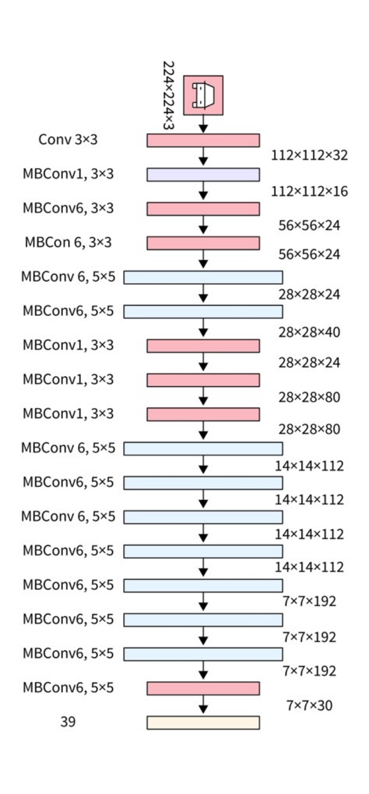
EfficientNet:



→ It is a product of 2 techniques:

1. Neural Architecture Search (NAS)
2. Compound Scaling

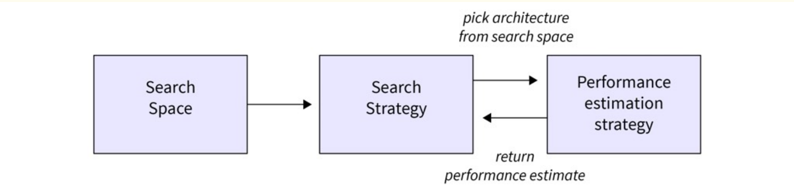
Process:

1. create an efficient baseline architecture using NAS.
2. use the compound scaling method to enhance the performance.

1. NAS:

→ finds efficient and optimized baseline model with better performance and several parameters keeping in mind.

→ a good baseline model is always desireable for scale up the performance.



→ It searches and evaluates many architectures in the search space & returns best model suited for the task.

2. Compound Scaling:

→ scaling is done on mainly 3 components.

- → width scaling
- → depth scaling
- → resolution scaling

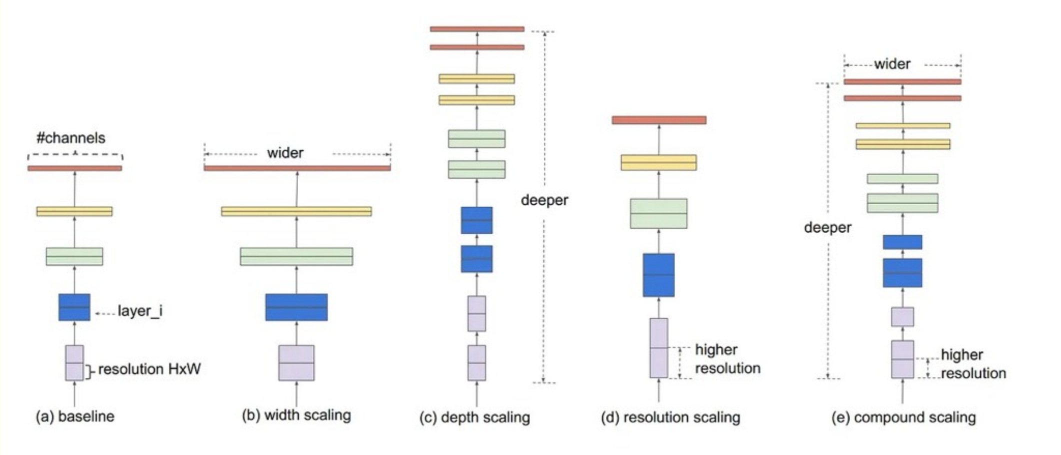
→ scaling of the co-ordinates of baseline model (w, d, r) in a balanced and co-ordinated manner.

→ scaling of the every dimension is derived from the compound coefficient = ϕ

→ goal is to find best exponents, that results best trade-off b/w model accuracy and computational efficiency.

→ $\phi \downarrow$ → more lightweight and resource-efficient model.

$\phi \uparrow$ → powerful but computation-expensive.



→ mathematically, it can be expressed-

depth : $d = \alpha^{\phi}$

width : $w = \beta^{\phi}$

resolution : $r = \gamma^{\phi}$

such that-

- $\alpha \beta^2 \gamma^2 \approx 2$
- $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$

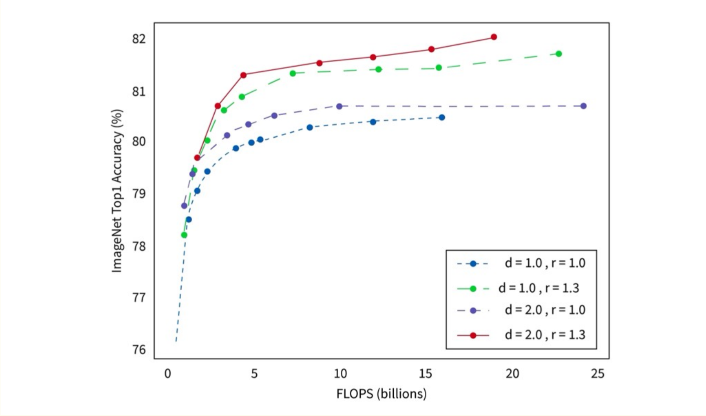
→ α, β, γ are chosen using grid search.

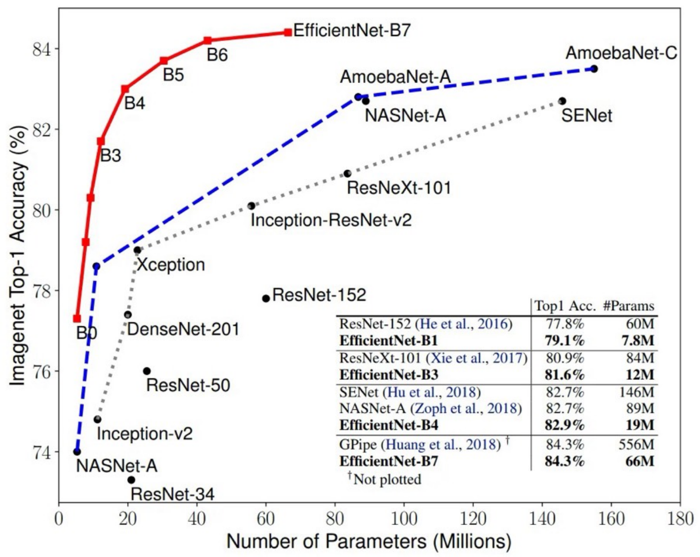
→ convolution operation (FLOPS) $\propto d, w^2, r^2$

↳ if we doubles the depth FLOPs will become 4 times.

→ In EfficientNet B0, $\phi = 1$ and $\alpha = 1.2, \beta = 1.1, \gamma = 1.15$

→ for different ϕ, we get models B1 to B7.





	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