University of Pittsburgh | School of Computing and Information

Proposal Defense

Doctor of Philosophy in Computer Science

"**Towards Efficient and Effective Visual Intelligence**" by **Sheng Li**

Date: April 3, 2026
Time: 11 a.m. – 1 p.m.
Place: 6106 Sennott Square, 210 S. Bouquet Street, Pittsburgh PA 15260

Committee:

- Xulong Tang, Associate Professor, Computer Science, University of Pittsburgh
- Youtao Zhang, Professor, Computer Science, University of Pittsburgh
- Stephen Lee, Assistant Professor, Computer Science, University of Pittsburgh
- Bo Yuan, Associate Professor, Electrical and Computer Engineering, Rutgers University

Abstract:
Deep learning has driven remarkable progress in visual intelligence, achieving strong performance across a wide range of tasks, from conventional visual recognition and segmentation to the emerging frontier of visual content generation. As models grow in scale, however, computational efficiency has become a fundamental challenge. Training large-scale models is already resource-intensive, and this challenge is further compounded by the difficulty of obtaining labeled data, as annotation is labor-intensive and often infeasible at scale. Self-supervised learning offers a compelling alternative by learning without labels, but often demands substantially higher (over 10x) computation than its supervised counterpart. The efficiency challenge becomes even more pronounced in video generative tasks, which require processing dense spatiotemporal token sequences across both spatial and temporal dimensions and therefore incur substantial computation cost.

This proposal studies how to build efficient and effective visual intelligence systems by identifying and exploiting structured redundancy. First, it investigates adaptive training strategies that selectively reduce unnecessary optimization for network layers that have already stabilized during model training, thereby reducing both computation and memory cost. Second, it studies efficient self-supervised learning by removing redundant and less informative regions across augmented views to reduce computation cost while preserving the semantic consistency needed for effective representation learning. Lastly, it develops efficient video generation methods that dynamically allocate computation according to both spatial and temporal content complexity, improving efficiency while preserving visual fidelity and coherent motion.