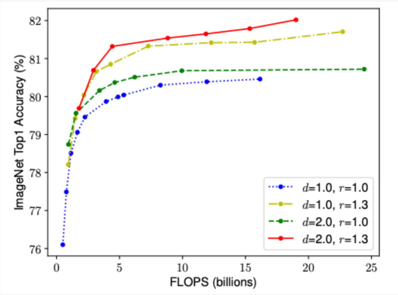
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks Summary

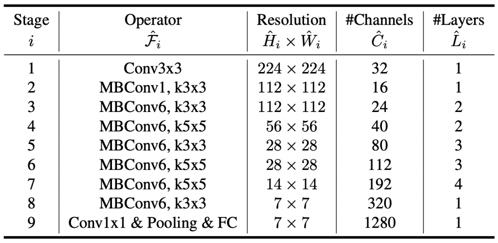
April 11, 2023

EfficientNet is a proposed neural architecture that scales a convolutional neural network uniformly by increasing its width (number of feature maps), depth (number of layers), and resolution (image resolution in pixels), using the method of compound scaling. The article presents a systematic approach to scaling CNNs, in contrast to the arbitrary methods used in previous studies. The paper discovered a certain dependence among the different scaling dimensions (width/depth and resolution) to increase overall accuracy. A balance between the different dimensions was achieved by scaling each of them with a fixed scaling coefficient. This scaling method also makes intuitive sense, as a larger input image requires more layers to expand the receptive field and more channels to capture intricate feature patterns. However, increasing only one of the three dimensions resulted in performance gains that eventually reached a plateau. The following figure displays the performance gains when applying compound scaling [1]:

Figure 1: Accuracy using different depth (d) and resolution (r). The points represent different widths (w) [1]

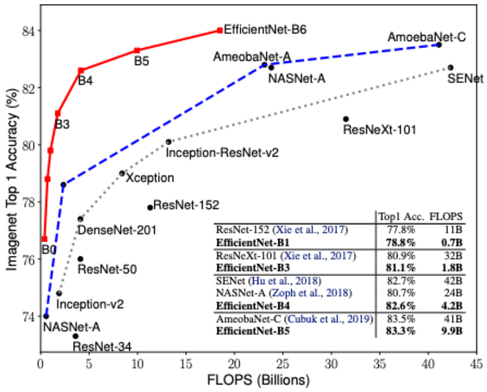


The following figue 2 represents the base configuration of EfficientNet-B0. Compound scaling is then applied to the base configuration to scale it up. Applying compound scaling to create EfficientNet-B7 resulted in 84.4% top-1 and 97.1% top-5 accuracy on ImageNet. Remarkably, this model is 8.4x smaller and 6.1x faster during inference than the best existing CNNs [1].



Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	28 × 28	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

Figure 2: EfficientNet-B0 architecture [1]



	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**

Figure 3: Accuracy of EfficientNet-B6 to conventional neural network architectures [1]

References

[1] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.