Center for Research in Computer Vision

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Taojiannan Yang
B.S., University of Science and Technology of China, 2017

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

29 June 2023, 11:15 A.M.
HEC 119

DISSERTATION COMMITTEE

Professor Chen Chen, *Chair*, *chen.chen@crcv.ucf.edu*
Professor Mubarak Shah, *shah@crcv.ucf.edu*
Professor Liqiang Wang, *liqiang.wang@ucf.edu*
Professor Mingjie Lin, *Mingjie.Lin@ucf.edu*

DISSERTATION RESEARCH IMPACT

Deep learning has achieved remarkable breakthroughs in various domains, including computer vision, natural language processing, and speech recognition. However, as deep learning models continue to grow in complexity and size, there is an increasing need to address the challenges of efficiency and scalability. Reducing the computational requirements, memory footprint, and energy consumption of deep learning models has profound implications for the deployment of intelligent systems on resource-constrained devices such as smartphones, wearables, and IoT devices.

This dissertation focuses on developing new representation learning methods to improve the efficiency and effectiveness of deep neural networks. The works proposed in this dissertation improve the efficiency of representation learning from multiple perspectives, such as training efficiency, inference efficiency and data efficiency. The methods also improve the performance, robustness and scalability of deep neural networks. The algorithms developed in this dissertation have been applied to various tasks and domains, including image classification, object detection, instance segmentation, video action recognition and action detection. The proposed methods largely reduce the computational cost of traditional deep neural networks while maintaining competitive performance with state-of-the-art models.

SELECTED PUBLICATIONS

Total Citations: 810, h-index: 14

1. **AIM: Adapting Image Models for Efficient Video Action Recognition.** T Yang, Y Zhu, Y Xie, A Zhang, C Chen, M Li. International Conference on Learning Representations (**ICLR**), 2023.
2. **Revisiting Training-free NAS Metrics: An Efficient Trainingbased Method**. T Yang, L Yang, X Jin, C Chen. Winter Conference on Applications of Computer Vision (**WACV**), 2023.
3. **HeatER: An Efficient and Unified Network for Human Reconstruction via Heatmap-based TransformE.** C Zheng, M Mendieta, T Yang, C Chen**.** IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2023
4. **MutualNet: Adaptive ConvNet via Mutual Learning from Different Model Configurations.** T Yang, S Zhu, M Mendieta, P Wang, R Balakrishnan, M Lee, T Han, M Shah, C Chen. IEEE Transaction on Pattern Analysis and Machine Intelligence (**TPAMI**), 2022
5. **Local Learning Matters: Rethinking Data Heterogeneity in Federated Learning.** M Mendieta, T Yang, P Wang, M Lee, Z Ding, C Chen. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2022.
6. **GradAug: A New Regularization Method for Deep Neural Networks.** T Yang, S Zhu, C Chen. Neural Information Processing Systems (**NeurIPS**), 2020.
7. **MutualNet: Adaptive ConvNet via Mutual Learning from Network Width and Resolution.** T Yang, S Zhu, C Chen, S Yan, M Zhang, A Willis. European Conference on Computer Vision (**ECCV**), 2020.

DISSERTATION

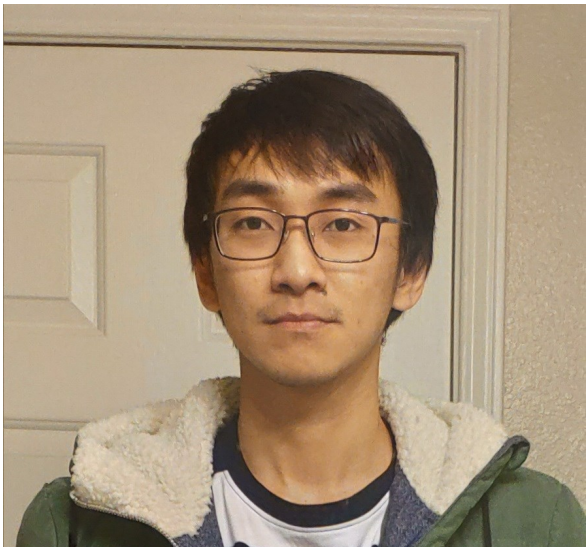
TOWARDS EFFICIENT AND EFFECTIVE REPRESENTATION LEARNING FOR IMAGE AND VIDEO UNDERSTANDING

Deep learning has demonstrated promising performance over various visual perception tasks including image classification, object detection, semantic segmentation and action recognition. However, deep neural networks are usually over-parameterized. They need to be trained on large-scale dataset and the training can take hundreds of GPU hours. During inference, large neural networks are also hard to be deployed on edge devices such as mobile phones and drones due to large memory cost and high computation complexity. In this dissertation, we propose methods to improve the efficiency of deep learning methods from multiple perspectives such as training efficiency, inference efficiency and data efficiency. Besides improving the model efficiency, our proposed methods also help deep neural networks learn better representations, which translates to better performance on various downstream tasks.

We first propose a new method to train an adaptive neural network that can run at different computation complexities during inference time. We highlight the importance of simultaneously considering network width and input resolution for efficient network design. The proposed method mutually learns from different network widths and input resolutions and enables one model to meet different resource budgets during inference. Our method outperforms traditionally neural networks on various tasks under different model complexities. It also bears the benefits of training and deploying only one model. Next, we extend this method to video understanding. Video understanding requires both spatial modeling and temporal modeling. Previous works proposed to process spatial and temporal dimensions asymmetrically for better performance and efficiency. Accordingly, we asymmetrically sample subnetworks, input resolutions and frames to do mutual training. After training, the adaptive network can run at different widths, resolutions and number of frames. We demonstrate its effectiveness and efficiency on multiple video understanding tasks including video recognition and action detection. Based on the adaptive mutual learning framework, we propose a new representation learning method for deep neural networks. Our idea is that a well-generalized network should provide consistent predictions for the same image with different augmentations, both for its subnetworks and for the network as a whole. Our method samples different sub-networks during training, feeds them with differently augmented samples, and pulls close their predictions. Our method demonstrates better performance than other state-of-the-art regularization and data augmentation methods on various network backbones and tasks. The improvement is more significant when labeled data is limited.

Besides learning methods, we also explore efficient neural architecture search methods to discover new architectures efficiently. Recent works have proposed training-free NAS metrics to accelerate the search process. However, we show that recent training-free NAS metrics are not fairly evaluated. Their performance is no better than the trivial number-of-parameter metric while being much more complicated to compute. Based on our observations, we proposed a new efficient training-based NAS method which outperforms previous methods with significantly smaller search cost. Our method is also more robust to different search spaces.

Finally, with the recent advancement in image foundation models, we propose an efficient finetuning method to adapt recent large-scale image foundation models to video understanding. Our proposed method freezes the pre-trained image model and only introduces few light-weight Adapters to tune the model. The proposed method largely saves the training cost of video models and achieves even better performance than traditional full finetuning. It also brings the benefit of data efficiency compared to full finetuning.



Taojiannan Yang

1994	Born in Sichuan, China
2013-2017	B.S., University of Science and Technology of China, Hefei, Anhui, China
2021	Research Intern, ByteDance Inc
2022	Applied Scientist Intern, Amazon Web Services