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

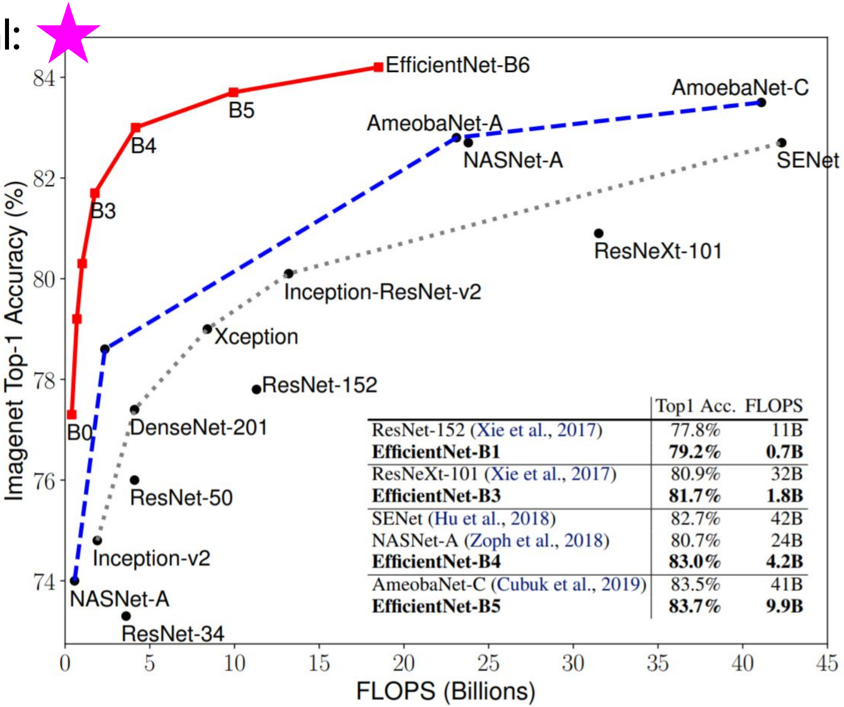
Google Research

Google Research

Towards Efficient Models

Find a **single** network architecture with high accuracy and low cost

Goal: ★



	Top1 Acc.	FLOPS
ResNet-152 (Xie et al., 2017)	77.8%	11B
EfficientNet-B1	**79.2%**	**0.7B**
ResNeXt-101 (Xie et al., 2017)	80.9%	32B
EfficientNet-B3	**81.7%**	**1.8B**
SENet (Hu et al., 2018)	82.7%	42B
NASNet-A (Zoph et al., 2018)	80.7%	24B
EfficientNet-B4	**83.0%**	**4.2B**
AmeobaNet-C (Cubuk et al., 2019)	83.5%	41B
EfficientNet-B5	**83.7%**	**9.9B**

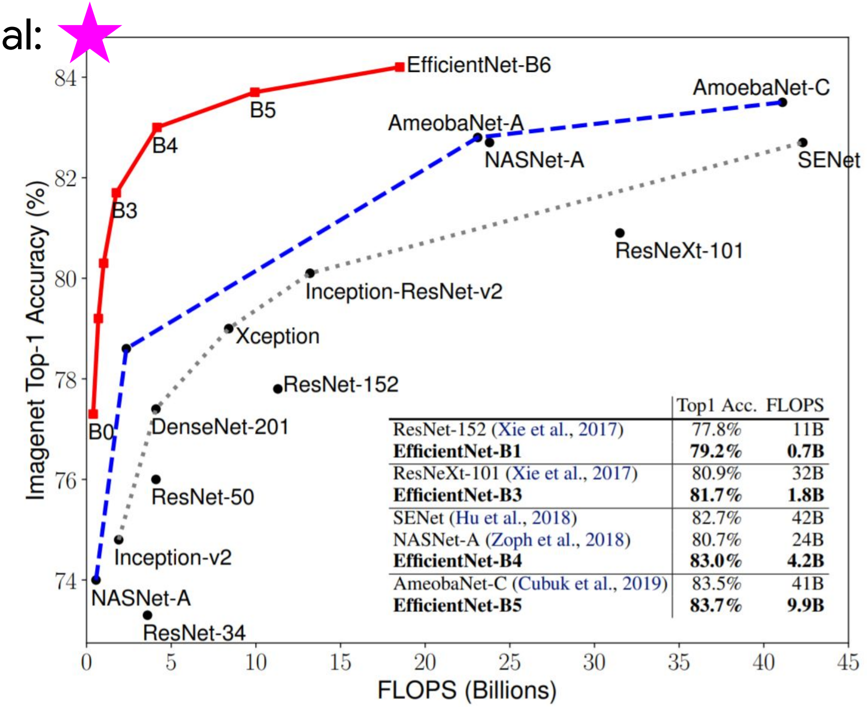
Photo Credit: EfficientNet [Tan et al. ICML 2019]

Designing better architectures is challenging

Find a **single** network architecture with high accuracy and low cost

- Expertise of downstream tasks
- Computational resources
- Engineering efforts

Goal: ★



	Top1 Acc.	FLOPS
ResNet-152 (Xie et al., 2017)	77.8%	11B
EfficientNet-B1	**79.2%**	**0.7B**
ResNeXt-101 (Xie et al., 2017)	80.9%	32B
EfficientNet-B3	**81.7%**	**1.8B**
SENet (Hu et al., 2018)	82.7%	42B
NASNet-A (Zoph et al., 2018)	80.7%	24B
EfficientNet-B4	**83.0%**	**4.2B**
AmeobaNet-C (Cubuk et al., 2019)	83.5%	41B
EfficientNet-B5	**83.7%**	**9.9B**

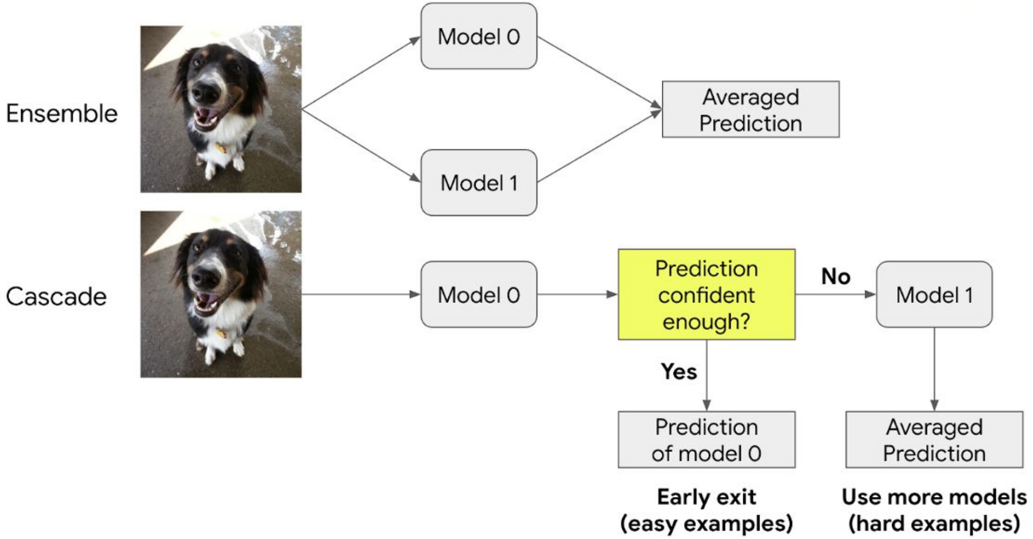
Photo Credit: EfficientNet [Tan et al. ICML 2019]

Google Research

How about Committee-based Models?

Committee-based models: model ensembles or cascades

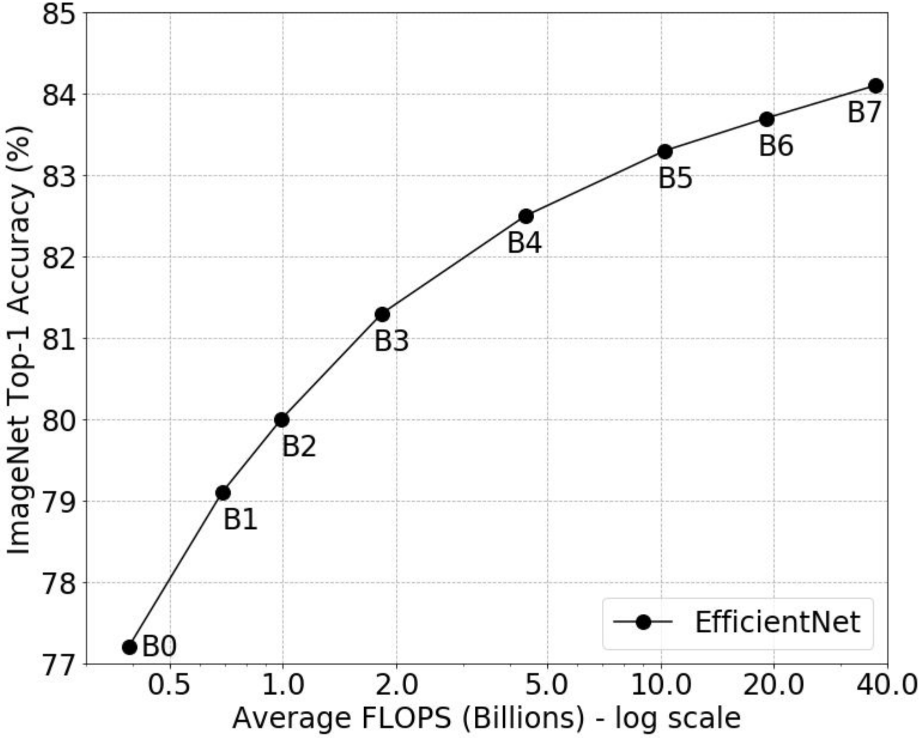
Committee: use **multiple** models



Google Research

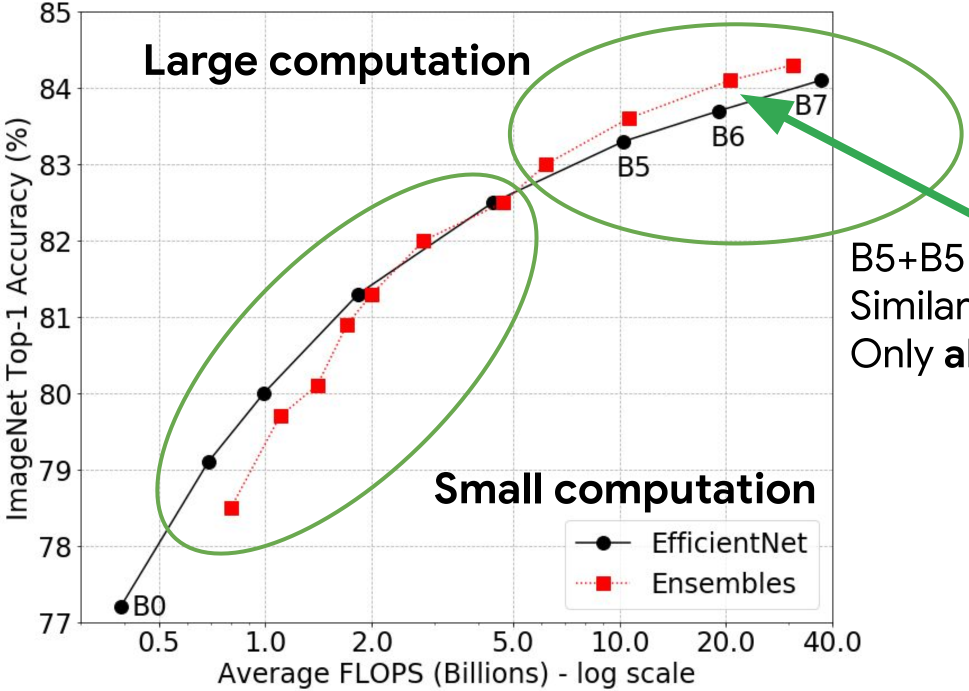
Model Ensembles vs. Single Models

When the total computation is fixed, which one will give higher accuracy?



Model Ensembles vs. Single Models

When the total computation is fixed, which one will give higher accuracy?

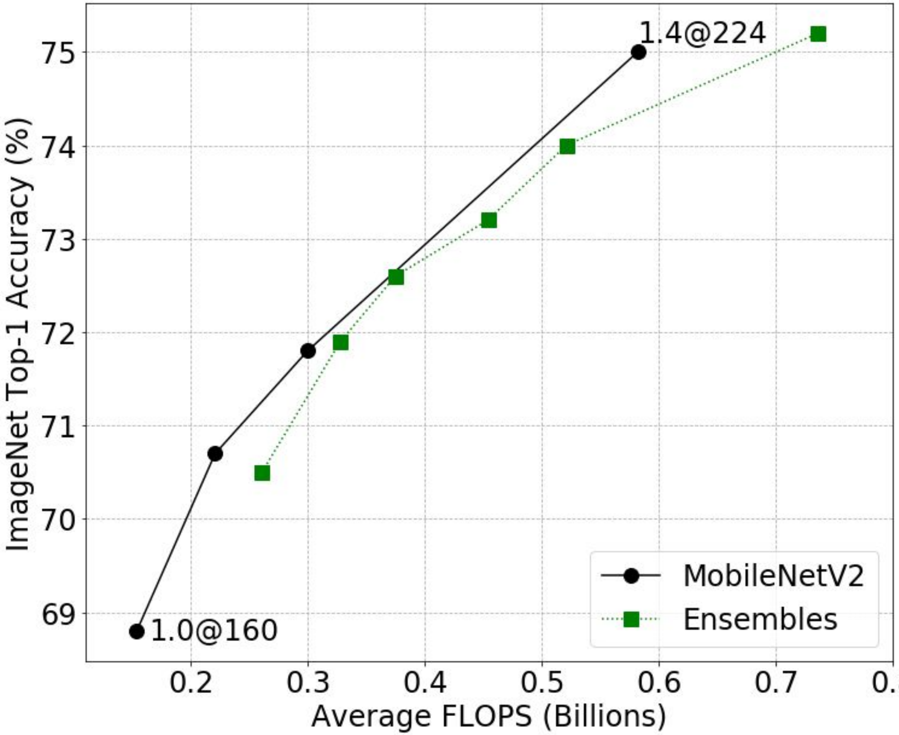


EfficientNet

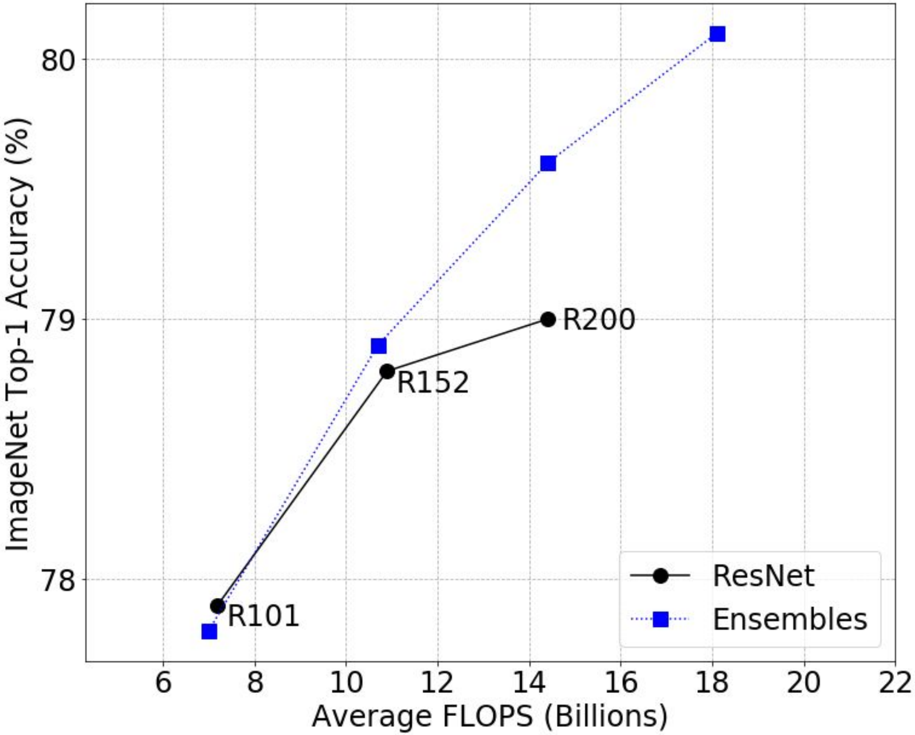
B5+B5 Ensemble
Similar accuracy to B7
Only **about half of the FLOPs**

Google Research

Model Ensembles vs. Single Models

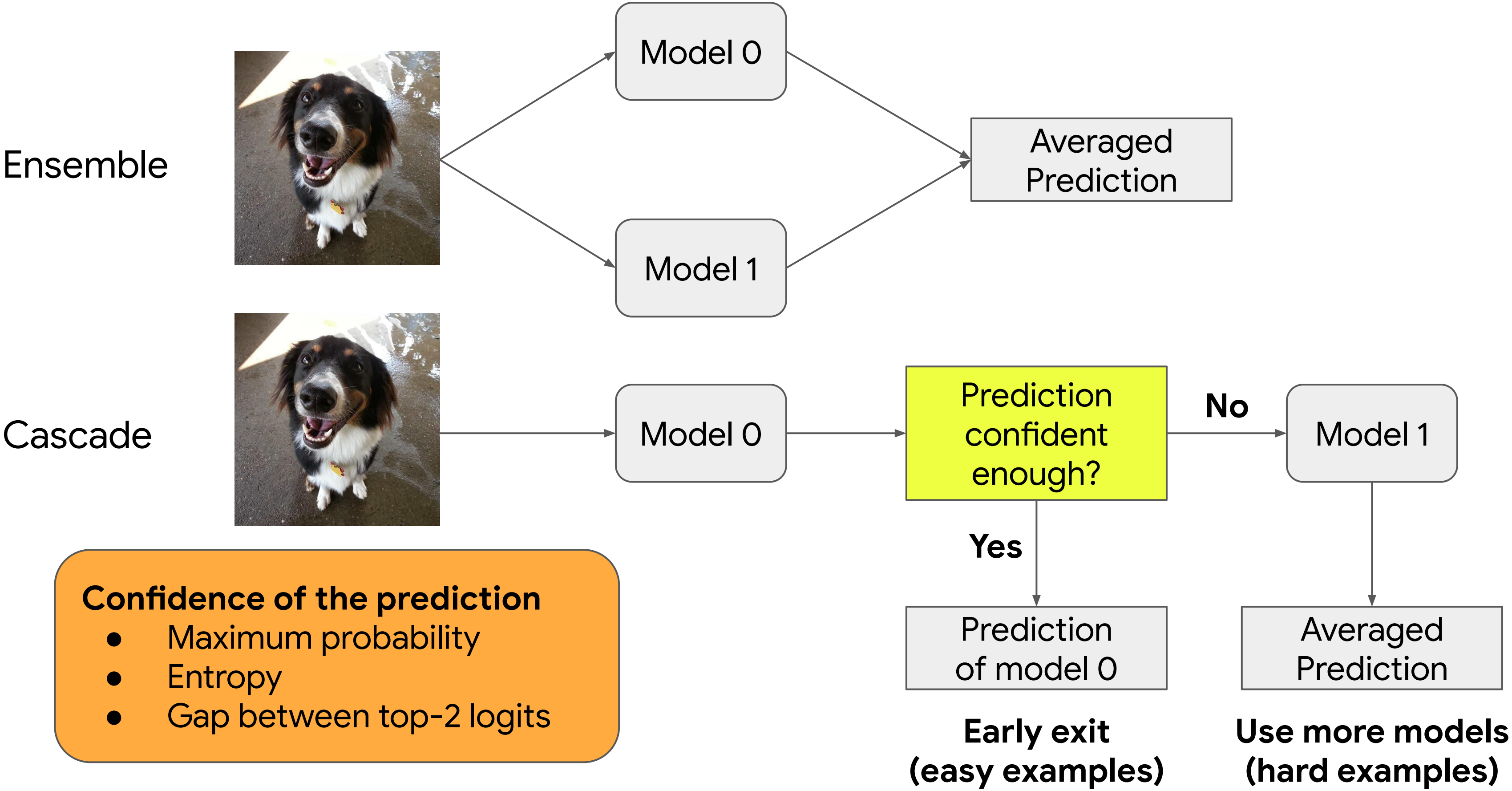


MobileNetV2 (Small computation)



ResNet (Large computation)

From Ensembles to Cascades



Ensemble

Model 0

Model 1

Averaged Prediction

Cascade

Model 0

Prediction confident enough?

No

Model 1

Yes

Prediction of model 0

Averaged Prediction

Early exit (easy examples)

Use more models (hard examples)

Confidence of the prediction

- Maximum probability
- Entropy
- Gap between top-2 logits

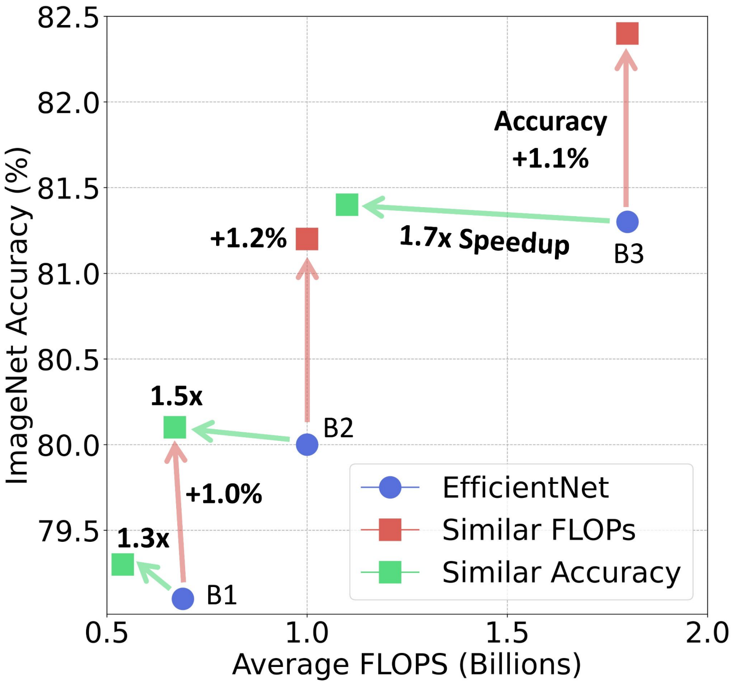
Build Cascades of Pre-trained Models

- **Step 1:** Prepare a pool of models
  - Could directly use off-the-shelf pre-trained models
  - No need to update the training pipeline or tune the architecture

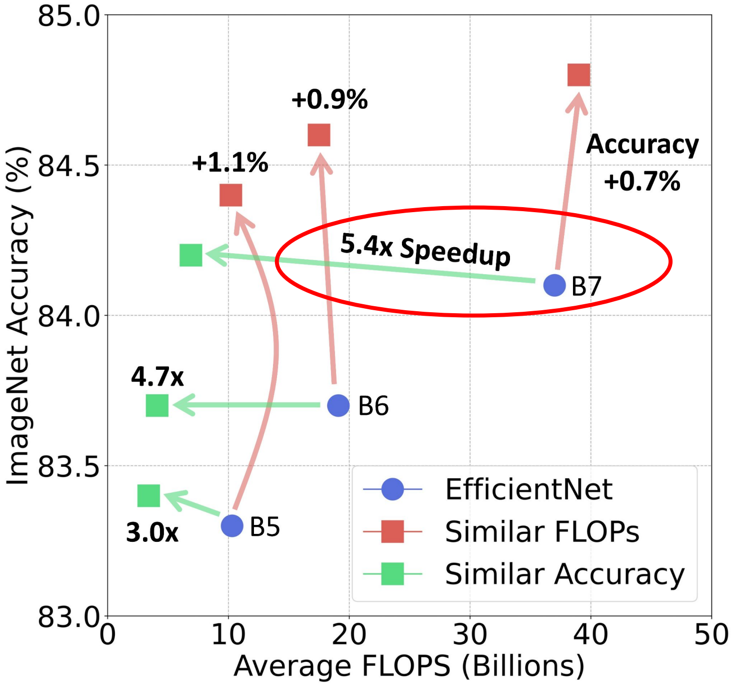
- **Step 2:** Try possible model combinations
  - Tune the confidence thresholds on held-out validation images
  - Only need the logits; no training required
  - Select the best cascade among all possible combinations

Google Research

Cascades outperform single models at **all** computation regimes



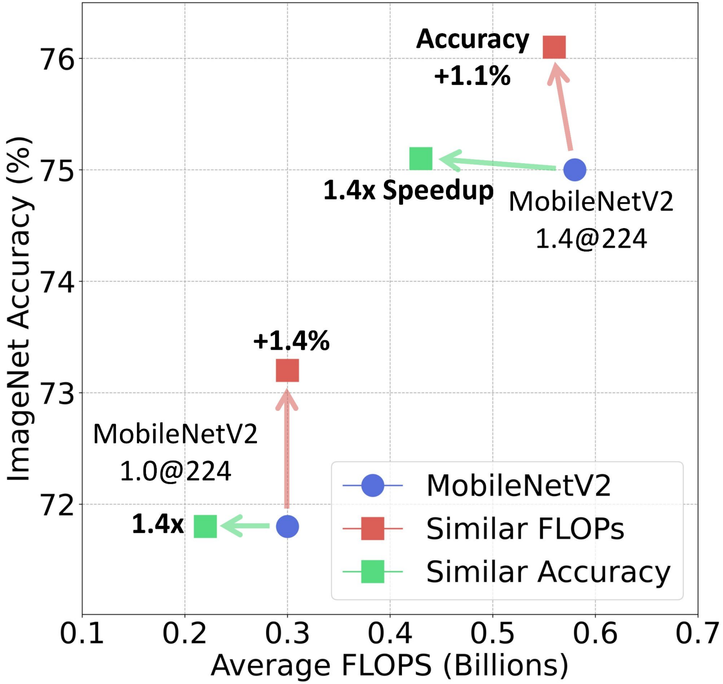
B1 to B3 (Small computation)



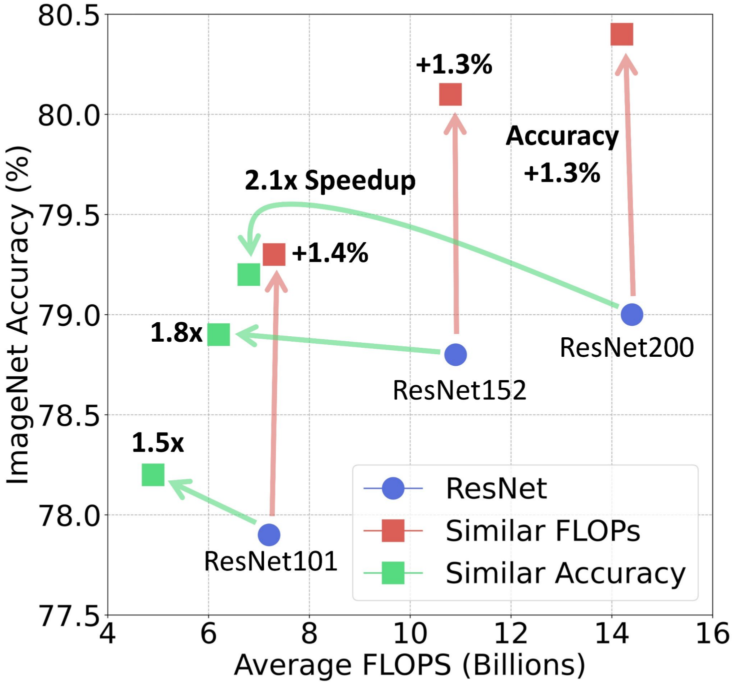
B5 to B7 (Large computation)

Google Research

Cascades outperform single models at **all** computation regimes



MobileNetV2 (Small computation)

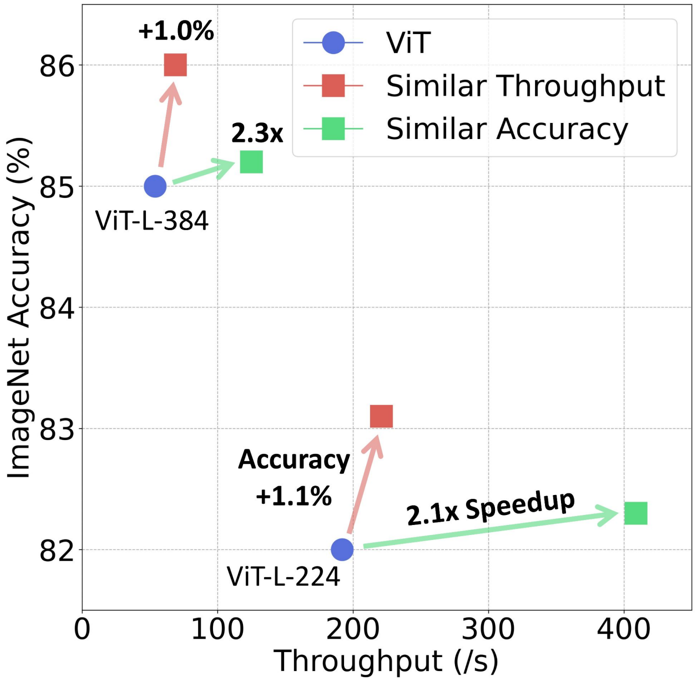


ResNet (Large computation)

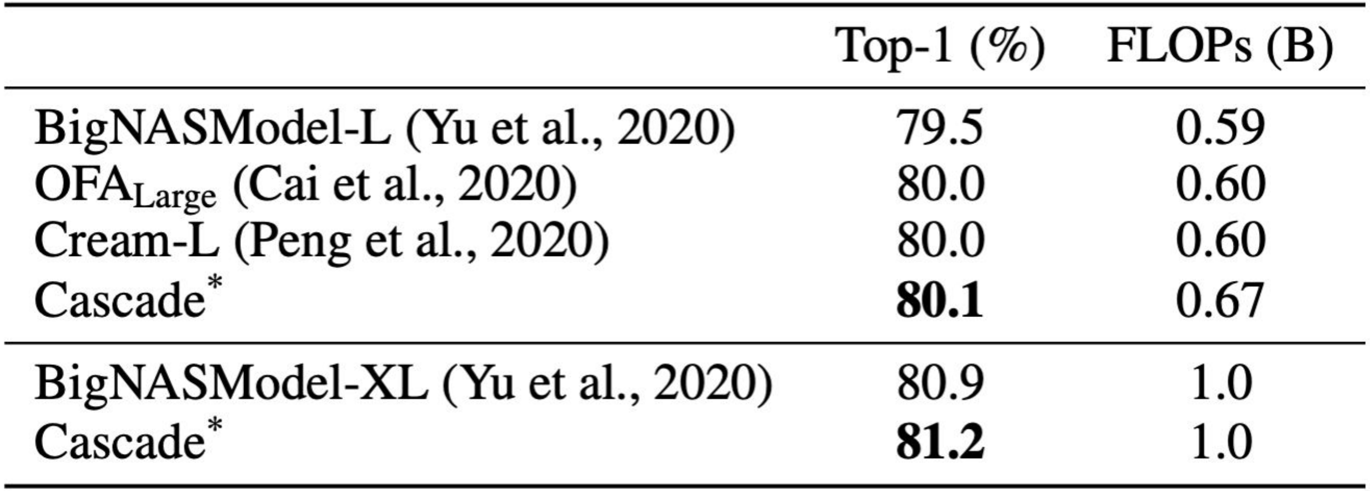
Google Research

Vision Transformer (ViT)

The benefit of cascades generalizes to Transformer architectures



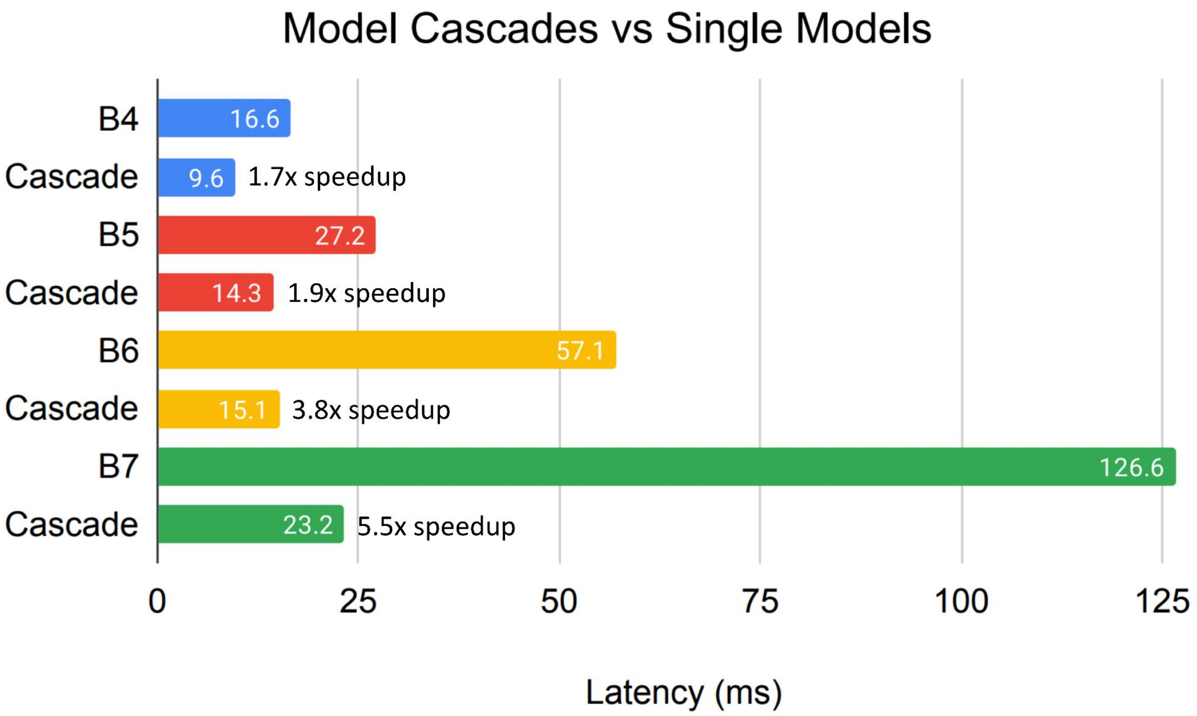
EfficientNet Cascades vs. SOTA NAS Methods



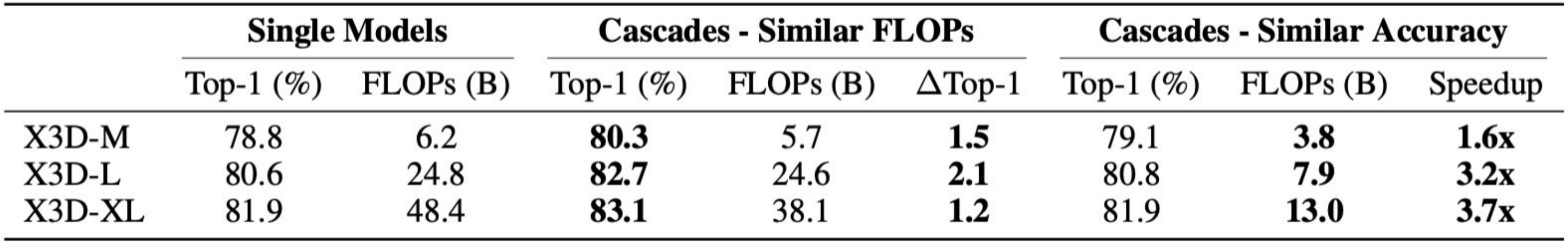
	Top-1 (%)	FLOPs (B)
BigNASModel-L (Yu et al., 2020)	79.5	0.59
OFA$_{\text{Large}}$ (Cai et al., 2020)	80.0	0.60
Cream-L (Peng et al., 2020)	80.0	0.60
Cascade*	**80.1**	0.67
BigNASModel-XL (Yu et al., 2020)	80.9	1.0
Cascade*	**81.2**	1.0

Google Research

Latency of Cascades



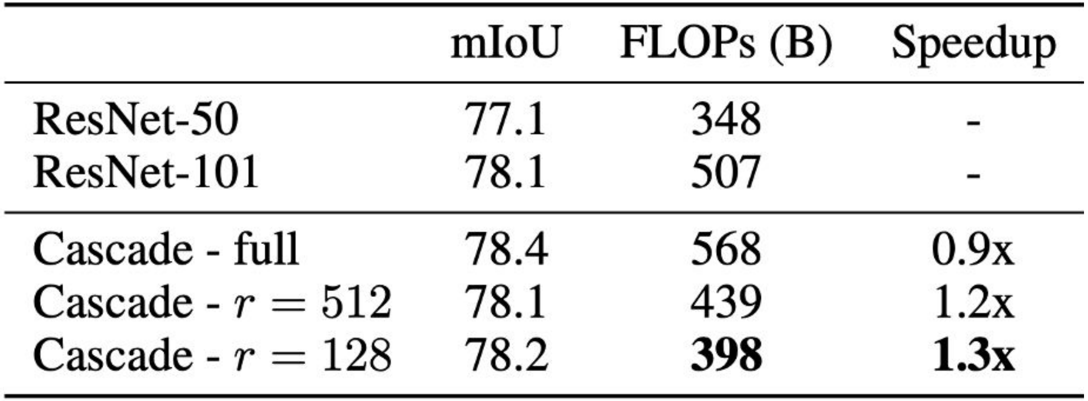
Video Classification on Kinetics-600



	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	**80.3**	5.7	**1.5**	79.1	**3.8**	**1.6x**
X3D-L	80.6	24.8	**82.7**	24.6	**2.1**	80.8	**7.9**	**3.2x**
X3D-XL	81.9	48.4	**83.1**	38.1	**1.2**	81.9	**13.0**	**3.7x**

Cascades of X3D Models

Semantic Segmentation on Cityscapes



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

Cascades of DeepLabv3 models

Wisdom of Committees

- A simple paradigm to improve efficiency without tuning the architecture
- Generalize to several architecture families and vision tasks
- Let's use and compare with committee-based models!