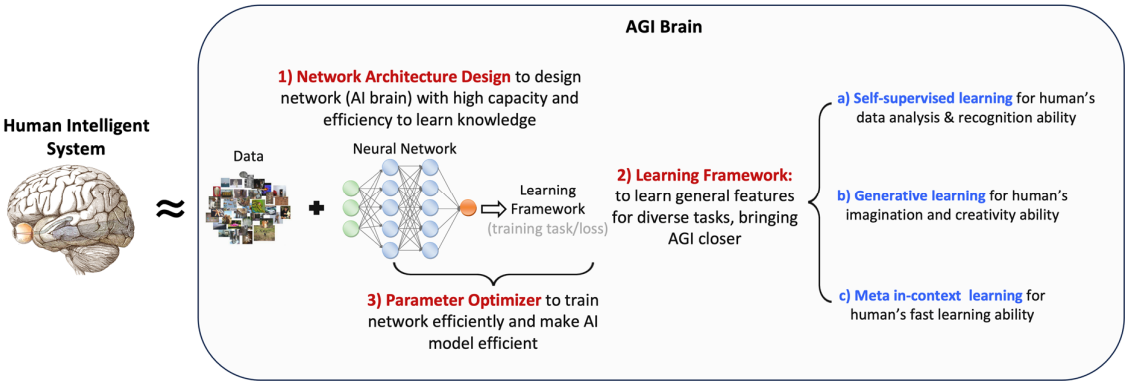
Research Statement

ZHOU Pan
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6808 5213; Email: panzhou@smu.edu.sg
28-December-2023

Background



My primary research interests are in the fields of machine learning, optimization, and computer vision, aiming at developing an efficient and effective artificial general intelligence (AGI) system to comprehend language, speech, and vision data and to solve a wide range of related tasks like a human. While intelligent language and speech systems have witnessed rapid advancements with the emergence of highly capable models like GPT and Wav2vec, intelligent vision system, particularly before 2019, suffers from much slow progress in developing generally capable models. This motivates me to study efficient and effective intelligent vision systems for addressing the limitations of AGI. Moreover, by tackling various vision tasks, e.g., image and video recognition and segmentation, intelligent vision system also arises in many real contexts, e.g., healthcare & medical diagnosis, automatic driving, and security monitoring, and thus is also highly desired in practice.

This intelligent vision system contains three main parts.

1) Network Architecture Design. It is to develop innovative network topology that has high capacity and efficiency to acquire vision knowledge, thus improving the vision model capacity of AGI.

2) Learning Framework. In this line, my research includes **a)** self-supervised learning that enables AI models to learn general knowledge from large-scale vision data and achieve human's strong recognition and analysis ability on the vision data; **b)** generative learning (e.g., diffusion models) which empowers AI models with human's imagination and creativity; **c)** meta in-context learning that aims to improve the few-shot / fast learning ability of AI models.

3) Parameter Optimizer. After having AI model and its learning framework, parameter optimizer aims to train AI models efficiently and reduce training costs, which is crucial to improve the efficiency and accessibility of intelligent vision system and AGI, as training large models with big data is costly, e.g., million-dollar training cost of GPT3.

Research Areas

With this goal in mind, my research works mainly focus on the following three aspects.

1) Parameter Optimizer

It aims to train AI models efficiently and thus makes AGI more efficient, as AI models to pursue AGI are often huge and need tons of data for costly training. In this line, I proposed a serial of works [1-7], among which two notable ones include a) Adan [1], a faster optimizer, and b) Win [2], a unified acceleration framework.

Faster Optimizer – Adan. Adan first computes an extrapolation point of the current parameter, and then uses the geometry curvature information (e.g., gradient) of the training objective at the extrapolation point to correct the first- and second-order moments in Adam for faster convergence. On non-convex problems, Adan finds an ϵ-accurate stationary point within $O(\epsilon^{-3.5})$ stochastic gradient complexity, and is faster than the popular optimizers, e.g., Adam. In practice, Adan is about 2× faster than the SoTA optimizers, e.g. SGD and Adam, while achieving higher or comparable performance on many networks, e.g., ViTs and MAE in the CV field, GPT2 and billion-scale LLaMA in the NLP field, UNet and ViTs in AIGC field, and MLPs in the RL field. Finally, Adan can use a large range of batch sizes, e.g., from 1k to 32k that fails most optimizers, and greatly accelerates model training with large batchsize.

Plug-and-play acceleration framework – Win. It integrates proximal point method with Nesterov acceleration to provide a simple and effective plug-and-play acceleration framework for Adan and other optimizers. Win can theoretically accelerate adaptive gradient algorithms, e.g., Adam, and empirically accelerates Adan, AdamW, Adam, and SGD by 1.5× speed with superior performance on the vision and language modelling tasks.

Significance. Adan and Win are mainly to reduce the network training cost in real scenarios. Thanks to its excellent performance, Adan has been used in multiple popular network training codebases, e.g., HuggingFace Timm (GitHub Star: 25k), OpenMMLab (GitHub Star: 24k), and Jittor of Tsinghua University (GitHub Star: 2.8k).

2) Learning Framework

In this line, I explored a) self-supervised learning [8–14] for human's strong recognition and analysis ability, b) generative learning [15–18] for human's good imagination and creativity ability, and c) meta in-context learning for human's fast learning ability.

a) Self-supervised learning (SSL). For SSL, it designs unsupervised training tasks to train models on the unlabeled data. Nowadays, it has shown big potential than supervised learning in learning highly qualified and well-transferable representations for handling various (vision) tasks like classification, detection and segmentation, and has achieved promising recognition and analysis ability.

In this line, I proposed a novel self-supervised multi-granular clustering learning framework which has three distinct approaches: PCL [8], SANE [9], and Mugs [10]. PCL clusters similar samples together to learn high-level clustering semantics that are critical for many vision tasks, e.g., classification. SANE learns higher-level semantics among data clusters via pulling similar clusters in PCL closer while pushing dissimilar ones apart, and can better distinguish different clusters. Lastly, Mugs explores multi-granular cluster structures in data via hierarchical clustering techniques, and thus captures multi-granular features, including coarse- and fine-grained features. So Mugs can better handle various vision tasks by selecting suitable features from the learnt multi-granular features through fine-tuning. These works address the limitations of previous SSL methods like MoCo that scatter all individual samples apart and cannot learn high-level clustering semantic features.

b) Generative learning. Regarding generative learning, it aims at allowing a machine to generate and edit input data with unparalleled flexibility, e.g., image synthesis and editing, and is a potential technique to achieve human's imagination and creativity ability. Both SSL and generative learning target on enabling AI models to mimic human's abilities for handling diverse tasks, thus bringing AGI closer.

My proposed MDT model [16] designs latent mask modeling to explicitly enhance its contextual relation learning among image regions, and better learns the overall image semantics, enjoying about 3× faster learning speed than previous SoTAs. Then, my EditAnything model [17] empowers users with high flexibility in generating and editing image, e.g., cross-image dragging (like try-on) and region-interactive editing (like hairstyle changing, object or person replacement). To edit an image region, EditAnything considers its contextual relations with other semantic regions, and edits all related regions together to achieve overall realistic results.

Significance. These two series of work have great practical significance, and achieve impressive performance on various visual analysis tasks (e.g., classification, object detection and segmentation), and visual synthesis tasks (e.g., image synthesis and editing), thus bringing human's analysis and creativity abilities closer. Notably, Mugs obtains the SoTA classification performance on the widely-used ImageNet dataset without extra data, and MDT achieves SoTA image synthesis performance on ImageNet. The novelties of these works are 1) high-level discriminative semantic learning in SSL: PCL (citation 670+), SANE and Mugs learn high-level semantic cluster structures, and address the limitations of previous SSL methods like MoCo in learning semantic features; and 2) contextual relation learning for generative learning: MDT and EditAnything generate and edit the image region by considering its contextual relations with other image regions for overall authenticity. With impressive performance and potential for real applications, EditAnything attains 2.6k GitHub stars.

c) Meta In-Context Learning. Both supervised and self-supervised (vision) models often suffer from unsatisfactory performance on few-shot vision tasks. So I studied meta-learning and prompt learning to aid a (pretained) model to quickly learn from a few amount of contextual data, which ultimately help to solve few-shot learning challenges in the AGI system.

Meta-learning learns prior knowledge from multi-tasks and adapts this knowledge to new environments with a few examples. My proposed meta-learning methods [19–21] only need gradient for training, and avoid Hessian matrix computation in existing methods, significantly enhancing its scalability on model and data. I also proved their fast learning ability on a few data, and applied them for medical diagnosis [22] and low-resource speech recognition [23]. To apply a self-supervised model to vision tasks, I designed a new prompt learning method [24] which uses a few data to adjust the task input so that the input can be proficiently processed by the pretrained model. These works have many real applications, especially for low-resource scenarios.

Significance. This series of work have great practical significance and achieve impressive performance on various few-shot learning tasks, such as medical diagnosis tasks for orphan diseases. So they can mimic human's fast learning ability to learn from a few examples, and apply the learnt knowledge to the tasks of interest.

3) Network Architecture

AI development has been largely driven by model architecture revolutions, from CNNs to Vision Transformers (ViTs). It is always worth exploring new architecture designs to lift the model capacity and efficiency for knowledge and feature acquisition. On this territory, I proposed a series of works [25-31], including a) manually designed networks, e.g., MetaFormer [25], IFormer [26] and CAFormer [27], and b) automatically designed network, such as PR-DARTS [28].

Manually designed networks. MetaFormer replaces complex attention in ViTs with a simple and efficient pooling operation, and surpasses strong CNN and ViT baselines, e.g. (RSB-)ResNet, ViT and PVT, on the image classification and segmentation tasks. This observation breaks the widely-accepted slogan that "attention is all you need". Lucas Beyer, the first author of ViT, also praised this work as "incredible" on Twitter. These results greatly deepen the understanding of the role of attention in ViTs.

Based on MetaFormer, IFormer and CAFormer aim to improve ViTs. IFormer finds that attention can well learn low-frequency information in data (e.g. object shape), but is poor at capturing high frequency that mainly conveys local information (e.g. texture) and is critical for many tasks. So IFormer grafts the advantages of convolution and max-pooling in capturing high-frequency information to attention for learning frequency-comprehensive features in visual data. CAFormer further discovers that learning local features first in shallow layers through convolution operations and then global semantics in deep layers through attention is more effective. CAFormer sets a new recording accuracy of 85.5% on ImageNet dataset under supervised settings without extra data.

Automatically designed network. I designed a theory-inspired PR-DARTS framework to automatically design effective network architectures, reducing the reliance on expert trial and error. PR-DARTS provides the first theory to show why previous network search methods (a.k.a. AutoML) often collapse due to selecting too many skip-connections, and then proposes a new method that can avoid previous collapse and thus automatically selects and combines various network operations, e.g. different sized pooling and convolution operations, to search more effective network architectures.

Significance. The insights of these works are 1) breaking the slogan that "attention is all you need", and deepening the understanding of attention in ViTs; 2) instead of pursuing novel attention variants, using the simple and efficient operations (e.g. pooling, convolution, attention) can still achieve impressive and even SoTA performance; 3) automatically selecting and combining various network operations to build network architectures greatly reduces the reliance on expert trial and error. Moreover, these works also have practical significance, since MetaFormer is used in Wechat for video recommendation and content audit because of its high efficiency.

Selected Publications and Outputs

[1]. Xingyu Xie*, Pan Zhou*, Huan Li, Zhouchen Lin, Shuicheng Yan. Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models, submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence. (* equal contribution)

[2]. Pan Zhou, Xingyu Xie, Shuicheng Yan. Win: Weight-Decay-Integrated Nesterov Acceleration for Adaptive Gradient Algorithms. International Conference on Learning Representations (ICLR), 2023 (oral)

[3]. Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E. Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning. Neural Information Processing Systems (NeurIPS), 2020

[4]. Pan Zhou, Hanshu Yan, Xiaotong Yuan, Jiashi Feng, Shuicheng Yan. Towards Understanding Why Lookahead Generalizes Better Than SGD and Beyond. Neural Information Processing Systems (NeurIPS), 2021

[5]. Pan Zhou and Xiaotong Yuan. Hybrid Stochastic-Deterministic Minibatch Proximal Gradient: Less-Than-Single-Pass Optimization with Nearly Optimal Generalization. International Conference on Machine Learning (ICML), 2020.

[6]. Pan Zhou, Xiaotong Yuan, Shuicheng Yan, Jiashi Feng. Faster first-order methods for stochastic non-convex optimization on Riemannian manifolds. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2), 459 – 472, 2021.

[7]. Pan Zhou, Xiaotong Yuan, Jiashi Feng. Efficient Stochastic Gradient Hard Thresholding. Neural Information Processing Systems (NeurIPS), 2018

[8]. Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven Hoi. Prototypical Contrastive Learning of Unsupervised Representations. International Conference on Learning Representations (ICLR), 2021

[9]. Pan Zhou, Caiming Xiong, Xiaotong Yuan, Steven Hoi. A Theory-Driven Self-Labeling Refinement Method for Contrastive Representation Learning. Neural Information Processing Systems (NeurIPS), 2021 (spotlight)

[10]. A multi-granular self-supervised learning framework A multi-granular self-supervised learning framework Pan Zhou*, Yichen Zhou*, Chenyang Si*, Weihao Yu, Teck Khim Ng, Shuicheng Yan (* equal contribution)

[11]. Jiachun Pan*, Pan Zhou*, Shuicheng Yan. Towards Understanding Why Mask Reconstruction Pretraining Helps in Downstream Tasks. International Conference on Learning Representations (ICLR), 2023 (* equal contribution)

[12]. Pan Zhou, Canyi Lu, Jiashi Feng, Zhouchen Lin, Shuicheng Yan. Tensor low-rank representation for data recovery and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5), 1718 – 1732, 2021.

[13]. Pan Zhou, Canyi Lu, Zhouchen Lin, Chao Zhang. Tensor factorization for low-rank tensor completion. IEEE transactions on image processing, 27(3), 1152-1163, 2017.

[14]. Pan Zhou, Jiashi Feng. Outlier-Robust Tensor PCA. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15]. Yue Wu, Pan Zhou, Andrew Gordon Wilson, Eric Xing, and Zhiting Hu. Improving GAN Training with Probability Ratio Clipping and Sample Reweighting. Neural Information Processing Systems (NeurIPS), 2020

[16]. Shanghua Gao, Pan Zhou, Ming-Ming Cheng, Shuicheng Yan. Masked Diffusion Transformer is a Strong Image Synthesizer. International Conference on Computer Vision (ICCV), 2023

[17]. Shanghua Gao, Zhijie Lin, Xingyu Xie, Pan Zhou*, Mingming Chen, and Shuicheng Yan. EditAnything via text command. In https: // github. com/ sail-sg/ EditAnything , 2023 (*corresponding author).

[18]. Zhongzhan Huang, Pan Zhou, Shuicheng Yan, Liang Lin. Towards More Stable Training of Diffusion Model via Scaling Network Long Skip Connection. Neural Information Processing Systems (NeurIPS), 2023

[19]. Pan Zhou, Xiaotong Yuan, Huan Xu, Shuicheng Yan, Jiashi Feng. Efficient Meta Learning via Minibatch Proximal Update. Neural Information Processing Systems (NeurIPS), 2019 (spotlight)

[20]. Pan Zhou, Yingtian Zou, XiaoTong Yuan, Jiashi Feng, Caiming Xiong, and Steven Hoi. Task Similarity Aware Meta Learning: Theory-inspired Improvement on MAML. International Conference on Uncertainty in Artificial Intelligence (UAI), 2021

[21]. Yu Bai, Minshuo Chen, Pan Zhou, Tuo Zhao, Jason D. Lee, Sham Kakade, Huan Wang, Caiming Xiong. How Important is the Train-Validation Split in Meta-Learning? International Conference on Machine Learning (ICML), 2021

[22]. Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen and Liang Lin. Graph-Evolving Meta-Learning for Low-Resource Medical Dialogue Generation. Association for the Advancement of Artificial Intelligence (AAAI), 2021

[23]. Guolin Zheng, Yubei Xiao, Ke Gong, Pan Zhou, Xiaodan Liang, and Liang Lin. Wav-BERT: Cooperative Acoustic and Linguistic Representation Learning for Low-Resource Speech Recognition. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021 (Findings)

[24]. Bowen Dong, Pan Zhou, Shuicheng Yan, Wangmeng Zuo. LPT: Long-tailed Prompt Tuning for Image Classification. International Conference on Learning Representations (ICLR), 2023

[25]. Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, Shuicheng Yan. MetaFormer is Actually What You Need for Vision. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022 (oral)

[26]. Chenyang Si*, Weihao Yu*, Pan Zhou, Yichen Zhou, Xinchao Wang, Shuicheng Yan. Inception Transformer. Neural Information Processing Systems (NeurIPS), 2022 (oral) (*equal contribution)

[27]. Weihao Yu, Chenyang Si, Pan Zhou, Mi Luo, Yichen Zhou, Jiashi Feng, Shuicheng Yan, and Xinchao Wang. MetaFormer Baselines for Vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[28]. Pan Zhou, Caiming Xiong, Richard Socher, and Steven Hoi. Theory-Inspired Path-Regularized Differential Network Architecture Search. Neural Information Processing Systems (NeurIPS), 2020 (oral)

[29]. Yuxuan Liang, Pan Zhou, Roger Zimmermann, Shuicheng Yan. DualFormer: Local-Global Stratified Transformer for Efficient Video Recognition. European Conference on Computer Vision (ECCV), 2022

[30]. Junbin Xiao, Pan Zhou, Tat-Seng Chua, Shuicheng Yan. Video Graph Transformer for Video Question Answering. European Conference on Computer Vision (ECCV), 2022

[31]. Pan Zhou, Jiashi Feng. Understanding Generalization and Optimization Performance of Deep CNNs. International Conference on Machine Learning (ICML), 2018

**See Full Publication List in Google Scholar Profile