Shixing Yu

Cockrell School of Engineering, The University of Texas at Austin
+1 737-288-6869 | shixingyu@utexas.edu | homepage: billysx.github.io

RESEARCH INTERESTS

- Machine Learning: Sparse Neural Network, Efficient Deep Learning (Pruning, Quantization, Knowledge Distillation), Federated Learning, Multitask-Learning
- Computer Vision: Video/Image Processing, Vision Transformers

EDUCATION

PEKING UNIVERSITY Beijing, China 09/2017–07/2021
School of Electronic Engineering and Computer Science
B.Sc., Data Science, Major GPA: **3.7**/ 4.0

THE UNIVERSITY OF TEXAS AT AUSTIN Texas, U.S. 09/2021–present
Cockrell School of Engineering
M.S., Electrical and Computer Engineering, GPA: **4.0**/ 4.0,
Advisor: Prof. Zhangyang(Atlas) Wang.

EXPERIENCE

Research Intern, advised by Dr. Jose Alvarez, *NVIDIA Corporation*
N:M fine-grained sparsity and general efficient deep learning 06/2022-now

- Explores wide ranges of model architecture to exploit the best speedup for fine-grained structured sparsity (Specially supported on NVIDIA's Ampere series GPU) and build an Ampere friendly network architecture that is best utilized on Ampere GPUs.
- Conducts reparameterization techniques to get rid of residual connections in networks – which emerges to be the general obstacle that burdens the performance of fine-grained structure sparsity.
- Rebuilds modules of Vision Transformer with efficient operations to make it accelerable by Ampere GPUs.
- Uses SAM to optimize the Neural Network to a smooth local minima to best exploit the network when exposed to pruning and other compression techniques and explores the robustness of the resulting network.

Research Assistant, advised by Prof. Atlas Wang, *The University of Texas at Austin*
Vision Transformer Compression 05/2021-09/2021

- Jointly leverages multiple compression means on ViT. The algorithm is designed to only require a specified global resource budget, and can automatically optimize the composition of different techniques.
- Formulates and solves the compression problem as a unified constrained optimization problem, which simultaneously learns model weights, layer-wise pruning ratios, and skip configurations, under a distillation loss and an overall budget constraint.
- Conducts experiments with several DeiT backbones on ImageNet, which consistently verify the effectiveness. For example, on DeiT-Tiny (with/without distillation tokens) yields around 50% FLOPs reduction, with little performance degradation (only 0.3%/0.9% loss compared to the baseline).
- The work is summarized in a paper accepted to ICLR 2022.

Federated Learning and Multi-task Learning 03/2022-06/2022

- Combines Federated Learning with Multi-task Learning to leverage both the data and label user may provide during practicing the application.
- Leverages the in-born privilege of Federated Learning which concludes and merges weights periodically, and uses the summarizing period to recognize positive/negative sets for each task to promote the performance for individual tasks.

Research Intern, supervised by Prof. Kurt Keutzer and Prof. Michael Mahoney, BAIR, University of California, Berkeley

Hessian Aware Pruning *06.2020-12/2020*

- Employed second-order analysis on network parameters and propose a novel and fast second-order based metric to find insensitive parameters in a NN model.
- Proposed a novel neural implant technique to alleviate accuracy degradation. Specifically, instead of pruning the entire insensitive kernels, spatial convolution channels are replaced with 1 × 1 pointwise convolution.
- Achieved 94.3% accuracy (< 0.1% degradation) on PreResNet29 (CIFAR-10), with only 31% parameters left. For ResNet50, achieved 75.1% top-1 accuracy (0.5% degradation) on ImageNet, with only half of the parameters left.
- The work is summarized in a paper accepted to WACV 2022.

Research Intern, supervised by Prof. Jiaying Liu, *WICT, Peking University*

Arbitrary Time Frame Interpolation *07/2019-06/2020*

- Used time step and optical flow (estimated by a SOTA method) to form a feature matrix and constructed a META block using these features to form a convolution kernel for every pixel on the target frame and constructed the target frame using these convolution kernels.
- Built a meta-learned kernel for temporal adjustment module to implement frame interpolation between two frames at arbitrary time step.
- Constructed a meta-learned local adjustment kernel from the optical flow estimated by the rescaled frames to instruct the refine of the final optical flow, in order to recognize large motion between frames.
- The work is summarized in a paper which wins **Best Paper Award** in ISCAS 2022 MSA-TC track.

PUBLICATIONS

- Yucong Liu, **Shixing Yu**, Tong Lin, *Regularizing Deep Neural Networks with Stochastic Estimators of Hessian Trace.* **In submission.**
- **Shixing Yu***, Tianlong Chen*, Jiayi Shen, Huan Yuan, Jianchao Tan, Sen Yang, Ji Liu, Zhangyang Wang, *Unified Vision Transformer Compression.* **ICLR 2022 Poster.**
- **Shixing Yu***, Zhewei Yao*, Amir Gholami*, Zhen Dong*, Michael Mahoney, Kurt Keutzer. *Hessian-Aware Pruning and Optimal Neural Implant.* **WACV 2022 Poster.**
- **Shixing Yu**, Yiyang Ma, Wei Xiang, Wenhan Yang, Jiaying Liu. *Meta-Interpolation: Time-Arbitrary Frame Interpolation via Dual Meta-Learning.* **ISCAS 2022 (MSA-TC Best Paper Award).**

AWARDS AND HONORS

Best Restored Video Quality in the Mobile Video Restoration Challenge @ 26th IEEE ICIP-19
Merit Student, Peking University, 2019, 2020 (top~10%)
DTZ scholarship, 2020 (top~5%)
1st Prize of EMC2 Model Compression Challenge in both Classification & Detection track @ EMC2 2020
Fellowship from the Cockrell School of Engineering at The University of Texas at Austin, 2021