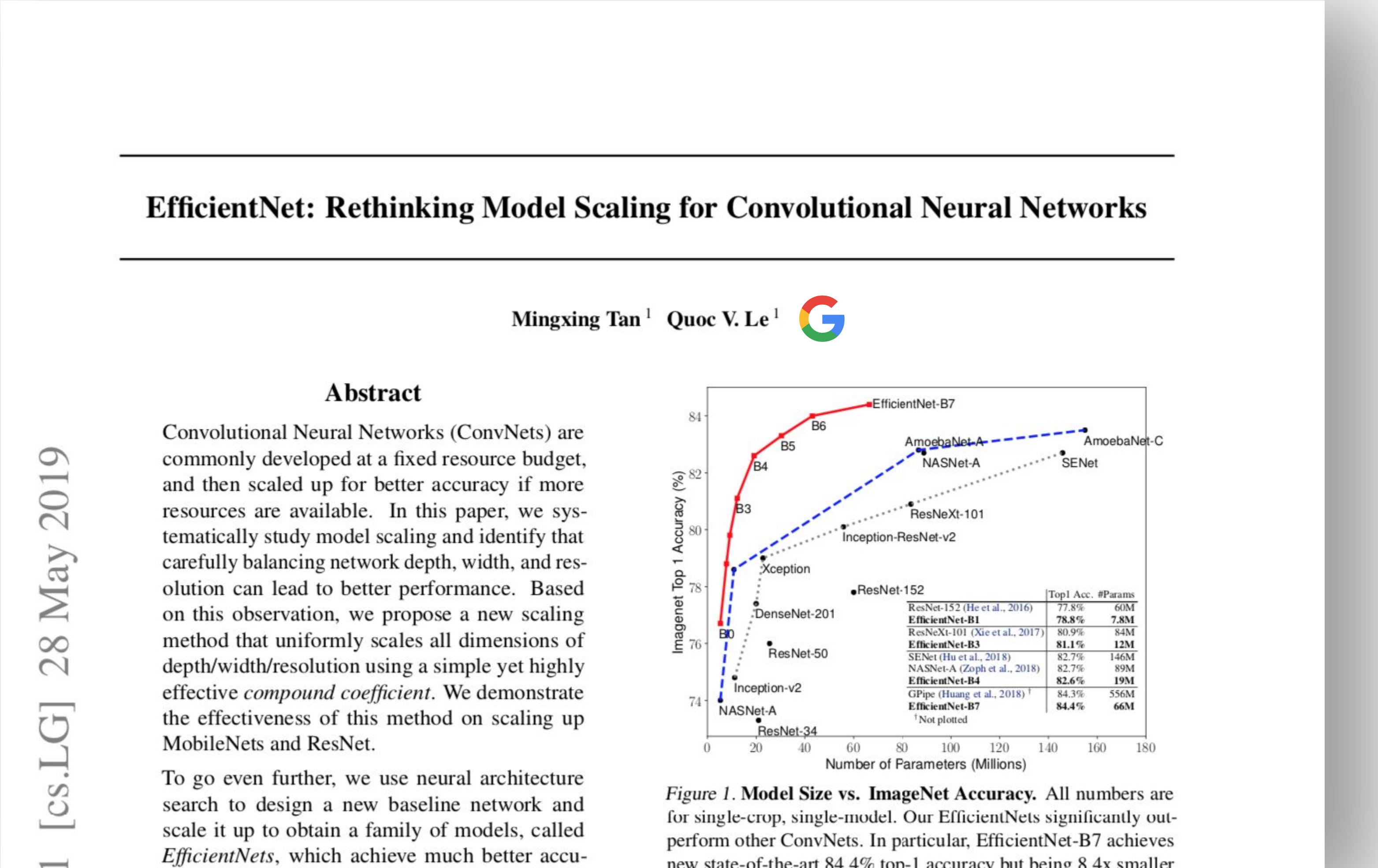


EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Mingxing Tan [1] Quoc V. Le [1]

Abstract

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are available. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective *compound coefficient*. We demonstrate the effectiveness of this method on scaling up MobileNets and ResNet.

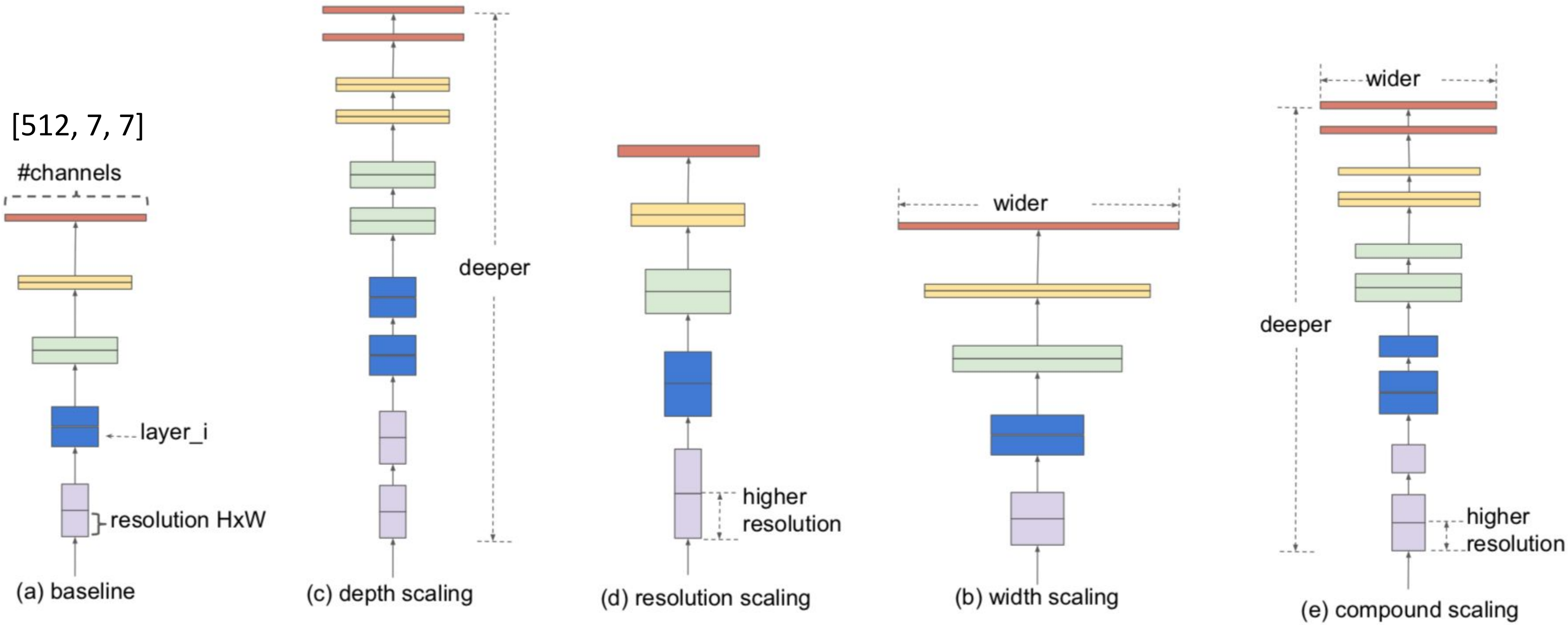
To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called *EfficientNets*, which achieve much better accu-

	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**78.8%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.1%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.6%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.4%**	**66M**

† Not plotted

Figure 1. **Model Size vs. ImageNet Accuracy.** All numbers are for single-crop, single-model. Our EfficientNets significantly outperform other ConvNets. In particular, EfficientNet-B7 achieves new state-of-the-art 84.4% top-1 accuracy but being 8.4x smaller

How to scale your baseline with more computational resources?



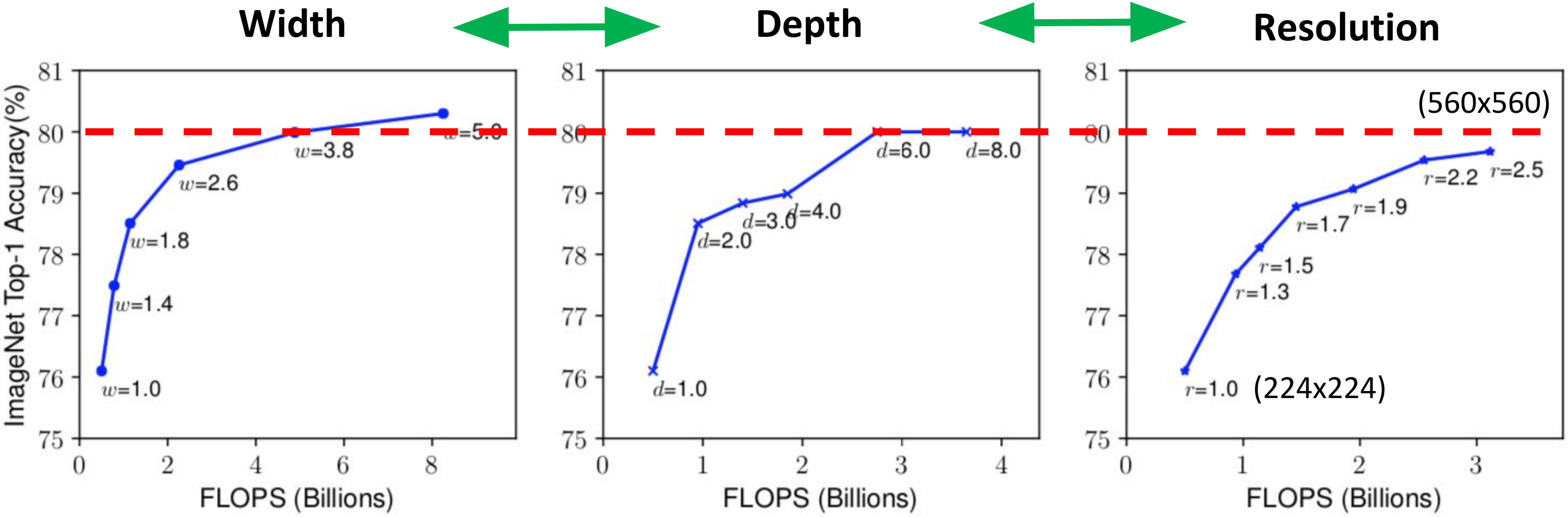
[512, 7, 7]

[3, 224, 224]

(ResNet)

(WideResNet)

Scaling width, depth, and resolution?



Width

Depth

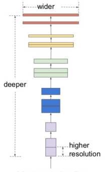
Resolution

(560x560)

(224x224)

Compound scaling and FLOP allocation

FLOPS in a CNN proportional to



depth
width^2
resolution^2

depth = α^{φ}
width = β^{φ}
resolution = γ^{φ}

Scale with compound coefficient φ

$$\boldsymbol{\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2} \textbf{ (constraint)}$$

Scaling will approx. increase FLOPS by

$$\boldsymbol{(\alpha \cdot \beta^2 \cdot \gamma^2)^{\varphi} = 2^{\varphi}}$$

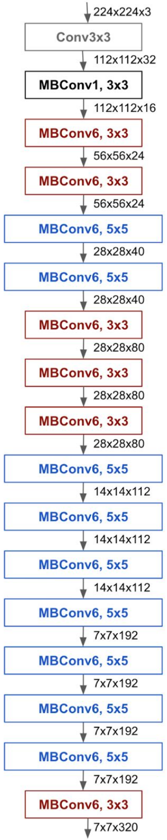
Scaling of architectures

Works on ResNets & Co. but was optimized for new architecture:

1. Neural architecture search optimizing accuracy and FLOPS to get EfficientNet-B0
2. Small grid search of α, β, γ
3. Fix α, β, γ and scale up baseline network with φ to obtain EfficientNet-B1 to B7

(Search for α, β, γ could be carried out directly on large networks but becomes very expensive.)

EfficientNet-B0 scaling



224x224x3

Conv3x3

112x112x32

MBConv1, 3x3

112x112x16

MBConv6, 3x3

56x56x24

MBConv6, 3x3

56x56x24

MBConv6, 5x5

28x28x40

MBConv6, 5x5

28x28x40

MBConv6, 3x3

28x28x80

MBConv6, 3x3

28x28x80

MBConv6, 3x3

28x28x80

MBConv6, 5x5

14x14x112

MBConv6, 5x5

14x14x112

MBConv6, 5x5

14x14x112

MBConv6, 5x5

7x7x192

MBConv6, 5x5

7x7x192

MBConv6, 5x5

7x7x192

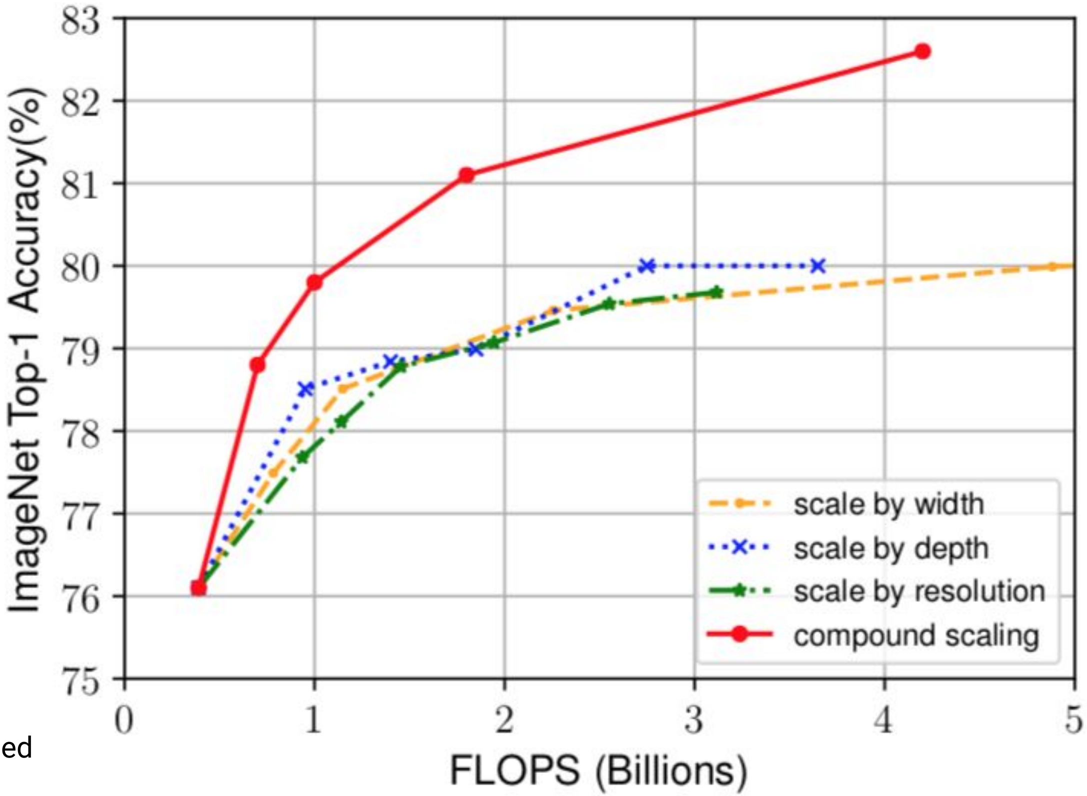
MBConv6, 5x5

7x7x192

MBConv6, 3x3

7x7x320

(MBConv = mobile inverted bottleneck convolution)



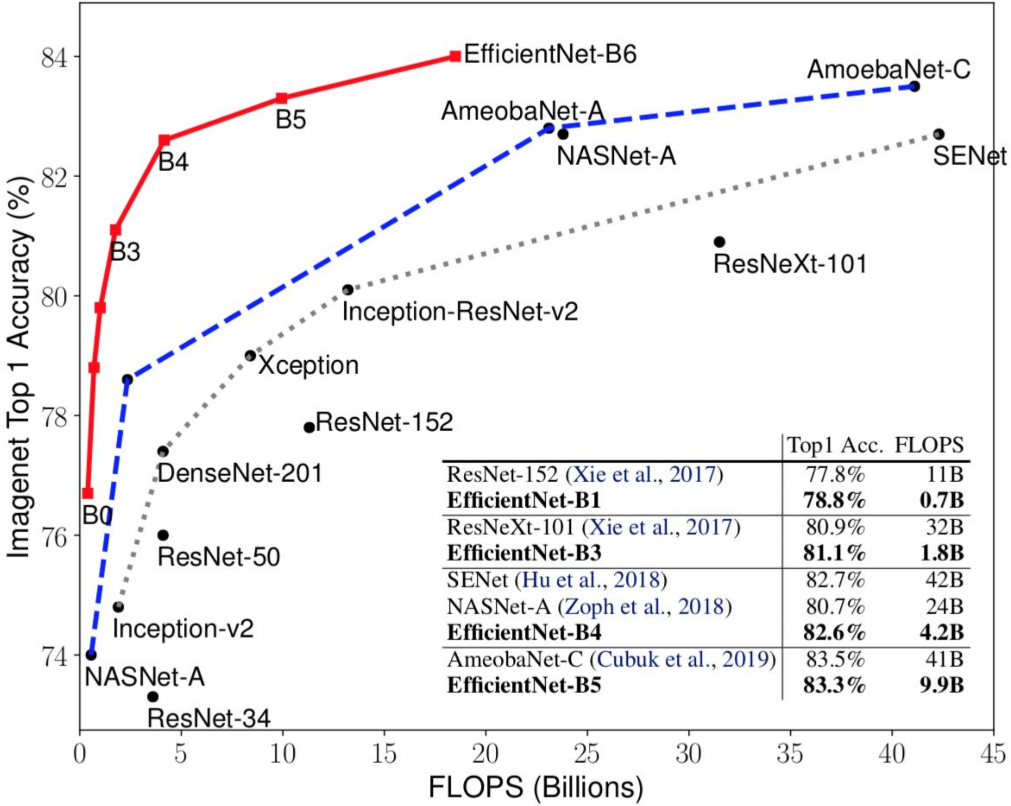
Results - EfficientNet-B7

- ImageNet SOTA
  84.4% top-1 / 97.1% top-5

- 8.4x smaller

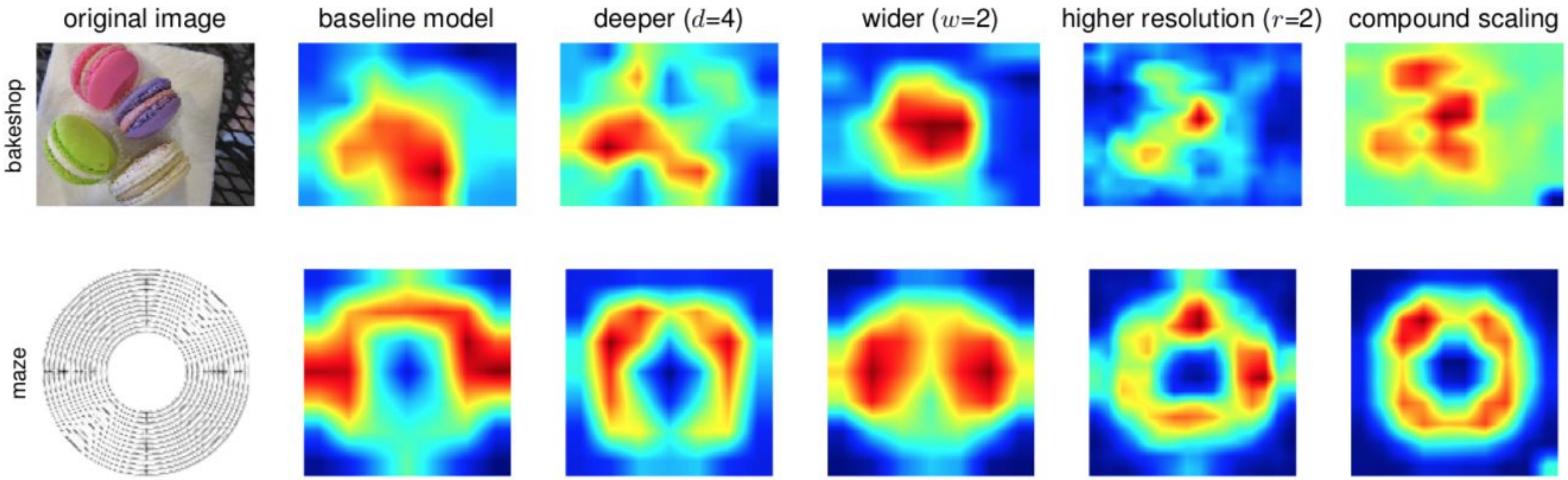
- 6.1x faster on inference

- transfers well to other datasets



	Top1 Acc.	FLOPS
ResNet-152 (Xie et al., 2017)	77.8%	11B
EfficientNet-B1	**78.8%**	**0.7B**
ResNeXt-101 (Xie et al., 2017)	80.9%	32B
EfficientNet-B3	**81.1%**	**1.8B**
SENet (Hu et al., 2018)	82.7%	42B
NASNet-A (Zoph et al., 2018)	80.7%	24B
EfficientNet-B4	**82.6%**	**4.2B**
AmeobaNet-C (Cubuk et al., 2019)	83.5%	41B
EfficientNet-B5	**83.3%**	**9.9B**

Results – Class activation maps (CAM)



Sources

(1) “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”, https://arxiv.org/abs/1905.11946

(2) EfficientNet Google AI blog post, https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html

(3) “Searching for MobileNetV3”, https://arxiv.org/abs/1905.02244

(4) “Squeeze-and-Excitation Networks”, https://arxiv.org/abs/1709.01507

(5) fast.ai forum thread, https://forums.fast.ai/t/efficientnet/46978

(6) Implementations, https://paperswithcode.com/paper/efficientnet-rethinking-model-scaling-for

Compiled by Michael M. Bishop (https://MichaelBishop.com)