Fang Chen

4368 Briggs Ln, Merced, CA 95348
fchen20@ucmerced.edu
(213)-284-1488
google scholar

Education

University of California, Merced *08/2023 - Present*
Ph.D. student, Computer Science *Advisor: Meng Tang*

University of Southern California *08/2021 - 05/2023*
M.Sc., Electrical Engineering *GPA: **3.83**/4.0*

Chongqing University of Posts and Telecommunications *09/2016 - 06/2020*
B.E., Digital Media Technology *GPA: **3.85**/4.0*

Professional Skills

Programming Language: Python, MATLAB, C/C++, JavaScript
Deep Learning Framework: PyTorch, TensorFlow, Keras
Parallel and Distributed Computation: CUDA C, PyCuda
Documentation Formatting: Latex

Work Experience

Graduate Technical Intern | Intel AI Lab, U.S. 06/2022 - 01/2023
Supervisor: *Anthony Sarah*

- ***Project 1: Simq-nas: Simultaneous quantization policy and neural architecture search [2]***
  - A **post-training Mixed Precision Quantization and Neural Architecture jointly aware Search (DyQ-NAS) method** was proposed for the deployment of extensive fully trained deep learning architectures on resource-constrained devices.
  - Proposed a **feasible quantization policy search method** to reduce the search space size of DyQ-NAS, and developed the **mixed precision quantization module** that enables the quantization policy to be customized and jointly searched by NAS methods.
  - Based on the observation and analysis to the performance of accuracy / latency predictors, **found and located a vital issue** that a set of different configures might generate the same subnetworks due to nonactivated parameters; and proposed a **masked encoding algorithm** for configuration parsing to address this issue.
  - The dynamic quantization module has been merged into main branch of **Intel Dynamic Neural Architecture Search Toolkit (DyNAS-T)**. Based on ImageNet dataset, the quantized subnetworks averagely achieve **75%** reduction of model size, **90%** reduction of inference time with only **3.75%** reduction of accuracy.

Research Experience

Ph.D. Research | University of California, Merced 08/2023 – *Present*
Advisor: *Meng Tang*

- ***Project 1: Residual Encoded Distillation for Peak Memory Reduction [1]***
  - We propose a distillation framework tailored for **reducing the peak memory** of convolutional neural networks, which allows **aggressive downsampling** of feature maps via pooling layers, while incurring a **negligible accuracy drop**.
  - We propose a **residual encoded distillation (RED)** block to align features between high-peak-memory teacher networks and low-peak-memory student networks, based on a **multiplicative gating mechanism** and **additive residual learning**.
  - For image classification tasks, our method **yields about 2$\times$ $\sim$ 3.2$\times$ reduction in measured peak memory** with a slight decrease in the classification accuracies for CNN based models. Additionally, our method **improves the accuracy of compact ViT based models**, when distilled from large CNNs.
  - We also show the versatility of our distillation method for image generation. For a U-Net based denoising diffusion probabilistic method, our method **reduces the theoretical peak memory by 4$\times$** while maintaining the fidelity and the diversity of synthesized images.

Research Assistant | Energy Efficient Secure Sustainable Computing Group of USC 12/2021 – 05/2023

Advisor: *Peter A. Beerel*

- ***Project 1: Self-Attentive Pooling for Efficient Deep Learning [3]***
  - A **non-local self-attentive pooling method** was proposed to address the issue that current pooling methods perform poorly in **aggressive feature aggregation**, of which the main purpose was to assign the pooling methods **with large pooling strides** but **without too much accuracy loss.**
  - Based on the analysis to activations, we hypothesized that the accuracy loss typically associated with aggressive down-sampling could be minimized by **considering both local and non-local information** during down-sampling.
  - Extensive experiments on standard **image recognition** (STL10, VWW, ImageNet) and **object detection** (Microsoft COCO) datasets with various backbone networks (MobileNetV2, MobileNetV3, ResNet-18, ResNeXt-18) demonstrated the superiority of our proposed mechanism over the state-of-the-art (SOTA) pooling techniques. For instance, we surpassed the **test accuracy** of existing methods on different variants of MobileNet-V2 on ImageNet by an average of $\sim$**1.2%**. With the aggressive down-sampling of the activations in the initial layers (providing up to **22x reduction** in **memory consumption**), our approach achieved **1.43%** higher **test accuracy** compared to SOTA techniques with iso-memory footprints.

Research Assistant | Key Laboratory of Signal and Information Processing of Chongqing 03/2019 - 06/2021

Advisor: *Chenqiang Gao*

- ***Project 1: Local Patch Network for Infrared Small Target Detection [4]***
  - A **local patch network with global attention** was proposed to eliminate the **extreme class-imbalance**, that the main challenge of small target detection, between sparse small target pixels and low-rank background pixels, through **leveraging global and local features** of infrared small targets.
  - Proposed an **attention module** to **suppress** most irrelevant **background pixels** from the **global view**, and a **local patch network (LPNet)** to **capture small targets** by viewing the attended feature maps patch by patch from the **local view**.
  - The proposed method outperformed the state-of-the-art methods on two widely used public datasets and one of our private datasets under **probability of detection** ($\sim$ **+3%**), **AUC** ($\sim$ **+7%**) and **f1-measure** ($\sim$ **+3%**) metrics.
- ***Project 2: Infrared Small-Dim Target Detection under Complex Backgrounds [5]***
  - Based on the idea widely used in traditional methods that treating the **small target** as the **noise item**, the challenge was to **distinguish** the small target from the ground-truth **noise distribution** of background.
  - Due to the ability of capturing **long-rang dependencies** of multi-head attention mechanism, a **Transformer and U-Net-like** skipped connection framework was proposed to capture the discriminative **differences** between **small target** and **global noise distribution** from complex backgrounds.
  - The proposed method outperformed the state-of-the-art methods on two widely used public datasets under **probability of detection** ($\sim$ **+3%**), **AUC** ($\sim$ **+8%**) and **f1-measure** ($\sim$ **+2%**) metrics, and was especially effective on **cross-scene generalization** and **anti-noise performance**.

Honors and Awards

10/2021	**Best Masters Poster Award** of the 11th Annual Research Festival by **USC Ming Hsieh Institute**
06/2020	**Outstanding Graduate** of Chongqing (Provincial Level, in top 0.1%)
11/2019	**Annual Progress Scholarship** in 2018-2019 Academic Year (in top 0.1%)
07/2019	**Silver Award** (**Rank 2** out of 300+ teams) and **Best Report** in **IEEE ISI World Cup 2019 (IWC 2019)**
11/2017	**Second Award** of Chongqing Division in China Undergraduate Mathematical Contest in Modeling

Publications

[1] **Fang Chen**, Gourav Datta, Mujahid Al Rafi, Hyeran Jeon, Meng Tang. ReDistill: Residual Encoded Distillation for Peak Memory Reductio. *Submitted to* ***The 39th Annual AAAI Conference on Artificial Intelligence****, 2025.*

[2] Sharath Nittur Sridhar, Maciej Szankin, **Fang Chen**, Sairam Sundaresan, Anthony Sarah. SimQ-NAS: Simultaneous Quantization Policy and Neural Architecture Search. *Accepted by* ***AAAI Edge Intelligence Workshop****, 2024.*

[3] **Fang Chen**, Gourav Datta, Souvik Kundu, and Peter Beerel. Self-attentive pooling for efficient deep learning. *In Proceedings of the* ***IEEE/CVF Winter Conference on Applications of Computer Vision****, pp. 3974-3983, 2023.*

[4] **Fang Chen**, Chenqiang Gao, Fangcen Liu, Yue Zhao, Yuxi Zhou, Deyu Meng, and Wangmeng Zuo. Local patch network with global attention for infrared small target detection. In ***IEEE Transactions on Aerospace and Electronic Systems***, vol. 58, no. 5, pp. 3979-3991, 2022.

[5] Fangcen Liu, Chenqiang Gao, **Fang Chen**, Deyu Meng, Wangmeng Zuo, and Xinbo Gao. Infrared small-dim target detection with transformer under complex backgrounds. In ***IEEE Transactions on Image Processing***, vol. 32, pp. 5921-5932, 2023.

[6] Fengshun Zhou, Chenqiang Gao, **Fang Chen**, Chaoyu Li, Xindou Li, Feng Yang, and Yue Zhao. Face anti-spoofing based on multi-layer domain adaptation. In ***IEEE International Conference on Multimedia Expo Workshops (ICMEW)***, pp. 192–197, 2019.