Thay đổi chiều sâu, rộng, độ phân giải (depth, height, resolution) có thể nâng cao hiệu suất mô hình.

Vd:

- ResNet-18 có thể scale thành ResNet-200 bằng cách thêm layer.
- GPipe đạt độ chính xác 84.3% trên tập ImageNet (top1 accuracy) bằng cách scale mô hình ban đầu lớn lên 4 lần.

Trước đây có nhiều cách để scale mô hình, thường là theo depth hay resolution. Paper để xuất một phương pháp scale gọi là **compound scaling method** (scale cả depth, height, resolution theo cùng một tỉ lệ).

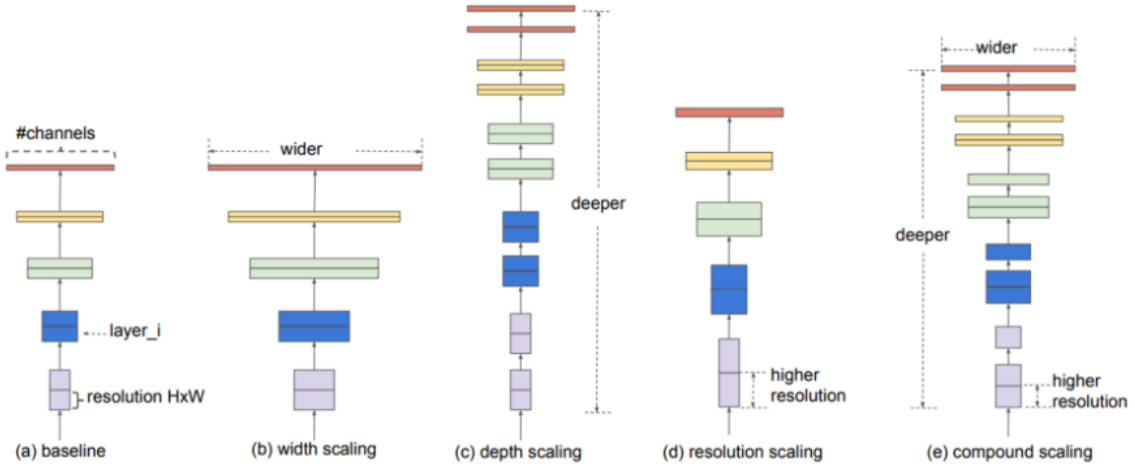
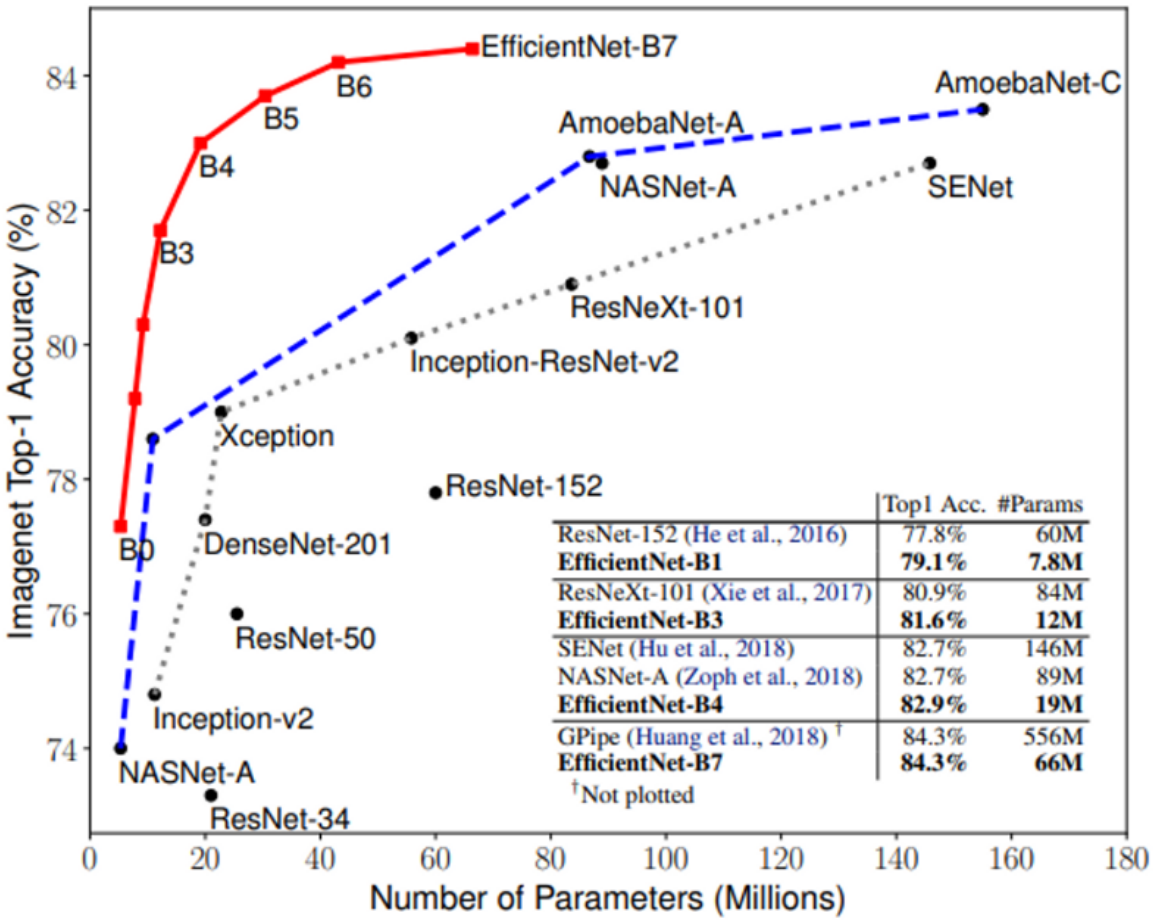


Figure 2. **Model Scaling.** (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.

Hiệu quả của việc scale model phụ thuộc khá nhiều vào baseline network. Nhóm tác giả sử dụng mô hình **Neural architecture search** để phát triển một baseline network mới và scale nó để tạo ra một họ mô hình gọi là **EfficientNets**.



	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

Họ EfficientNets có độ chính xác cao hơn so với các ConvNet hiện tại.

Bình thường một ConvNet có thể biểu diễn dưới dạng:

$$\mathcal{N} = \mathcal{F}_k \odot \ldots \odot \mathcal{F}_2 \odot \mathcal{F}_1(X_1) = \bigodot_{j=1\ldots k} \mathcal{F}_j(X_1)$$

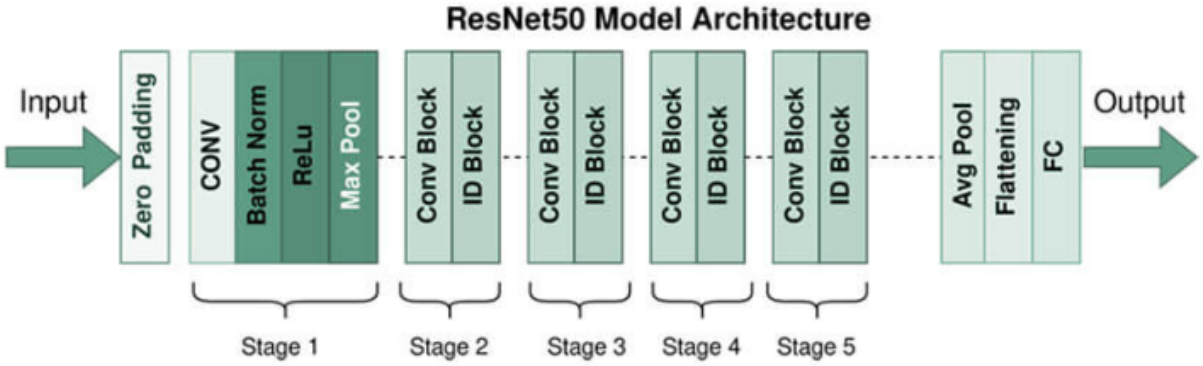
Trong đó:

N: ConvNet

X_i: input tensor

F_j: phép toán tại layer j

ConvNet cũng thường được chia thành nhiều stage và các layer trong từng stage thường có cấu trúc tương tự nhau.



Do đó ConvNet có thể được biểu diễn thành

$$\mathcal{N} = \bigodot_{i=1\ldots s} \mathcal{F}_i^{L_i}\left(X_{\langle H_i, W_i, C_i\rangle}\right)$$

Trong đó:

$F_i^{L_i}$ biểu thị layer F_i lặp lại L_i lần ở stage i

$< H_i, W_i, C_i >$ là shape của tensor X

Mục tiêu chính mà bài báo muốn thực hiện đó là scale model để maximize model accuracy với bất kì ràng buộc nào về tài nguyên.

$$\begin{aligned} \max_{d,w,r} \quad & Accuracy\left(\mathcal{N}(d,w,r)\right) \\ s.t. \quad & \mathcal{N}(d,w,r) = \bigodot_{i=1\ldots s} \hat{\mathcal{F}}_i^{d\cdot\hat{L}_i}\left(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\right) \\ & \text{Memory}(\mathcal{N}) \leq \text{target\_memory} \\ & \text{FLOPS}(\mathcal{N}) \leq \text{target\_flops} \end{aligned} \tag{2}$$

với d, w, r là các hệ số scale.

Vấn đề ở công thức (2) bên trên là các hệ số d, w, r tối ưu phụ thuộc vào nhau và giá trị của chúng thay đổi với tài nguyên tính toán khác nhau.

Nhóm tác giả đã thực hiện 2 khảo sát:

Khảo sát 1: scale chiều rộng, sâu hay độ phân giải có làm tăng accuracy nhưng tới 1 mức nào đó nó sẽ bão hòa

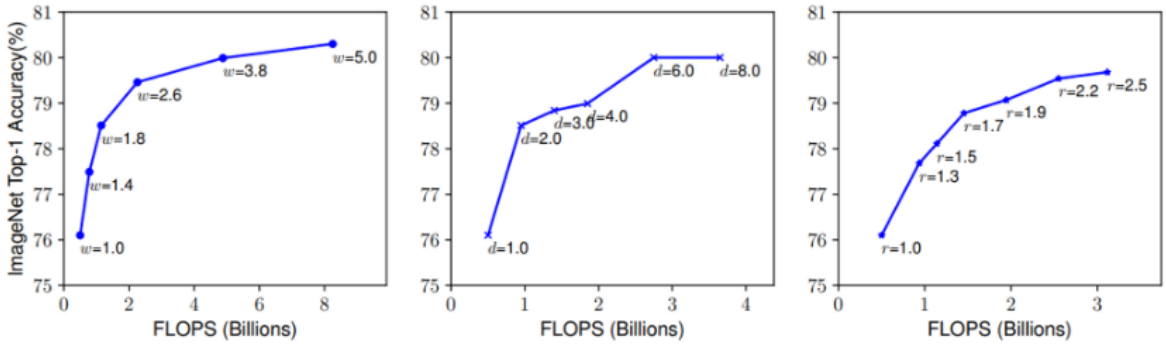


Figure 3. **Scaling Up a Baseline Model with Different Network Width (w), Depth (d), and Resolution (r) Coefficients.** Bigger networks with larger width, depth, or resolution tend to achieve higher accuracy, but the accuracy gain quickly saturate after reaching 80%, demonstrating the limitation of single dimension scaling. Baseline network is described in Table 1.

Khảo sát 2: để có được hiệu suất tốt, cần phải cân bằng các chiều depth, height, resolution khi scale

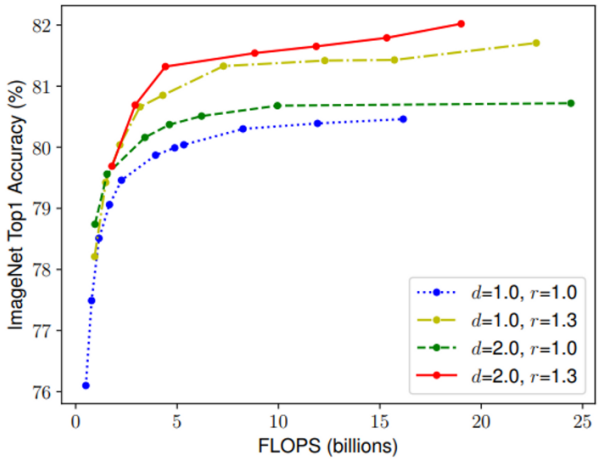


Figure 4. **Scaling Network Width for Different Baseline Networks**. Each dot in a line denotes a model with different width coefficient (w). All baseline networks are from Table 1. The first baseline network (d=1.0, r=1.0) has 18 convolutional layers with resolution 224x224, while the last baseline (d=2.0, r=1.3) has 36 layers with resolution 299x299.

Qua 2 khảo sát trên, nhóm tác giả đã đề xuất sử dụng một hệ số Φ để scale các chiều của ConvNet một cách có quy tắc.

$$\begin{aligned} \text{depth: } & d = \alpha^\phi \\ \text{width: } & w = \beta^\phi \\ \text{resolution: } & r = \gamma^\phi \\ \text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned} \tag{3}$$

Trong đó;

Φ là hệ số do người dùng tự cho để kiểm soát việc có thêm bao nhiêu tài nguyên có sẵn cho việc scale model (với mỗi Φ mới tổng FLOPS sẽ tăng xấp xỉ 2^Φ)

α, β, γ quyết định việc phân chia tài nguyên tính toán vào network depth, width, resolution

Kiến trúc mạng EfficientNet-B0 tương tự như MnasNet

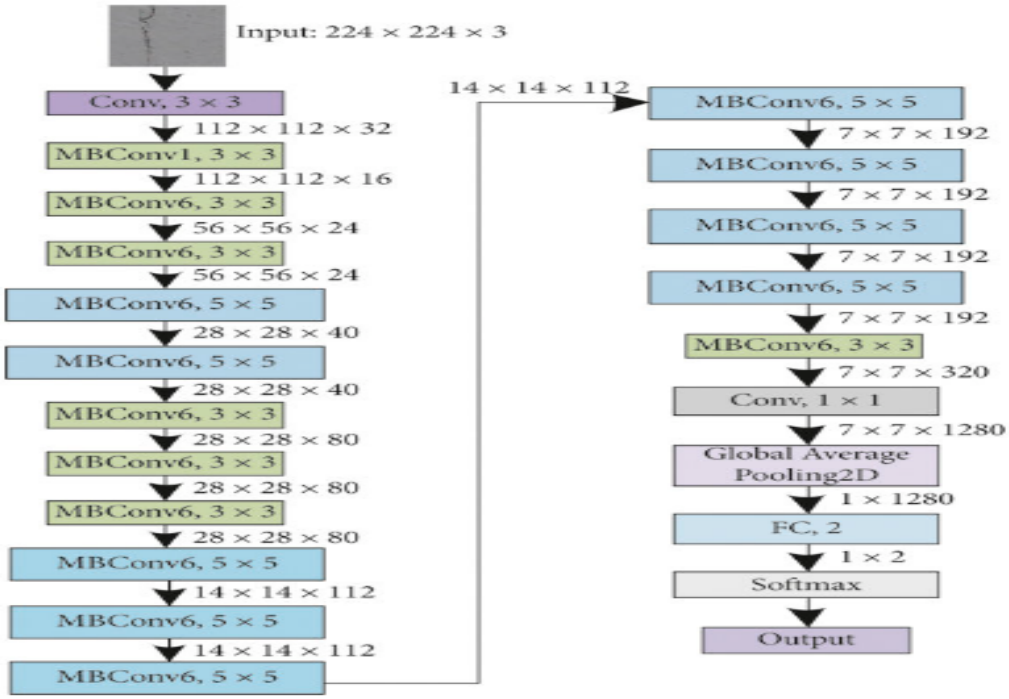


Table 1. **EfficientNet-B0 baseline network** – Each row describes a stage i with $\hat{L}_i$ layers, with input resolution $\langle \hat{H}_i, \hat{W}_i \rangle$ and output channels $\hat{C}_i$. Notations are adopted from equation 2.

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

từ baseline EfficientNet-B0, thực hiện **compound scaling method** qua 2 bước:

Bước 1: cho Φ = 1 (giả sử tài nguyên tính toán tăng gấp đôi), sau đó khảo sát để tìm α, β, γ qua công thức (2) và (3). Theo thực nghiệm best value cho EfficientNet-B0 là α = 1.2, β =1.1, γ = 1.1 (với ràng buộc α · β^2· γ^2 ≈ 2)

Bước 2: giữ α, β, γ làm hằng số và sau đó thay đổi Φ để có được EfficientNet-B1 → EfficientNet-B7 (chi tiết bảng 2)

Table 2. **EfficientNet Performance Results on ImageNet** (Russakovsky et al., 2015). All EfficientNet models are scaled from our baseline EfficientNet-B0 using different compound coefficient ϕ in Equation 3. ConvNets with similar top-1/top-5 accuracy are grouped together for efficiency comparison. Our scaled EfficientNet models consistently reduce parameters and FLOPS by an order of magnitude (up to 8.4x parameter reduction and up to 16x FLOPS reduction) than existing ConvNets.

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	-	-

We omit ensemble and multi-crop models (Hu et al., 2018), or models pretrained on 3.5B Instagram images (Mahajan et al., 2018).