Yiding Wang

yiding.wang@connect.ust.hk | hi@yidingwang.xyz | yidingwang.xyz | Hong Kong

EDUCATION

Hong Kong University of Science and Technology **May 2023 (Expected)**
Ph.D. in Computer Science Hong Kong

- Thesis: Towards Efficient Deep Learning Systems with Learning-Based Optimizations. Advisor: Prof. Kai Chen.
- Research interests: Building and optimizing ML (training & inference) systems for CV & NLP applications.

Shanghai Jiao Tong University **July 2017**
B.Eng. in Instrument Science (part of Electrical Engineering). Shanghai, China

SELECTED PUBLICATIONS

Tabi: An Efficient Multi-Level Inference System for Large Language Models.
Yiding Wang, Kai Chen, Haisheng Tan, Kun Guo. **EuroSys 2023**

- Tabi is a model-less inference system that reduces the inference latency of text classification tasks that require Transformer-based large language models (LLMs, e.g., RoBERTa-large) by 21%-40%.
- Using ML techniques including confidence calibration, Transformer's attention weights, token pruning, and model ensembling. Serving diverse workloads with per-query feedback instead of one-model-fits-all.

Egeria: Efficient DNN Training with Knowledge-Guided Layer Freezing.
Yiding Wang, Decang Sun, Kai Chen, Fan Lai, Mosharaf Chowdhury. **EuroSys 2023**

- Egeria accurately freezes the converged layers and saves their computation and communication costs. It accelerates DL training by 19%-43% without sacrificing accuracy.
- Using knowledge distillation and transfer learning techniques. Proposing the training plasticity metric to quantify layers' training progress since different layers converge differently during training.

Enabling Edge-Cloud Video Analytics for Robotic Applications.
Yiding Wang, Weiyan Wang, Duowen Liu, Xin Jin, Junchen Jiang, Kai Chen. **INFOCOM 2021** and **TCC 2022**

- Runespoor is an edge-cloud video analytics system that preserves and augments the detail in compressed data to manage the accuracy loss of critical classes over bandwidth-constrained networks.
- Using ML techniques like super-resolution (SR) fine-tuned for video analytics tasks to augment details tailored for the accuracy of inference tasks when reconstructing high-resolution frames from compressed data.

INDUSTRIAL EXPERIENCE

Intel **Aug 2016 – Jan 2017**
Software Engineer Intern Shanghai, China

- Implementing Torch-like APIs such as confusion matrix for Intel BigDL, a Spark-based DL framework.
- Developing a chord recognition application with MLP and Autoencoder using BigDL for customers.

OTHER

- **Technical Skills**: Python, PyTorch, HuggingFace Transformers, ML Systems, Computer Vision, NLP
- **Teaching Assistants:** Operating System (Spring 2020), Introduction to Computing with Java (Fall 2018)
- **Languages**: English (fluent), Mandarin (native), Cantonese (beginner)
- **Awards**: HKUST Ph.D. Studentship (2017-2023), ACM/IEEE/HKUST Conference Grants (2018-2021)
- **References**: Prof. Kai Chen (HKUST), Prof. Mosharaf Chowdhury (UMich), Prof. Junchen Jiang (UChicago)