Revisiting ResNets: Improved Training and Scaling Strategies

Irwan Bello [1] **William Fedus** [1] **Xianzhi Du** [1] **Ekin D. Cubuk** [1] **Aravind Srinivas** [2] **Tsung-Yi Lin** [1]
Jonathon Shlens [1] **Barret Zoph** [1]

Abstract

Novel computer vision architectures monopolize the spotlight, but the impact of the model architecture is often conflated with simultaneous changes to training methodology and scaling strategies. Our work revisits the canonical ResNet (He et al., 2015) and studies these three aspects in an effort to disentangle them. Perhaps surprisingly, we find that training and scaling strategies may matter more than architectural changes, and further, that the resulting ResNets match recent state-of-the-art models. We show that the best performing scaling strategy depends on the training regime and offer two new scaling strategies: (1) scale model depth in regimes where overfitting can occur (width scaling is preferable otherwise); (2) increase image resolution more slowly than previously recommended (Tan & Le, 2019). Using improved training and scaling strategies, we design a family of ResNet architectures, ResNet-RS, which are 1.7x - 2.7x faster than EfficientNets on TPUs, while achieving similar accuracies on ImageNet. In a large-scale semi-supervised learning setup, ResNet-RS achieves 86.2% top-1 ImageNet accuracy, while being 4.7x faster than EfficientNet-NoisyStudent. The training techniques improve transfer performance on a suite of downstream tasks (rivaling state-of-the-art self-supervised algorithms) and extend to video classification on Kinetics-400. We recommend practitioners use these simple revised ResNets as baselines for future research.

1. Introduction

The performance of a vision model is a product of the architecture, training methods and scaling strategy. However, research often emphasizes architectural changes. Novel architectures underlie many advances, but are often simultaneously introduced with other critical – and less publicized – changes in the details of the training methodology and hyperparameters. Additionally, new architectures enhanced by modern training methods are sometimes compared to older architectures with dated training methods (e.g. ResNet-50 with ImageNet Top-1 accuracy of 76.5% (He et al., 2015)). Our work addresses these issues and empirically studies the impact of *training methods* and *scaling strategies* on the popular ResNet architecture (He et al., 2015).

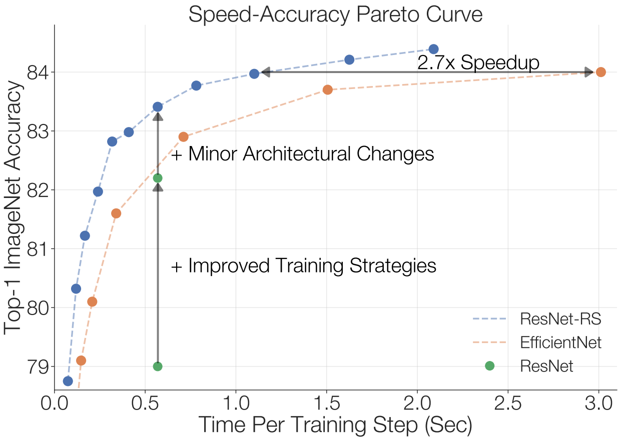

Figure 1. **Improving ResNets to state-of-the-art performance.** We improve on the canonical ResNet (He et al., 2015) with modern training methods (as also used in EfficientNets (Tan & Le, 2019)), minor architectural changes and improved scaling strategies. The resulting models, **ResNet-RS**, outperform EfficientNets on the speed-accuracy Pareto curve with speed-ups ranging from **1.7x - 2.7x** on TPUs and **2.1x - 3.3x** on GPUs. ResNet (•) is a ResNet-200 trained at 256×256 resolution. Training times reported on TPUs.

We survey the modern training and regularization techniques widely in use today and apply them to ResNets (Figure 1). In the process, we encounter interactions between

[1] Google Brain [2] UC Berkeley. Correspondence to: Irwan Bello and Barret Zoph <{ibello,barretzoph}@google.com>.

* Code and checkpoints available in TensorFlow:
https://github.com/tensorflow/models/tree/master/official/vision/beta
https://github.com/tensorflow/tpu/tree/master/models/official/resnet/resnet_rs