### EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks - [Paper](https://arxiv.org/pdf/1905.11946.pdf), [Code](https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet) (TF)

Paper Publication Date: May, 2019

Summary Notes by: [Ashvini Jindal](https://www.linkedin.com/in/ashvini-jindal-26653262/)

* **Contribution**:
  + Scaling ConvNets across each of depth, width, and resolution based on novel **Compound Scaling** technique. Existing approaches generally scale along one or more dimensions in non-systematic but not all three.
  + Proposed model is 8x smaller and 6x faster on inference, achieving the same performance as the SOTA [GPipe](https://arxiv.org/abs/1811.06965) Model.
* **Advantage**: Much smaller models with faster inference and better performance (accuracy).
* Based on NAS, obtain a family of models called EfficientNets.
* EfficientNet variants: EfficientNet-B0 (Smallest model) to EfficientNet-B7 (Largest model)
* How to find EfficientNet-B0:
  + First fix ɸ = 1 and find find 𝛼, 𝛽, and 𝛾 for EfficientNet-B0 architecture using grid search with
* Obtain EfficientNet-B1 to EfficientNet-B7:
  + Change only ɸ value with (above) fixed 𝛼, 𝛽, and 𝛾 parameters. Larger ɸ value corresponds to a model with more parameters leading to various EfficientNet variants: see equation (2) below.
* Intuitively, ɸ acts as user specified coefficient that controls how many resources are available (see equation (3) and (2) below)
* EfficientNet-B0 is simply a scaled-ConvNet where model architecture is **similar** to [**MnasNET**](https://arxiv.org/abs/1807.11626).
* Biggest model (EfficientNet-B7) has **66M** parameters compared to the SOTA model (GPipe) with **560M** parameters without any performance dip on ImageNet dataset.
* Model transfers well on other 8 transfer learning classification datasets and achieves SOTA for 5 out of 8.



