Vision Transformers in 2022: An Update On Tiny-imagenet

董宇坤 PB21000237

dyk2021@mail.ustc.edu.cn

2023 年 7 月 18 日

论文概要信息

- 作者:Ethan M. Huynh,University of California, San Diego
- 发表时间:21 May 2022
- 数据集:tiny-imagenet
- 论文长度: 共 6 页

0. 摘要 (Abstract)

近期的图像转换器技术有了重大进步，并在一定程度上缩小了与传统 CNN 架构之间的差距。标准的处理流程是先在像 ImageNet-21k 这样的大型数据集上进行训练，然后再在 ImageNet-1k 上进行微调。微调之后，研究者们通常会考虑在像 CIFAR-10/100 这样的小型数据集上进行迁移学习的性能，但往往会忽略 Tiny ImageNet 数据集。本文为视觉转换器在 Tiny ImageNet 上的性能提供了更新。包括了视觉转换器 (Vision Transformer,ViT)，数据高效图像转换器 (Data Effificient Image Transformer,DeiT)，图像转换器中的类别注意力 (Class Attention in Image Transformer,CaiT) 以及 Swin 转换器 (Swin Transformers)。此外,Swin 转换器以 91.35% 的验证准确率超越了当前的最优结果。

remark

迁移学习意味着训练该模型需要调用预训练模型；摘要提到了四种转换器 (transformer)，训练这四种转换器并且比较他们的性能是此次论文复现的主要目标。

1. 简介 (Introduction)

ViT 论文 (Dosovitskiy 等人,2020 年) 展示了可以将转换器应用于图像分类任务。然而,ViT 是在 JFT-300M 数据集 (Sun 等人,2017 年) 上预训练的, 这是 Google 的内部数据集, 包含了 3 亿张图片。因此, 训练效率和数据可用性的问题就显而易见了。DeiT(Touvron 等人,2020 年) 对此做出了回应, 表明了缓解 Transformer 数据匮乏性质的一种方法是通过严格的培训计划和知识蒸馏 (knowledge distillation)。因此, 可以使用 ImageNet-21k(Ridnik 等人, 2021 年) 来训练视觉转换器, 并在 ImageNet-1k(Russakovsky 等人,2014 年) 上进一步微调。像 CaiT(Touvron 等人,2021 年) 和 Swin(Liu 等人,2021b 年) 等后续的图像转换器, 都紧密遵循了 DeiT 所制定的蓝图。

除了 ImageNet-1k, 这些研究还在 CIFAR-10 和 CIFAR-100(Krizhevsky,2009 年) 上进行迁移学习测试。然而, 每一篇论文都没有包含 Tiny ImageNet(Le & Yang,2015 年)。Tiny ImageNet 是 ImageNet-1k 的一个子集, 包含了 100,000 张图片和 200 个类别, 最初在斯坦福大学的一个计算机视觉课程中被介绍出来。自从它的诞生以来, 很少有论文在他们的基准测试中使用这个数据集。

1. 简介 (续)

也就是说,Lee 等人 (2021 年) 做过一项研究, 他们提出了修改视觉转换器的方法, 以提高在 Tiny ImageNet 上从零开始训练的准确性。但实际上, 当涉及到准确性时, 迁移学习是一种更常见且更强大的技术。因此, 没有现代研究对 Tiny ImageNet 上的视觉转换器进行评估。本文将填补这个空白, 并报告使用与 DeiT 类似的训练制度的 ViT,DeiT,CaiT, 和 Swin 转换器的准确性。

2. 实验设置 (Experimental Setting)

- 所有的视觉转换器都来自于 timm 库 (Wightman,2019)
- 使用 Nvidia RTX 3070(8GB 内存) 和 8 核 CPU 训练每个模型
- 每个 epoch 训练的时间在 10 到 60 分钟之间来选择

Model	ImageNet-1k	CIFAR-100	CIFAR-10
ViT-L/16	87.08	94.04	99.38
DeiT-B/16-D	85.43	91.40	99.20
CaiT-M/36	86.05	93.10	99.40
Swin-L/4	87.15	-	-

表: Results of ViT, DeiT, CaiT, and Swin on ImageNet-1k, CIFAR-100, and CIFAR-10

2.1 数据增强 (Data Augmentation)

作者参考了 DeiT 的的相关工作, 使用了数据增强技术, 要点和参考的来源如下:

- 分别以 0.8 和 1.0 的概率使用 Mixup(Zhang 等人,2017) 和 Cutmix(Yun et al. ,2019)
- 以概率 0.25 使用 Random Erasing(Zhong 等人.,2017)
- (Touvron,2020) 在训练和测试中使用尺寸调整为 384x384 分辨率的完整图像并加以使用双三次插值算法 (bicubic interpolation)

2.2 正则化和优化器 (Regularization and Optimizer)

正则化

使用标签平滑 (Label Smoothing) 正则化方法, 参数 $\epsilon = 0.1$, 随机深度取 0.1。
训练每个模型采用 batch size=128,epochs=30。由于图像分辨率为 384x384,8GB 的显存不足以将模型和批量加载到 GPU 内存中。因此, 需要使用梯度累加 (gradient accumulation) 来训练 batch size 为 128 的模型。

优化器

选择的优化器是 AdamW, 初始学习率为 10^{-3}, 余弦衰减和权重衰减为 0.05。

3. 文章的结果 (Results)

Model	Tiny ImageNet	#params	FLOPs
ViT-L/16	86.43	304M	190.7B
CaiT-S/36	86.74	68M	48.0B
DeiT-B/16-D	87.29	87M	55.5B
Swin-L/4	**91.35**	196M	103.9B

表: Analysis of ViT, DeiT, CaiT, and Swin accuracy and training efficiency on Tiny ImageNet.The highest validation accuracy during training is reported.

- Swin 转换器以 91.35% 的验证准确率超越了当前的最优结果。将一个窗口应用于多头自注意力 (MSA) 和一个平移窗口应用于 MSA 已经被证明是有效的。
- DeiT 达到了可观的 87.29% 的准确度，同时以大幅度的优势训练得最快。知识蒸馏 (knowledge distillation) 的力量显而易见，因为它比 ViT-L 表现得更好，并且训练速度以很大的优势领先。
- CaiT-M/36 模型可能会超过 DeiT 模型。尽管参数个数和 FLOPs 的数量较少,CaiT 的训练时间却是最长的。

3(*). 复现结果

在训练精度上，我的训练结果与论文的结果高度相近，可以说是高度复现，论文的真实性与可再现性得到了基本的证实。

model	(paper)acc@1	(my)acc@1	(my)acc@5
ViT-L	86.43	**86.52**	95.82
CaiT-S36	86.74	86.66	96.61
DeiT-B distilled	87.29	82.52	94.16
Swin-L	91.35	91.32	98.04

表: 复现结果与论文对比

3(*). 复现结果

DeiT 模型的准确率不够理想，怀疑是训练进程多次中断，断断续续调出来而导致的。也有可能是，消融实验得到的一些最优参数和最优方法的选择不一定适用于所有的模型。注意到此时模型 DeiT 使用的 batch size 为 64，因此我又换了 batch size 重新实验，得到以下结果：

DeiT-B training batch size	(paper)acc@1	(my)acc@1	(my)acc@5
32	-	81.79	93.84
64	87.29	82.15	94.02
128	-	82.11	93.91

表: 复现结果与论文对比 (续)

3(*). 复现结果

Model	(my)acc@1	storage space	time per epoch
ViT-L/16	86.52	1.13GB	22:56
CaiT-S/36	86.66	781MB	23:10
DeiT-B/16-D	82.52	330MB	9:20
Swin-L/4	91.35	761MB	20:30

表: 模型性能对比

3.1 参数个数和 FLOPs 的影响并非均等

CaiT 模型显示，参数数量和 FLOPs 并不能准确反映模型的效率。尽管 CaiT-S/36 有最少的参数数量和 FLOPs，但它的工作量最低，训练速度最慢。相反，层次的数量和工作量之间存在一种趋势。嵌入的大小是另一个需要考虑的因素，但考虑到 CaiT 的嵌入尺寸较小，层次的数量似乎是主要的决定因素。

Model	#layers	Embedding Size	Throughput (images/sec)
ViT-L/16	24	1024	31.5
CaiT-S/36	36	368	24.0
DeiT-B/16-D	12	768	83.1
Swin-L/4	18	192	36.0

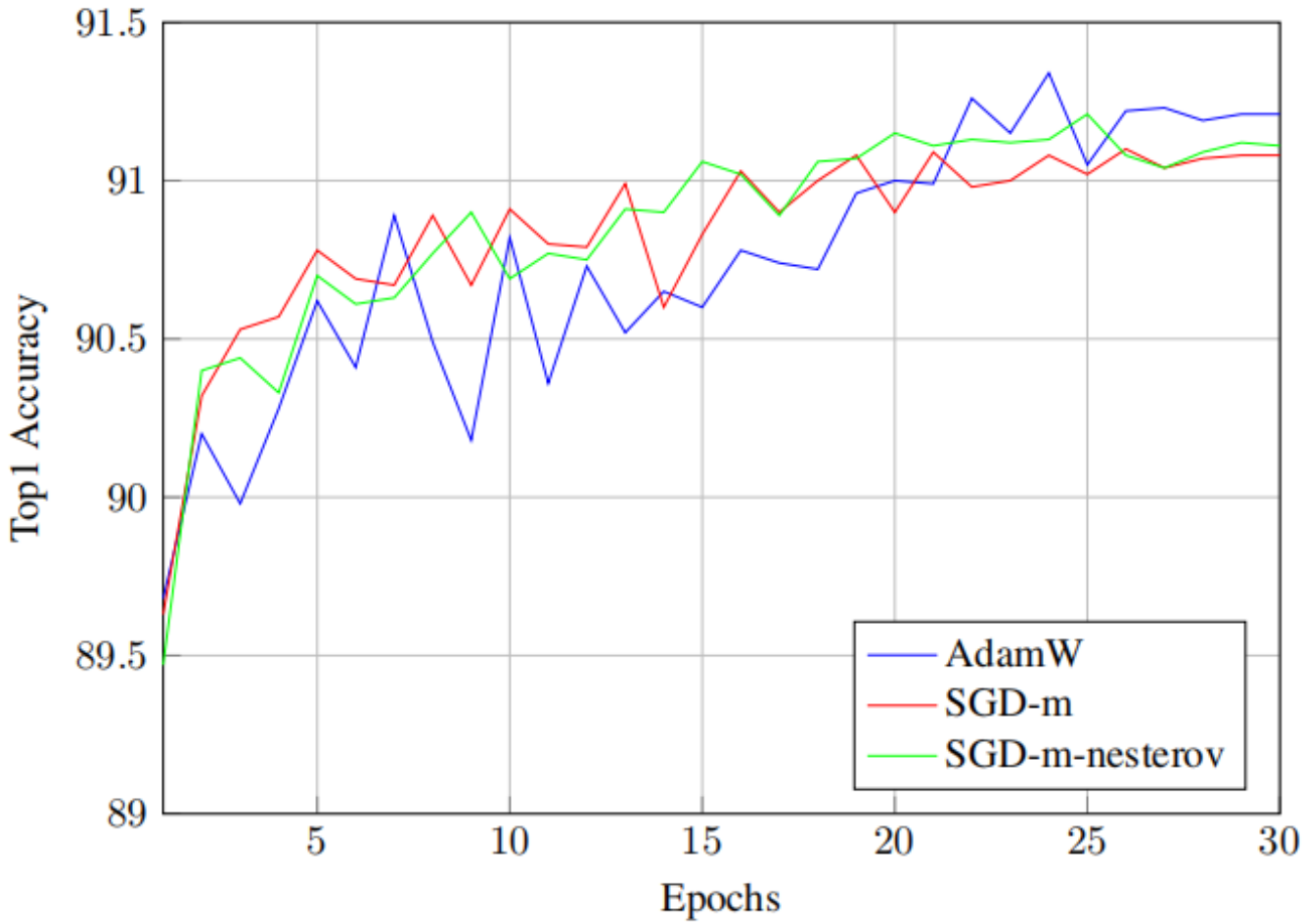
表: Comparison of various model size metrics with throughput.

4.1. 训练过程讲解之超参数

- 优化器方面: 对于 AdamW, 作者尝试了学习率和权重衰减的组合, 范围分别为 $[3.10^{-3}, 10^{-3}, 7.10^{-4}, 5.10^{-4}]$ 和 $[0.2, 0.05, 0.01]$。实验表明 10^{-3} 的学习率和 0.05 的权重衰减效果最好。
- 作者也考虑了扰动优化器, 如 SAM(Foret 等人,2020),ASAM(Kwon 等人,2021) 和 PUGD(Tseng 等人,2021)。初步测试显示,SAM,ASAM 和 PUGD 将一个 epoch 的训练时间增加了 85%, 而前五个 epoch 的准确率仅在 60-70% 左右。相比之下,AdamW 在第一个 epoch 后的准确率为 89%。因此, 作者决定不使用这些优化器进行训练。

4.1. 训练过程讲解之超参数

- 此外, 作者还测试了 SGD, 学习率为 10^{-2}, 权重衰减为 10^{-5}, 动量为 0.9, 包括有或没有 nesterov 动量两种情况。

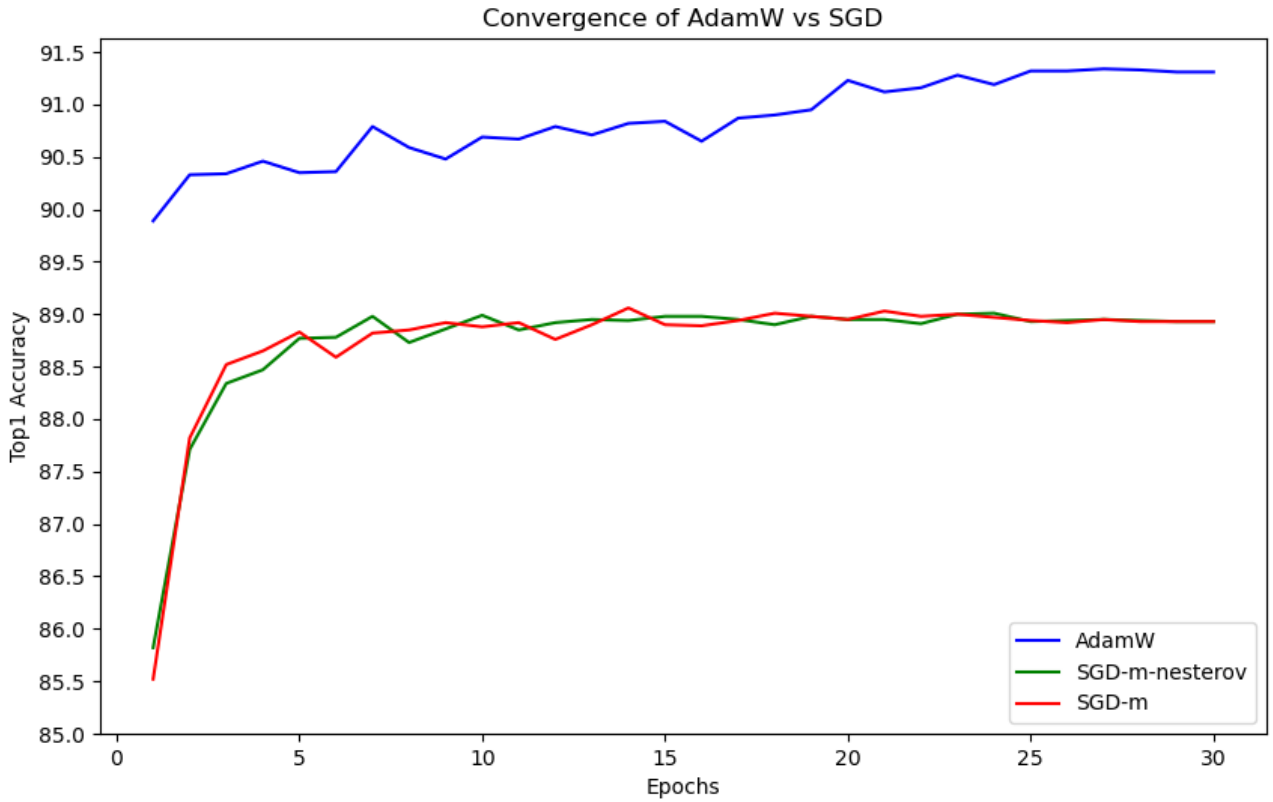


4.1. 训练过程讲解之超参数 (续)

上图显示,SGD 比 AdamW 收敛更快, 这可能是因为作者为 SGD 使用了更高的学习率。另一方面, AdamW 比 SGD 变化更大, 这可能是由于 AdamW 的自适应性质。无论如何, 就准确率而言, SGD 均不如 AdamW, 但是 nesterov 动量的表现优于普通动量 (91.21% 对比 91.1%)。

4.1(*). 复现结果

复现了 AdamW 和 SGD with nesterov momentum 以及 SGD without nesterov momentum 三种情况,acc@1 随 epoch 的变化图表如下:



4.2. 消融实验 (Ablation Study)

本节描述了移除各种数据增强和正则化技术，以及尝试不同技术对于在 Tiny ImageNet 上训练 Swin 的影响，以确定最优的训练设置。

- 默认训练设置：自动增强 (Cubuk 等人,2018)、图像裁剪和模型 EMA
- 尝试了两种裁剪方式，随机重设裁剪 (RRC) 和简单随机裁剪 (SRC) (Touvron 等人,2022)
- 模型 EMA 表明，在某些情况下，它可以提高精度，在某些情况下，它可能会降低精度。Liu 等人 (2021b) 报告说模型 EMA 没有帮助，但是仍值得在数据集级别上进行尝试。

4.2. 消融实验 (续)

表: Comparison of the effects of various data augmentation and regularization techniques. RA stands for Rand-Augment.

Change	**Accuracy**
Default	91.34
Remove RandAugment	**91.35**
Remove Rand-Erasing	91.28
Add Simple Random Crop	91.26
Remove Stochastic Depth	91.25
Remove Label Smoothing	91.25
Replace RA with AutoAugment	91.25
Remove Mixup	91.11
Add Random Resized Crop	91.06
Remove CutMix	91.04
Add Model EMA	90.83

4.2(*). 复现消融实验

表: Comparison of the effects of various data augmentation and regularization techniques. RA stands for Rand-Augment.

Change	**(my)Accuracy**	**(paper)Accuracy**
Default(no RandAugment)	91.32	91.35
Add RandAugment	91.01	91.34
Remove Rand-Erasing	91.03	91.28
Add Simple Random Crop	-	91.26
Remove Stochastic Depth	-	91.25
Remove Label Smoothing	91.20	91.25
Replace RA with AutoAugment	-	91.25
Remove Mixup	91.14	91.11
Add Random Resized Crop	-	91.06
Remove CutMix	91.28	91.04
Add Model EMA	-	90.83

4.3. 消融实验的结果 (Ablation Results)

作者详细讨论了在 Swin 模型上使用不同的数据增强和正则化技术的影响:

- RandAugment 是唯一降低训练精度的技术，作者选择移除它
- 尽管 AutoAugment 和 Model EMA 在实验中表现出精度下降，但作者认为对 AutoAugment 进行一些参数调整可能会有所帮助
- 其他的裁剪方法在实验环境中没有表现出有效性，可能是因为在预训练和微调阶段已经进行了裁剪

4.3. 消融实验的结果 (续)

作者给出了默认的训练设置配置如下:

- AdamW, 学习率为 10^{-3}, 权重衰减为 0.05
- 30 epochs, 余弦衰减, and batch size=128
- Mixup
- CutMix
- 随机深度为 0.1
- 标签平滑为 0.1

5. 结论 (Conclusion)

论文表明,视觉变换器 (vision transformers) 在 Tiny ImageNet 上的迁移性能很好, 这是合理的, 因为它是 ImageNet-1k 的一个子集。两个突出的架构是 DeiT 和 Swin。DeiT 在训练最快的同时, 报告了一项相当可观的准确性, 而 Swin 达到了最先进的准确性, 比以前提高了 0.33%。未来可以在更多的视觉变换器上进行工作。SwinV2(Liu 等人,2021a) 改进了 Swin, 并使用自我监督学习技术和后规范化等手段扩大了 Swin。另一个是 MiniViT (Zhang 等人,2022), 它使用了权重共享和权重蒸馏的组合, 极大地减少了参数数量, 并提高了视觉变换器的准确性。

Thank You!