PMLR Proceedings of Machine Learning Research

Volume 97JMLR MLOSS FAQ Submission Format RSS

[edit]

EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks

Mingxing Tan, Quoc Le

Proceedings of the 36th International Conference on Machine Learning, PMLR 97:6105-6114, 2019.

Abstract

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are given. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient. We demonstrate the effectiveness of this method on MobileNets and ResNet. To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called EfficientNets, which achieve much better accuracy and efficiency than previous ConvNets. In particular, our EfficientNet-B7 achieves stateof-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet (Huang et al., 2018). Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flower (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer parameters.

Cite this Paper

BibTeX

@InProceedings{pmlr-v97-tan19a, title = {{E}fficient{N}et: Rethinking Model Scaling for Convolutional Neural Networks}, author = {Tan, Mingxing and Le, Quoc}, booktitle = {Proceedings of the 36th International Conference on Machine Learning}, pages = {6105--6114}, year = {2019}, editor = {Chaudhuri, Kamalika and Salakhutdinov, Ruslan}, volume = {97}, series = {Proceedings of Machine Learning Research}, month = {09--15 Jun}, publisher = {PMLR}, pdf = {http://proceedings.mlr.press/v97/tan19a/tan19a.pdf}, url = {https://proceedings.mlr.press/v97/tan19a.html}, abstract = {Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are given. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient. We demonstrate the effectiveness of this method on MobileNets and ResNet. To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called EfficientNets, which achieve much better accuracy and efficiency than previous ConvNets. In particular, our EfficientNet-B7 achieves stateof-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet (Huang et al., 2018). Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flower (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer parameters.} }

Copy to ClipboardDownload

Endnote

%0 Conference Paper %T EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks %A Mingxing Tan %A Quoc Le %B Proceedings of the 36th International Conference on Machine Learning %C Proceedings of Machine Learning Research %D 2019 %E Kamalika Chaudhuri %E Ruslan Salakhutdinov %F pmlr-v97-tan19a %I PMLR %P 6105--6114 %U https://proceedings.mlr.press/v97/tan19a.html %V 97 %X Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are given. In this paper, we systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance. Based on this observation, we propose a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient. We demonstrate the effectiveness of this method on MobileNets and ResNet. To go even further, we use neural architecture search to design a new baseline network and scale it up to obtain a family of models, called EfficientNets, which achieve much better accuracy and efficiency than previous ConvNets. In particular, our EfficientNet-B7 achieves stateof-the-art 84.4% top-1 / 97.1% top-5 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet (Huang et al., 2018). Our EfficientNets also transfer well and achieve state-of-the-art accuracy on CIFAR-100 (91.7%), Flower (98.8%), and 3 other transfer learning datasets, with an order of magnitude fewer parameters.

Copy to ClipboardDownload

APA

Tan, M. & Le, Q.. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:6105-6114 Available from https://proceedings.mlr.press/v97/tan19a.html.

Copy to ClipboardDownload

Related Material

- Download PDF
- Code

This site last compiled Wed, 17 Nov 2021 22:46:05 +0000

Github Account