

# 推理速度提升5.1倍，參數減少88%：谷歌提出新型CNN網絡EfficientNet（附代碼）

© 2019-05-31

卷積神經網絡（CNN）通常以固定成本開發，然後再按比例放大，從而在獲得更多資源時可以達到更高的準確率。例如，ResNet ([https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg\\_edit&action=edit&type=10&appmsgid=503279373&isMul=1&token=2067326396&lang=zh\\_CN](https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg_edit&action=edit&type=10&appmsgid=503279373&isMul=1&token=2067326396&lang=zh_CN)) 可以通過增加網絡層數，從 ResNet-18 擴展到 ResNet-200。近期 GPipe 將基線 CNN 擴展了 4 倍，從而在 ImageNet 數據集上達到了 84.3% 的 top-1 準確率。模型縮放的通常做法是任意增加 CNN 的深度或寬度，或者使用更大的輸入圖像分辨率進行訓練和評估。儘管這些方法確實可以改進準確率，但它們通常需要大量手動調參，且通常獲得的是次優性能。那麼，我們是否可以尋找更好的 CNN 擴展方法，來獲得更高的準確率和效率呢？

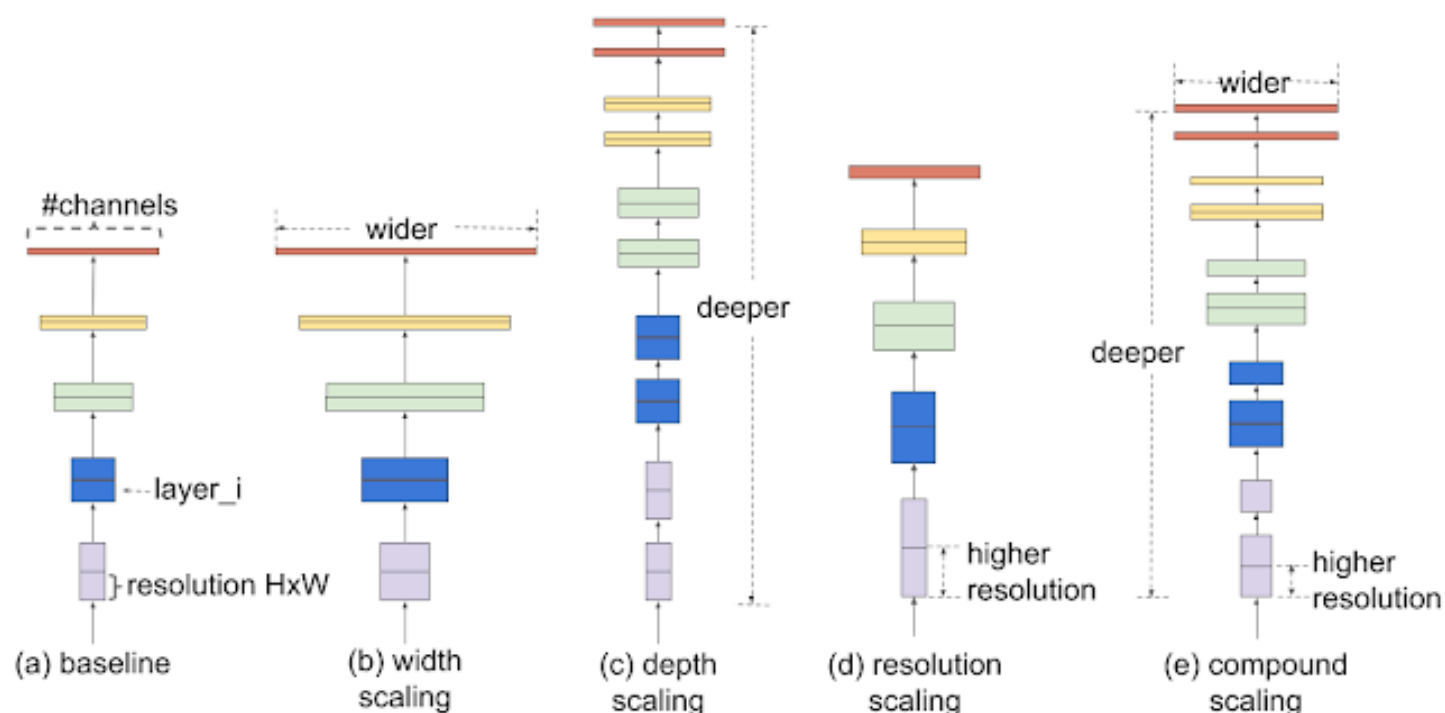
谷歌研究人員在一篇 ICML 2019 論文《EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks》中，提出了一種新型模型縮放方法，該方法使用一種簡單但高效的複合係數（compound coefficient）以更加結構化的方式擴展 CNN。與任意擴展網絡維度（如寬度、深度、分辨率）的傳統方法不同，該新方法使用固定的一組縮放係數擴展每個維度。受益於該方法和 AutoML ([https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg\\_edit&action=edit&type=10&appmsgid=503279373&isMul=1&token=2067326396&lang=zh\\_CN](https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg_edit&action=edit&type=10&appmsgid=503279373&isMul=1&token=2067326396&lang=zh_CN)) 的最新進展，谷歌開發出了一系列模型——EfficientNets，該模型的準確率超越了當前最優模型，且效率是後者的 10 倍（模型更小，速度更快）。

- 論文鏈接：<https://arxiv.org/pdf/1905.11946.pdf>

## 複合模型縮放：擴展 CNN 的更好方法

為了理解網絡縮放的效果，谷歌研究人員系統地研究了縮放模型不同維度的影響。雖然縮放單個維度可以改善模型性能，但研究人員發現平衡網絡的所有維度（寬度、深度和圖像分辨率）和可用資源才能最優地提升整體性能。

該複合縮放方法的第一步就是執行網絡搜索，尋找固定資源限制下基線模型不同縮放維度之間的關係。這決定了每個維度的恰當縮放係數。第二步是應用這些係數，將基線網絡擴展到目標模型大小或目標計算成本。



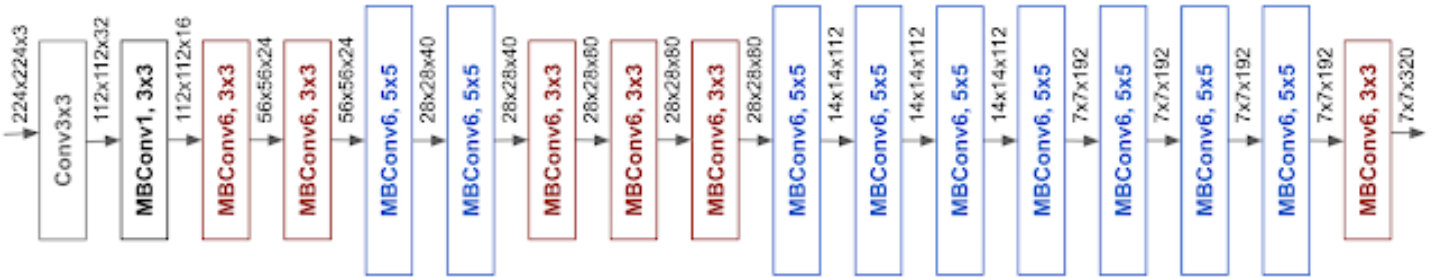
不同縮放方法對比。傳統縮放方法 (b)-(d) 任意縮放模型的單個維度，而谷歌提出的新型複合縮放方法則不同，它擴展模型的所有維度。

相比於傳統模型縮放方法，該複合縮放方法可持續改善模型的準確率和效率，如 MobileNet 的 ImageNet 準確率提升了 1.4%，ResNet 的準確率提升了 0.7%。

### EfficientNet 架構

模型縮放的效果嚴重依賴基線模型。因此，為了進一步提升性能，谷歌研究人員使用 AutoML MNAS 框架執行神經架構搜索，從而開發出一種新型基線模型，該模型可以優化準確率和效率。

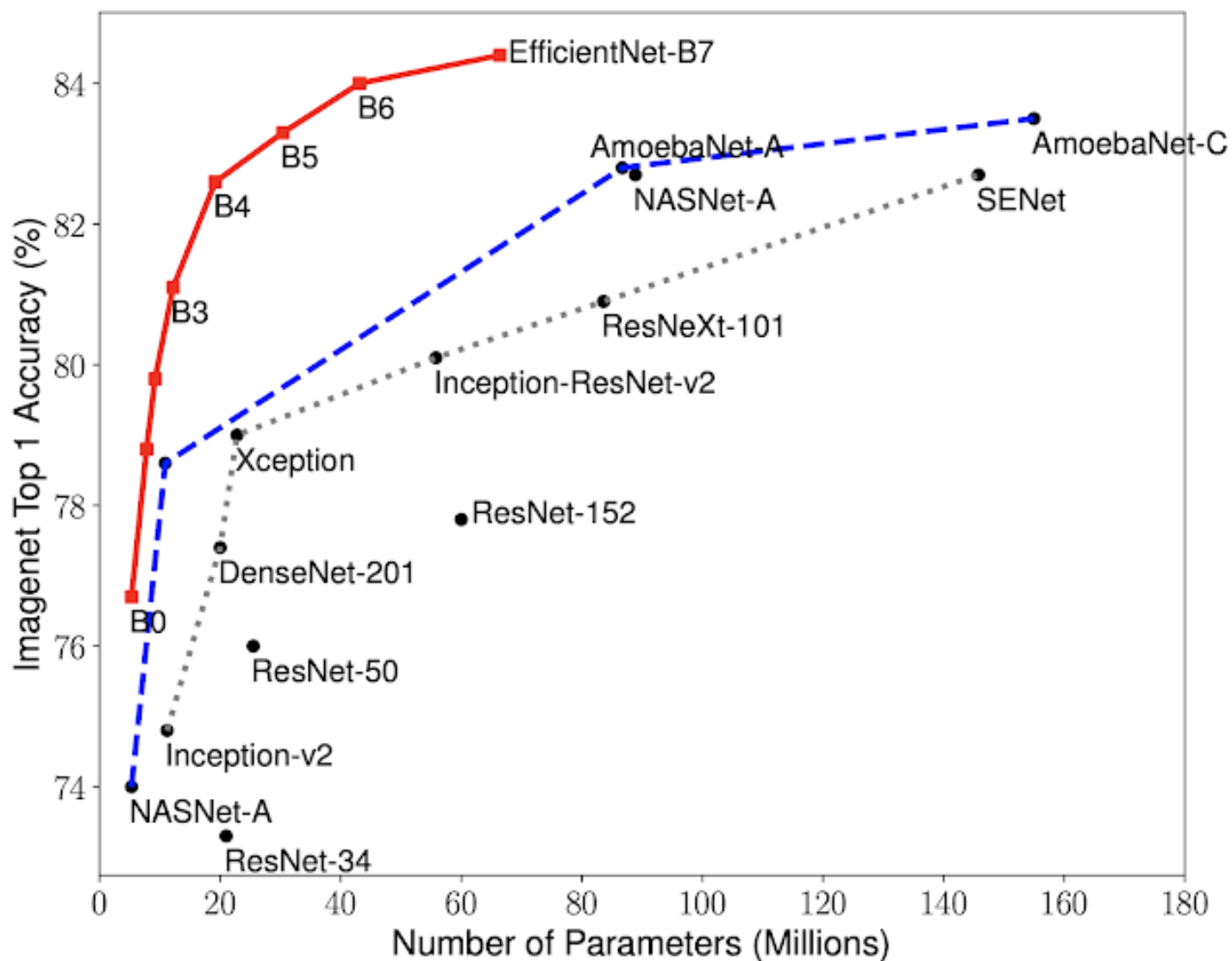
該基線模型使用 mobile inverted bottleneck convolution ( MBConv )，類似於 MobileNetV2 和 MnasNet，但是由於 FLOP 預算增加，該模型較大。於是，研究人員繼續縮放該基線模型，得到一組模型——EfficientNets。



基線模型 *EfficientNet-B0* 的架構簡單、乾淨，這使得它易於擴展和泛化。

### EfficientNet 性能

研究人員在 ImageNet 數據集上對比了 EfficientNets 和已有 CNN 模型。EfficientNet 模型要比已有 CNN 模型準確率更高、效率更高，其參數量和 FLOPS 都下降了一個數量級。例如，在高準確率的模式中，EfficientNet-B7 在 ImageNet 上獲得了當前最優的 84.4% top-1 / 97.1% top-5 準確率，CPU 推斷速度是 Gpipe 的 6.1 倍，而後者的大小是 EfficientNet-B7 的 8.4 倍。與現在廣泛使用的 ResNet-50 相比，EfficientNet-B4 使用類似的 FLOPS 取得的 top-1 準確率比 ResNet-50 高出 6.3% ( ResNet-50 76.3%，EfficientNet-B4 82.6% )。



模型大小 vs. 準確率。

EfficientNet-B0 是通過 AutoML MNAS 開發出的基線模型，Efficient-B1 到 B7 是擴展基線模型後得到的網絡。EfficientNet 顯著優於其他 CNN。具體來說，EfficientNet-B7 取得了新的 SOTA 結果：84.4% top-1 / 97.1% top-5 準確率，且其大小遠遠小於之前的最優 CNN 模型 GPipe（後者的模型大小是 EfficientNet-B7 的 8.4 倍），速度是 GPipe 的 6.1 倍。EfficientNet-B1 的參數量遠遠小於 ResNet-152，但速度是後者的 5.7 倍。

**Table 2. EfficientNet Performance Results on ImageNet (Russakovsky et al., 2015).** All EfficientNet models are scaled from our baseline EfficientNet-B0 using different compound coefficient  $\phi$  in Equation 3. ConvNets with similar top-1/top-5 accuracy are grouped together for efficiency comparison. Our scaled EfficientNet models consistently reduce parameters and FLOPS by an order of magnitude (up to 8.4x parameter reduction and up to 16x FLOPS reduction) than existing ConvNets.

Model	Top-1 Acc.	Top-5 Acc.	#Params	Ratio-to-EfficientNet	#FLOPS	Ratio-to-EfficientNet
<b>EfficientNet-B0</b>	<b>76.3%</b>	<b>93.2%</b>	<b>5.3M</b>	<b>1x</b>	<b>0.39B</b>	<b>1x</b>
ResNet-50 (He et al., 2016)	76.0%	93.0%	26M	4.9x	4.1B	11x
DenseNet-169 (Huang et al., 2017)	76.2%	93.2%	14M	2.6x	3.5B	8.9x
<b>EfficientNet-B1</b>	<b>78.8%</b>	<b>94.4%</b>	<b>7.8M</b>	<b>1x</b>	<b>0.70B</b>	<b>1x</b>
ResNet-152 (He et al., 2016)	77.8%	93.8%	60M	7.6x	11B	16x
DenseNet-264 (Huang et al., 2017)	77.9%	93.9%	34M	4.3x	6.0B	8.6x
Inception-v3 (Szegedy et al., 2016)	78.8%	94.4%	24M	3.0x	5.7B	8.1x
Xception (Chollet, 2017)	79.0%	94.5%	23M	3.0x	8.4B	12x
<b>EfficientNet-B2</b>	<b>79.8%</b>	<b>94.9%</b>	<b>9.2M</b>	<b>1x</b>	<b>1.0B</b>	<b>1x</b>
Inception-v4 (Szegedy et al., 2017)	80.0%	95.0%	48M	5.2x	13B	13x
Inception-resnet-v2 (Szegedy et al., 2017)	80.1%	95.1%	56M	6.1x	13B	13x
<b>EfficientNet-B3</b>	<b>81.1%</b>	<b>95.5%</b>	<b>12M</b>	<b>1x</b>	<b>1.8B</b>	<b>1x</b>
ResNeXt-101 (Xie et al., 2017)	80.9%	95.6%	84M	7.0x	32B	18x
PolyNet (Zhang et al., 2017)	81.3%	95.8%	92M	7.7x	35B	19x
<b>EfficientNet-B4</b>	<b>82.6%</b>	<b>96.3%</b>	<b>19M</b>	<b>1x</b>	<b>4.2B</b>	<b>1x</b>
SENet (Hu et al., 2018)	82.7%	96.2%	146M	7.7x	42B	10x
NASNet-A (Zoph et al., 2018)	82.7%	96.2%	89M	4.7x	24B	5.7x
AmoebaNet-A (Real et al., 2019)	82.8%	96.1%	87M	4.6x	23B	5.5x
PNASNet (Liu et al., 2018)	82.9%	96.2%	86M	4.5x	23B	6.0x
<b>EfficientNet-B5</b>	<b>83.3%</b>	<b>96.7%</b>	<b>30M</b>	<b>1x</b>	<b>9.9B</b>	<b>1x</b>
AmoebaNet-C (Cubuk et al., 2019)	83.5%	96.5%	155M	5.2x	41B	4.1x
<b>EfficientNet-B6</b>	<b>84.0%</b>	<b>96.9%</b>	<b>43M</b>	<b>1x</b>	<b>19B</b>	<b>1x</b>
<b>EfficientNet-B7</b>	<b>84.4%</b>	<b>97.1%</b>	<b>66M</b>	<b>1x</b>	<b>37B</b>	<b>1x</b>
GPipe (Huang et al., 2018)	84.3%	97.0%	557M	8.4x	-	-

We omit ensemble and multi-crop models (Hu et al., 2018), or models pretrained on 3.5B Instagram images (Mahajan et al., 2018).

EfficientNet 在 ImageNet 上的性能。

**Table 4. Inference Latency Comparison –** Latency is measured with batch size 1 on a single core of Intel Xeon CPU E5-2690.

Acc. @ Latency		Acc. @ Latency	
ResNet-152	77.8% @ 0.554s	GPipe	84.3% @ 19.0s
EfficientNet-B1	78.8% @ 0.098s	EfficientNet-B7	84.4% @ 3.1s
<b>Speedup</b>	<b>5.7x</b>	<b>Speedup</b>	<b>6.1x</b>

推斷延遲對比。


儘管 EfficientNets 在 ImageNet 上性能優異，但要想更加有用，它們應當具備遷移到其他數據集的能力。谷歌研究人員在 8 個常用遷移學習 (https://mp.weixin.qq.com/cgi-bin/appmsg?t=media/appmsg\_edit&action=edit&type=10&appmsgid=503279373&isMul=1&token=2067326396&lang=zh\_CN) 數據集上評估了 EfficientNets，結果表明 EfficientNets 在其中的 5 個數據集上達到了當前最優的準確率，且參數量大大減少，這表明 EfficientNets 具備良好的遷移能力。

EfficientNets 能夠顯著提升模型效率，谷歌研究人員希望 EfficientNets 能夠作為未來計算機視覺任務的新基礎。因此，研究人員開源了 EfficientNet 模型。

- EfficientNet 源代碼和 TPU 訓練腳本參見：  
https://github.com/tensorflow/tpu/tree/master/models/official/efficientnet

參考鏈接：<https://ai.googleblog.com/2019/05/efficientnet-improving-accuracy-and.html>

文章來源：機器之心 (https://www.jiqizhixin.com/articles/2019-05-30-12)

 喜歡這篇文章嗎？快分享吧！