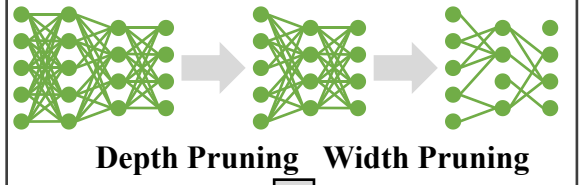
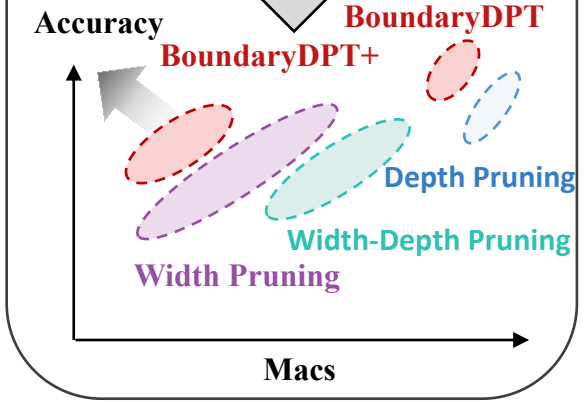
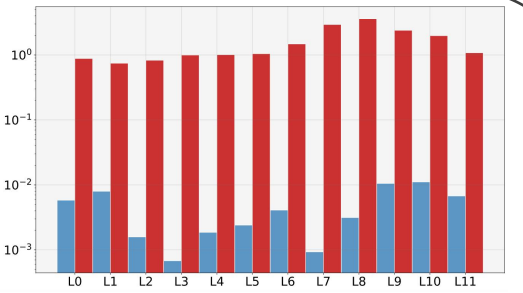
Goals

Our goal is offer an **enhanced accuracy-speedup Pareto frontier** for Vision Transformer by making full use of the sparsity

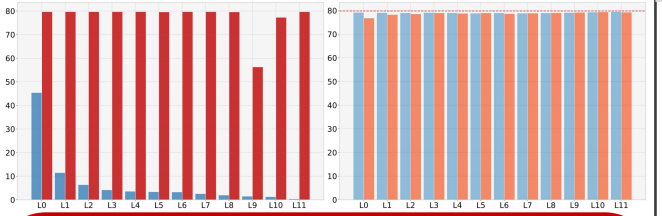




Key Insights

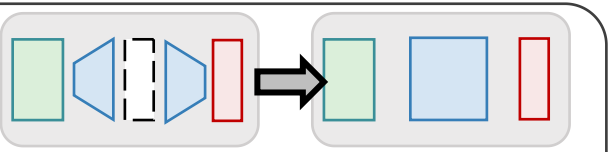


Gradient Disparity: Significant gradient differences between self-attention and activation layers

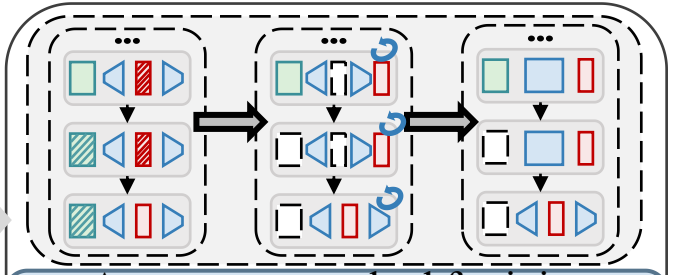


Recovery Asymmetry: Activation layer pruning causes larger accuracy drops but recovers quickly after fine-tuning, while self-attention layers exhibit the opposite behavior

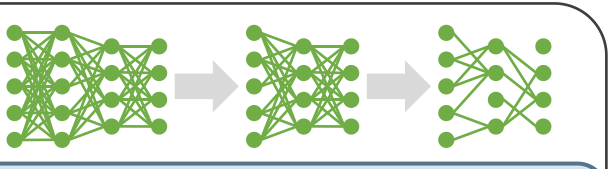
Contributions



The first to identify and mitigate activation redundancy in ViT



A two-stage method for joint pruning to push the boundary of depth pruning



combined with width pruning for extreme compression, BoundaryDPT+ sets a new sota record in ViT pruning

Results