Title: EfficientNet

URL: https://en.wikipedia.org/wiki/EfficientNet

PageID: 78149844

Categories: Category:Artificial neural networks, Category:Computer vision, Category:Google software, Category:Machine learning

-----

EfficientNet is a family of convolutional neural networks (CNNs) for computer vision published by researchers at Google AI in 2019. Its key innovation is compound scaling , which uniformly scales all dimensions of depth, width, and resolution using a single parameter.

EfficientNet models have been adopted in various computer vision tasks, including image classification , object detection , and segmentation .

Compound scaling

EfficientNet introduces compound scaling , which, instead of scaling one dimension of the network at a time, such as depth (number of layers), width (number of channels), or resolution (input image size), uses a compound coefficient φ {\displaystyle \phi } to scale all three dimensions simultaneously. Specifically, given a baseline network, the depth, width, and resolution are scaled according to the following equations: depth multiplier: d = α φ width multiplier: w = β φ resolution multiplier: r = γ φ {\displaystyle {\begin{aligned}{\text{depth multiplier: }}d&=\alpha ^{\phi }\\{\text{width multiplier: }}w&=\beta ^{\phi }\\{\text{resolution multiplier: }}r&=\gamma ^{\phi }\end{aligned}}} subject to α · β 2 · γ 2 ≈ 2 {\displaystyle \alpha \cdot \beta ^{2}\cdot \gamma ^{2}\approx 2} and α ≥ 1 , β ≥ 1 , γ ≥ 1 {\displaystyle \alpha \geq 1,\beta \geq 1,\gamma \geq 1} . The α · β 2 · γ 2 ≈ 2 {\displaystyle \alpha \cdot \beta ^{2}\cdot \gamma ^{2}\approx 2} condition is such that increasing φ {\displaystyle \phi } by a factor of φ 0 {\displaystyle \phi _{0}} would increase the total FLOPs of running the network on an image approximately 2 φ 0 {\displaystyle 2^{\phi _{0}}} times. The hyperparameters α {\displaystyle \alpha } , β {\displaystyle \beta } , and γ {\displaystyle \gamma } are determined by a small grid search . The original paper suggested 1.2, 1.1, and 1.15, respectively.

Architecturally, they optimized the choice of modules by neural architecture search (NAS), and found that the inverted bottleneck convolution (which they called MBConv) used in MobileNet worked well.

The EfficientNet family is a stack of MBConv layers, with shapes determined by the compound scaling. The original publication consisted of 8 models, from EfficientNet-B0 to EfficientNet-B7, with increasing model size and accuracy. EfficientNet-B0 is the baseline network, and subsequent models are obtained by scaling the baseline network by increasing φ {\displaystyle \phi } .

Variants

EfficientNet has been adapted for fast inference on edge TPUs and centralized TPU or GPU clusters by NAS.

EfficientNet V2 was published in June 2021. The architecture was improved by further NAS search with more types of convolutional layers. It also introduced a training method, which progressively increases image size during training, and uses regularization techniques like dropout , RandAugment , and Mixup. The authors claim this approach mitigates accuracy drops often associated with progressive resizing.

See also

Convolutional neural network

SqueezeNet

MobileNet

You Only Look Once

References

External links

EfficientNet: Improving Accuracy and Efficiency through AutoML and Model Scaling (Google AI Blog)

v

t

e

Google

Google Brain

Google DeepMind

AlphaGo (2015)

Master (2016)

AlphaGo Zero (2017)

AlphaZero (2017)

MuZero (2019)

Fan Hui (2015)

Lee Sedol (2016)

Ke Jie (2017)

AlphaGo (2017)

The MANIAC (2023)

AlphaFold (2018)

AlphaStar (2019)

AlphaDev (2023)

AlphaGeometry (2024)

AlphaGenome (2025)

Inception (2014)

WaveNet (2016)

MobileNet (2017)

Transformer (2017)

EfficientNet (2019)

Gato (2022)

Quantum Artificial Intelligence Lab

TensorFlow

Tensor Processing Unit

Assistant (2016)

Sparrow (2022)

Gemini (2023)

BERT (2018)

XLNet (2019)

T5 (2019)

LaMDA (2021)

Chinchilla (2022)

PaLM (2022)

Imagen (2023)

Gemini (2023)

VideoPoet (2024)

Gemma (2024)

Veo (2024)

DreamBooth (2022)

NotebookLM (2023)

Vids (2024)

Gemini Robotics (2025)

" Attention Is All You Need "

Future of Go Summit

Generative pre-trained transformer

Google Labs

Google Pixel

Google Workspace

Robot Constitution

Category

Commons

v

t

e

Differentiable programming

Information geometry

Statistical manifold

Automatic differentiation

Neuromorphic computing

Pattern recognition

Ricci calculus

Computational learning theory

Inductive bias

IPU

TPU

VPU

Memristor

SpiNNaker

TensorFlow

PyTorch

Keras

scikit-learn

Theano

JAX

Flux.jl

MindSpore

Portals Computer programming Technology

Computer programming

Technology