Google Research
Carnegie Mellon University

Wisdom of Committees: An Overlooked Approach To Faster and More Accurate Models

Xiaofang Wang, Dan Kondratyuk, Eric Christiansen, Kris M. Kitani, Yair Alon (prev. Movshovitz-Attias), Elad Eban

ICLR

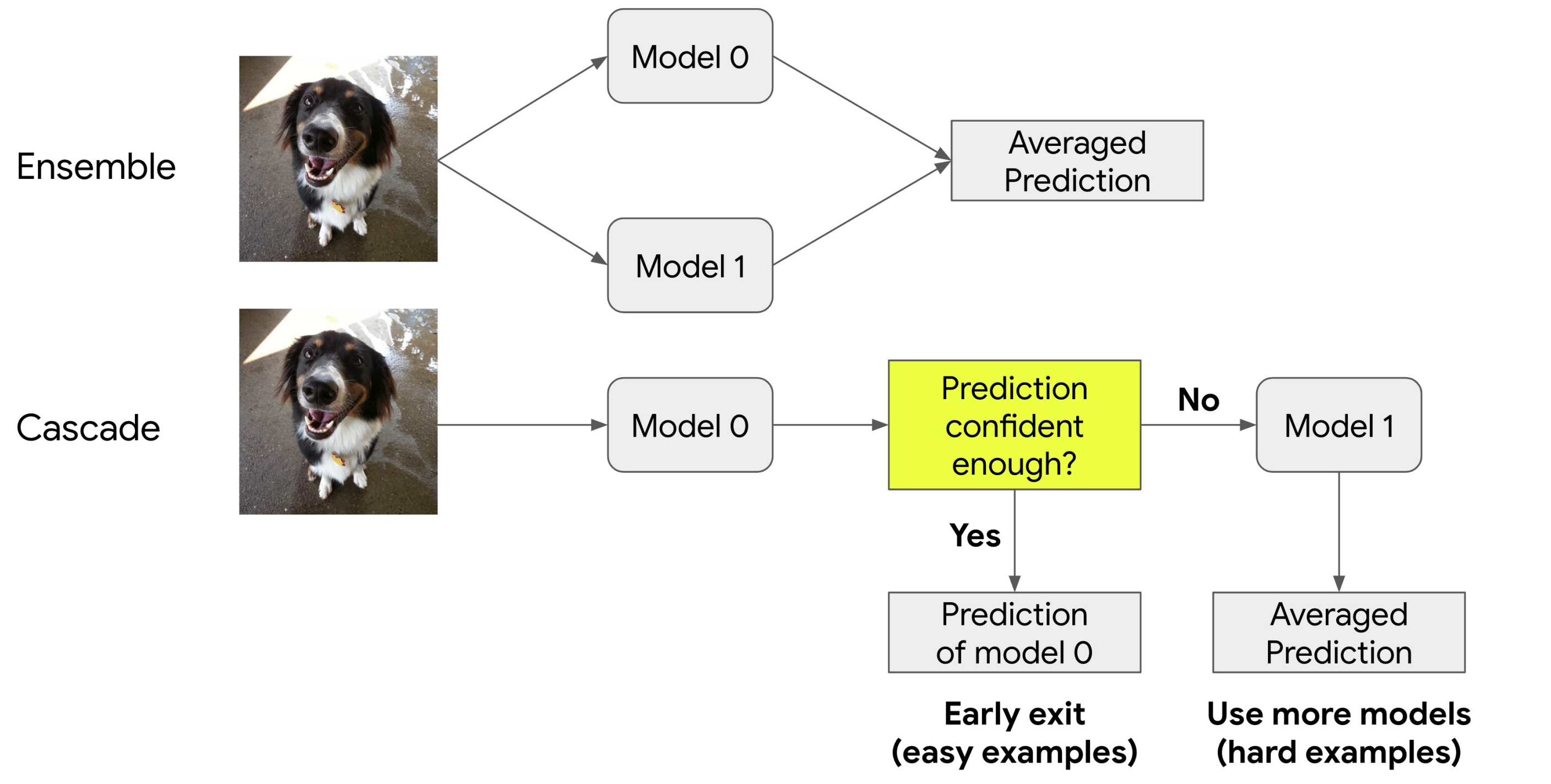
Blog Paper

Towards Efficient Models

- **Common practice:** find a **single** network architecture with high accuracy and low cost
  - Designing better architectures is highly challenging

Committee-based Models

- **Committee-based models:** Model ensembles or cascades
  - **Committee:** use **multiple** models
  - Well-known techniques but rarely considered when developing efficient neural network models
- **Our work: committee-based models are more efficient and accurate than SOTA architectures**
  - A comprehensive analysis; not inventing new techniques
  - Keep everything simple to highlight the practical benefit

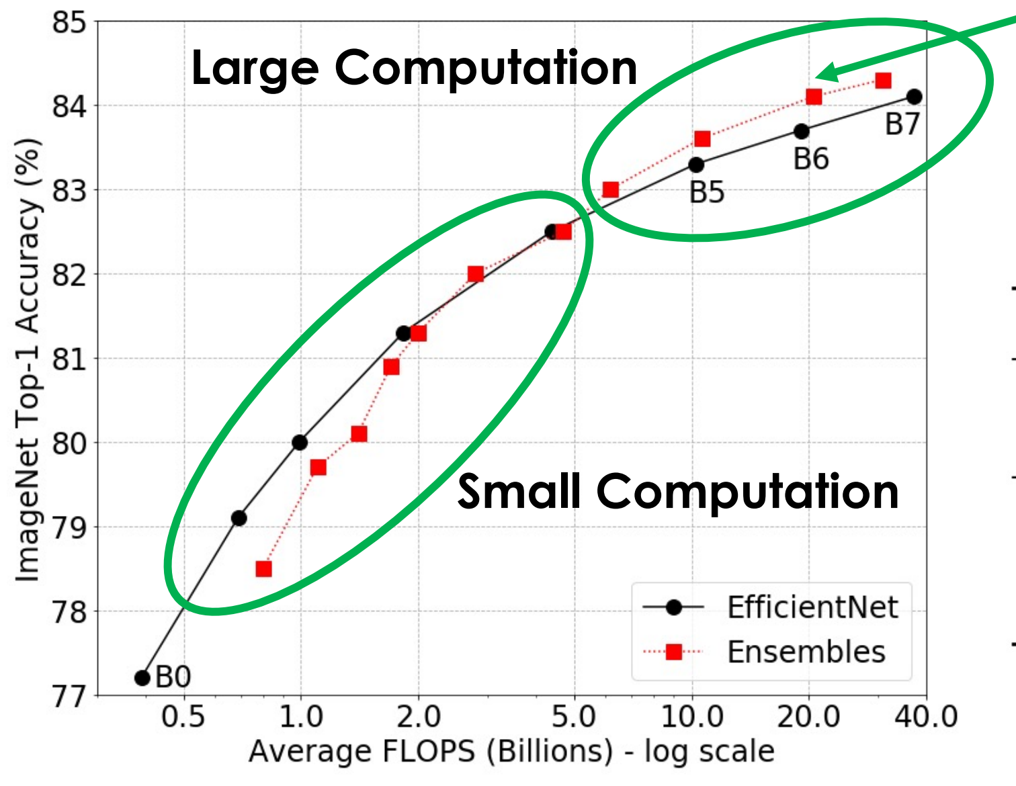


Ensemble
Model 0
Model 1
Averaged Prediction

Cascade
Model 0
Prediction confident enough?
No
Model 1
Averaged Prediction
Yes
Prediction of model 0
Early exit (easy examples)
Use more models (hard examples)

Model Ensembles vs. Single Models

- **When the total computation is fixed, which one is better?**
- Ensembles: average predictions of pre-trained models
- **Ensembles are better at large computation regime**

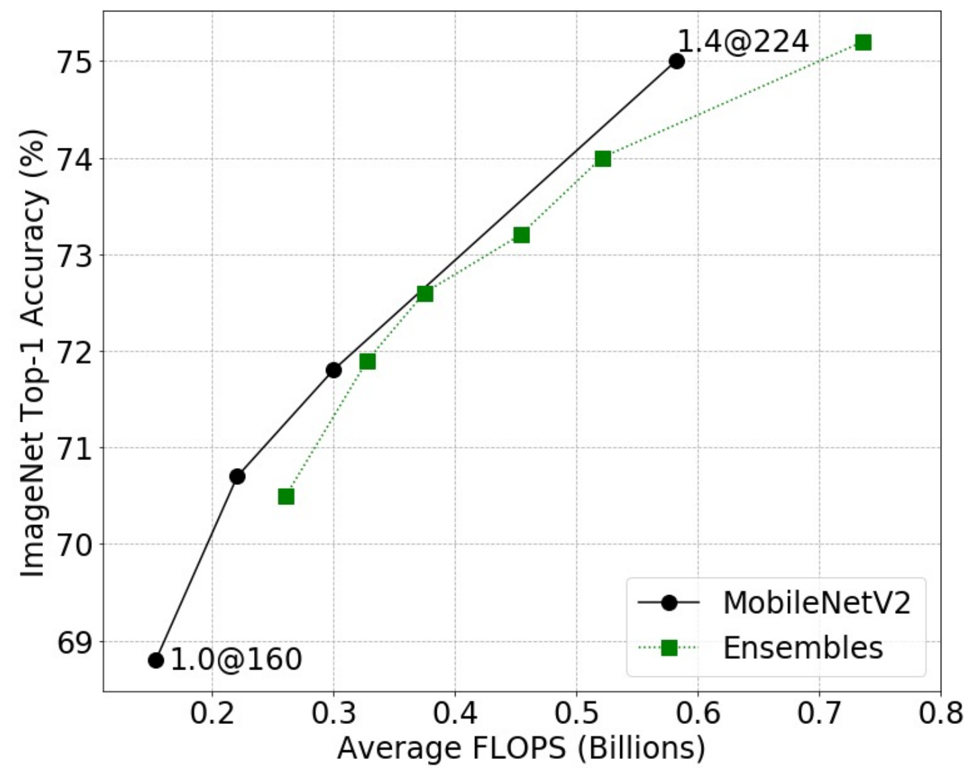


B5+B5 Ensemble
Similar accuracy to B7
Only about half of the FLOPs

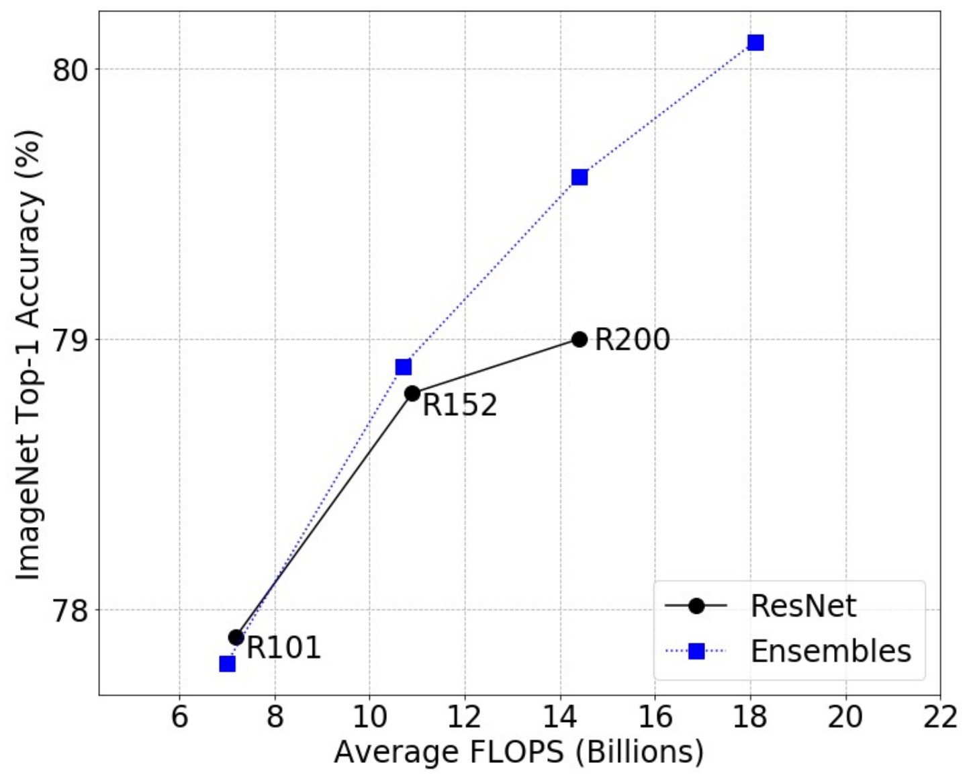
EfficientNet

	Top-1 (%)	FLOPs (B)	Training
B6	83.7	19.1	128
B3+B4+B4	83.6	10.6	88
B7	84.1	37	160
B5+B5	84.1	20.5	96
B5+B5+B5	84.3	30.8	144

Ensembles are also often faster to train

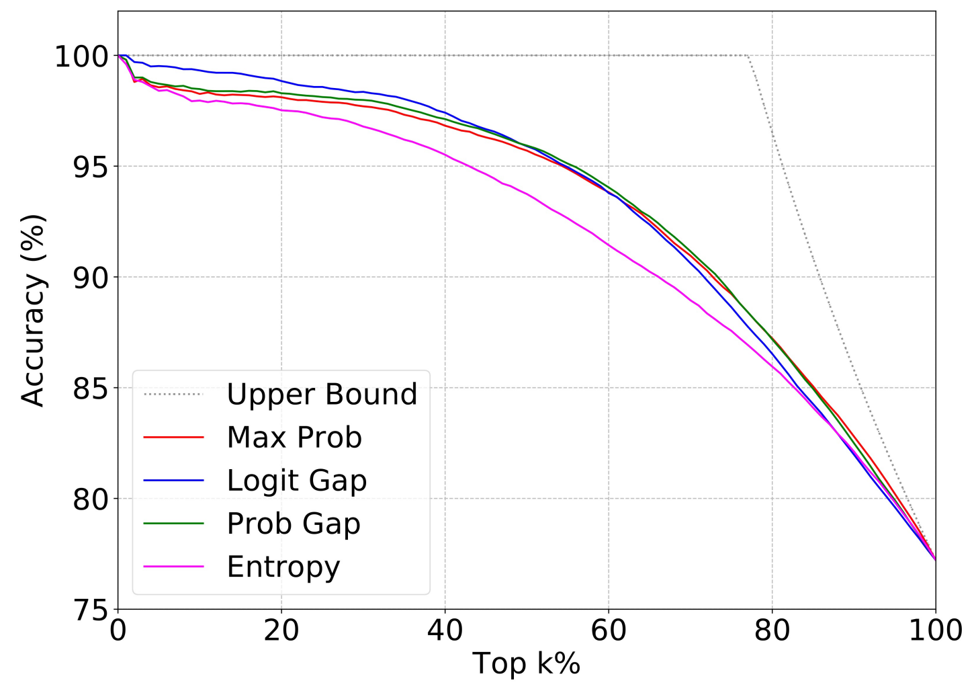


MobileNetV2 (Small computation)

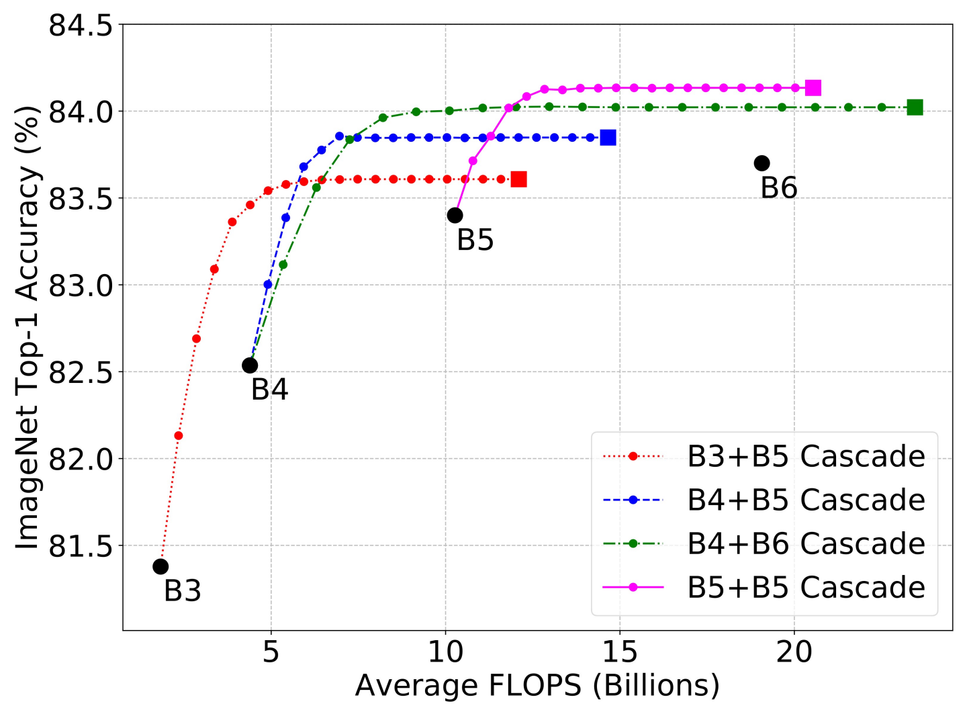


ResNet (Large computation)

Confidence Metrics in Model Cascades

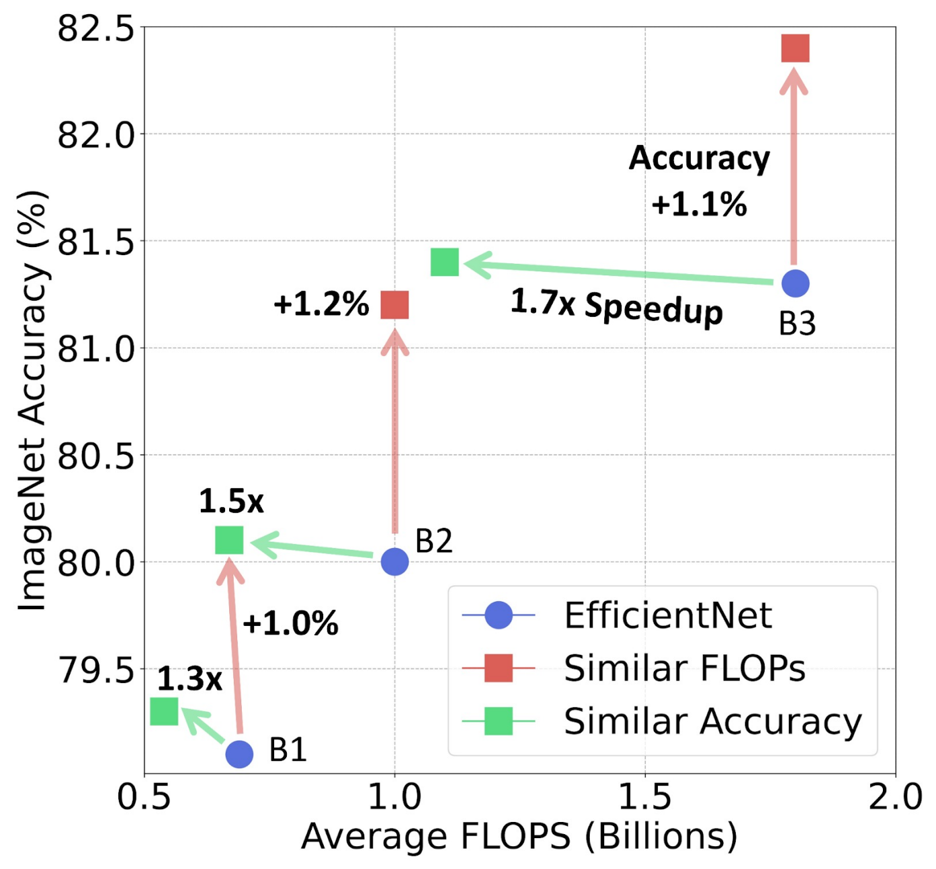


Different Confidence Metrics

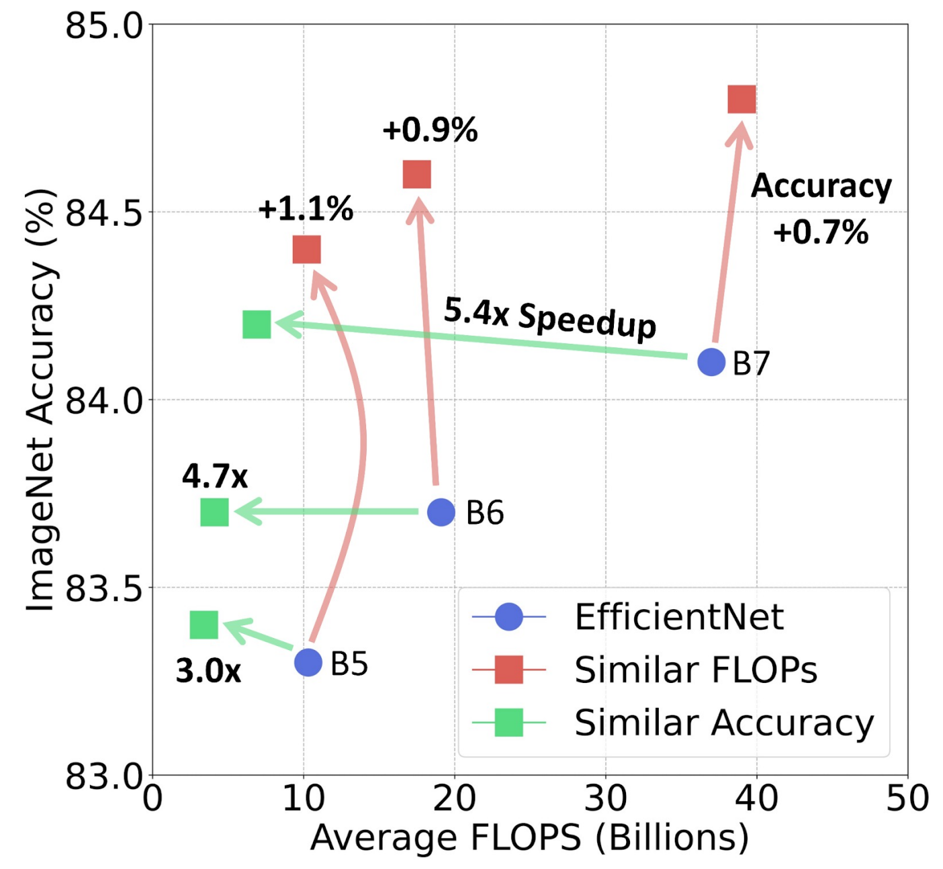


Different Confidence Thresholds

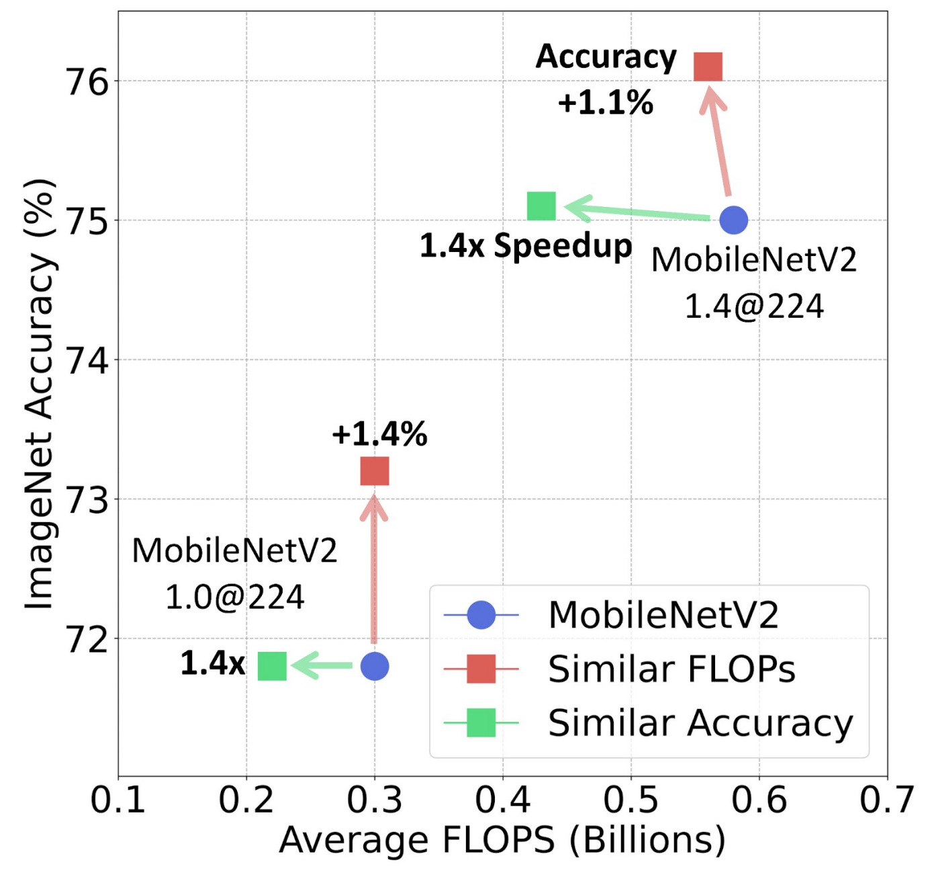
Cascades of CNN Models



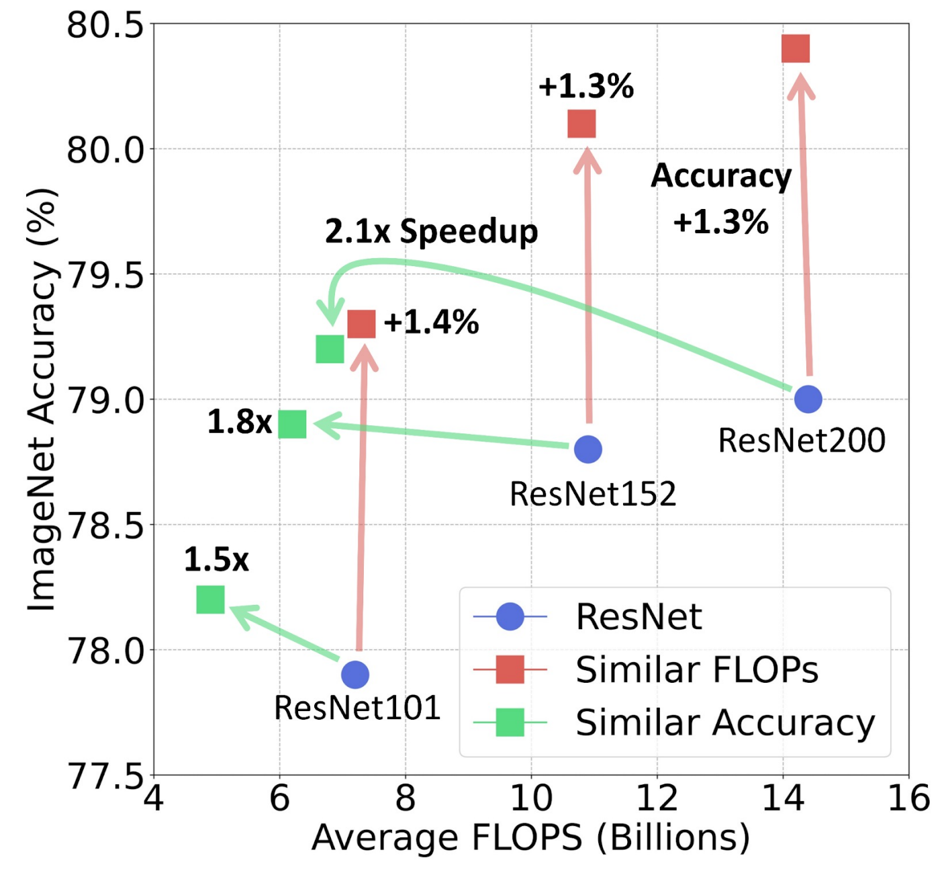
B1-B3 (Small computation)



B5-B7 (Large computation)



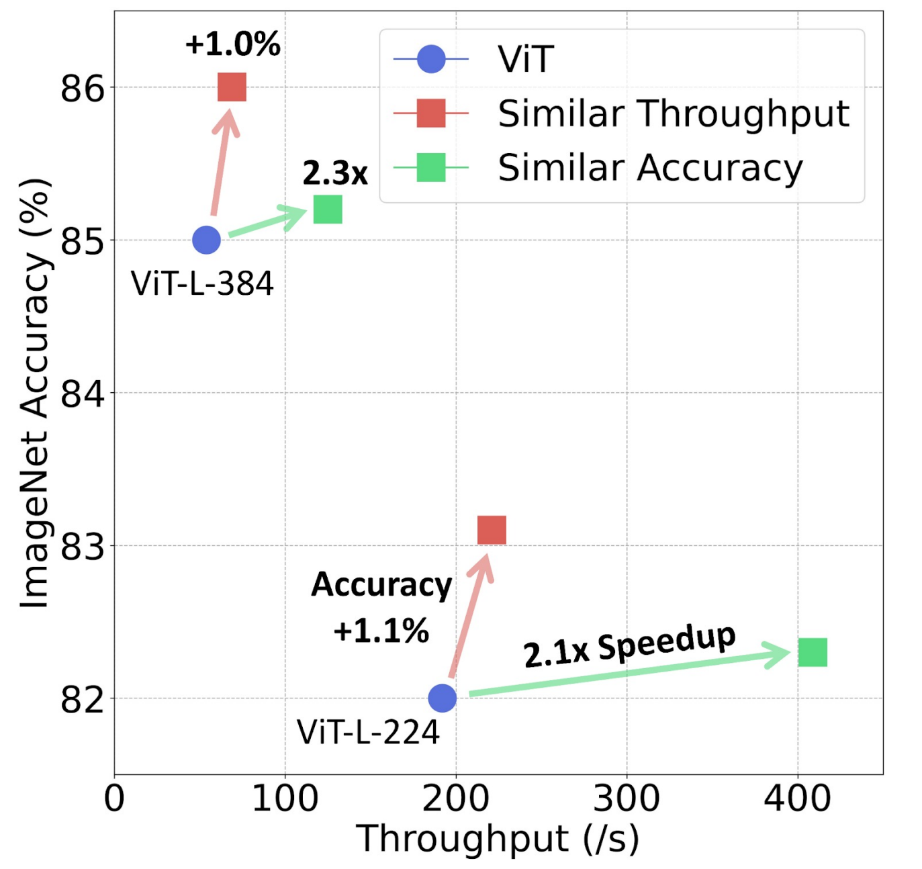
MobileNetV2 (Small computation)



ResNet (Large computation)

Cascades outperform single models at all computation regimes

Cascades of Vision Transformer Models



The benefit of cascades generalizes to ViT models

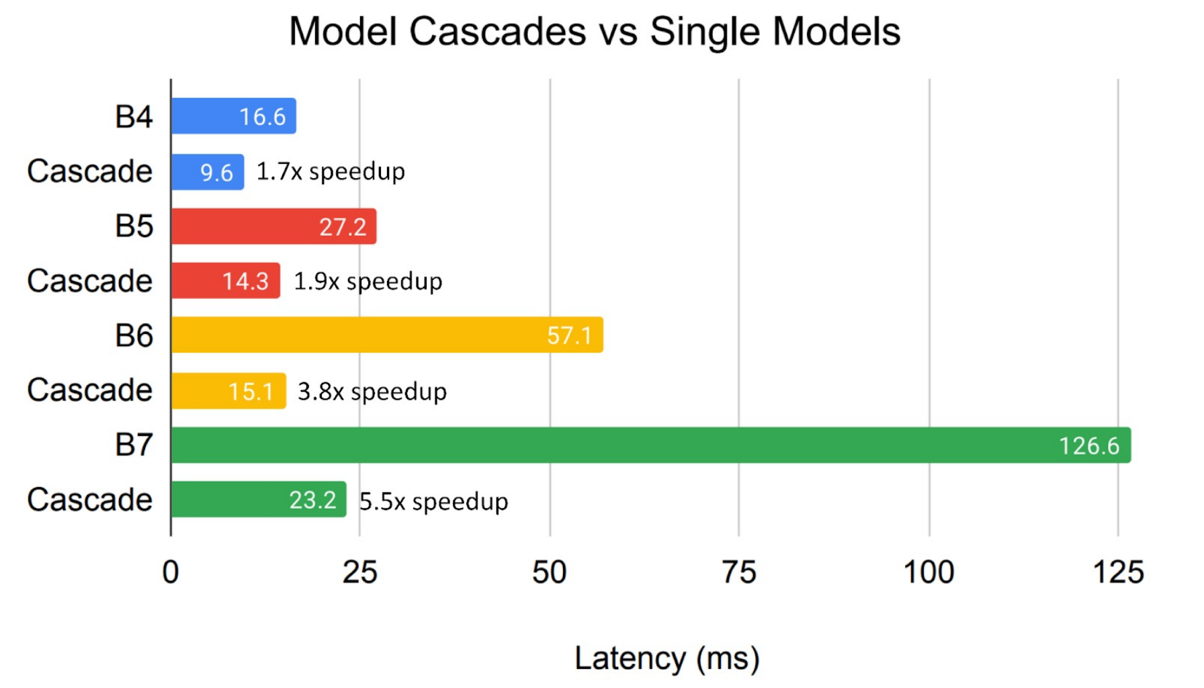
Comparison with SOTA NAS Methods

	Top-1 (%)	FLOPs (B)
BigNASModel-L (Yu et al., 2020)	79.5	0.59
$\text{OFA}_{\text{Large}}$ (Cai et al., 2020)	80.0	0.60
Cream-L (Peng et al., 2020)	80.0	0.60
Cascade*	**80.1**	0.67
BigNASModel-XL (Yu et al., 2020)	80.9	1.0
Cascade*	**81.2**	1.0

Worst-case Guarantee

	Top-1 (%)	Average-case FLOPS (B)	Worst-case FLOPS (B)	Average-case Speedup
B5	83.3	10.3	10.3	
w/o	83.4	3.4	14.2	3.0x
with	83.3	3.6	**9.8**	2.9x
B6	83.7	19.1	19.1	
w/o	83.7	4.1	25.9	4.7x
with	83.7	4.2	**15.0**	4.5x

Latency of Model Cascades



Beyond Image Classification

	Single Models		Cascades - Similar FLOPs			Cascades - Similar Accuracy		
	Top-1 (%)	FLOPs (B)	Top-1 (%)	FLOPs (B)	ΔTop-1	Top-1 (%)	FLOPs (B)	Speedup
X3D-M	78.8	6.2 × 30	**80.3**	5.7 × 30	**1.5**	79.1	**3.8 × 30**	**1.6x**
X3D-L	80.6	24.8 × 30	**82.7**	24.6 × 30	**2.1**	80.8	**7.9 × 30**	**3.2x**
X3D-XL	81.9	48.4 × 30	**83.1**	38.1 × 30	**1.2**	81.9	**13.0 × 30**	**3.7x**

Video Classification on Kinetics-600 (X3D)

	mIoU	FLOPs (B)	Speedup
ResNet-50	77.1	348	-
ResNet-101	78.1	507	-
Cascade - full	78.4	568	0.9x
Cascade - $s = 512$	78.1	439	1.2x
Cascade - $s = 128$	78.2	**398**	**1.3x**

Semantic Segmentation on Cityscapes (DeepLabV3)

Wisdom of Committees

- A simple paradigm to boost efficiency without tuning architectures
- Generalize to several architecture families and vision tasks
- Practitioners: use committee-based models!
- Researchers: an overlooked design space for efficient models
  - Better confidence functions?
  - Better training technique for ensembles / cascades?
  - More tasks, e.g., object detection?