YOUNG SCIENTIST AWARD 2025

Wang Xinchao

Presidential Young Professor
National University of Singapore

"For advancing machine learning techniques that train compact AI models using limited resources, while achieving the capabilities of larger AI systems."

+++

Recent advances in machine learning have driven significant breakthroughs across diverse artificial intelligence (AI) applications, transforming many aspects of daily life. However, these achievements have also introduced substantial challenges. Advanced models now depend on massive datasets, increasingly complex architectures, and training processes that can span weeks or months on clusters of thousands of GPUs. Such demands hinder the deployment of AI on resource-constrained platforms such as edge devices and mobile systems, and limit the ability of researchers with modest resources to tailor models to their needs. They also raise critical concerns regarding environmental impact and long-term sustainability.

Assistant Professor Wang Xinchao's research addresses these challenges through efficient machine learning. His work streamlines training and inference while designing compact, high-performance architectures. By lowering computational and financial barriers, his innovations enable smaller laboratories, startups and individual researchers to train competitive models using limited hardware, and facilitate deployment on platforms with strict computational and memory constraints. Optimising model size, speed and efficiency not only advances accessibility but also contributes to energy conservation and sustainable AI practices. The significance of these contributions has been recognised through prestigious honours, including the Institute of Electrical and Electronics Engineers (IEEE) AI's *10 to Watch* and the National University of Singapore's *Young Researcher Award*.

Asst Prof Wang's research spans three interconnected domains: efficient strategies, efficient models and efficient data.

In efficient strategies, he has developed methods to derive smaller, faster and more effective models from pre-trained networks. A prominent contribution is DepGraph, the world's most widely adopted structural pruning scheme, and the first fully automated approach to tracing neuron interdependencies in neural networks. Where previous methods required laborious manual tracing, DepGraph reduces the task to three lines of code, eliminating days of effort while maintaining full flexibility. Its open-source implementation, Torch-Pruning, has become the most popular structural pruning library in the community, with over 290,000 downloads, and has been integrated into NVIDIA's commercial products, underscoring its industrial impact.

In efficient models, his focus is on designing new network architectures that deliver state-of-the-art performance at reduced computational cost and model size. A notable breakthrough is MetaFormer, one of the most widely adopted efficient transformer backbones. By replacing the computationally intensive self-attention mechanism with a simple pooling operation, MetaFormer achieves competitive or superior performance while challenging the conventional view that self-attention is the defining component of transformer architectures. Its efficiency and versatility have led to broad adoption across applications, influencing the design of lightweight, high-performance models.

In efficient data, Asst Prof Wang has advanced training efficiency through innovative approaches to data representation. His pioneering work on Dataset Distillation (DD) condenses large datasets into smaller, representative synthetic sets that preserve or enhance model performance while reducing computational burden. He introduced Dataset Factorization, the first decomposition-based approach to DD, which overcame the limitations of earlier methods that treated synthetic samples independently. By factorising a dataset into bases and a hallucination network that together generate synthetic samples, Dataset Factorization captures inter-sample coherence and achieves state-of-the-art performance. This paradigm has since become a dominant framework in DD, inspiring numerous follow-up studies and applications.

By lowering barriers to AI development, enabling deployment on constrained platforms, and reducing energy consumption, Asst Prof Wang’s work advances the goal of making AI more accessible, sustainable and impactful. In doing so, it reinforces Singapore’s standing as a global hub for cutting-edge AI research and innovation, while fostering a dynamic and internationally competitive scientific ecosystem.