Chi-Heng **Lin**

Senior Research Engineer @ Samsung Research America

(+1) 413-362-2903 | clin354@gatech.edu | www.chihenglin.com | uldyssian2008 | chihenglin | Google Scholar

Research Interests/Summary

As a Senior Machine Learning Research Engineer with over two years of industry experience, I specialize in on-device AI and natural language processing, with a strong focus on the practical applications of large language models (LLMs). I am passionate about advancing artificial intelligence in real-world scenarios by addressing important industrial challenges through interdisciplinary machine learning techniques. My work has led to successful industry patents and several publications in prestigious machine learning venues. I am deeply committed to continuous learning and professional growth, embracing opportunities to expand my expertise and perspectives in the rapidly evolving field of artificial intelligence.

Educations

Georgia Institute of Technology — *Atlanta, GA, USA*
Ph.D. in Electrical and Computer Engineering. GPA: 4.0/4.0 — *Sep. 2017 - Dec. 2022*
Advisor: Dr. Eva L. Dyer

Columbia University — *New York, NY, USA*
M.A. in Statistics. GPA: 4.1/4.3 — *Sep. 2015 - Dec. 2016*

National Taiwan University — *Taipei, Taiwan*
B.S. & M.S. in Electrical Engineering. GPA: 3.8/4.0 — *July. 2007 - Sep. 2013*

Experiences

Samsung Research America — *Mountain View, CA, USA*
Senior Research Engineer (on-device AI/ML) @ Samsung AI center — *Jan. 2023 - Present*

- **Efficient LLM Compression.** Implemented a low-compute compression scheme to fit LLMs onto edge devices, such as cell phones. Our innovative technique partitions transformer models into modules and applies tailored matrix decompositions to each module separately. The results demonstrate that our approach can efficiently run with a **single GPU** for model sizes up to 13B. For larger models, such as Llama-2 70B, it **reduces model size by 30% while increasing throughput by 29%, with a negligible performance drop of just 3%**—achieved without recovery fine-tuning or backward propagation.
- **LLM Inference Acceleration.** Developed a cost-efficient multi-token prediction scheme to enhance LLM inference time and simultaneously reduce compute costs. The method leverages a lightweight bi-gram table to calibrate the joint probability distribution of a speculative decoding algorithm. Our solution achieves a significant **inference speedup of 2.0× to 2.5×, while reducing computation FLOPs by up to 66%**.
- **State-Space LLM Architecture Design.** Designed and developed a high-performance language model tailored for on-device applications. The solution leverages hybrid architectural designs combining **transformers** and **state-space models**. This innovative architecture achieves a notable **1.25× prefill speedup** compared to pure transformer models, while maintaining accuracy. The model is planned for deployment in the upcoming **Galaxy S26** series.

Ambarella Corporation — *Santa Clara, CA, USA*
Algorithm Engineer Intern (Self-driving algorithm group) — *Jan. 2022 - Apr. 2022*

- **Pedestrian Detection.** Developed a pedestrian detection algorithm for self-driving vehicles using a **CNN-LSTM** hybrid model. The model integrates mixed features from keypoints, bounding boxes, and images captured by cameras and LiDAR. Trained on mixed datasets: PIE, JAAD, and Argoverse, it employs **multi-task learning** to improve the robustness across domains. Our final method achieved **85% classification accuracy** for pedestrian crossing on multi-domains.

Georgia Institute of Technology — *Atlanta, GA, USA*
Research Assistant @ Neural Data Science Lab — *Apr. 2022 - Aug. 2022*

- **Data Augmentation.** Established a theoretical framework for analyzing the generalization effects of data augmentation, drawing analogies to classical ridge regression. The framework highlights improved generalization and richer characteristics for augmentations such as random crop and random noise. Building on this, we developed an **augmentation strategy that achieves performance comparable to a well-tuned ridge regressor, without the need for parameter tuning.**
- **Optimal Transport.** Developed a domain adaptation method based on a low-rank optimal transport algorithm. Our algorithm factors the transport plan into low-rank matrices. The low-rank enhancement improves interpretability and achieves a **10% increase in classification accuracy on the MNIST-USPS domain adaptation task**.
- **Bayesiam Optimization.** Developed a **fast and cost-efficient hyperparameter tuning** algorithm for a neuroimaging system. Our innovation integrates the moving-cost bandit algorithm with Bayesian optimization that can dynamically adjust exploration and exploitation based on the moving-cost. The algorithm substantially **reduces the overall tuning time by up to 75%, decreasing it from 5.6 hours to just 1.4 hours** compared to standard Bayesian optimizations.

Technical Skills

Python, Pytorch, R, MATLAB, Mathematica, LaTeX, C++, Linux, macOS, Git

AI/ML Publications

Preprint	**C.-H. Lin**, S. Gao, J. S. Smith, A. Patel, S. Tuli, Y. Shen, H. Jin, Y.-C. Hsu *"MoDeGPT: Modular Decomposition for Large Language Model Compression"*, 2024.
NeurIPS	Z. Chen, **C.-H. Lin**, R. Liu, J. Xiao, and E. L. Dyer. *"Your contrastive learning problem is secretly a distribution alignment problem"*, 2024.
NeurIPS	S. Gao, **C.-H. Lin**, T. Hua, Z. Tang, Y. Shen, H. Jin, Y. Hsu. *"DISP-LLM: Dimension-independent structural pruning for large language models"*, 2024.
NAACL (Findings)	**C.-H. Lin**, S. Tuli, J. S. Smith, Y.-C. Hsu, Y. Shen, H. Jin, *"SLiM: Speculative Decoding with Hypothesis Reduction"*, 2024.
NAACL	S. Tuli, **C.-H. Lin**, Y.-C. Hsu, N. Jha, Y. Shen, H. Jin. *"Dynamo: Accelerating language model inference with dynamic multi-token sampling"*, 2024.
ICML	C. Kaushik*, R. Liu*, **C.-H. Lin**, A. Khera, M. Jin, W. Ma, V. Muthukumar, E. L. Dyer. *"Balanced data, imbalanced spectra: Unveiling class disparities with spectral imbalance"*, 2024.
JMLR	**C.-H. Lin**, C. Kaushik, E. L. Dyer*, V. Muthukumar*. *"The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective"*, 2024.
ICML	M. Azabou, V. Ganesh, S. Thakoor, **C.-H. Lin**, L. Sathidevi, R. Liu, M. Valko, P. Veličković, E. L. Dyer. *Half-hop: A graph upsampling approach for slowing down message passing*, 2023.
ICML (Spotlights)	J. -K. Wang, **C.-H. Lin**, A. Wibisono, B. Hu. *Provable acceleration of heavy ball beyond quadratics for a class of polyak-Łojasiewicz functions when the non-convexity is averaged-out*, 2022
NeurIPS (Orals)	R. Liu, M. Azabou, M. Dabagia, **C.-H. Lin**, M. Gheshlaghi Azar, K. Hengen, M. Valko, E. L. Dyer. *"Drop, swap, and generate: A self-supervised approach for generating neural activity"*, 2021
ICML (Spotlights)	**C.-H. Lin**, M. Azabou, E. L. Dyer. *"Making transport more robust and interpretable by moving data through a small number of anchor points"*, 2021
ICML (Spotlights)	J.-K. Wang, **C.-H. Lin**, J. D. Abernethy. *"A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network"*, 2021
UAI	**C.-H. Lin**, J. D. Miano, E. L. Dyer. *"Bayesian optimization for modular black-box systems with switching costs"*, 2021.
NeurIPS (Workshops)	M. Azabou, M. G. Azar, R. Liu, **C.-H. Lin**, E. Johnson, K. Bhaskaran-Nair, M. Dabagia, B. AvilaPires, L. Kitchell, K. B. Hengen, W. Gray Roncal, M. Valko, E. L. Dyer. (*"Mine your own view: a self-supervised approach for learning representations of neural activity"*, 2021.
NeurIPS (Workshops)	M. Azabou, M. Dabagia, R. Liu, **C.-H. Lin**, K. B. Hengen, E. L. Dyer. *"Using self-supervision and augmentations to build insights into neural coding"*, 2021.
ICLR	J.-K. Wang, **C.-H. Lin**, J. D. Abernethy. *"Escaping saddle points faster with stochastic momentum"*, 2020.

Patents

- **C.-H. Lin**, S. Gao, Y.-C. Hsu, J. S. Smith, A. Patel, S. Tuli, Y. Shen, H. Jin, Z. Tang, T. Hua. ***"Large Language Model Compression."***, 2024.
- S. Tuli, **C.-H. Lin**, Y.-C. Hsu, Y. Shen, H. Jin. ***"DynaMo: Why Predict Just One Token at a Time?"***, 2024.
- J. S. Smith, Y.-C. Hsu, **C.-H. Lin**, S. Tuli, G. M. H. Jeelani, Y. Shen, H. Jin ***"Efficient self-speculative decoding architecture for increasing LLM inference throughput."***, 2024.

Honors & Awards

Samsung Best Paper Award — *Mountain View, CA, USA*
Samsung Research America — *2024*

DEaS-TRIAD Research Scholarship — *Atlanta, GA, USA*
Georgia Institute of Technology — *2020*

M&H Bourne Fellowship — *Atlanta, GA, USA*
Georgia Institute of Technology — *2017*

Davis Fellowship (Two times) — *New York, NY, USA*
Columbia University — *2016*

Academic Services

Reviewer of NeurIPS, ICML, ICLR, AAAI