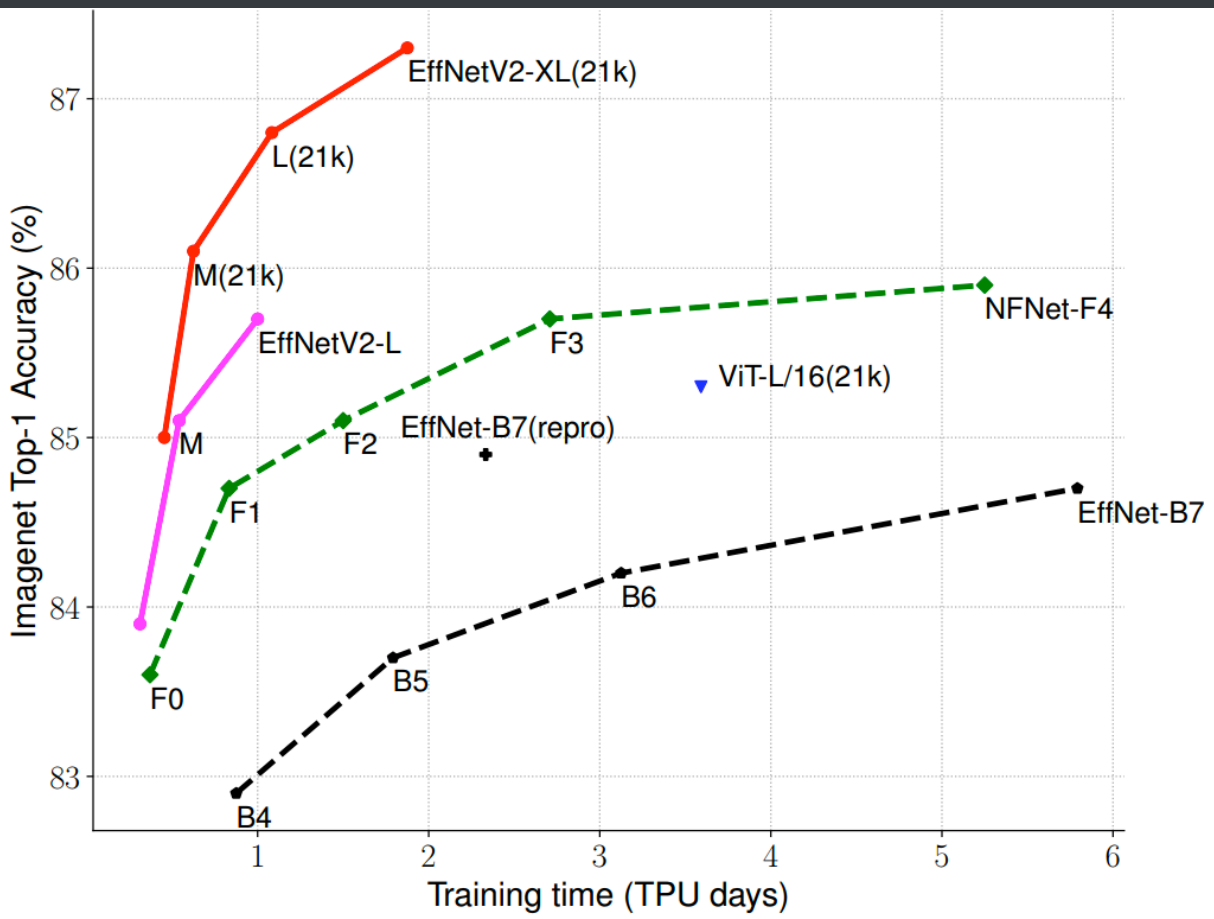
EfficientV2

- 速度：NAS(neural architecture search)、复合缩放技术
  - NAS：定义一个网络结构搜索空间，结合特定的搜索策略，找出空间内的最优解
  - 复合缩放策略：定义一个统一的缩放系数，缩放网络结构的每一个维度，得到不同大小的模型结构
- 模型：引入了MBConv(mobilenet conv,mobile inverted bottleneck block),SE
  - 基本组件
    - MBConv降低参数量以及FLOPs
    - SE：attention机制，学习每个channel的权重，并作用与输入特征
  - 变化：
    - Fused-MBConv，3x3的卷积替代1x1 pointwise conv + 3x3 depthwise conv,适当的增加一些FLOPs，换取精度的提升
    - 采用了更小的扩张比例
    - 3x3的kernel更小
    - 移除了部分层
- 训练速度
  - 训练图片的尺寸和推理图片尺寸保持一致
  - 训练图片的尺寸比先前小30%
- 训练策略
  - 渐进式学习
    - 训练初期使用小尺寸图片，结合弱正则，加速模型收敛
    - 逐渐增加输入图片的尺寸，并线性增加正则的系数（作者认为，图像尺寸的改变对模型的精度有较大的影响，原因在于不平衡的正则化因子，大图片需要更强的正则防止过拟合），比如:
      - mixup，增加两张图片混合的比例
      - dropout，增加dropout的比例
      - 随机增强的量纲增加
- 训练效率



(a) Training efficiency.

	EfficientNet (2019)	ResNet-RS (2021)	DeiT/ViT (2021)	EfficientNetV2 (ours)
Top-1 Acc.	84.3%	84.0%	83.1%	83.9%
Parameters	43M	164M	86M	24M

(b) Parameter efficiency.

- 模型表现（imageNet）

	Model	Top-1 Acc.	Params	FLOPs	Infer-time(ms)	Train-time (hours)
ConvNets & Hybrid	EfficientNet-B3 (Tan & Le, 2019a)	81.5%	12M	1.9B	19	10
	EfficientNet-B4 (Tan & Le, 2019a)	82.9%	19M	4.2B	30	21
	EfficientNet-B5 (Tan & Le, 2019a)	83.7%	30M	10B	60	43
	EfficientNet-B6 (Tan & Le, 2019a)	84.3%	43M	19B	97	75
	EfficientNet-B7 (Tan & Le, 2019a)	84.7%	66M	38B	170	139
	RegNetY-8GF (Radosavovic et al., 2020)	81.7%	39M	8B	21	-
	RegNetY-16GF (Radosavovic et al., 2020)	82.9%	84M	16B	32	-
	ResNeSt-101 (Zhang et al., 2020)	83.0%	48M	13B	31	-
	ResNeSt-200 (Zhang et al., 2020)	83.9%	70M	36B	76	-
	ResNeSt-269 (Zhang et al., 2020)	84.5%	111M	78B	160	-
	TResNet-L (Ridnik et al., 2020)	83.8%	56M	-	45	-
	TResNet-XL (Ridnik et al., 2020)	84.3%	78M	-	66	-
	EfficientNet-X (Li et al., 2021)	84.7%	73M	91B	-	-
	NFNet-F0 (Brock et al., 2021)	83.6%	72M	12B	30	8.9
	NFNet-F1 (Brock et al., 2021)	84.7%	133M	36B	70	20
	NFNet-F2 (Brock et al., 2021)	85.1%	194M	63B	124	36
	NFNet-F3 (Brock et al., 2021)	85.7%	255M	115B	203	65
	NFNet-F4 (Brock et al., 2021)	85.9%	316M	215B	309	126
	LambdaResNet-420-hybrid (Bello, 2021)	84.9%	125M	-	-	67
	BotNet-T7-hybrid (Srinivas et al., 2021)	84.7%	75M	46B	-	95
	BiT-M-R152x2 (21k) (Kolesnikov et al., 2020)	85.2%	236M	135B	500	-
Vision Transformers	ViT-B/32 (Dosovitskiy et al., 2021)	73.4%	88M	13B	13	-
	ViT-B/16 (Dosovitskiy et al., 2021)	74.9%	87M	56B	68	-
	DeiT-B (ViT+reg) (Touvron et al., 2021)	81.8%	86M	18B	19	-
	DeiT-B-384 (ViT+reg) (Touvron et al., 2021)	83.1%	86M	56B	68	-
	T2T-ViT-19 (Yuan et al., 2021)	81.4%	39M	8.4B	-	-
	T2T-ViT-24 (Yuan et al., 2021)	82.2%	64M	13B	-	-
	ViT-B/16 (21k) (Dosovitskiy et al., 2021)	84.6%	87M	56B	68	-
	ViT-L/16 (21k) (Dosovitskiy et al., 2021)	85.3%	304M	192B	195	172
ConvNets (ours)	**EfficientNetV2-S**	83.9%	22M	8.8B	24	7.1
	EfficientNetV2-M	85.1%	54M	24B	57	13
	EfficientNetV2-L	85.7%	120M	53B	98	24
	EfficientNetV2-S (21k)	84.9%	22M	8.8B	24	9.0
	EfficientNetV2-M (21k)	86.2%	54M	24B	57	15
	EfficientNetV2-L (21k)	86.8%	120M	53B	98	26
	EfficientNetV2-XL (21k)	87.3%	208M	94B	-	45

We do not include models pretrained on non-public Instagram/JFT images, or models with extra distillation or ensemble.