatinal Conf. on Cloud Engineering (IC2E)*, Orlando, FL, pp. 385-391, Apr 2018.

L. Ong, C. K. Ho, **F. Lim**, "The single-uniprior index-coding problem: The single-sender case and the multi-sender extension," *IEEE Trans. Info Theory*, vol. 62, no. 6, pp 3165-3182, Jun 2016

F. Lim, V. Stojanovic, "On U-statistics and compressed sensing I: non-asymptotic average-case analysis," *IEEE Trans. Signal Proc.*, vol. 61, no. 10, pp. 2473-2485, May 2013.

F. Chen, **F. Lim**, O. Abari, A. Chandrakasan, V. Stojanovic, "Energy-aware design of compressed sensing systems for wireless sensors under performance and reliability constraints," *IEEE Trans. Circuits and Sys. I*, vol. 60, no.3, pp. 650-661, Mar 2013.

F. Lim, M. Hagiwara, "Linear programming upper bounds on permutation code sizes from coherent configurations related to the Kendall-tau distance metric," in *Proc IEEE International Symp. Inform. Theory. (ISIT)*, Cambridge, MA, July 2012.

F. Lim, M. Fossorier, A. Kavcic, "Code automorphisms and permutation decoding of certain Reed-Solomon Binary Images," in *IEEE Trans. on Inform. Theory*, vol. 56, no. 10, pp. 5253-5273, Oct 2010

F. Lim, B. Wilson, R.Wood, "Analysis of shingle-write readback using magnetic-force microscopy," *IEEE Trans. on Magn*, vol 46, no 6, pp 1548 - 1551, May 2010.

F. Lim, A. Kavcic, "Optimal precompensation for partial erasure and non-linear transition shift in magnetic recording using dynamic programming," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, St Louis, MO, Jan 2005.

S. Cao, A. De Caro, K. Elkhiyaoui, **Y. C. F. Lim**, "Consent-based data management", *US Patent P201806722*, published 02-01-2022.

T. Inagaki, Y. Ueda, M. Ohara, **Y. C. F. Lim**, C. H. Suen, V. Ramakrishna, T. Nakaike, "Identifying software and hardware bottlenecks", *US Patent P201703267*, published 04-06-2021.

P. Jayachandran, A. Kumar, **Y. C. F. Lim**, V. Ramakrishna, "Anonymous consent and data sharing on blockchain", *US Patent P201702435*, published 08-04-2020.

E.Ragnoli, **Y. C. F. Lim**, A. De Caro, V. Ramakrishna, "Offloaded chaincode execution for a database", *US Patent P201802546*, published 09-04-2021.

Y. C. F. Lim, K. Jeong, Q. Zuo, K. Nguyen, S. Yang, "Systems and Methods for Efficient Targeted Symbol Flipping", *US Patent 20150303943*, published 10-22-2015.