Dan Fu

Contact

E-mail: danfu@cs.stanford.edu
Homepage: https://www.danfu.org/

Research Interests

My research interests are in the intersection between machine learning and systems. I'm particularly interested in developing more efficient algorithms and architectures for ML, along with hardware-aware systems solutions to make them practically effective.

Education

2018–	PhD in Computer Science, Stanford University Advisors: Christopher Ré, Kayvon Fatahalian
2014–2018	AB/SM in Computer Science, Harvard University Thesis: *Design of Influencing Agents for Flocking in Low-Density Settings*

Experience

2018–	PhD Student, **Stanford University**
2022–	Academic Partner, **Together AI**
2020–	Founder/Podcast Host, **Stanford MLSys Seminar** 16,000+ subscribers, 30,000+ monthly views
Summer 2019	Research Intern, **Argo AI**
Summer 2016, 2017	Software Engineering Intern, **Google**
Summer 2015	Field Engineering Intern, **Tamr**
Summer 2014	Software Engineering Intern, **Interactive Intelligence**
Summer 2013	Development Intern, **DyKnow**

Awards

May 2024	**Stanford Data Science Open Source Software Prize 2024** Inaugural open source software prize awarded for FlashAttention
December 2023	**Best Poster: Efficient Natural Language and Speech Processing Workshop at NeurIPS 2023** FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores
December 2023	**Oral Presentation: NeurIPS 2023** Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture
July 2023	**Oral Presentation: ICML 2023** Hyena Hierarchy: Towards Larger Convolutional Language Models High-throughput Generative Inference of Large Language Models with a Single GPU
May 2023	**Top-25% / Spotlight: ICLR 2023** Hungry Hungry Hippos: Towards Language Modeling with State Space Models
August 2022	**Best Student Paper Runner Up: UAI 2022** Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision
July 2022	**NDSEG Award for Exemplary Impact and Relevance to DoD Research Objectives** Awarded for research presented at 2022 NDSEG conference
July 2022	**Best Paper: Hardware Aware Efficient Training Workshop at ICML 2022** FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness
February 2022	**Best Paper: AIBSD Workshop at AAAI 2022**

	The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning
2019-2022	**United States Department of Defense NDSEG Fellow**
2019-2020	**Brown Institute for Media Innovation Magic Grant**
Fall 2017, Spring 2018	**Harvard Derek Bok Center Certificate of Distinction in Teaching**
2014	**Presidential Scholar**
2012	**Siemens Research Competition National Runner-Up**

Publications

[1] **Benchmarking and Building Long-Context Retrieval Models with LoCo and M2-BERT**
International Conference on Machine Learning (ICML) (2024) and *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)* (2024)
Jon Saad-Falcon, **Daniel Y. Fu**, Simran Arora, Neel Guha, Christopher Ré

[2] **FlashFFTConv: Efficient Convolutions for Long Sequences with Tensor Cores**
International Conference on Learning Representations (ICLR) (2024)
Workshop on Efficient Natural Language and Speed Processing at NeurIPS (2023), **Best Poster**
Daniel Y. Fu*, Hermann Kumbong*, Eric Nguyen, and Christopher Ré

[3] **Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture**
Advances in Neural Information Processing Systems (NeurIPS) (2023), **Oral Presentation**
Daniel Y. Fu, Simran Arora, Jessica Grogan, Isys Johnson, Sabri Eyuboglu, Armin W. Thomas, Benjamin F. Spector, Michael Poli, Atri Rudra, and Christopher Ré

[4] **Laughing Hyena Distillery: Extracting Compact Recurrences from Convolutions**
Advances in Neural Information Processing Systems (NeurIPS) (2023)
Stefano Massaroli*, Michael Poli*, **Daniel Y. Fu***, Hermann Kumbong, David W. Romero, Rom Nishijima Parnichkun, Aman Timalsina, Quinn McIntyre, Beidi Chen, Atri Rudra, Ce Zhang, Christopher Ré, Stefano Ermon, and Yoshua Bengio

[5] **High-throughput Generative Inference of Large Language Models with a Single GPU**
International Conference on Machine Learning (ICML) (2023), **Oral Presentation**
Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, **Daniel Y. Fu**, Zhiqiang Xie, Beidi Chen, Clark Barrett, Joseph E. Gonzalez, Percy Liang, Christopher Ré, Ion Stoica, Ce Zhang

[6] **Hyena Hierarchy: Towards Larger Convolutional Language Models**
International Conference on Machine Learning (ICML) (2023), **Oral Presentation**
Michael Poli, Stefano Massaroli, Eric Nguyen, **Daniel Y. Fu**, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, Christopher Ré

[7] **Simple Hardware-Efficient Long Convolutions for Sequence Modeling**
International Conference on Machine Learning (ICML) and *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models (ME-FoMo)* (2023)
Daniel Y. Fu*, Elliot L. Epstein*, Eric Nguyen, Armin W. Thomas, Michael Zhang, Tri Dao, Atri Rudra, Christopher Ré

[8] **Hungry Hungry Hippos: Towards Language Modeling with State Space Models**
International Conference on Learning Representations (ICLR) (2023), **Notable Top-25% / Spotlight Presentation**
Daniel Y. Fu*, Tri Dao*, Khaled K. Saab, Armin W. Thomas, Atri Rudra, Christopher Ré

[9] **FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness**
Neural Information Processing Systems (NeurIPS) (2022)
Sparsity in Neural Networks: Advancing Understanding and Practice (2022), **Oral Presentation**
Hardware Aware Efficient Training Workshop at ICML (2022), **Best Paper**
Tri Dao, **Daniel Y. Fu**, Stefano Ermon, Atri Ruda, Christopher Ré

[10] **Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision**
Conference on Uncertainty in Artificial Intelligence (UAI) (2022), **Best Student Paper Runner Up**
Mayee F. Chen*, **Daniel Y. Fu***, Dyah Adila, Michael Zhang, Frederic Sala, Kayvon Fatahalian, Christopher Ré

[11] **Perfectly Balanced: Improving Transfer and Robustness in Supervised Contrastive Learning**
International Conference on Machine Learning (ICML) (2022)
Workshop version: **The Details Matter: Preventing Class Collapse in Supervised Contrastive Learning**
Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD) at AAAI (2022), **Best Paper**
Mayee F. Chen*, **Daniel Y. Fu***, Avanika Narayan, Michael Zhang, Zhao Song, Kayvon Fatahalian, Christopher Ré

[12] **TABi: Type-Aware Bi-encoders for End-to-End Entity Retrieval**
Findings of the Association for Computational Linguistics: ACL (2022)
Megan Leszczynski, **Daniel Y. Fu**, Mayee F. Chen, Christopher Ré

[13] **Harmonizing Attention: Attention Map Consistency For Unsupervised Fine-Tuning**
Bridging the Gap: From Machine Learning Research to Clinical Practice Workshop at NeurIPS (2021)
Ali Mirzazadeh, Florian Dubost, Maxwell Pike, Krish Maniar, **Daniel Y. Fu**, Khaled K Saab, Christopher Lee-Messer, Daniel Rubin

[14] **Analyzing Who and What Appears in a Decade of US Cable TV News**
ACM SigKDD Conference on Knowledge Discovery & Data Mining (KDD) (2021)
James Hong, Will Crichton, Haotian Zhang, **Daniel Y. Fu**, Jacob Ritchie, Jeremy Barenholtz, Ben Hannel, Xinwei Yao, Michaela Murray, Geraldine Moriba, Maneesh Agrawala, Kayvon Fatahalian

[15] **Beyond the Pixels: Exploring the Effect of Video File Corruptions on Model Robustness**
ECCV 2020 Workshop on Adversarial Robustness in the Real World (2020)
Trenton Chang, **Daniel Y. Fu**, Sharon Yixuan Li, Christopher Ré

[16] **Fast and Three-rious: Speeding Up Weak Supervision with Triplet Methods**
International Conference on Machine Learning (ICML) (2020)
Daniel Y. Fu*, Mayee F. Chen*, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, Christopher Ré

[17] **Multi-Resolution Weak Supervision for Sequential Data**
Neural Information Processing Systems (NeurIPS) (2019)
Frederic Sala, Paroma Varma, Jason Fries, **Daniel Y. Fu**, Shiori Sagawa, Saelig Khattar, Ashwini Ramamoorthy, Ke Xiao, Kayvon Fatahalian, James R. Priest, Christopher Ré

[18] **Video Event Specification Using Programmatic Composition**
AI Systems Workshop at SOSP (2019), **Oral Presentation**
Daniel Y. Fu, Will Crichton, James Hong, Xinwei Yao, Haotian Zhang, Anh Truong, Avanika Narayan, Maneesh Agrawala, Christopher Ré, Kayvon Fatahalian

[19] **Influencing Flock Formation in Low-Density Settings**
International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2018)
Daniel Y. Fu, Emily S. Wang, Peter M. Krafft, Barbara J. Grosz

Journal Publications

[20] **Orexinergic Neurotransmission in Temperature Responses to Methamphetamine and Stress: Mathematical Modeling as a Data Assimilation Approach**
PLOS ONE, May 20, 2015
Abolhassan Behrouzvaziri, **Daniel Y. Fu**, Patrick Tan, Yeonjoo Yoo, Maria V. Zaretskaia, Daniel E. Rusyniak, Yaroslav I. Molkov, Dmitry V. Zaretsky

[21] **Chaos and Robustness in a Single Family of Genetic Oscillatory Networks**
PLOS ONE, March 25, 2014
Daniel Y. Fu, Patrick Tan, Alexey Kuznetsov, Yaroslav I. Molkov

Unpublished Work and Preprints

[22] **Rekall: Specifying Video Events using Compositions of Spatiotemporal Labels**
arXiv preprint arXiv:1910.02993, October 2019
Daniel Y. Fu, Will Crichton, James Hong, Xinwei Yao, Haotian Zhang, Anh Truong, Avanika Narayan, Maneesh Agrawala, Christopher Ré, Kayvon Fatahalian

[23] **Automatic Parallelization of Sequential Programs**
arXiv preprint arXiv:1809.07684, July 2018
Peter Kraft, Amos Waterland, **Daniel Y. Fu**, Anitha Gollamudi, Shai Szulanski, Margo Seltzer

Invited Talks

2024	**The Unreasonable Power of Synthetics for Efficient Machine Learning**
	Young Professional Symposium at MLSys 2024, Santa Clara, CA
2024	**Hardware-Aware Efficient Primitives for Machine Learning**
	Duke University, Durham, NC
	University of Michigan, Ann Arbor, MI
	Cornell Tech, New York, NY
	Yale University, New Haven, CT
	UCLA, Los Angeles, CA
	NYU New York, NY
	UT Austin, Austin, TX
	Harvard University, Cambridge, MA
	Purdue University, West Lafayette, IN
	UCSD, San Diego, CA
	Cornell University, Ithaca, NY
	Northwestern University, Evanston, IL
	University of Wisconsin-Madison IDEA Seminar, Madison, WI
	Google, Mountain View, CA
	Apple, Virtual
2023	**Efficient Sub-Quadratic Architectures for Machine Learning**
	Bangkok AI Hack 2023, Thailand
2023	**Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture**
	SystemX Conference, Stanford, CA
2023	**Hungry Hungry Hippos**
	Neural Notes Podcast, Virtual
	AI Pub Deep Papers Podcast, Virtual
	IBM Research, Virtual
2023	**Perspectives in Generative AI Panel**
	Pear VC Firm, Menlo Park, CA
2022-2023	**FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness**
	Google, Mountain View, CA
	Stanford CS 217 Guest Lecture, Stanford, CA
	MLOps World: Machine Learning in Production, Virtual
	Meta PyTorch Performance Team, Virtual
	MosaicML, Virtual
	Google N2Formal Team, Mountain View, CA
2022	**Improving Transfer and Robustness of Supervised Contrastive Learning**
	Stanford MLSys Seminar, Stanford, CA
	KitWare Vision Research Group, Virtual
2019	**Rekall: Modeling Concepts in Video with Compositions of Spatiotemporal Labels**
	Intel, Bend, OR
	Stanford Graphics Group, Stanford, CA
	Stanford Vision Lab, Stanford, CA
	Stanford DAWN Retreat, Menlo Park, CA

Open-Source Artifacts

2023	**FlashFFTConv: Efficient Convolutions for Long Sequences** A library for fast exact convolutions optimized for GPU. Now being used in production by Together AI to train and serve long-sequence convolutional language models. https://github.com/HazyResearch/flash-fft-conv
2023	**Monarch Mixer BERT Models** A suite of BERT models trained with Monarch Mixer, from 80M to 341M parameters, supporting sequence lengths up to 8K. Now being served by Together AI and in use by MongoDB. https://github.com/HazyResearch/m2
2023	**RedPajama-1T** A trillion+ token dataset for training large language models, mimicking the data gathering process from Llama-1. So far downloaded 1 million+ times. https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T
2023	**Safari: Convolutions for Sequence Modeling** A training repository for gated convolution models on language, images, and long-sequence data. Used to train H3, Hyena, and M2. https://github.com/HazyResearch/safari
2023	**Hungry Hungry Hippos (H3) Models** A suite of hybrid attention + gated SSM architectures trained on language modeling, up to 2.7B parameters. https://github.com/HazyResearch/H3
2022	**FlashAttention** Fast and memory-efficient exact attention with IO-Awareness. Now integrated into PyTorch and used in every major AI research lab in industry. https://github.com/HazyResearch/flash-attention
2020	**FlyingSquid** A fast algorithm for weak supervision without SGD using method-of-moments estimation. In use at Snorkel AI. https://github.com/HazyResearch/flyingsquid
2019	**Rekall: Compositional Video Event Specification** A library for analyzing video data using compositions of image labels. Once used at Argo AI for event mining. https://github.com/scanner-research/rekall

Teaching Experience

Fall 2021, Winter/Spring 2022, Fall 2023	**Instructor**, CS 528: Machine Learning Systems Seminar Stanford University
Winter 2023	**Interviewer**, CS 324: Advances in Foundation Models Stanford University
Spring 2018	**Teaching Fellow**, CS 152: Programming Languages Harvard University
Fall 2015, 2016, 2017	**Teaching Fellow**, CS 61: Systems Programming and Machine Organization Harvard University

Mentorship

2024	**Aaryan Singhal** (Stanford CS undergrad)
2023	**Jon Saad-Falcon** (Stanford CS PhD rotator)
2023	**Hermann Kumbong** (Stanford CS MS)
2023	**Pranav Vaid** (Stanford CS undegrad/coterm)
2022-2023	**Elliot Epstein** (Stanford ICME PhD rotator)
2020	**Heidi Chen** (Stanford CS undergrad, now Google)
2019-2020	**Trenton Chang** (Stanford undergrad/coterm in American Studies, now EECS PhD at University of Michigan)
2019-2022	**Avanika Narayan** (Stanford CS undergard/coterm, now CS PhD at Stanford)
2020-2021	**Manasi Ganti** (high school student, now CS undergrad at University of Washington)

Service

2020–	**Creator/Host: Stanford MLSys Seminar Series**
ICML 2023, 2024	**Workshop Organizer: Efficient Systems for Foundation Models (ES-FoMo)**
2020-2023	**Stanford PhD Admissions Committee**

Referee/program committee member for NeurIPS 2024, ICLR 2023 Blog Post Track, NeurIPS 2023, ICML 2023, ICLR 2023, NeurIPS 2022, ICML 2022 (**top 10% reviewer, session chair**), ICLR 2022, NeurIPS 2021, ICML 2021 Workshop on ML for Data, Workshop on Weakly Supervised Learning @ ICLR 2021.

Last updated: June 13, 2024[*]

[*]CV template inspired by Neel Guha and Christopher Morris.