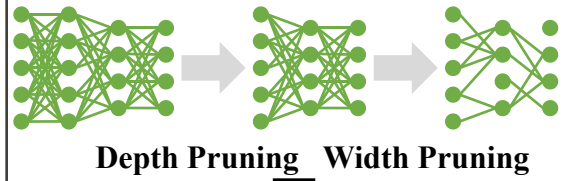
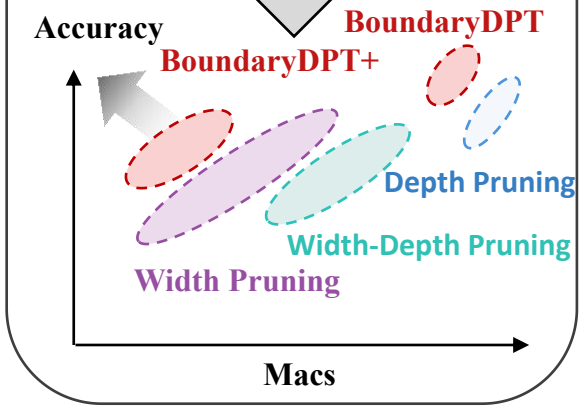
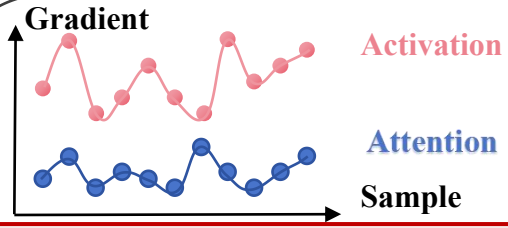
Goals

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

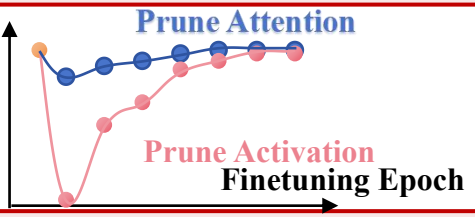




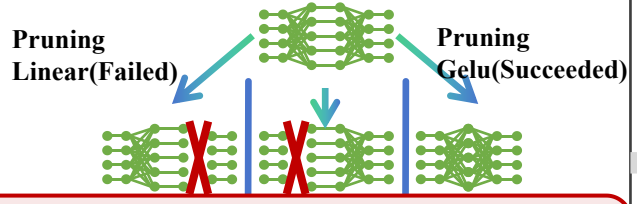
Key Insights



Gradient Disparity: Large gradient differences between attention and activation

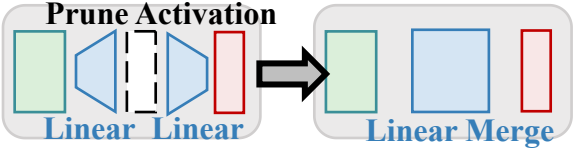


Recovery Asymmetry: Activation pruning hurts acc more but recovers quickly, while Attention shows the opposite.

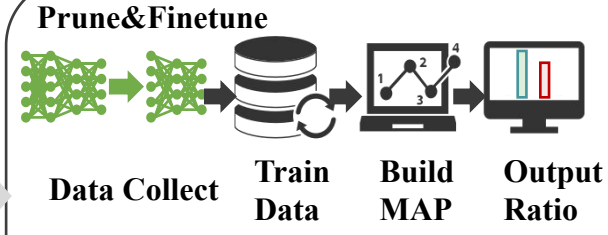


Dimensions Mismatch: Pruning attention and linear together breaks tensor compatibility, making jointly depth-pruned ViTs infeasible.

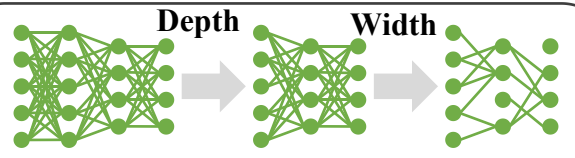
Contributions



The first to identify and mitigate activation redundancy in ViT

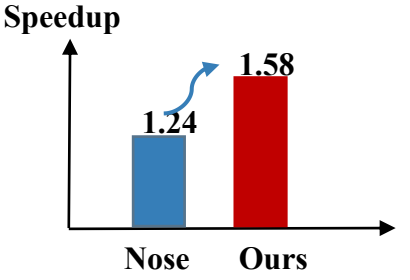


A two-stage method featuring a model accuracy predictor to manage heterogeneity.

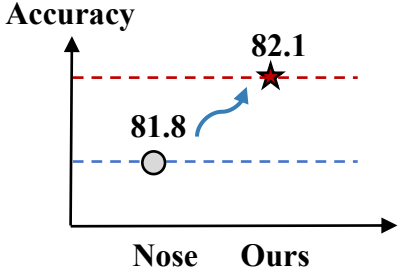


combined with width pruning for extreme compression, BoundaryDPT+ sets a new sota record in ViT pruning

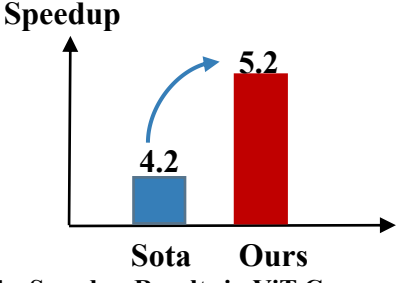
Results



The Speedup Results in DepthPruning



The Accuracy Results in DepthPruning



The Speedup Results in ViT Compression