ParеCO: Pareto-aware Channel Optimization for Slimmable Neural Networks

Carnegie Mellon University
Electrical & Computer ENGINEERING

.AI

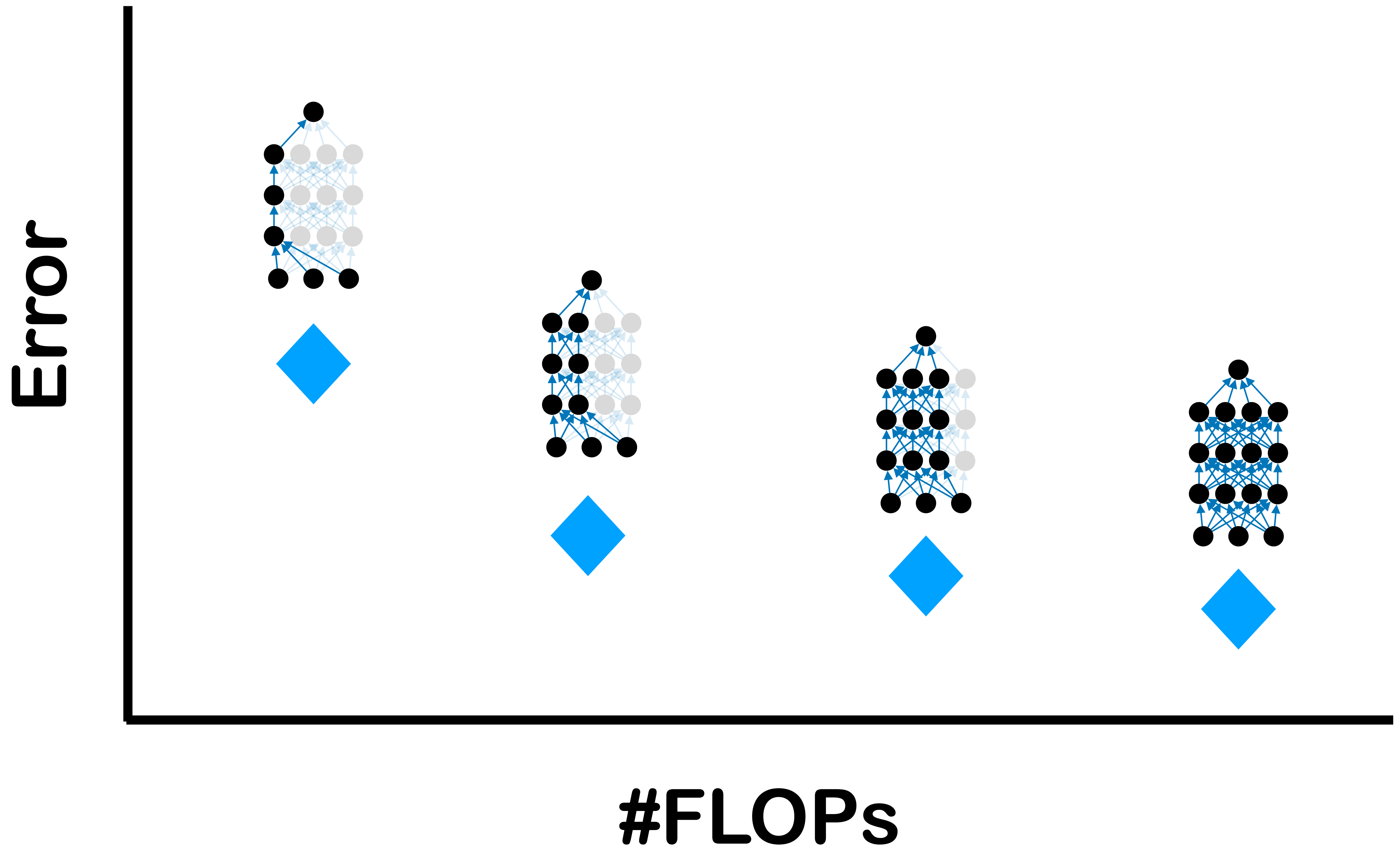
The University of Texas at Austin
Electrical and Computer Engineering
Cockrell School of Engineering

Ting-Wu (Rudy) Chin

Ari S. Morcos

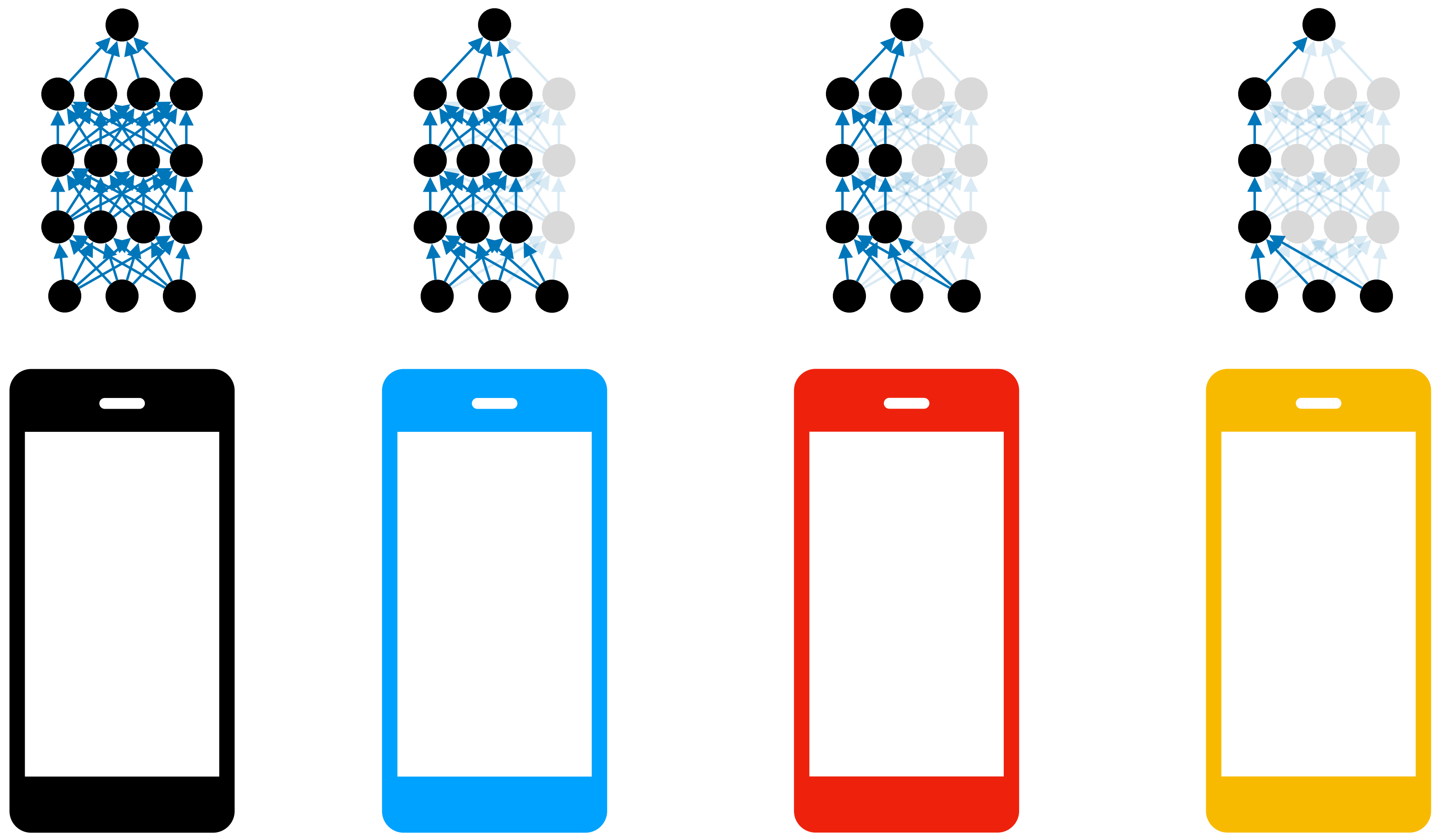
Diana Marculescu

Slimmable Neural Networks

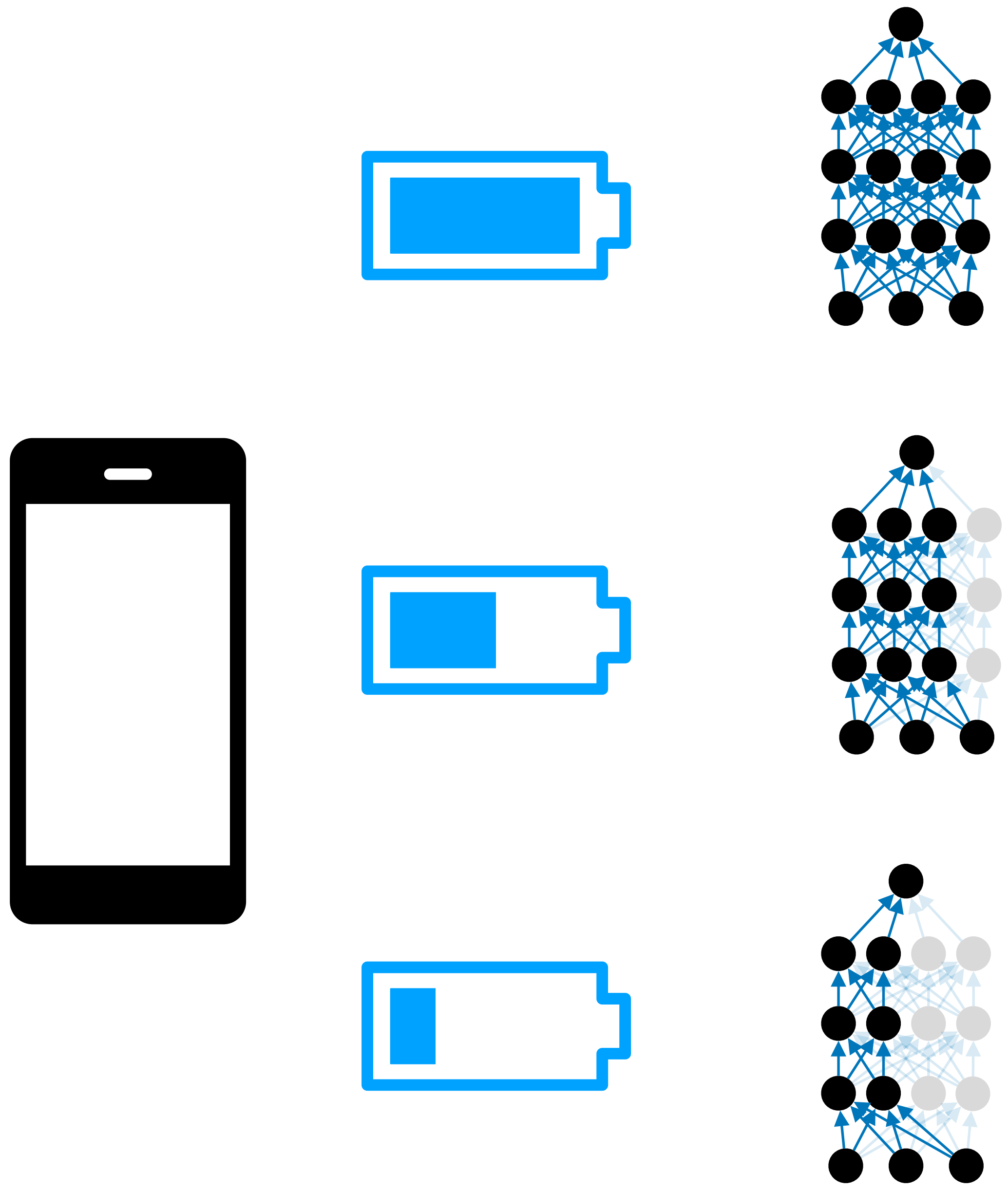


One set of weights, multiple networks on the trade-off front!

Why Slimmable Neural Networks?

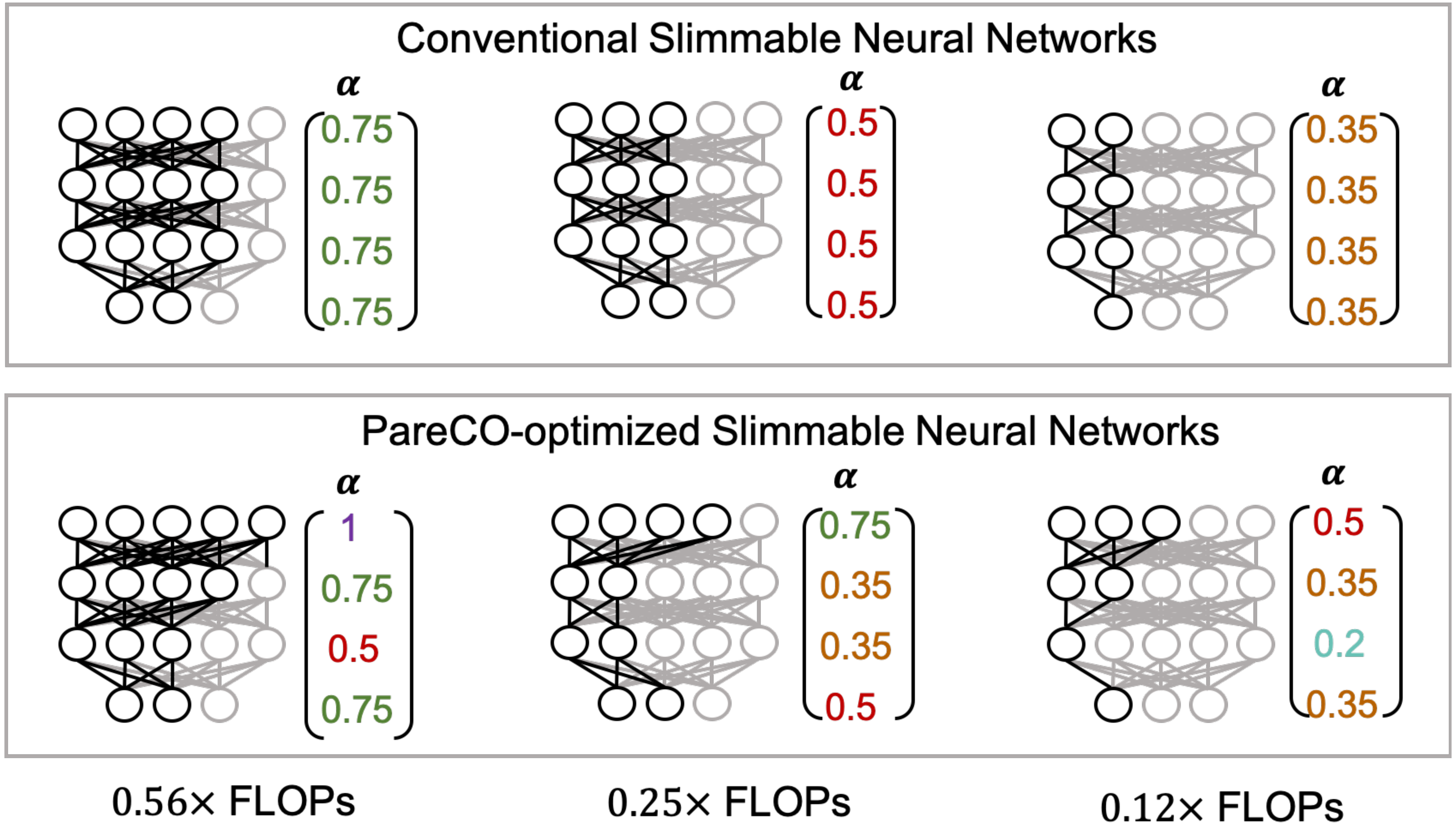


Reduce model maintenance cost



Runtime optimization

The Gap



Conventional Slimmable Neural Networks

$$\boldsymbol{\alpha} \begin{pmatrix} 0.75 \\ 0.75 \\ 0.75 \\ 0.75 \end{pmatrix} \qquad \boldsymbol{\alpha} \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \qquad \boldsymbol{\alpha} \begin{pmatrix} 0.35 \\ 0.35 \\ 0.35 \\ 0.35 \end{pmatrix}$$

ParetoCO-optimized Slimmable Neural Networks

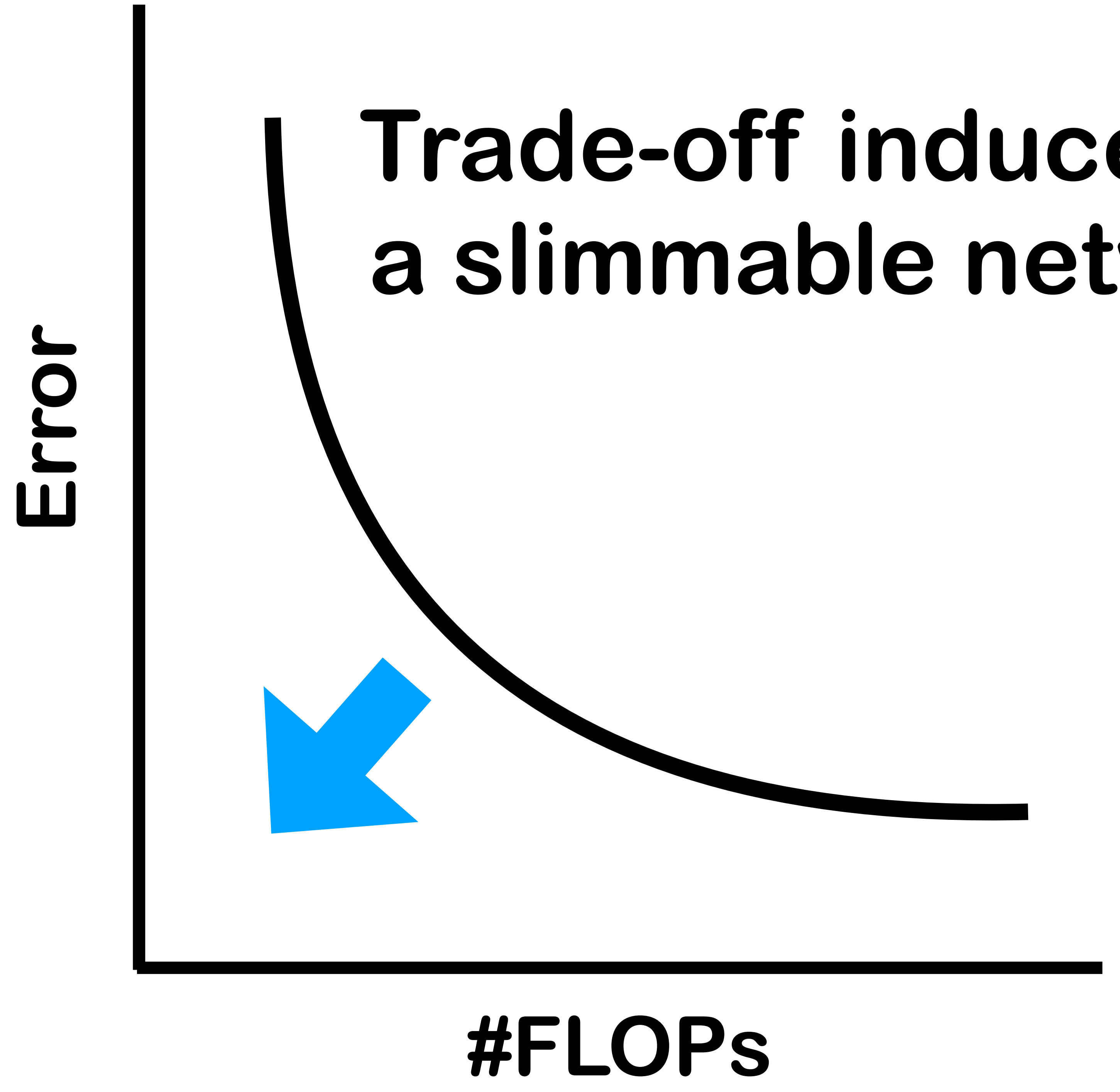
$$\boldsymbol{\alpha} \begin{pmatrix} 1 \\ 0.75 \\ 0.5 \\ 0.75 \end{pmatrix} \qquad \boldsymbol{\alpha} \begin{pmatrix} 0.75 \\ 0.35 \\ 0.35 \\ 0.5 \end{pmatrix} \qquad \boldsymbol{\alpha} \begin{pmatrix} 0.5 \\ 0.35 \\ 0.2 \\ 0.35 \end{pmatrix}$$

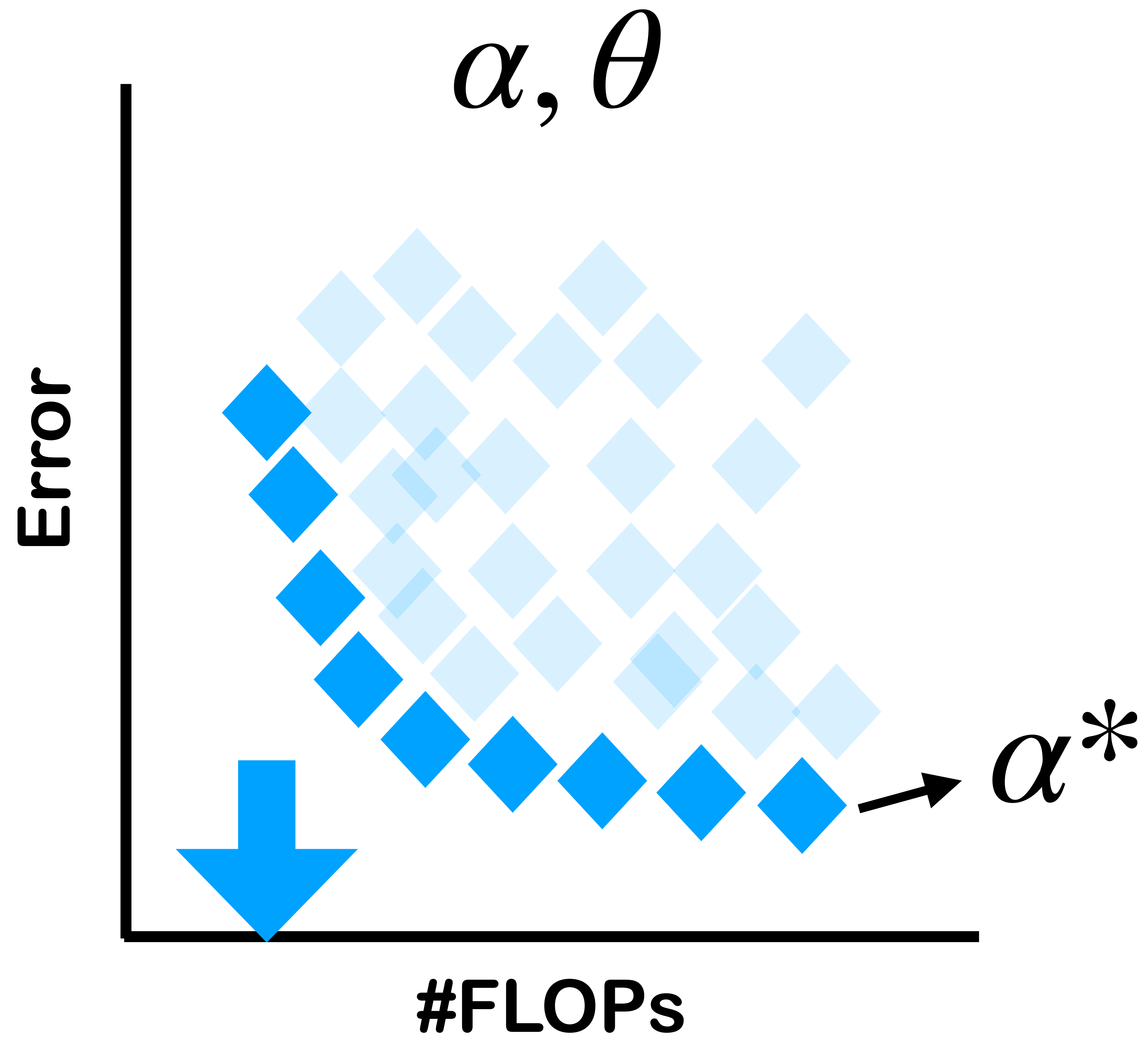
0.56× FLOPs

0.25× FLOPs

0.12× FLOPs

How can we optimize slimmable neural networks with flexible widths?


Trade-off induced by a slimmable network



The objective of our problem

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

$$\textbf{s.t. } \alpha^* = \arg\min T_{\lambda}(\alpha; \theta, x, y)$$

A Flexible Framework for Multi-Objective Bayesian Optimization using Random Scalarizations

Biswajit Paria
MLD, Carnegie Mellon University
bparia@cs.cmu.edu

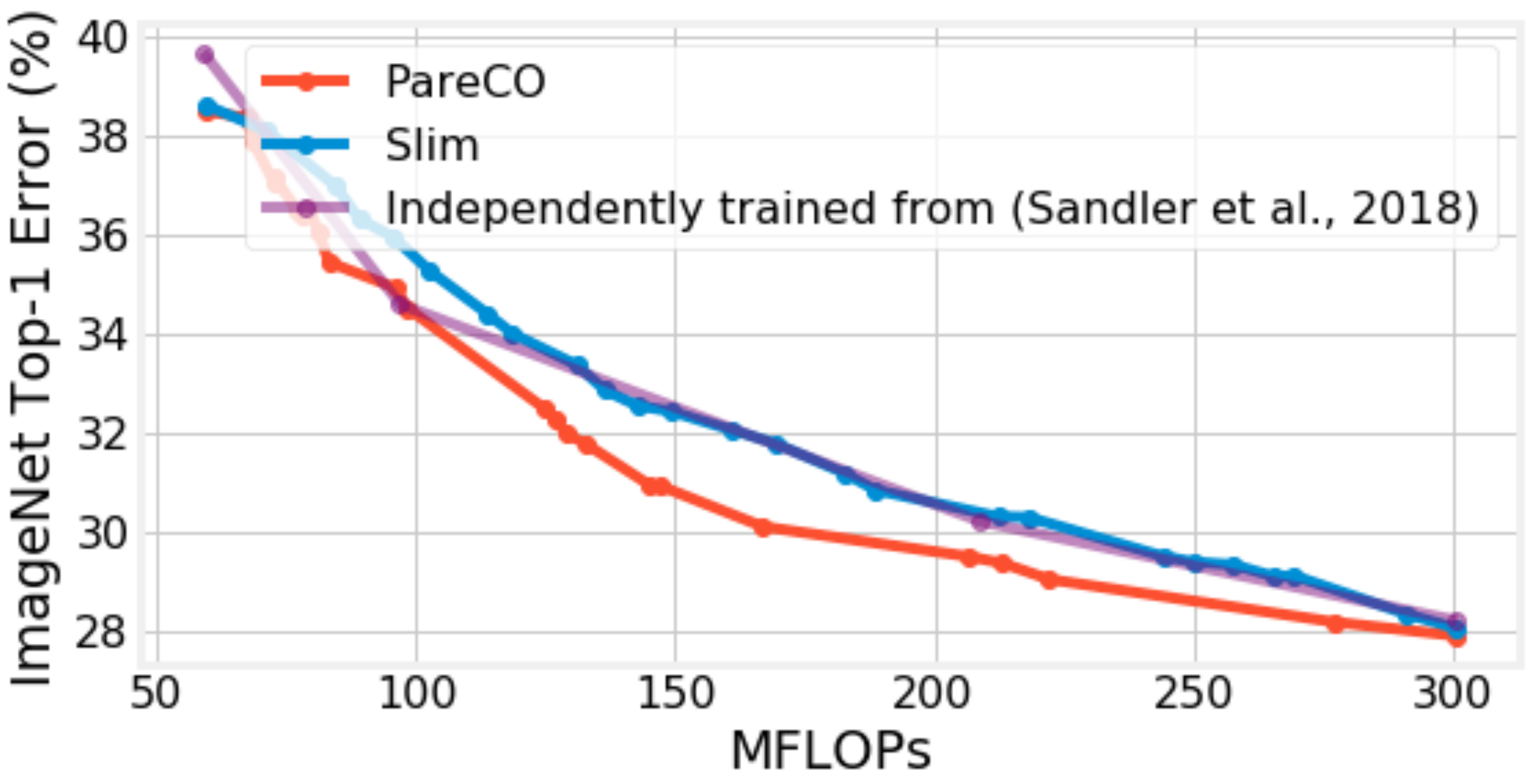
Kirthevasan Kandasamy*
EECS, UC Berkeley
kandasamy@eecs.berkeley.edu

Barnabás Póczos
MLD, Carnegie Mellon University
bapoczos@cs.cmu.edu

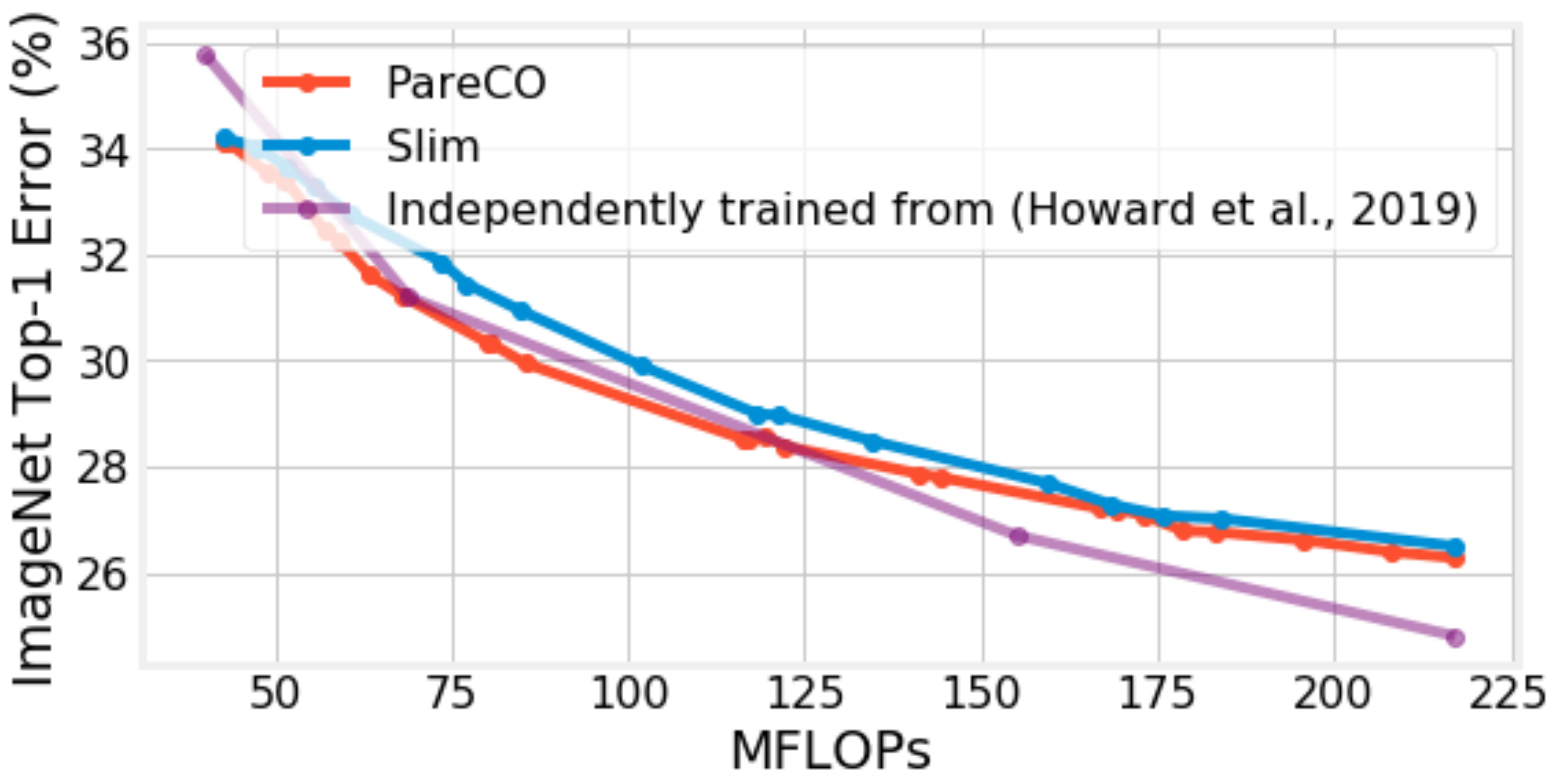
Augmented Tchebyshev Scalarization

ImageNet: Compared to conventional slimmable neural networks

MobileNetV2



MobileNetV3



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