PareCO: Pareto-aware Channel Optimization for Slimmable Neural Networks

Ting-Wu (Rudy) Chin

Ari S. Morcos

Diana Marculescu
Slimmable Neural Networks

One set of weights, multiple networks on the trade-off front!
Why Slimmable Neural Networks?

Reduce model maintenance cost

Runtime optimization
The Gap

Conventional Slimmable Neural Networks

PareCO-optimized Slimmable Neural Networks

0.56× FLOPs  0.25× FLOPs  0.12× FLOPs
How can we optimize slimmable neural networks with flexible widths?
The objective of our problem

\[ \min_{\theta} \mathbb{E}_{x,y} \mathbb{E}_{\lambda} L_{CE}(\theta; x, y, \alpha^*) \]

\textbf{s.t.} \quad \alpha^* = \arg \min T_\lambda (\alpha; \theta, x, y)
ImageNet: Compared to conventional slimmable neural networks

MobileNetV2

MobileNetV3

Independently trained from (Sandler et al., 2018)

Independently trained from (Howard et al., 2019)
Takeaways

• Optimizing the layer-wise channel counts for the sub-networks in slimmable neural networks allows for better trade-off between prediction error and FLOPs

• This work provides a principled formulation and a practical algorithm for optimizing the layer-wise channel counts for slimmable neural networks