PhD

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

2020/03/10 10:16 PM

Introduces a new methodology of scaling CNN architectures using a compound scaling coefficient ϕ, that scales

Depth = L

Width = C

Resolution = H x W

With ratios such that :

Depth = L . d

Width = C . w

Resolution = H . r x W . r

And,

$d = \alpha$

$w = \beta$

$r = \gamma$

Under the constraints, $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ *and*, $\alpha \geq 1, \beta \geq 1, \gamma \geq 1$

Because $\alpha \cdot \beta^2 \cdot \gamma^2 \propto FLOPs$ and with these constraints flops increase in the order of 2 as ϕ increases.

Intuitively, ϕ represents the amount of more computation resources available, through which the model can be scaled accordingly. [α, β, γ calculated using a small grid search]

The base Architecture named - EfficientNet-B0 formulated using a Network architecture search on the constraint - $ACC(m) \times [FLOPS(m)/T]^{\omega}$, where ACC(m) → accuracy of model m, FLOPS(m) → FLOPS of model m, T → Target FLOPS, ω → hyperparameter for tradeoff b/w accuracy and flops.

MBConv Block used (Inverted Residual Block from MobileNet-V2)