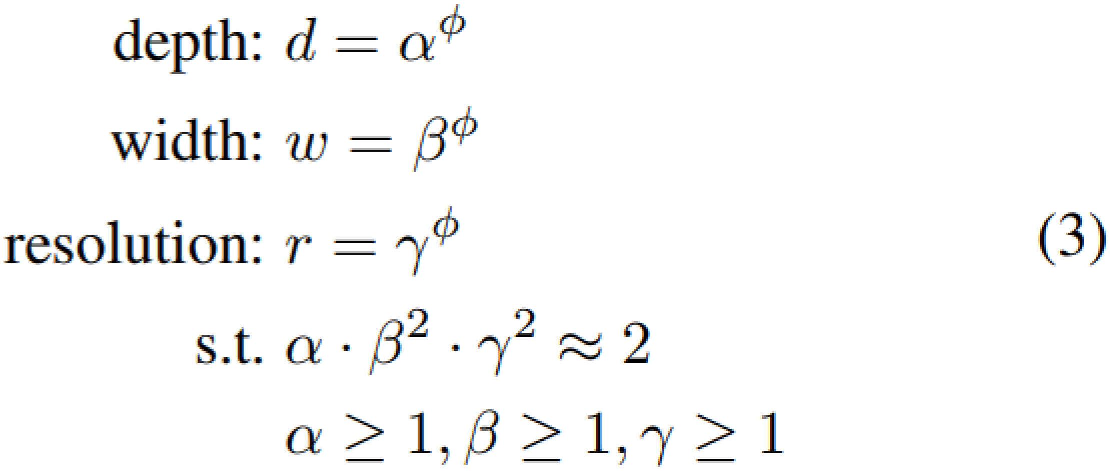
Notes for “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”

When data scientists tweak models, they tweak the width of layers (number of channels in layers), depth of layers, and resolution of images. However, there lacks a systematic method to make these tweaks to maximize performance. This article talks about the scaling coefficients to perform the most beneficial tweaks to the baseline model. These scaling coefficients can produce models with state-of-the-art performance and significant reduction in parameters.

Neural architecture search is a method to autonomously produce neural network architectures to outperform or be on par with hand-crafted

FLOPS = floating point operations per second.

The constraints for this method are: the baseline architecture is not changed; all layers are scaled uniformly with constant ratio.



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

Where α, β, γ are constants that can be determined by a small grid search. Intuitively, φ is a user-specified coefficient that controls how many more resources are available for model scaling, while α, β, γ specify how to assign these extra resources to network width, depth, and resolution respectively.

Notably, the FLOPS of a regular convolution op is proportional to d, w^2, r^2. Since convolution ops usually dominate the computation cost in ConvNets, scaling a ConvNet with equation 3 will approximately increase total FLOPS by$(\alpha \cdot \beta^2 \cdot \gamma^2)^{\phi}$. The paper constraints $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ such that for any new ϕ, the total FLOPS will approximately increase by 2^{ϕ}.

Steps to use this approach:

1. First fix $\varphi = 1$, assuming twice more resources available, and do a small grid search of α, β, γ based on Equation 2 and 3.
2. Then fix α, β, γ as constants and scale up baseline network with different φ using Equation 3.

EfficientNet Training information is listed under “ImageNet Results for EfficientNet".

References

- https://arxiv.org/pdf/1905.11946.pdf