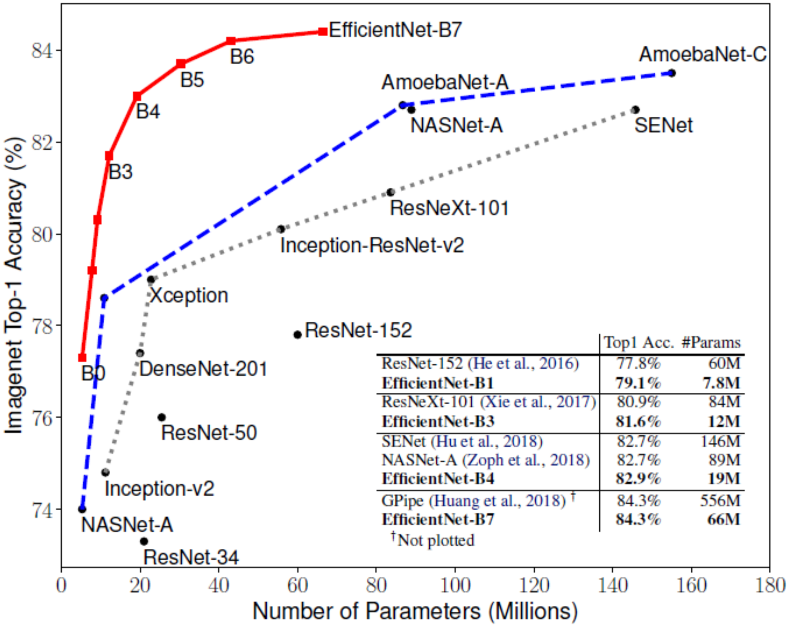
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks



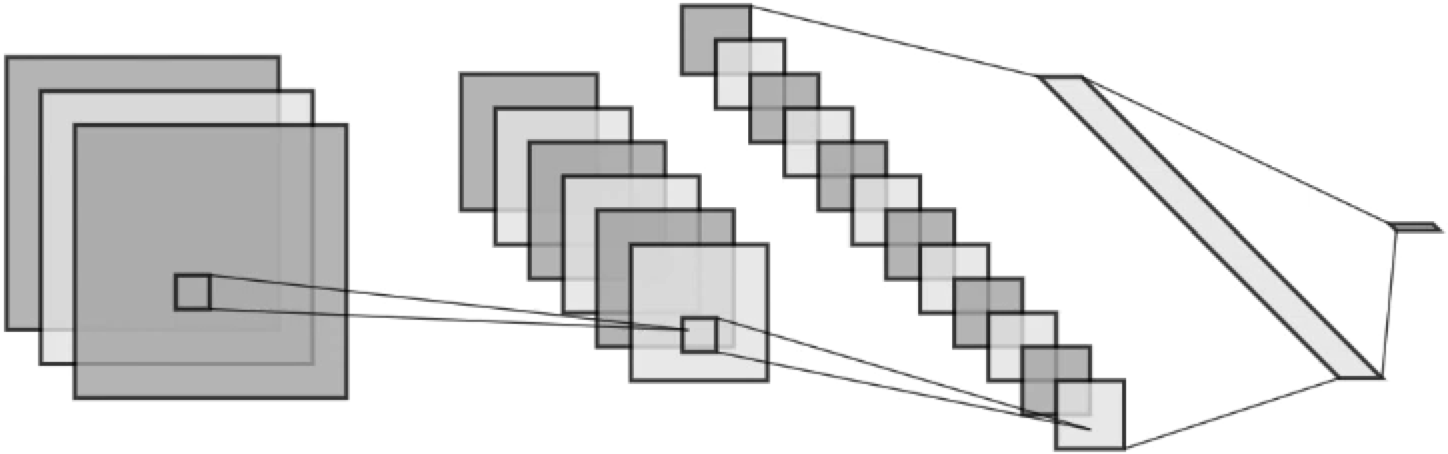
Pere Gilabert
February 2021

Index

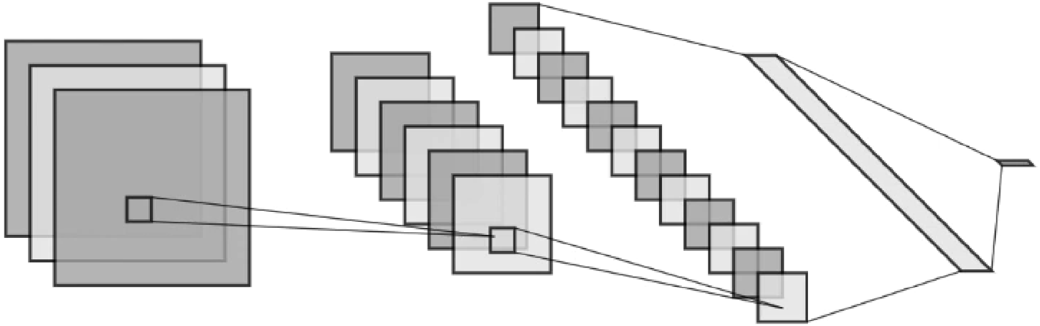
1. How to scale
2. Architecture
3. Results

1. How to scale

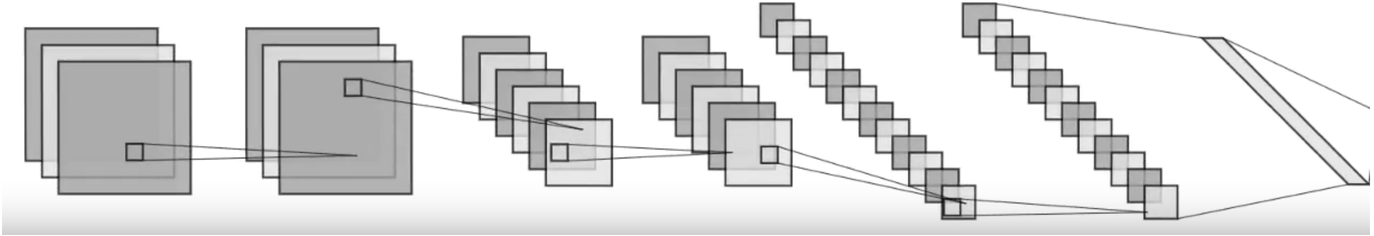
1. How to scale



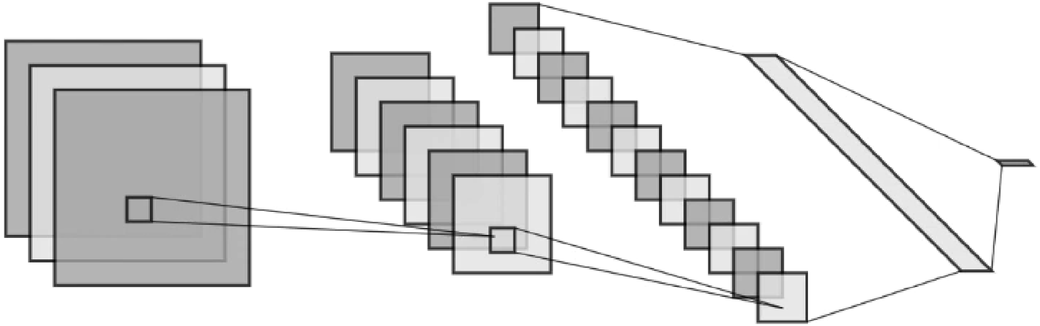
1. How to scale



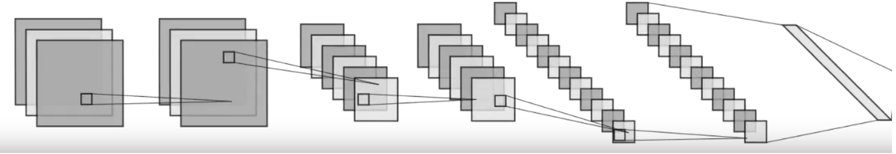
1. Add new layers



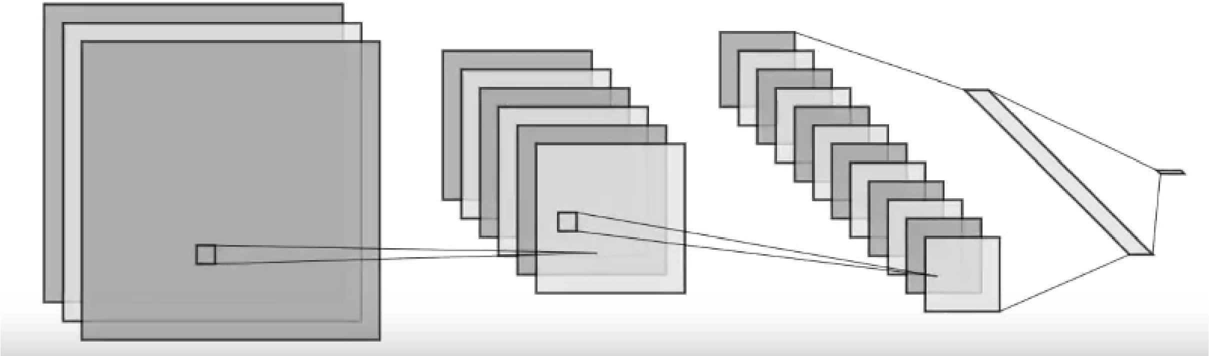
1. How to scale



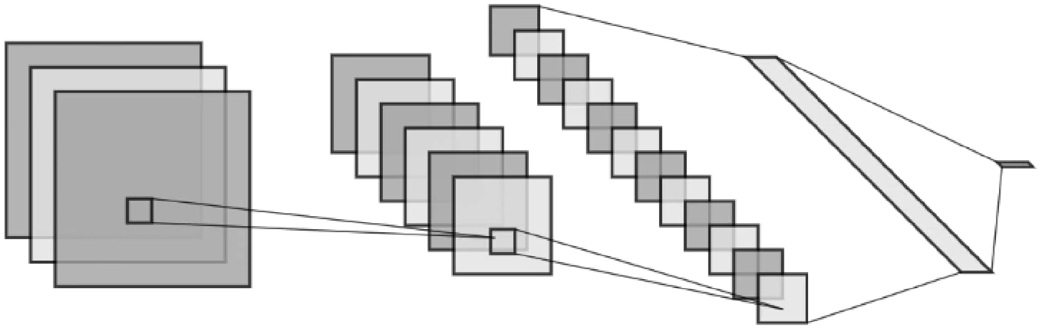
1. Add new layers



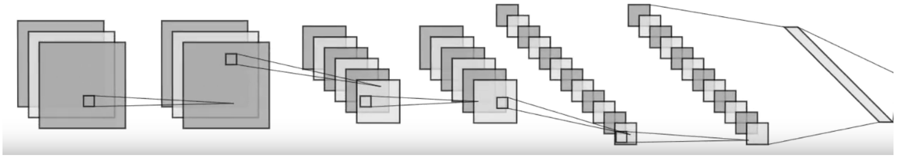
2. Increase height and width



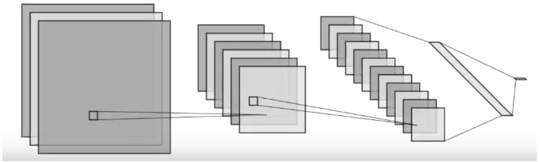
1. How to scale



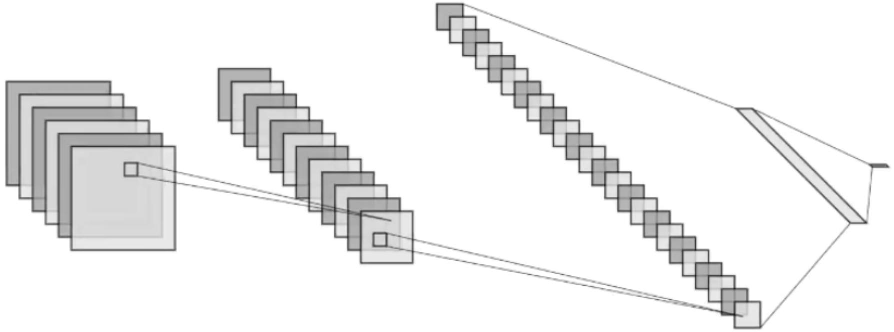
1. Add new layers



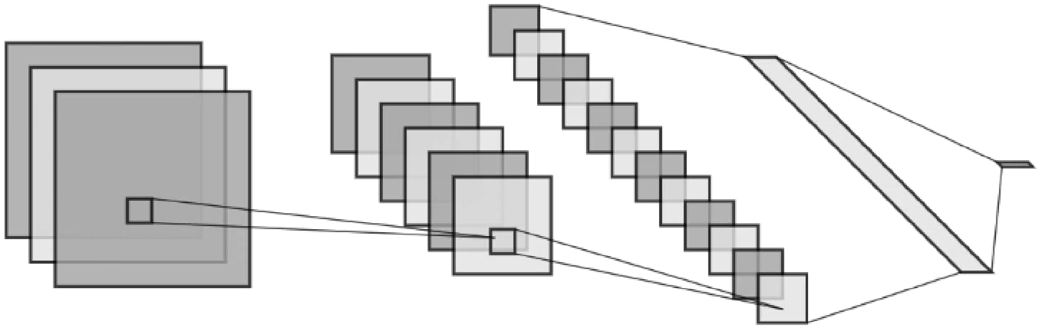
2. Increase height and width



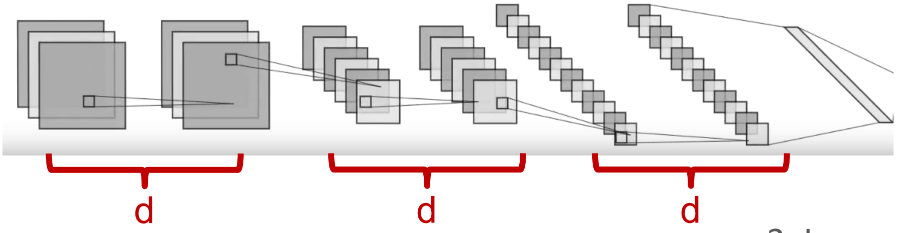
3. Increase the number of channels



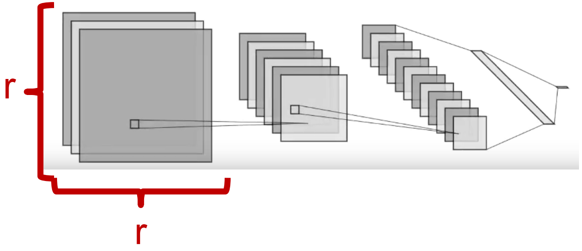
1. How to scale



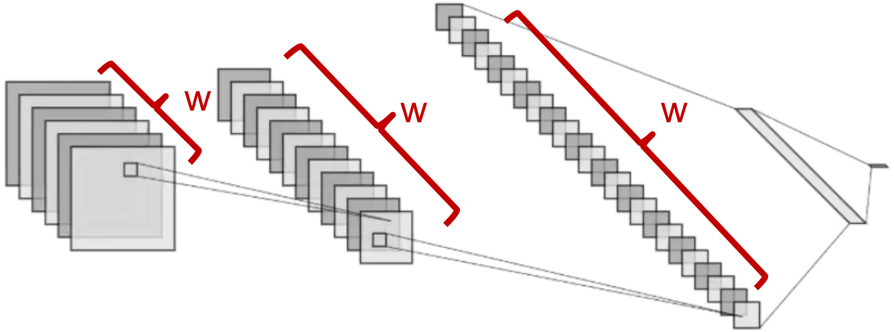
1. Add new layers **(d=2)**



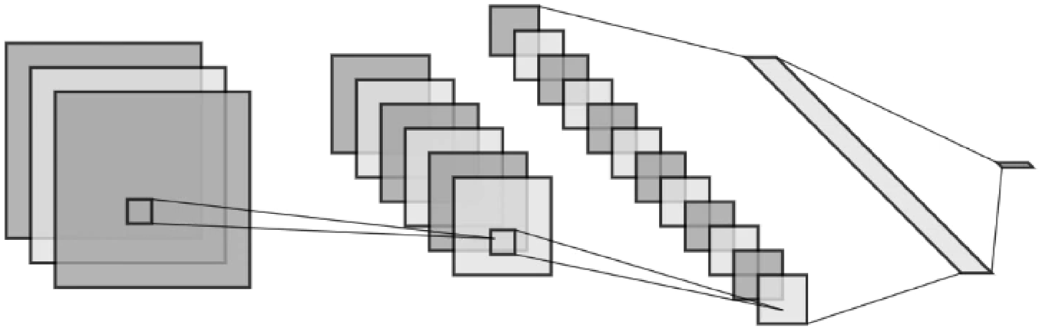
2. Increase height and width **(r=2)**



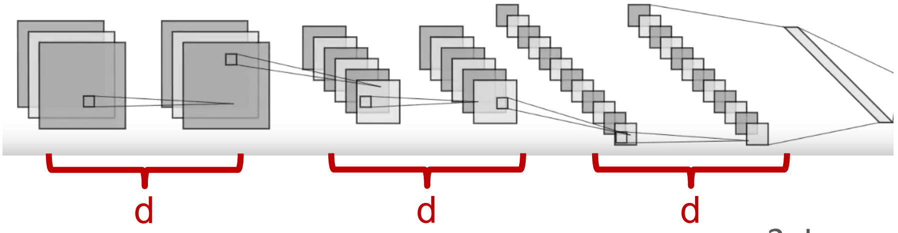
3. Increase the number of channels **(w=2)**



1. How to scale

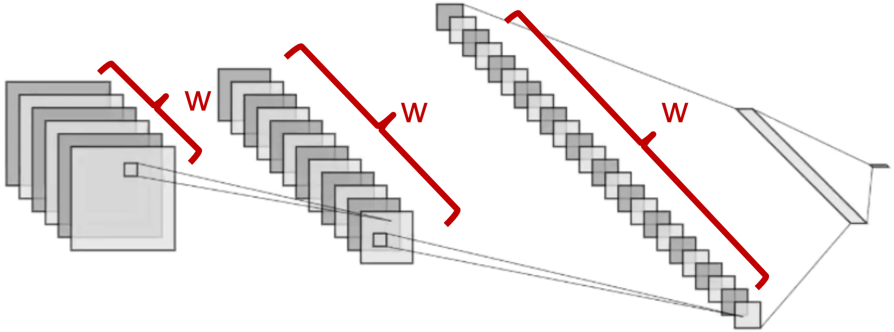


1. Add new layers **(d=2)**

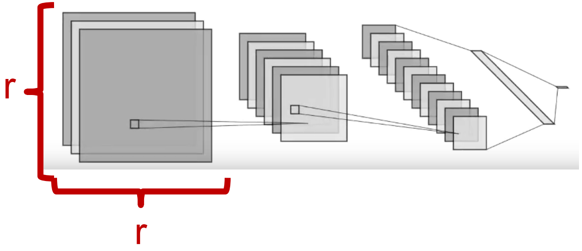


$$\max_{d,w,r}\ Accuracy(Network(d,w,r))$$
$$s.t.\ Hardware\ requiered < Hardware\ capabilities$$
$$d \geq 1, w \geq 1, r \geq 1$$

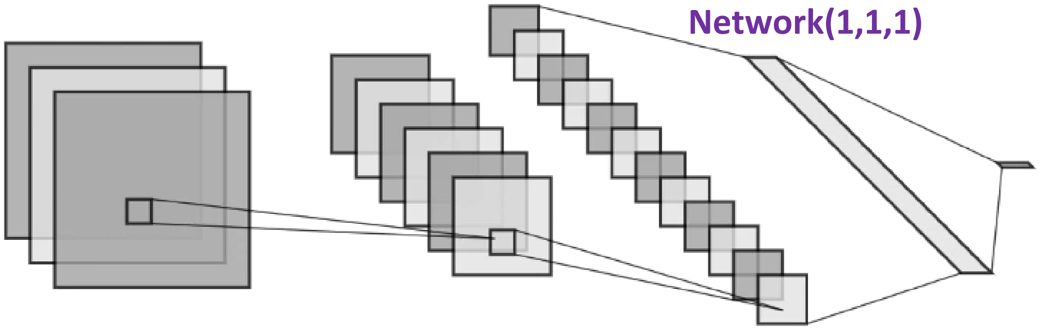
3. Increase the number of channels **(w=2)**



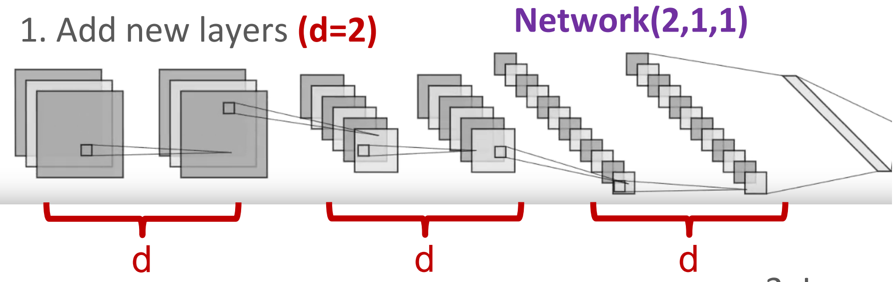
2. Increase height and width **(r=2)**



1. How to scale

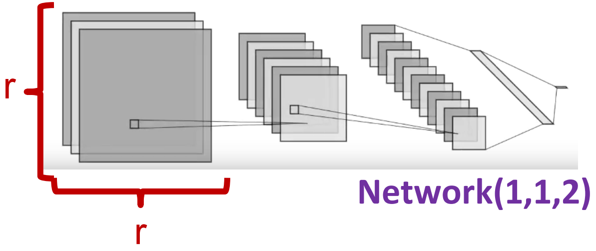


1. Add new layers **(d=2)**

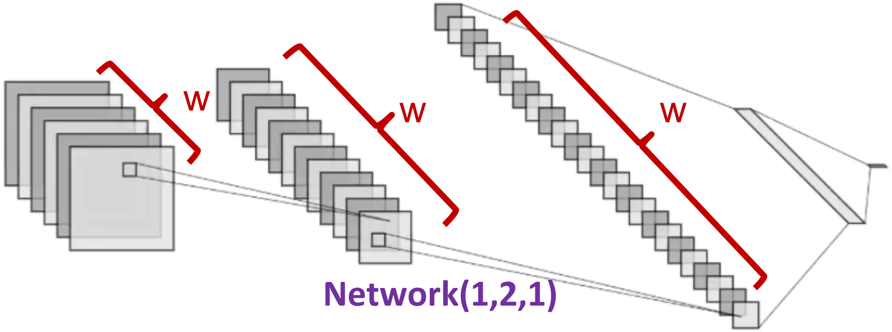


$$\max_{d,w,r} \; Accuracy(Network(d,w,r))$$
$$s.t. \; Hardware\ requiered < Hardware\ capabilities$$
$$d \geq 1, w \geq 1, r \geq 1$$

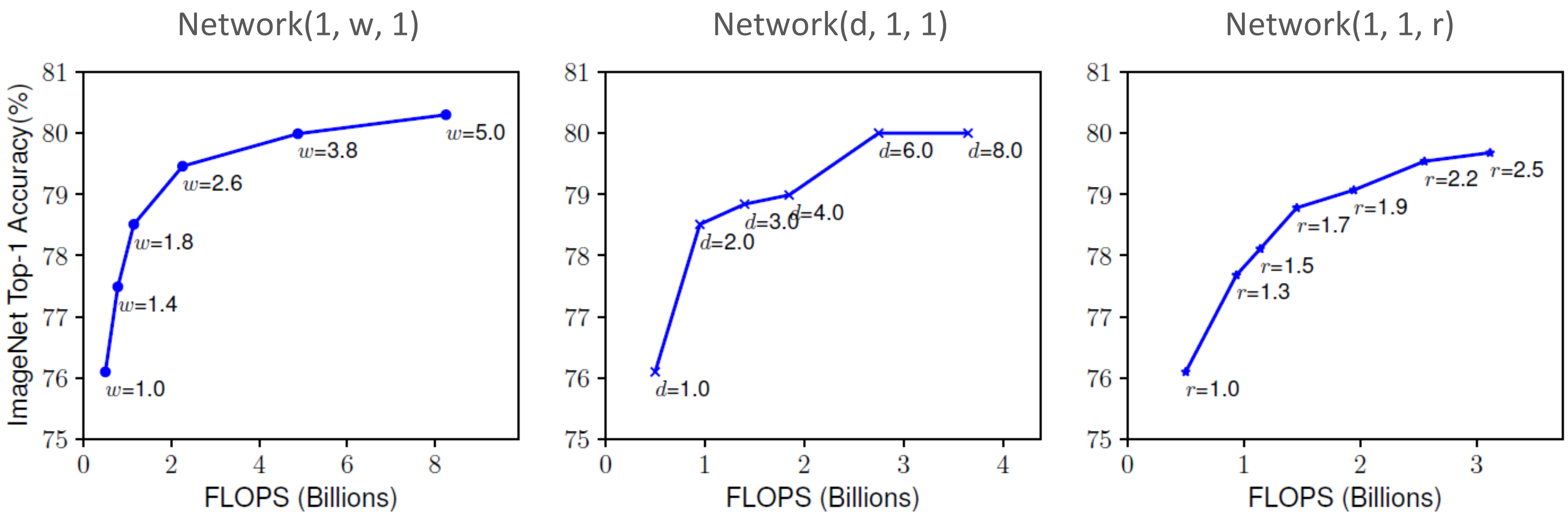
2. Increase height and width **(r=2)**



3. Increase the number of channels **(w=2)**



1. How to scale



FLOPS: Float Operations Per Second

1. How to scale

Hypothesis: Scale d, r, w evenly. We can define $\alpha,\ \beta,\ \gamma$ s.t.

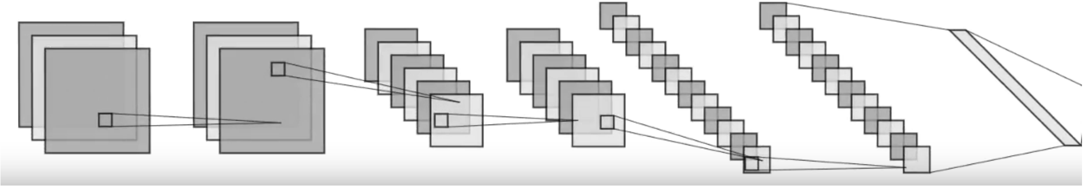
$$d = \alpha^{\phi}$$
$$w = \beta^{\phi}$$
$$r = \gamma^{\phi}$$

$$\max_{d,w,r} \ Accuracy(Network(d,w,r))$$
$$s.t.\ Hardware\ requiered < Hardware\ capabilities$$
$$d \geq 1, \qquad w \geq 1, \qquad r \geq 1$$

$$\max_{d,w,r} \ Accuracy(Network(\alpha,\beta,\gamma))$$
$$s.t.\ Hardware\ requiered < Hardware\ capabilities$$
$$\alpha \geq 1, \qquad \beta \geq 1, \qquad \gamma \geq 1$$

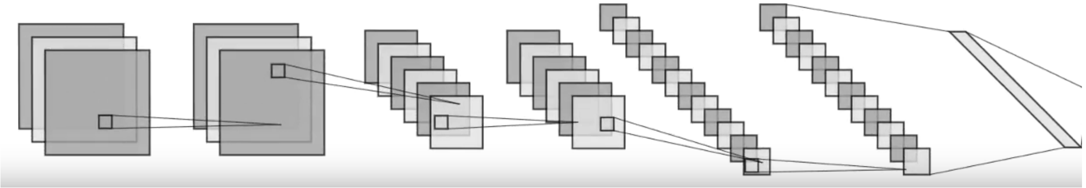
We can find the best parameters $\alpha,\ \beta,\ \gamma$ using grid search

1. How to scale

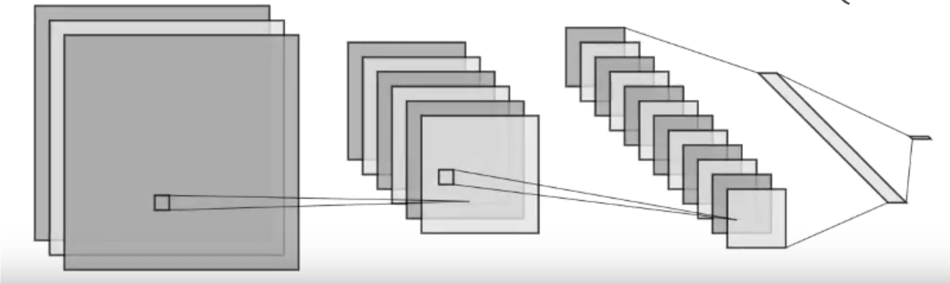


$$FLOPS\big(Network(d,w,r)\big) \propto d \cdot FLOPS\big(Network(1,1,1)\big)$$

1. How to scale

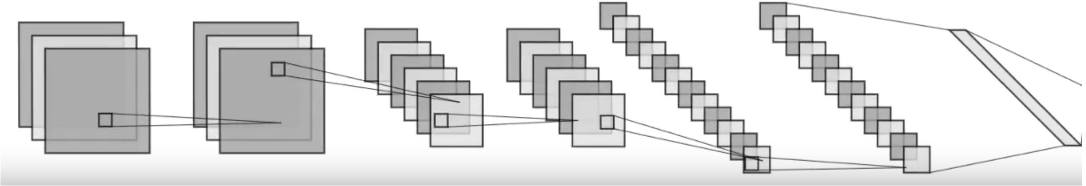


$$FLOPS(Network(d,w,r)) \propto d \cdot FLOPS(Network(1,1,1))$$

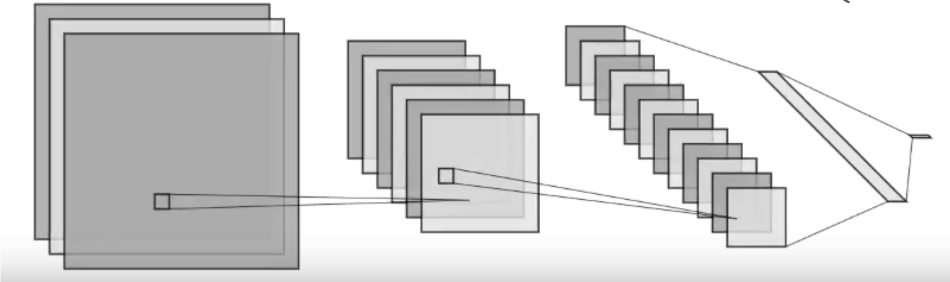


$$FLOPS(Network(d,w,r)) \propto r^2 \cdot FLOPS(Network(1,1,1))$$

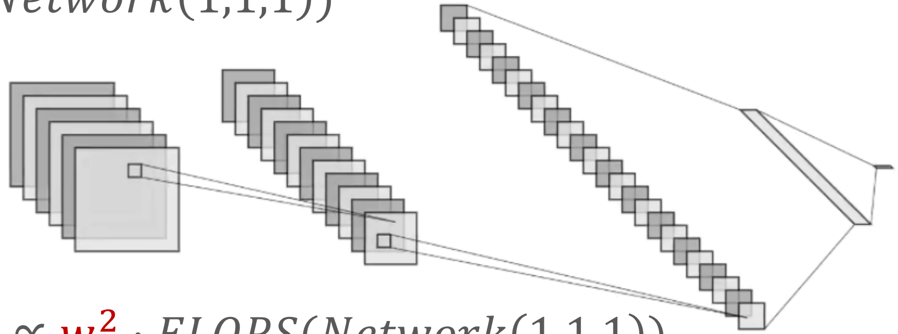
1. How to scale



$$FLOPS\big(Network(d,w,r)\big) \propto d \cdot FLOPS\big(Network(1,1,1)\big)$$



$$FLOPS\big(Network(d,w,r)\big) \propto r^2 \cdot FLOPS(Network(1,1,1))$$



$$FLOPS\big(Network(d,w,r)\big) \propto w^2 \cdot FLOPS(Network(1,1,1))$$

1. How to scale

$$FLOPS\big(Network(d, w, r)\big) \propto dw^2r^2 \cdot FLOPS\big(Network(1,1,1)\big)$$

1. How to scale

$$FLOPS\big(Network(d,w,r)\big) \propto dw^2r^2 \cdot FLOPS\big(Network(1,1,1)\big)$$

$$FLOPS\big(Network(d,w,r)\big) = \alpha\beta^{2\phi}\gamma^{2\phi} \cdot FLOPS\big(Network(1,1,1)\big)$$

$$\frac{FLOPS\big(Network(d,w,r)\big)}{FLOPS\big(Network(1,1,1)\big)} = (\alpha\beta^2\gamma^2)^\phi = 2^\phi$$

1. How to scale

$$FLOPS\big(Network(d,w,r)\big) \propto dw^2r^2 \cdot FLOPS\big(Network(1,1,1)\big)$$

$$FLOPS\big(Network(d,w,r)\big) = \alpha\beta^{2\phi}\gamma^{2\phi} \cdot FLOPS\big(Network(1,1,1)\big)$$

$$\frac{FLOPS\big(Network(d,w,r)\big)}{FLOPS\big(Network(1,1,1)\big)} = (\alpha\beta^2\gamma^2)^\phi = 2^\phi$$

$$\max_{d,w,r} \; Accuracy(Network(\alpha,\beta,\gamma))$$

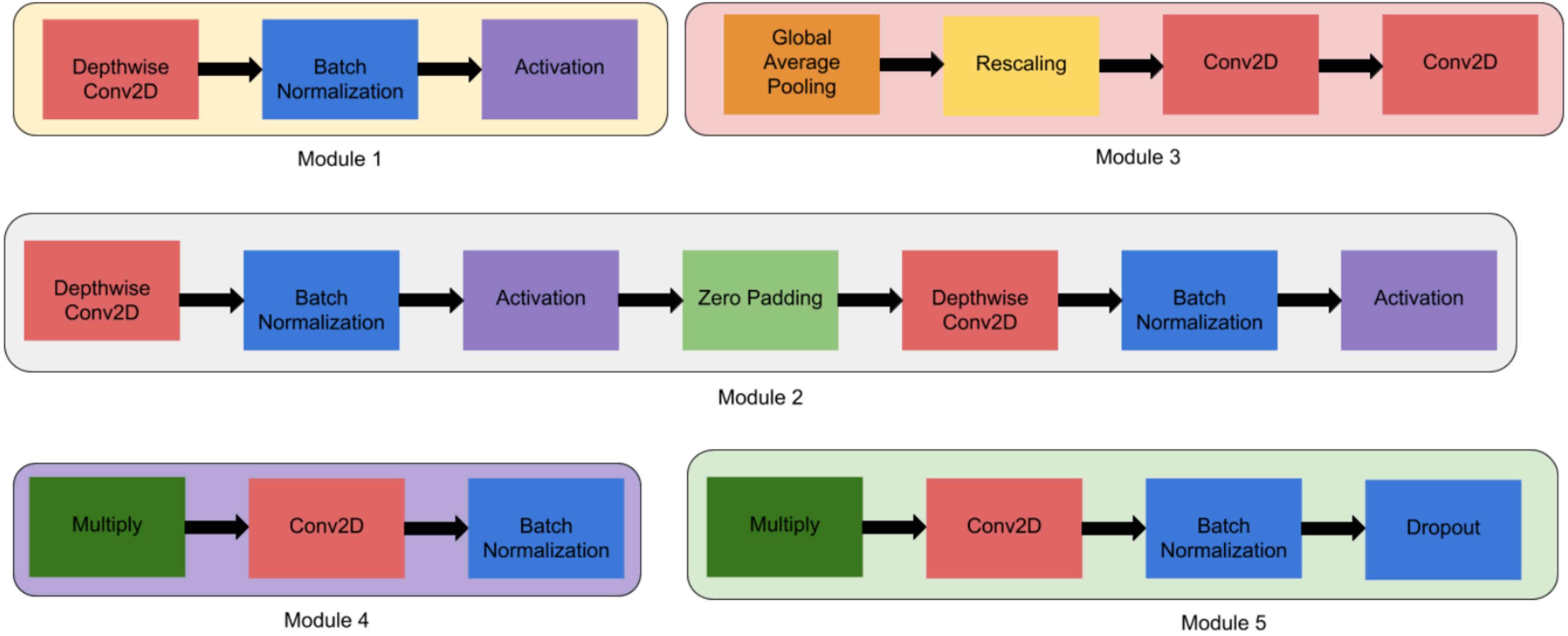
$$\alpha\beta^2\gamma^2 \approx 2$$

$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

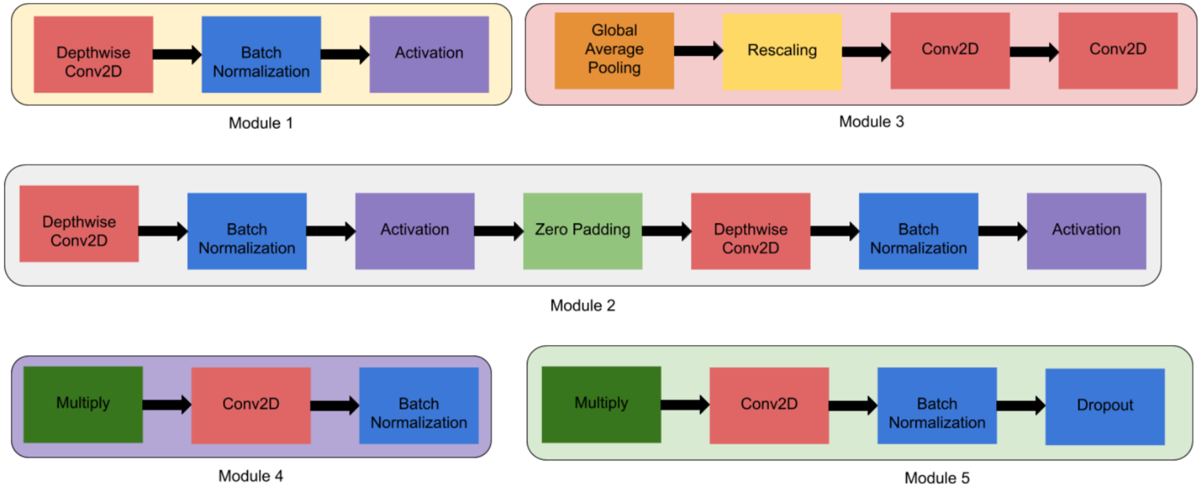
Given a network, ϕ is used to define EfficientNet B0 (ϕ=1), …, B7 (ϕ=8)

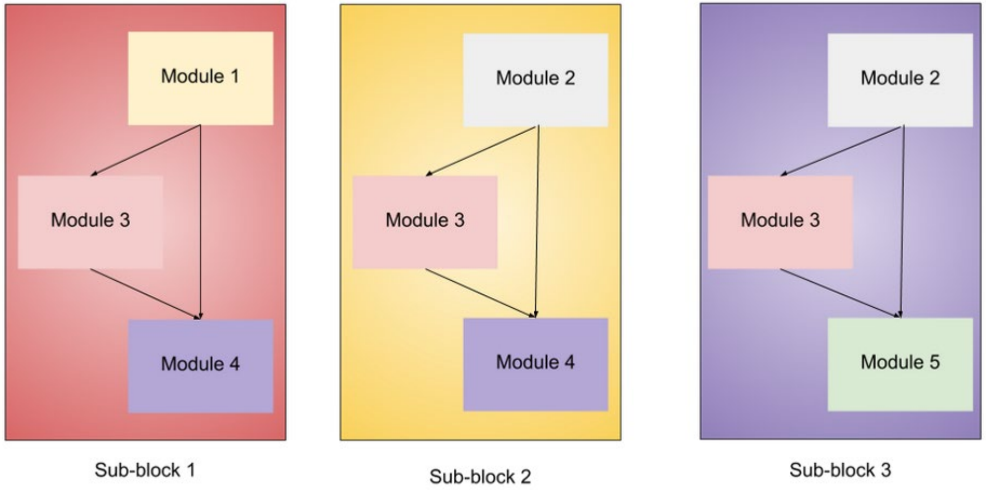
2. Architecture

2. Architecture

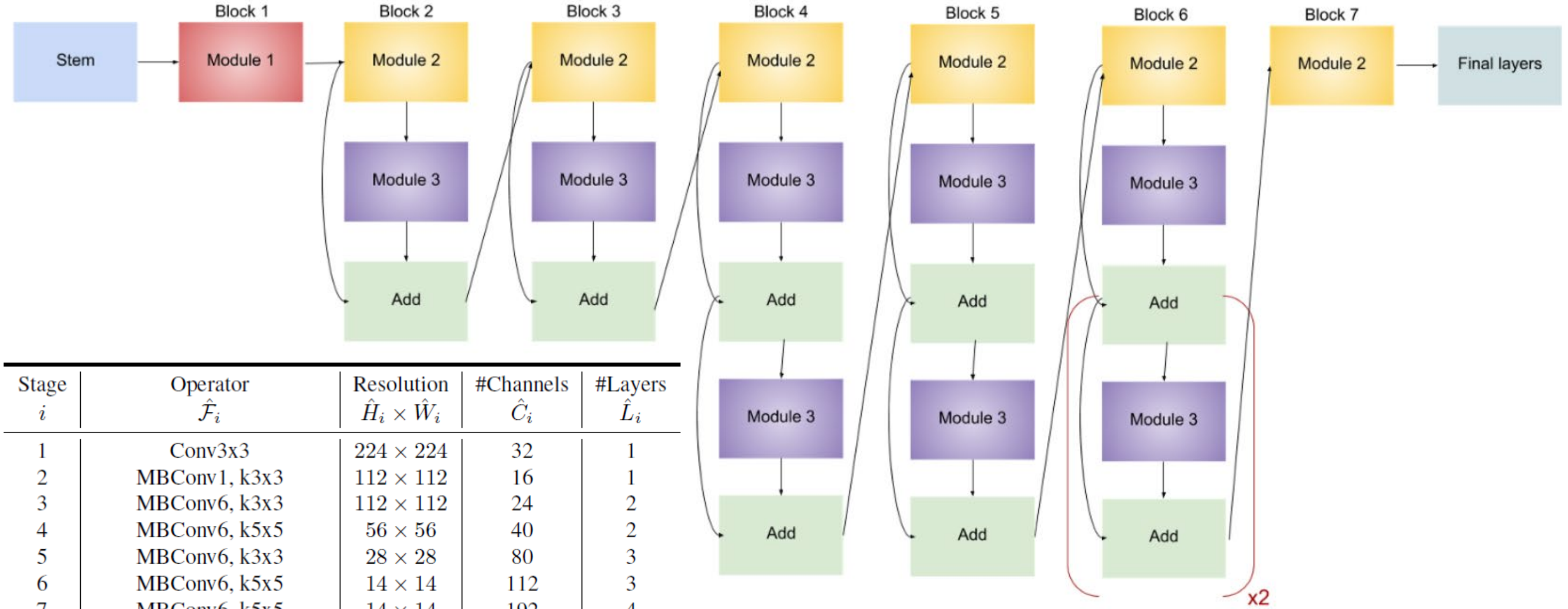


2. Architecture





2. Architecture



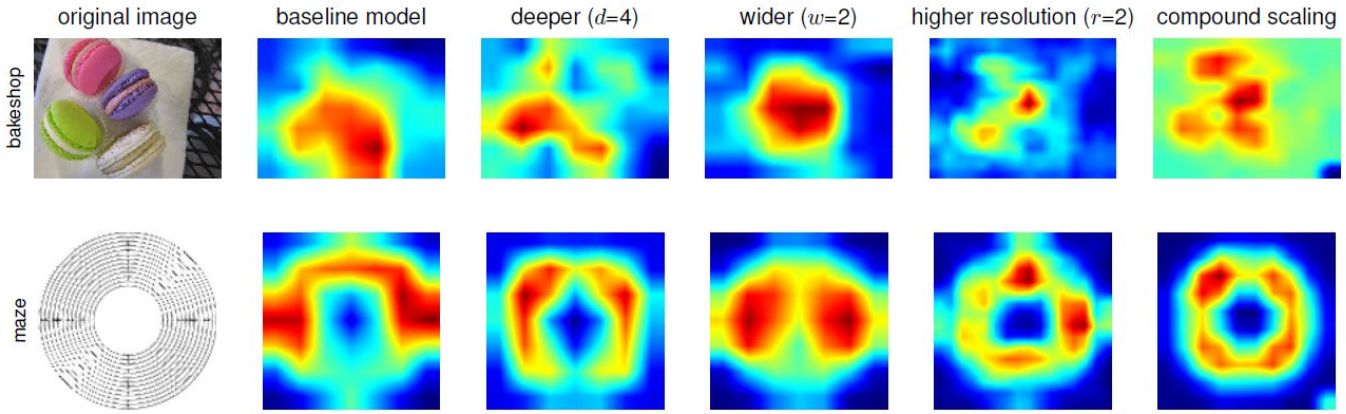
Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14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

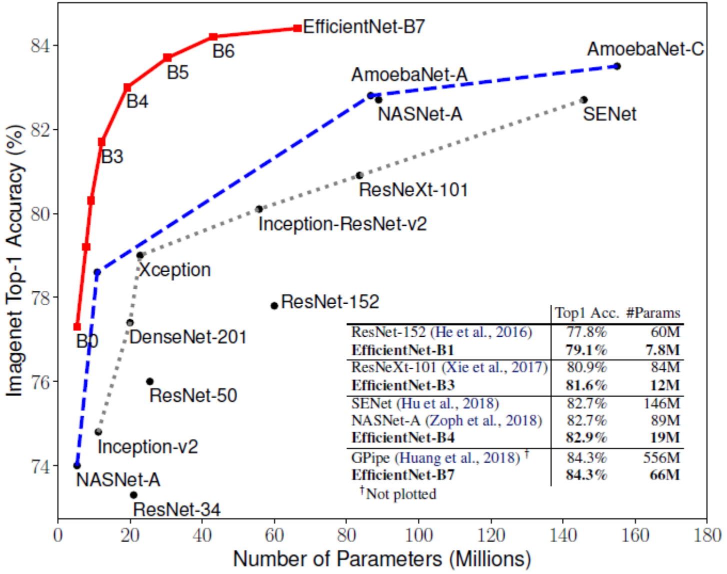
MBConv: mobile inverted bottleneck (https://arxiv.org/abs/1801.04381v4)

3. Results

Model	Top-1 Acc.	Top-5 Acc.	#Params	Ratio-to-EfficientNet	#FLOPs	Ratio-to-EfficientNet
EfficientNet-B0	**77.1%**	**93.3%**	**5.3M**	**1x**	**0.39B**	**1x**
ResNet-50 (He et al., 2016)	76.0%	93.0%	26M	4.9x	4.1B	11x
DenseNet-169 (Huang et al., 2017)	76.2%	93.2%	14M	2.6x	3.5B	8.9x
EfficientNet-B1	**79.1%**	**94.4%**	**7.8M**	**1x**	**0.70B**	**1x**
ResNet-152 (He et al., 2016)	77.8%	93.8%	60M	7.6x	11B	16x
DenseNet-264 (Huang et al., 2017)	77.9%	93.9%	34M	4.3x	6.0B	8.6x
Inception-v3 (Szegedy et al., 2016)	78.8%	94.4%	24M	3.0x	5.7B	8.1x
Xception (Chollet, 2017)	79.0%	94.5%	23M	3.0x	8.4B	12x
EfficientNet-B2	**80.1%**	**94.9%**	**9.2M**	**1x**	**1.0B**	**1x**
Inception-v4 (Szegedy et al., 2017)	80.0%	95.0%	48M	5.2x	13B	13x
Inception-resnet-v2 (Szegedy et al., 2017)	80.1%	95.1%	56M	6.1x	13B	13x
EfficientNet-B3	**81.6%**	**95.7%**	**12M**	**1x**	**1.8B**	**1x**
ResNeXt-101 (Xie et al., 2017)	80.9%	95.6%	84M	7.0x	32B	18x
PolyNet (Zhang et al., 2017)	81.3%	95.8%	92M	7.7x	35B	19x
EfficientNet-B4	**82.9%**	**96.4%**	**19M**	**1x**	**4.2B**	**1x**
SENet (Hu et al., 2018)	82.7%	96.2%	146M	7.7x	42B	10x
NASNet-A (Zoph et al., 2018)	82.7%	96.2%	89M	4.7x	24B	5.7x
AmoebaNet-A (Real et al., 2019)	82.8%	96.1%	87M	4.6x	23B	5.5x
PNASNet (Liu et al., 2018)	82.9%	96.2%	86M	4.5x	23B	6.0x
EfficientNet-B5	**83.6%**	**96.7%**	**30M**	**1x**	**9.9B**	**1x**
AmoebaNet-C (Cubuk et al., 2019)	83.5%	96.5%	155M	5.2x	41B	4.1x
EfficientNet-B6	**84.0%**	**96.8%**	**43M**	**1x**	**19B**	**1x**
EfficientNet-B7	**84.3%**	**97.0%**	**66M**	**1x**	**37B**	**1x**
GPipe (Huang et al., 2018)	84.3%	97.0%	557M	8.4x	-	-

3. Results





	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Thanks!