EFFICIENT KNOWLEDGE TRANSFER FROM VISION FOUNDATION MODELS FOR DEPLOYABLE EDGE AI IMAGE CLASSIFICATION

ABSTRACT

The convergence of Deep Learning and large-scale unlabeled data has led to the rise of Foundation Models—self-supervised architectures like Vision Transformers (ViTs), DINOv2, and Masked Autoencoders (MAE)—that achieve exceptional image classification performance. However, their massive parameter counts and computational demands hinder deployment on resource-constrained Edge AI devices.

This research aims to bridge this gap by developing an efficient **compression and distillation framework** for transferring the semantic richness of self-supervised ViT foundation models to lightweight, edge-optimized networks. The proposed approach introduces: (1) **Feature-centric compression** using quantization and sparsity-aware pruning guided by self-supervised alignment metrics; (2) **Self-supervised knowledge distillation**, transferring both global and patch-level representations from large ViT models (e.g., DINOv2-Giant) to compact student networks; and (3) **Hardware-aware optimization** for real-time inference and energy-efficient deployment.

The expected outcome is a scalable, low-latency vision framework that preserves near-foundation-level accuracy while drastically reducing computational footprint. This work establishes a pathway toward high-performance, energy-efficient image classification on edge platforms, enabling practical adoption of advanced Deep Learning in real-world vision applications.

Keywords: Deep Learning, Vision Transformers, Self-Supervised Learning, Model Compression, Knowledge Distillation, Edge AI, Image Classification