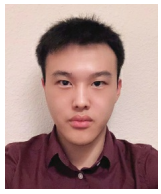
# Dacheng Li

(858)465-0617 | dacheng2@cs.cmu.edu
Homepage: dacheng-li.info



## EDUCATION

**Carnegie Mellon University** Dec 2021 - Feb 2023

- Research Assistant at Machine Learning Department; Supervisors: *Prof. Eric P. Xing* and *Prof. Hao Zhang*.
- Research: *MPCFormer: fast, performant and private Transformer inference with MPC*.

**Carnegie Mellon University** Aug 2020 - Dec 2021

Master of Science in Machine Learning

- GPA: 3.95/4,0; Supervisors: *Prof. Eric P. Xing* and *Prof. Hao Zhang*.
- Research: *AMP: Automatically Finding Model Parallel Strategies with Heterogeneity Awareness*.

**University of California, San Diego** Sep 2016 - Mar 2020

Double Majors in Computer Science and Mathematics

- Major GPA: 3.99/4.0; Supervisor: *Prof. Zhuowen Tu*;
- Research: *Dual Contradistinctive generative autoencoder.*

## PUBLICATION

- **Li, Dacheng**, Hongyi Wang, Eric P. Xing, and Hao Zhang. "AMP: Automatically Finding Model Parallel Strategies with Heterogeneity Awareness." (**NeurIPS 2022**)
- **Li, Dacheng***, Rulin Shao*, Hongyi Wang*, Han Guo, Eric P. Xing, Hao Zhang, "MPCFormer: fast, performant and private Transformer inference with MPC." (Under Submission to **ICLR 2023**)
- Bian, Song, **Dacheng Li**, Hongyi Wang, Eric P. Xing, Shivaram Venkataraman. "Does compressing activations help model parallel training?" (Under submission to **MLSys 2023**)
- Parmar, Gaurav*, **Dacheng Li*,** Kwonjoon Lee*, and Zhuowen Tu. "Dual contradistinctive generative autoencoder." (**CVPR 2021**) * denotes equal contribution

## RESEARCH EXPERIENCE

**Automatically Finding Model Parallel Strategies with Heterogeneity Awareness (Project Lead)** Feb 2021 - Nov 2022

*NeurIPS 2022 Paper* | *Code* | *Presentation* (Supervisors: *Prof. Eric P. Xing*, *Prof. Hao Zhang*)

- Developed an **automatic** procedure and a **cost model** to find good **model-parallel** strategies specifically for **heterogeneous** distributed clusters and deep learning models.
- Matched expert-designed strategies in no hetergeneous setup; found **1.54x** and **1.77x** faster strategies when heterogeneity exists in the cluster and model.

**Fast, performant and private Transformer inference (Project Lead)** Feb 2022 - Present

*Paper* (Supervisors: *Prof. Eric P. Xing*, *Prof. Hao Zhang*)

- Developed a framework that speeds up privacy-preserving Transformer inferene by using MPC-friendly approximations and Knowledge distillation(**KD**) on Secure Multi-Party computation(**MPC**) systems.
- Achieved **5.9x** speedup with BERT-LARGE with the same ML performance on the IMDb dataset; Achieved **2.2x** speedup with ROBERTA-BASE on the GLUE benchmark with 97% the ML performance .

**Activation compression in model-parallel training (Project Co-lead)** Apr 2022 - Present

(Supervisor: *Prof. Shivaram Venkataraman*)

- Conducted the first empirical study with *160* settings on the utility of activation compression in model parallelism, on **pruning**-based, **learning**-based and **quantization**-based compression algorithms.
- Developed a performance analysis and proposed a list of desirable properties on compression algorithms, training hyper-parameters and hardware when using communication compression to speed up model-parallel training.

**Dual Contradistinctive Autoencoders (Project Co-Lead)** Feb 2020 - Nov 2020

*CVPR 2021 Paper* / *Code* (Supervisor: Prof. *Zhuowen Tu*)

- Developed a general-purpose Variational Autoencoder model (**VAE**) by intergrating a set-level objective (**Generative Adversarial** loss) and an instance-level objective (**contrastive learning**).
- Achieved State-of-the-art(SOTA) Fréchet Inception Distance, Inception score, and perceptual distance in Cifar-10 and CelebA dataset; Achieved SOTA image editing and latent space interpolation quality in CelebA dataset.

## OPEN-SOURCE CONTRIBUTIONS

**NCCL backend for collective operations in Ray**

*Code* (Collaborator: *Prof. Hao Zhang*)

- Developed the **NCCL** backend for all-to-all and P2P **collective** operations with multi-GPU and multi-stream support.
- The code has been intergrated in *Ray* (a popular ML training framework with **22.6k** stars), and has been widely used as the default GPU implementation for collective operation calls in distributed ML training.

**Distributed training implementation in Spacy**

*Code* (Collaborator: *Prof. Hao Zhang*)

- Implemented data-parallel ML training using the above developed NCCL backend in *Spacy*.
- Achieved **5.22x** speedup when scaled to 16 workers compared to the native Spacy implementation.

**Resource scheduling policy support for AdaptDL**

Collaborator: *Dr. Aurick Qiao*

- Developed a speedup constraint that reduces the re-allocation frequency by **75%**. (*Code*)
- Developed an asynchronous job detector to shorten the user waiting time by up to **60** seconds. (*Code*).
- Integrated with *AutoDist*, a single-job optimazation system to enable job-level and cluster-level co-optimization. Demonstrated a running demo to the *CASL* open source community (*Code*).
- Codes have been intergaeted in *AdaptDL* (OSDI'2021 best system winner)

## START-UP

**AILink Technology Corporation** Feb 2019 - Sep 2020

*Link* (**Co-founder**)

- Developed the **recommender system** backend using ML-based methods , e.g., clustering and collaborative filtering.
- Deployed the system in our Android App, UniLink, which has been used by more than **200** real students to find roommates.
- Co-founded the company AiLink and raised **$100k** funding.

## TEACHING EXPERIENCE

**Undergraduate Teaching Assistant** Sep 2019 - Mar 2020

- Fall 2019 CSE 158 - Recommender Systems (with *Prof. Julian McAuley*), UC San Diego
- Spring 2020 CSE 30 - Data Structure and algorithms (with *Prof. Marina Langlois*), UC San Diego

## TECHNICAL STRENGTH

- Programming Languages: Python, Java, C++, R, Matlab, Haskell;
- ML frameworks: PyTorch, Tensorflow, DeepSpeed, Megatron-LM, Ray, Spacy, HuggingFace, Autodist, AdaptDL.
- BigData+Database: sklearn, Scipy, NumPy, Pandas, Jupyter, AWS EC2, Kubernetes, Docker, MongoDB,
- Web+Mobile: React Native, HTML/CSS/JS, Android.