Tri Dao

Curriculum Vitae

Work Experience

6/24–present **Assistant Professor in Computer Science**, *Princeton University*.

7/23–present **Founding Chief Scientist**, *Together AI*.

Education

9/16–06/23 **PhD in Computer Science**, *Stanford University*.
Advisor: Christopher Ré, Stefano Ermon.

1/18–1/19 **MS in Statistics**, *Stanford University*.

9/14–6/16 **MS in Computer Science**, *Stanford University*.

9/12–6/16 **BS in Mathematics**, *Stanford University*.

Research Interests

Machine learning and systems, with a focus on efficient training and long-range context:
- Efficient Transformer training and inference.
- Sequence models with long-range memory.
- Structured sparsity for compact deep learning models.

Publications

Tri Dao* and Albert Gu*. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning (ICML)*, 2024.

Albert Gu* and **Tri Dao***. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Poli*, Stefano Massaroli*, Eric Nguyen, Daniel Y Fu, **Tri Dao**, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning (ICML)*, 2023. **Oral**.

Daniel Y. Fu*, Elliot L Epstein*, Eric Nguyen, Armin W Thomas, Michael Zhang, **Tri Dao**, Atri Rudra, and Christopher Ré. Simple hardware-efficient long convolutions for sequence modeling. In *International Conference on Machine Learning (ICML)*, 2023.

Tri Dao*, Daniel Y. Fu*, Khaled K. Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations (ICLR)*, 2023. **Spotlight**.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, **Tri Dao**, Beidi Chen, Percy Liang, Christopher Ré, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. In *Advances in Neural Information Processing Systems*, 2022. **Oral**.

June Wang, Binhang Yuan, Luka Rimanic, Yongjun He, **Tri Dao**, Beidi Chen, Percy Liang, Christopher Ré, and Ce Zhang. Fine-tuning language models over slow networks using activation compression with guarantees. In *Advances in Neural Information Processing Systems*, 2022.

Michael Poli, Stefano Massaroli, Federico Berto, Jinkyoo Park, **Tri Dao**, Christopher Ré, and Stefano Ermon. Transform once: Efficient operator learning in frequency domain. In *Advances in Neural Information Processing Systems*, 2022.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, **Tri Dao**, Stephen Baccus, and Christopher Ré. S4ND: Modeling images and videos as multidimensional signals with state spaces. In *Advances in Neural Information Processing Systems*, 2022.

Tri Dao, Beidi Chen, Nimit Sohoni, Arjun Desai, Michael Poli, Jessica Grogan, Alexander Liu, Aniruddh Rao, Atri Rudra, and Christopher Ré. Monarch: Expressive structured matrices for efficient and accurate training. In *International Conference on Machine Learning (ICML)*, 2022. **Outstanding Paper runner-up**.

Chenlin Meng, Linqi Zhou, Kristy Choi, **Tri Dao**, and Stefano Ermon. ButterflyFlow: Building invertible layers with butterfly matrices. In *International Conference on Machine Learning (ICML)*, 2022.

Tri Dao*, Beidi Chen*, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, and Christopher Ré. Pixelated butterfly: Simple and efficient sparse training for neural network models. In *International Conference on Learning Representations (ICLR)*, 2022. **Spotlight**.

Beidi Chen*, **Tri Dao***, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, **Tri Dao**, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34, 2021.

Nicholas Roberts, Mikhail Khodak, **Tri Dao**, Liam Li, Christopher Ré, and Ameet Talwalkar. Rethinking neural operations for diverse tasks. In *Advances in Neural Information Processing Systems*, 2021.

Jared Q Davis*, Albert Gu*, Krzysztof Choromanski, **Tri Dao**, Christopher Ré, Chelsea Finn, and Percy Liang. Catformer: Designing stable transformers via sensitivity analysis. In *International Conference on Machine Learning (ICML)*, 2021.

Tri Dao, Govinda M Kamath, Vasilis Syrgkanis, and Lester Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations (ICLR)*, 2021.

Beidi Chen, Zichang Liu, Binghui Peng, Zhaozhuo Xu, Jonathan Lingjie Li, **Tri Dao**, Zhao Song, Anshumali Shrivastava, and Christopher Ré. Mongoose: A learnable LSH framework for efficient neural network training. In *International Conference on Learning Representations (ICLR)*, 2021. **Oral**.

Albert Gu*, **Tri Dao***, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. In *Advances in neural information processing systems (NeurIPS)*, 2020. **Spotlight**.

Tri Dao, Nimit Sohoni, Albert Gu, Matthew Eichhorn, Amit Blonder, Megan Leszczynski, Atri Rudra, and Christopher Ré. Kaleidoscope: An efficient, learnable representation for all structured linear maps. In *The International Conference on Learning Representations (ICLR)*. 2020. **Spotlight**.

Avner May, Jian Zhang, **Tri Dao**, and Christopher Ré. On the downstream performance of compressed word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, 2019. **Spotlight**.

Jonathan Kuck, **Tri Dao**, Hamid Rezatofighi, Ashish Sabharwal, and Stefano Ermon. Approximating the permanent by sampling from adaptive partitions. In *Advances in Neural Information Processing Systems (NeurIPS) 32*, 2019.

Jonathan Kuck, **Tri Dao**, Shengjia Zhao, Burak Bartan, Ashish Sabharwal, and Stefano Ermon. Adaptive hashing for model counting. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*. 2019.

Tri Dao, Albert Gu, Matthew Eichhorn, Atri Rudra, and Christopher Ré. Learning fast algorithms for linear transforms using butterfly factorizations. In *The International Conference on Machine Learning (ICML) 36*. 2019. **Full oral presentation**.

Tri Dao, Albert Gu, Alexander J Ratner, Virginia Smith, Christopher De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *The International Conference on Machine Learning (ICML) 36*. 2019.

Jian Zhang, Avner May, **Tri Dao**, and Christopher Ré. Low-precision random Fourier features for memory-constrained kernel approximation. In *The International Conference on Artificial Intelligence and Statistics (AISTATS) 22*. 2019.

Anna T Thomas, Albert Gu, **Tri Dao**, Atri Rudra, and Christopher Ré. Learning compressed transforms with low displacement rank. In *Advances in Neural Information Processing Systems (NeurIPS) 31*. 2018.

Tri Dao, Christopher M De Sa, and Christopher Ré. Gaussian quadrature for kernel features. In *Advances in Neural Information Processing Systems (NeurIPS) 30*. 2017. **Spotlight**.

Research Adoption

FlashAttention.

- Integrated into many ML frameworks (Pytorch, Huggingface's transformers and diffusers, Microsoft's DeepSpeed, Nvidia's Megatron-LM, MosaicML's Composer) that benefit a large audience of researchers and practitioners.
- Used by many organizations, in both research and production, to speed up the training and inference of the most widely used models such as large language models and diffusion models (e.g., at Meta, Microsoft Azure, Nvidia, OpenAI, Stability AI, Adept, Colossal-AI, SambaNova Systems).

See this page for a partial list of places where FlashAttention is being used.

Industry Experience

10/22 – 06/23 **Adept AI**, PhD Fellow (part-time), San Francisco, CA.

- Develop multi-modal Transformers to model users' interaction with browser tools.
- Speed up large-scale Transformer distributed training and inference.
- Train large language models on long context (16K).

6/20 – 9/20 **Microsoft Research**, Research Intern, Cambridge, MA.

- Developed a novel loss function for knowledge distillation that improves the performance of the student model.

6/16 – 9/16 **Citadel Securities**, Quantitative Researcher, Chicago, IL.

- Developed a novel feature generation method to analyze large quantitative trading datasets.
- Built a black-box optimization system for state-of-the-art quantitative trading strategies.

6/14 – 9/14 **Google**, Software Engineering Intern, Mountain View, CA.

- Designed machine learning algorithms to find best advertisements for each Ad Group.

Talks

06/22-09/22 FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, Hardware Aware Efficient Training Workshop (**Best Paper award**), Sparsity in Neural Networks Workshop (**Oral**), Adept AI, MosaicML, Google Brain, FAIR, Pytorch team, MedAI seminar, Stanford AHA Agile Hardware Center, Stanford NLP, Stanford CRFM, Stanford SystemX conference.

07/22 Monarch: Expressive Structured Matrices for Efficient and Accurate Training, ICML (**Outstanding Paper runner-up**), Baltimore.

05/22 Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models, ICLR (**Spotlight**), Virtual.

04/20 Kaleidoscope: An Efficient, Learnable Representation For All Structured Linear Maps, ICLR (**Spotlight**), Virtual.

06/19 Learning Fast Algorithms for Linear Transforms Using Butterfly Factorizations, ICML (**Full oral presentation**), Long Beach.

06/19 A Kernel Theory of Modern Data Augmentation, ICML, Long Beach.

12/17 Gaussian Quadrature for Kernel Features, NeurIPS (**Spotlight**), Long Beach.

Teaching

1/20 – 3/20 CS 228: Probabilistic Graphical Models, Teaching Assistant, Stanford University.

4/19 – 6/19 CS 229: Machine Learning, Teaching Assistant, Stanford University.

1/16 – 3/16 EE 364A: Convex Optimization I, Teaching Assistant, Stanford University.

9/15 – 12/15 EE 103: Intro to Matrix Methods, Teaching Assistant, Stanford University.

Awards

2022 International Conference on Machine Learning (ICML) 2022, **Outstanding Paper runner-up**.

2022 Hardware Aware Efficient Training Workshop 2022, **Best Paper award**.

2016 Sterling Award for top 25 graduating seniors in School of Humanities and Sciences, Stanford University.

2015 Tau Beta Pi, Stanford University.

Service

Area Chair: COLM 2024.

Reviewer: NeurIPS, ICML, ICLR, AISTATS, JMLR, NeurIPS 2019 **best reviewers**, ICML 2019 **best reviewers**.