arXiv:2305.04526v2 [cs.CV] 9 Jul 2023

CrAFT：面向高效视觉任务适配的压缩感知微调

Jung Hwan Heo[1,2] Seyedarmin Azizi[1] ArashFayyazi[1] MassoudPedram[1]南加州大学[2]NAVERCloud{johnheo,seyedarm,fayyazi,mnazemi,pedram}@usc.edu

摘要

在基础模型时代，迁移学习已成为一种广受欢迎的任务适配方法。然而，许多基础模型需要大量存储和计算资源，使得开箱即用的部署并不实际。剪枝和量化等训练后压缩技术有助于降低部署成本。不幸的是，由此造成的性能下降限制了这些技术的可用性和优势。为弥合这一性能差距，我们提出了***CrAFT***，这是一种简单的微调框架，可实现有效的训练后网络压缩。在 *CrAFT* 中，用户只需采用默认的微调计划，并结合尖锐度最小化目标，即可同时促进任务适配并提升模型对压缩的友好度。与在预训练期间应用的传统尖锐度最小化技术不同，*CrAFT* 方法带来的训练开销可以忽略不计，因为使用单个图形处理器进行微调只需不到几分钟或几小时。我们在多种目标任务上，针对基于卷积和基于注意力的视觉基础模型，验证了 *CrAFT* 的有效性。*CrAFT* 是一种通用工具，能够显著提升一次性剪枝和训练后量化的效果。代码将公开发布。

1. 引言

卷积神经网络（卷积神经网络） [19, 33, 42, 58] 和视觉变换器（视觉变换器） [7, 9, 18, 41, 60] 在图像分类 [19], 目标检测 [59], 语义分割 [26], 等多种视觉任务中取得了巨大的实证成功。凭借最先进的预训练技术，自监督方法（如 MAE） [64]以及十亿参数规模的视觉基础模型 [7]在 ImageNet-1k 上的 Top-1 准确率分别可达 90.94% 和 89.5%。然而，高昂的计算成本阻碍了神经网络在资源受限环境中的开发和部署。

通信请联系 johnheo@usc.edu

不断扩大的深度学习革命背后，模型压缩研究一直在悄然推进，其目标是在不影响性能的情况下，将大型模型压缩成更小巧的模型。研究者已开发出多种方法，例如剪枝 [14, 17, 21, 43, 67], 量化 [1, 5, 50, 68], 和知识蒸馏 [2, 24, 44, 62]。得益于这一领域的巨大进展，GPT-3 [4]和 LLaMA [61] 等大语言模型可以通过一次性剪枝（无需重新训练）压缩至原大小的 50%，且准确率损失微乎其微 [15, 57]。

这里的一个趋势是，绝大多数模型压缩研究都集中于训练后方法，而没有考虑训练过程中实现更优压缩方案的潜力。这主要是由于预训练成本极其高昂，使得研究人员难以开展此类高成本项目。目前存在少数以提高可压缩性为目标的预训练方法，但它们仅限于剪枝 [47, 54] 或量化 [38, 39],，因此不适合作为通用压缩框架。鉴于训练后压缩的简洁性和有效性，我们旨在将注意力转向相反的方向，并提出以下问题：什么样的**通用且有效的**训练方法能够最终**提升**训练后压缩的效果?

本文提出了一种简单而有效的微调框架，其灵感来源于最小化尖锐度这一特性。为实现这一目标，我们将模型压缩问题分解为可压缩性和泛化能力两个方面，并将其归结为对损失景观的统一理解。我们展示了该方法在多种无需训练的压缩方法中的有效性，包括一次性权重剪枝、INT8/INT6 量化以及仅权重量化。

- 我们对损失尖锐度与可压缩性之间的关系进行了详细分析。我们首次证明，经过尖锐度最小化的模型具有较低的内在秩和较平滑的激活范围。这促使我们在微调过程中采用尖锐度最小化目标，以实现更好的训练后压缩。
- 我们提出了**CrAFT**，这是一种压缩感知微调框架，可打造**经过压缩且适配任务的**视觉模型。由于我们的方法与

现有压缩技术正交，因此我们展示了采用 CrAFT 可显著提升准确率。

- 我们通过将该框架应用于广泛的网络骨干架构，在各种迁移学习任务（例如 Flowers、Cars 和 Food）上对视觉变换器和卷积神经网络进行了全面测试。CrAFT 在剪枝和量化方面分别可将准确率最高提升 **12%**和 **3.3%**。

2. 是什么让压缩变得困难?

预备知识。 近期的研究表明，尖锐度感知最小化（SAM）技术能够在语言建模和图像分类任务中优化泛化能力，同时提升可压缩性 [23, 49]。然而，我们对这种可压缩性的来源还不甚了解。在本节中，我们使用 Timm 库中现成的 ViT-B 模型，探究剪枝和量化面临的可压缩性挑战的来源 [63]，并观察 SAM 如何缓解这些挑战。我们从损失几何的角度统一审视这些压缩难题 $\lambda_{\max}$，从而为 CrAFT 方法提供依据。在这里，$\lambda_{\max}$ 表示损失的最坏曲率，它等价于海森矩阵的最大特征值（损失曲面的凸性）。

2.1. 剪枝难度

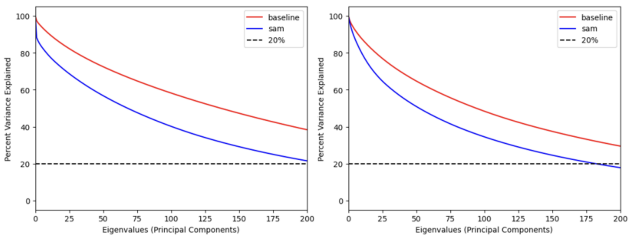


图 1。**特征值分解**。使用 SAM 优化器训练的 ViT-B 得到的权重具有更低的内在维度。对全部 12 个块取平均后可以看到，前 200 个主成分即可构成 768 个完整通道大小中 80% 或更多的权重方差（熵）；这一现象同时出现在前馈网络（左）和 QKV 注意力矩阵（右）中。

稀疏度是一种长期存在的正则化策略，旨在寻找能够在训练集之外实现良好泛化的最简单解。因此，剪枝任务可以被表述为一个约束优化问题，用于寻找参数效率更高且不会降低性能的稀疏子网络。

鉴于大多数网络都存在过度参数化，LoRA[28]使用少量参数（低秩适配器）对大型模型进行微调。类似地，我们通过主成分分析量化参数效率。

如图所示1，SAM优化器生成的权重具有更低的固有秩。具体而言，前馈网络权重的信息熵有 80% 可由 200 个主成分解释，而使用基线 SGD 优化器时，相同数量的主成分只能解释 40% 的信息。对于 QKV 注意力权重，其秩甚至更低。由此可知，显式地最小化损失景观的尖锐度，从根本上降低了剪枝的难度。例如，在一次性剪枝场景中，当稀疏度为 50% 时，与使用 ADAM 优化的基线模型相比，SAM 的准确率高出 3.6%（表1）。

表 1。**SAM 的可压缩性**。与 ADAM 基线相比，使用 SAM 预训练的 ViT 在各种模型压缩方案下始终表现出更强的鲁棒性。ADAM 还额外使用了较强的数据增强（AugReg [56]），从而在 ImageNet-1k 预训练 Top-1 准确率上取得更高性能。加速比在 NVIDIA A6000 图形处理器上测量。

ViT-B	ADAM	SAM
$\lambda_{\max}$	739	21
预训练准确率	84.58	80.28
权重剪枝（50% 稀疏度）	-4.78	**-1.11**
令牌剪枝（1.18×加速比）	-4.30	**-1.04**
令牌合并（1.90×加速比）	-4.11	**-0.95**
A8W8 均匀量化	-72.05	**-0.15**

2.2. 量化难度

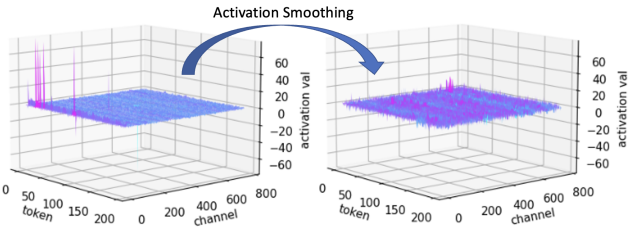


图 2. **激活平滑**。普通 ViT-B（左）存在激活异常值（幅值较大），而 ViT-B-SAM（右）的分布则更加平滑。动态范围更小时，经 SAM 训练的模型量化难度更低。

大语言模型（LLM）在某些通道中存在激活异常值，动态范围因此增大，使静态量化难以应用 [8, 66]。类似地，在图 2中，我们观察到普通 ViT-B 存在“激活异常值问题”。第一个通道的范围从 -60 到 60，而 ViT-B-SAM 的激活分布则更加平滑。在 A8W8量化场景下（表 2），SAM 能显著提升量化鲁棒性，**损失仅为 0.15%**，而 ADAM 则使其下降了 72.05%!

我们承认，ViT 也可以采用混合精度量化等针对量化的解决方案 [8] 或均衡化方法 [50, 66]。不过，我们旨在

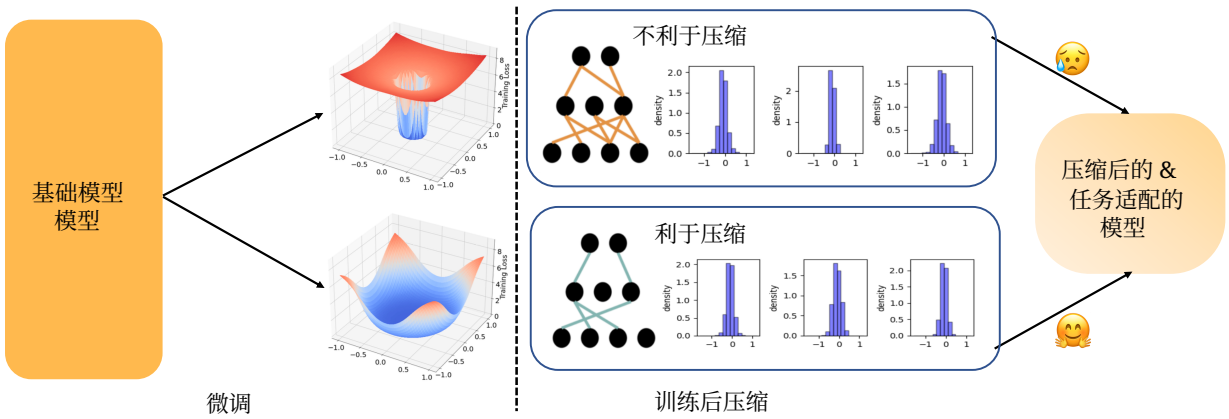


图 3. **CrAFT 是一种压缩感知微调框架**，可实现充分压缩并适配任务的视觉基础模型。通过将网络压缩问题转化为泛化问题，CrAFT 在微调期间应用尖锐度最小化，以获得损失景观更平坦（$\lambda_{\max}$ 较小）的网络。凭借更高的固有稀疏性和更平滑的激活值，它的表现优于现有训练后压缩方法，例如一次性剪枝和训练后量化。

兼顾不同压缩方案的通用性，因此我们重点将 $\lambda_{\max}$ 作为核心压缩难度指标进行最小化。

2.3. 优化难度

如前所述，尽管 SAM 能有效降低损失曲面的尖锐程度，但由于需要进行二阶计算，训练速度会显著减慢，最多可达 2× [10, 11] 。这可能会阻碍将预训练技术用于压缩，因为即使是很小的开销，也会导致大规模预训练场景的预算大幅增加。我们将其视为一种优化难度，并有意避免采用昂贵的优化策略。

总而言之，理想的网络压缩方法应当满足以下条件：(i) 能够适用于不同的网络骨干和压缩方案；(ii) 不会产生显著的训练开销，从而阻碍实际应用；(iii) 能够有效降低 $\lambda_{\max}$ ，以促进训练后压缩的便捷性。基于这些设计标准，我们现提出 CrAFT 框架。

3. 基于尖锐度最小化的微调

本节介绍尖锐度感知最小化（SAM），以及我们如何将其用于微调视觉基础模型，从而在几乎不增加训练开销的情况下，得到相对平坦的损失景观。由于这种微调与训练后压缩技术正交，因此可以显著提升准确率。总体而言，我们关注以下问题：位于更平坦极小值处的模型是否确实更容易进行压缩?

已有研究表明，SAM 不仅能在语言建模和图像分类任务中优化泛化能力，也能提升可压缩性 [23, 49]。然而，这种可压缩性的来源尚未得到充分理解。在本节中，我们将探究剪枝和量化面临的可压缩性挑战的来源，并分析 SAM 如何缓解这些挑战。我们从损失几何的视角统一这些困难 $\lambda_{\max}$，为我们提出的 CrAFT 框架奠定基础。

3.1. 尖锐度感知最小化

损失函数相对于权重的海森矩阵体现了损失曲面的凸性和尖锐度。因此，我们可以通过取海森矩阵的最大特征值来获取损失的最坏曲率，后文将其记作 $\lambda_{\max}$。

尖锐度感知最小化（SAM）并非寻找局部最小损失对应的特定位置，而是旨在找到一个训练损失始终较低的区域。该区域应同时具有较低的损失值和较低的曲率。此目标可表示为如下的极小极大优化问题：

$$\min_{w} \max_{||\epsilon||_2 \leq \rho} \mathcal{L}(w+\epsilon) \tag{1}$$

内层优化问题的目标是在半径为 ρ 的欧几里得球内寻找扰动 ϵ。其目的是通过一阶泰勒展开，在该球内寻找最优扰动，从而最大化扰动后的损失 $\mathcal{L}(w+\epsilon)$，具体如下：

$$\begin{aligned}\hat{\boldsymbol{\epsilon}} &= \arg\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\mathbf{w}+\boldsymbol{\epsilon}) \\ &\approx \arg\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(\mathbf{w}) + \boldsymbol{\epsilon}^\top \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) \\ &\approx \rho \frac{\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})}{\|\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})\|_2}.\end{aligned} \tag{2}$$

将方程 2代回方程 1后，可得到以下优化问题：

$$\min_{\mathbf{w}} \mathcal{L}(\mathbf{w} + \hat{\epsilon}) \tag{3}$$

作为最后一步，SAM 根据梯度调整模型权重：

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w})|_{\mathbf{w}+\hat{\epsilon}} \tag{4}$$

3.2. 构建压缩友好性

参考文献 [25] 提出了这样一种观点：更平坦的局部极小值可能有助于提升神经网络的泛化性能。随后，许多研究进一步探索了损失几何与泛化之间的联系；例如，尖锐度感知最小化器（SAM） [13]在训练过程中显式地最小化损失曲面的尖锐度。然而，由于需要额外的二阶计算，训练成本（和时间）可能会增加 2× [11]。

我们选择原始SAM优化器作为微调阶段尖锐度最小化方法的代表，并评估曲面曲率损失与可压缩性之间的关系。我们主要关注训练后量化 [69] 情境下的可压缩性。此外，我们还探索一次性剪枝方法，例如简单的基于幅值的剪枝 [17] 和激活感知剪枝 [57]。

4. 实验

4.1. 设置

数据集。 按照迁移学习的常用评估设置 [30], ，我们使用 ImageNet [55] 作为预训练数据集，然后将来自 Ti mm [63] 的公开预训练模型检查点迁移到 7 个下游目标分类任务，包括 Flowers [51], 、Cars [31]、Pets [53], 、Aircraft [46], 、Food [3], 、CIFAR10 [32], 和 CI- FAR 100 [32]。我们的评估大多使用 ViT-B@IMNET-21k，该模型使用 ImageNet-21k 作为预训练数据集。本文提及的所有其他模型均在 ImageNet-1k 数据集上进行预训练。

模型。 为了评估 CrAFT 是否能够在不同网络架构之间实现泛化，我们采用了基于注意力的模型和基于卷积的模型，包括 ViT-B、EfficientNet-B0、ResNet101 和 MobileNet-V2。同样，鉴于基于注意力的模型在大规模预训练中的日益普及及其有效性，大部分实验仍采用 ViT-B 进行 [9, 18]。由于基础模型架构的可扩展性是构建强大基础模型的必要条件，因此我们主要通过针对基于 Transformer 的基础模型来充分验证我们的方法。

训练细节。我们对整个网络进行微调，而不采用冻结特征提取器等传统方法。由于压缩将应用于整个网络，我们旨在通过尖锐度最小化目标更新所有参数，以提升可压缩性。我们将更高参数效率的微调（参数高效微调，PEFT）方案留待未来研究，因为本文的目标是建立一个强有力的基线。此外，与语言基础模型相比，视觉基础模型的内存消耗相对较低。例如，本文中最大的模型 ViT-B 拥有 8600 万个参数，而如今流行的语言基础模型 LLaMa [61] 最小的模型也有 70 亿个参数。

模型以 Adam 优化器作为基线，并在 CrAFT 框架中使用 SAM 优化器进行微调，训练五个周期。在单张 NVIDIA A6000 图形处理器上，批量大小为 8。针对每个数据集分别调整初始学习率，并采用余弦调度 [45] 进行学习率衰减。为简单起见，在实验部分的所有对比方法和数据集中，我们均假设批量大小为 8；CIFAR 实验除外，其批量大小设为 64。

4.2. 主要结果

在所有实验中，我们将 CrAFT 与 ADAM 微调模型进行比较。微调后，我们比较两种微调方法在量化和剪枝方面的表现。

4.2.1 训练后量化

基线方法。 作为一种强基线，我们评估了 BasePTQ，它采用对称均匀量化 [70]。更复杂的方案 PTQ4ViT [69], 通过考虑注意力模块和前馈模块中的激活分布，采用双重均匀量化；同时使用海森矩阵引导的缩放来最大限度地降低量化误差。

迁移学习结果。 与表 2一致，CrAFT 的表现始终优于 BasePTQ 和 PTQ4ViT 方法，准确率也高出许多。对于 Oxford Pets 数据集，采用 6 位仅权重量化可将准确率提升 3%，使其与全精度相比仅 $<$ 1% 的性能下降。

我们观察到，PTQ4ViT 在基于 INT 的量化中通常表现更强，此时权重和激活值都会被量化。这是因为其方法包含针对激活值的技巧，而这些技巧未必能提升仅权重量化场景下的表现。因此，BasePTQ 成为一个强有力的基线，在 Flowers-102 和斯坦福汽车等某些下游任务中表现更优。不过，PTQ4ViT 在其他领域表现更出色，这得益于其基于海森矩阵的量化尺度搜索，该方法利用了额外的二阶信息。

对于 4 位的更极端仅权重量化，CrAFT 的效果更加显著。与其 ADAM 对应方法相比，它始终能大幅提升性能。

表 2. **任务适配后的训练后量化**。 ViT-B@IMNET-21k 微调至各类迁移学习数据集后的结果。作为简单基线，我们评估了 BasePTQ，它采用对称均匀量化 [70]。更复杂的 PTQ4ViT [69] 采用双均匀量化和海森矩阵引导的缩放。我们报告绝对 Top-1 准确率（%），并在右侧括号中列出相较于全精度的准确率下降幅度。WxAy 表示量化配置，其中 x 和 y 分别表示权重和激活值的位数。与标准 ADAM 微调相比，CrAFT 能持续且往往显著提升训练后量化的效果。

下游任务	方法	FP32	W8A8	W6A6	W6A16	W5A16	W4A16
Flowers-102	BasePTQ	98.29	97.90 (-0.39)	90.11 (-8.18)	97.58 (-0.71)	94.37 (-3.92)	23.5 (-74.79)
	+ CrAFT	99.13	98.94 (-0.19)	97.02 (-2.11)	**98.80 (-0.33)**	**98.03 (-1.10)**	69.48 (-29.66)
	PTQ4ViT	98.29	98.02 (-0.27)	92.99 (-5.30)	95.85 (-2.44)	89.04 (-9.25)	66.69 (-31.60)
	+ CrAFT	99.13	**99.11 (-0.02)**	**97.35 (-1.78)**	98.39 (-0.74)	96.24 (-2.89)	**85.71 (-13.43)**
斯坦福汽车	BasePTQ	84.37	82.95 (-1.42)	63.45 (-20.92)	81.53 (-2.84)	70.39 (-13.99)	1.72 (-82.65)
	+ CrAFT	84.77	84.12 (-0.65)	71.42 (-13.35)	**82.86 (-1.91)**	**77.11 (-7.66)**	3.37 (-81.40)
	PTQ4ViT	84.37	82.95 (-1.42)	63.45 (-20.92)	81.53 (-2.84)	70.39（-13.98）	1.72 (-82.65)
	+ CrAFT	84.77	**84.47 (-0.30)**	**79.70 (-5.07)**	82.32 (-2.44)	76.93 (-7.84)	**49.55 (-35.22)**
Oxford Pets	BasePTQ	92.50	90.41 (-2.09)	73.56 (-18.94)	90.71 (-1.79)	86.29（-6.21）	10.44 (-82.06)
	+ CrAFT	93.02	91.52 (-1.49)	84.47 (-8.56)	91.82（-1.20）	**89.45（-3.57）**	37.31 (-55.71)
	PTQ4ViT	92.50	91.82（-0.68）	87.98 (-4.52)	88.31 (-4.20)	79.75 (-12.75)	46.33 (-46.17)
	+ CrAFT	93.02	**92.78 (-0.24)**	**89.99 (-3.02)**	**92.09（-0.92）**	89.23（-3.78）	**74.27 (-18.75)**
Aircraft	BasePTQ	81.21	80.35（-0.86）	69.69 (-11.51)	79.39（-1.82）	73.87（-7.34）	19.23（-61.98）
	+ CrAFT	81.24	80.29（-0.95）	74.38 (-6.86)	**79.78 (-1.46)**	75.99 (-5.24)	36.06 (-45.18)
	PTQ4ViT	81.21	80.56 (-0.65)	75.31 (-5.90)	78.76 (-2.45)	75.24 (-5.96)	57.61 (-23.60)
	+ CrAFT	81.24	**81.43 (+0.19)**	**78.33 (-2.90)**	79.54 (-1.70)	**77.47 (-3.77)**	**68.41 (-12.83)**
Food	BasePTQ	89.78	89.11 (-0.67)	73.92 (-15.86)	87.69 (-2.08)	78.71 (-11.07)	2.82 (-86.95)
	+ CrAFT	90.80	90.26 (-0.54)	81.67（-9.13）	89.24（-1.56）	84.53（-6.27）	11.09（-79.70）
	PTQ4ViT	89.78	88.89（-0.89）	83.43（-6.35）	87.91（-1.87）	80.72（-9.06）	57.90（-31.88）
	+ CrAFT	90.80	**90.45（-0.35）**	**87.35（-3.45）**	**89.48（-1.32）**	**86.08（-4.71）**	**75.57（-15.23）**
CIFAR-100	BasePTQ	90.78	89.03（-1.75）	37.74 (-53.04)	86.84（-3.94）	71.11 (-19.67)	1.72（-89.06）
	+ CrAFT	89.74	88.89 (-0.85)	78.77 (-10.97)	**88.13 (-1.61)**	81.70 (-8.04)	10.06 (-79.68)
	PTQ4ViT	90.78	89.79 (-0.99)	79.14 (-11.64)	87.04 (-3.74)	74.08 (-16.70)	32.21 (-58.57)
	+ CrAFT	89.74	**89.23 (-0.51)**	**85.40 (-4.34)**	87.47 (-2.27)	**83.99 (-5.75)**	**68.76 (-20.98)**
CIFAR-10	BasePTQ	97.36	**97.14** (-0.22)	87.78（-9.58）	94.17（-3.19）	38.20 (-59.16)	10.44（-86.92）
	+ CrAFT	97.30	97.12 **(-0.18)**	91.50 (-5.80)	95.51 (-1.79)	55.88 (-41.42)	10.02 (-87.28)
	PTQ4ViT	97.36	97.11 (-0.25)	89.72 (-7.64)	95.27（-2.09）	67.10（-30.62）	12.5（-84.86）
	+ CrAFT	97.30	97.05 (-0.25)	**94.69（-2.61）**	**96.23（-1.07）**	**79.73（-17.57）**	**23.5 (-73.80)**

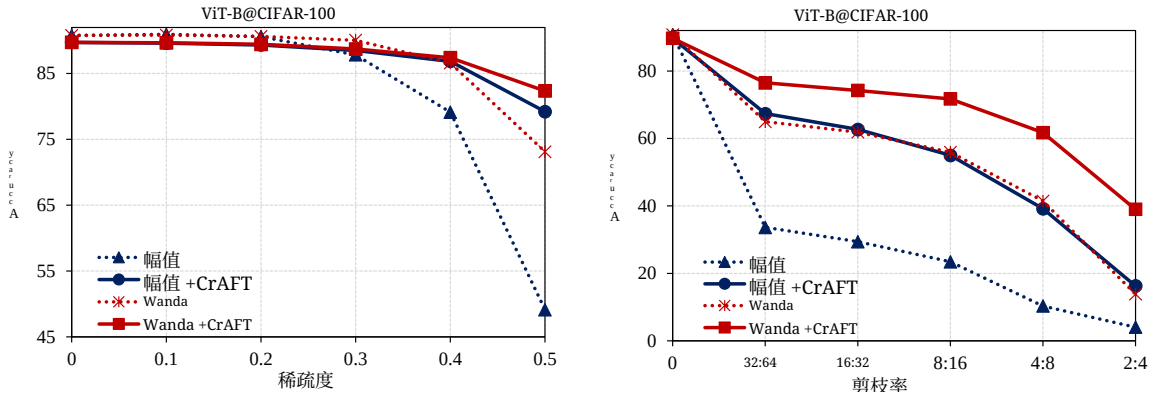
表 3. **各种基础模型**。 基于注意力和基于卷积的网络上的训练后量化结果。@ 表示其预训练数据集。所有模型均在 CIFAR-100 上进行微调。评估在 BasePTQ [70] 方法上进行。CrAFT 对压缩的友好性能够很好地推广到各种架构。

骨干网络	方法 FP32 W8A8 W6A16		
ViT-B@IMNET-21k	ADAM **90.78**54.91		79.05
	+CrAFT {		**85.76**
EfficientNet-B0@IMNET-1k	ADAM	84.75 77.32	78.39
	+CrAFT {		**82.53**
ResNet101@IMNET-1k	ADAM	83.82 73.32	82.66
	+CrAFT {		**82.96**
MobileNet-V2@IMNET-1k	ADAM	77.43 65.34	61.10
	+CrAFT {		**71.24**

差距幅度从最低的 1%（参见 Cars、BasePTQ）到最高的 46%（参见 Flowers、BasePTQ）不等。然而，我们认识到，当数据集规模较小时（例如 Flowers-102），这些指标可能不够稳定。因此，规模较大的数据集（如 Food）呈现出适中的准确率差距，而 CrAFT 的准确率则提升了 18%。

最令人惊讶的是，即使经过量化，CrAFT 也确实能够带来准确率的绝对提升。对于 Aircraft 数据集上的 W8A8 PTQ4ViT 量化，其准确率提升了约 0.2%。我们推测这可能是某种正则化效应所致。

架构泛化结果。 如表3所示，CrAFT 能够在各种基础模型架构上泛化其效果。使用 BasePTQ 进行均匀对称量化时，它在权重/激活和权重



(a) 非结构化剪枝　　(b) N:M 块剪枝

图 4. **一次性权重剪枝**。 在 CIFAR100 上对经过微调的 ViT-B@IMNET-21k 进行实验的结果。作为简单的基线方法，我们采用基于幅值的剪枝。更复杂的方法 Wanda [57] 采用激活感知剪枝，以考虑激活异常值。我们报告了在非结构化剪枝和 N:M 块剪枝（稀疏度固定为 50%）两种情况下评估得到的绝对 Top-1 准确率（%）。CrAFT 始终能够提升这两种方法的准确率，在 16:32 配置下最高提升 12% 。

仅限配置。其最显著的提升体现在 ViT-B 骨干网络上：W8A8 设置会使 ADAM 微调后的模型准确率骤降至 54%，而 CrAFT 则能将准确率维持在 81%。这一结果也与表 1 中的观察结果一致，其中 A8W8 同样导致了 -72% 的性能下降。这一现象可能归因于激活异常值问题，如图 2 所示，SAM 通过间接平滑激活表面，缩小了动态范围，从而降低了量化难度。

4.2.2 一次性剪枝

基线。 作为强有力的基线，我们评估了基于幅值的剪枝 [14, 17], ，它会从网络中移除幅值最低的权重。一种更复杂的方法是 Wanda [57],，它利用激活幅值重新缩放权重重要性。这种方法的前提是，激活值较大的通道更为重要；因此，应提高这些通道的权重重要性（即进行放大）。

非结构化剪枝结果。 如图 4a 所示，CrAFT 相较于基于幅值的方法和激活感知（Wanda）方法，均取得了稳定的性能提升，且差距不容忽视。在高稀疏度区域（>40%），这种差异尤为显著：CrAFT 相比 Wanda 和基于幅值的方法，分别提升约 9% 和惊人的 30%。

块剪枝结果。 与非结构化剪枝相比，块剪枝是一项更困难的任务，原因有二：a) 它通常要求每个块（或组）的稀疏度达到 50%，而如图 4a所示，这属于性能大幅下降区间；b) 它存在局部性约束，即在每个块内确定最不重要的权重（这可能并非全局最优解）。因此，CrAFT 的优势更加显著，如图 4b所示。各 N:M 设置的难度从左到右依次递增，因为组大小越大，局部性约束越弱。

令人惊讶的是，在基于幅值的设置中，ADAM 基线从最简单的 32:64 设置开始，准确率就显著下降，而 CrAFT 的准确率则能高出 34 个百分点。另一个有趣的观察是，Wanda 与 magnitude+CraFT 的帕累托曲线高度吻合。请注意，这两条曲线分别基于两个单独微调的网络进行剪枝，但性能似乎仍然高度一致。我们可以推测，这两种设置下的稀疏网络相似；CrAFT 带来的潜在收益，相当于 Wanda 通过激活感知的权重缩放所获得的收益。

5. 分析

本节将分析我们的方法的实用性，以及其在微调过程中最小化尖锐度的行为，以便更深入地理解其内在机制。所有实验均采用默认的 ViT-B @IMNET-21k 模型进行。

5.1. 训练开销

如第2.3节前文所述，尖锐度最小化目标（在我们的例子中为原始SAM优化器）会显著增加训练开销。鉴于这一优化挑战，我们试图量化与ADAM基线相比，使用CrAFT带来的实际墙钟时间开销。如表4所示，CrAFT使微调所需的总时长增加约 2× 。尽管这可能意味着

表 4. **端到端（E2E）微调时间分析** 迁移学习实验。布局对应于表 2，因此我们建议读者交叉参考每种微调配置的运行时间。正如预期，CrAFT 中的尖锐度最小化目标增加了约 2× 的额外开销。不过，端到端（E2E）时间仍低于在单个图形处理器上运行几分钟或几小时，对于微调视觉基础模型而言，这是可控的负载。

下游任务	方法	端到端（E2E）时间（分钟）	额外开销
Flowers-102	ADAM	1.5	1.0×
	CrAFT	3.5	2.3×
斯坦福汽车	ADAM	10.9	1.0×
	CrAFT	22	2.0×
Oxford Pets	ADAM	6.3	1.0×
	CrAFT	14	2.2×
Aircraft	ADAM	10.4	1.0×
	CrAFT	20	2.0×
Food	ADAM	85.2	1.0×
	CrAFT	172	2.0×
CIFAR	ADAM	29	1.0×
	CrAFT	57	1.9×

在大规模预训练场景中，成本相当高昂；得益于快速微调方案，我们只需在单个图形处理器上额外花费几分钟或几小时即可完成。

5.2. 微调中的尖锐度表现

表5. **损失曲面的最大曲率**（$\lambda_{\max}$）ViT-B@IMNET-21K微调前后的对比。微调前等同于预训练后收敛的基础模型。较低的 $\lambda_{\max}$ 表示损失更平坦，这与可压缩性相关。与 ADAM 相比，CrAFT 始终实现更平坦的损失，为更优异的量化和剪枝性能奠定了基础。

下游任务	微调前	ADAM	CrAFT
斯坦福汽车	125.63	103.97	**84.22**
Oxford Pets	284.22	135.6	**61.05**
Aircraft	172.72	142.93	**109.35**

可压缩性与训练模型的泛化能力相关。因此，在整个微调过程中量化这些特性至关重要。如表 5所示，相对于下游数据集，预训练基础模型所处的位置相对接近尖锐极小值。这意味着，如果直接压缩基础模型，并以零样本方式在目标任务上进行评估，性能将大幅下降。

值得庆幸的是，微调完成后，模型能够有效地找到一个更平坦的极小值，同时对训练后压缩也更加稳健。在比较 ADAM 和我们的 CrAFT 方法时，后者在 Oxford Pets 数据集上将尖锐度额外降低了最高达 2× 。这是因为尖锐度最小化目标能够帮助模型在适应目标任务的同时，明确地最小化尖锐度。

6. 相关工作

6.1. 模型压缩

6.1.1 网络剪枝

权重剪枝。 可追溯至《最优脑损伤》的研究 [35], 研究表明，神经网络包含冗余参数，剪除这些参数并不会损失其性能。此方法的迭代版本由[14, 17] 迭代幅值剪枝（IMP）方法提出，该方法通过循环重复训练和剪枝过程，以获得稀疏子网络。

为提高硬件效率，结构化剪枝 [21,22, 36, 48] 旨在以牺牲剪枝粒度为代价，换取计算规律性。我们的工作采用一次性方法对非结构化权重剪枝进行基准测试，因为这些方法无需昂贵的重新训练。

6.1.2 网络量化

训练后量化。 量化的目标是通过降低网络数值表示的位宽（包括权重、激活值和梯度），来减少内存占用和计算成本。一种简单的基线方法是线性量化，即将数值线性映射到量化桶中。训练后量化（PTQ）利用一些校准数据（例如，一小部分训练数据）来近似预训练模型的激活分布，从而确定裁剪范围和缩放因子 [16]。随后，可以根据激活值的统计信息对模型进行静态量化。更复杂的方法则采用基于熵的 [52] 方法或分段线性函数 [12]。

由于网络组件之间的量化难度可能有所不同，另一类研究关注量化粒度（逐张量、逐通道、逐令牌），或在不同层上进行多比特量化 [20, 66]。在我们的工作中，我们对 INT8、INT6 和仅权重量化进行了基准测试。

6.2. 视觉架构

基于卷积的。 随着残差网络 [19] 的出现，卷积神经网络已成为计算机视觉研究中的主流方法。此后，人们通过逐点 [29] 和逐深度 [6] 等高效卷积原语开发出了更复杂的网络。然而，这种参数效率使压缩变得困难。因此，我们选择 ResNet20 [19] 和 MobileNetV1 [27] 网络作为

首选模型，用于测试不同压缩难度下的基于卷积的架构。

基于Transformer的。视觉变换器 [9]已成为卷积神经网络的一种流行替代方案。具体而言，自注意力机制充当词元之间的消息传递模块，增强了表达能力并降低了归纳偏置。更复杂的视觉变换器，例如Swin变换器 [41], 使用分层骨干网络，在局部窗口上施加注意力。由于Transformer模型在较大的数据集上效果最佳，我们在其上开展可压缩性研究，但由于训练成本高昂，不测试架构变化。

6.3. 尖锐度感知最小化

近年来，研究人员日益关注损失景观的曲率，以及它与深度神经模型泛化能力之间的联系。Wu等人 [65]的一项重要研究表明，宽阔的损失盆地意味着更简单的解。经验观察还发现，直接针对平坦极小值进行优化，可以提升模型在各种监督学习任务（包括视觉任务）上的泛化性能。

文献 [25] 提出，较平坦的局部极小值可能有助于提升神经网络的泛化性能。此后，许多研究探索了损失几何与泛化之间的联系；例如，尖锐度感知最小化器（SAM） [13]在训练过程中显式地最小化损失曲面的尖锐度。然而，由于需要额外的二阶计算，训练成本（和时间）可能会增加 2× [11]。近期，研究人员提出了更高效的替代方案 [10,11,34,40]，以降低训练开销。我们选择原始SAM优化器作为尖锐度最小化方法的代表，评估损失曲面曲率与可压缩性之间的关系。

6.4. 压缩感知训练

由于计算复杂度高，压缩感知训练在模型压缩研究中仍是一个探索严重不足的领域。我们介绍了几项尝试从头开始（预训练）或基于预训练基础模型（微调）构建可压缩模型的研究工作。

6.4.1 预训练

剪枝。受传统剪枝方法需要进行多轮计算成本高昂的重新训练这一事实启发，一些研究尝试通过采用稀疏友好型预训练方法，以一次性 方式进行剪枝（即无需任何微调）。例如，CrAM [54] 是一种压缩感知最小化器，它利用梯度上升和近似梯度，在训练过程中针对造成最大损失的扰动进行优化。另一种方法是SFW [47], ，它使用随机通过将稀疏度表示为 K-稀疏多面体约束，采用 Frank-Wolfe 算法求解约束优化问题。

量化。SAQ [38]是一种感知尖锐度的量化方案，它认识到，由于低精度训练具有离散性，损失曲面中存在尖锐的不连续点。为抑制不稳定的梯度，该方案提出了一个 SAM风格损失项，通过平滑损失曲面来稳定训练。类似地，其他研究通过闭式归一化因子来应对训练不稳定性，例如 [37]。这两种方法都着眼于量化感知训练中的可训练性挑战，而不一定关注训练后量化模型的可压缩性。在我们的工作中，我们通过在全精度模型上采用标准 SAM 进行预训练，并应用量化感知后训练，克服了可训练性挑战。

6.4.2 微调

探索网络压缩微调的研究更是寥寥无几。与我们的工作最接近的是自然语言处理领域的研究，其中探索了 BERT 等基础模型，以观察尖锐度最小化对压缩的影响 [49]。

然而，我们的贡献使我们的工作在根本上有所不同，具体体现在：a）我们纳入了**实证分析**，以阐明为何尖锐度最小化有利于剪枝（第2.1节）和量化（第2.2节）；b）我们首次探索了视觉基础模型，这类模型具有**视觉专用压缩技术**，例如 PTQ4ViT。将这些技术与我们的 CrAFT 框架相结合时，能够取得出色的效果。

7. 结论

本文提出了 CrAFT，以弥合传统视觉基础模型的性能差距；这些模型在微调时并未考虑可压缩性。通过深入分析压缩难题及其与损失尖锐度之间的关系，我们提出在微调过程中最小化尖锐度。我们展示了将 CrAFT 与一次性剪枝、训练后量化等现有压缩技术结合所带来的显著实证优势。我们希望 CrAFT 能启发未来研究，从泛化能力（或对抗扰动的鲁棒性）的角度研究压缩。

局限性与未来工作。我们采用 SAM优化器的原始实现，以考察此类尖锐度最小化方法在模型压缩中的基线性能。由于对于更大的基础模型，训练开销很快就会变得难以承受，我们认为，CrAFT 的参数效率更高的版本（例如 LoRA [28]）可能有助于扩展基础模型。

References

[1] Ron Banner、Yury Nahshan 和 Daniel Soudry。用于快速部署的卷积网络训练后4位量化。《神经信息处理系统进展》，第32卷，2019年。1[2] Lucas Beyer、Xiaohua Zhai、Am´elie Royer、Larisa Markeeva、Rohan Anil 和 Alexander Kolesnikov。知识蒸馏：优秀的教师耐心且始终如一。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10925–10934页，2022年。1[3] Lukas Bossard、Matthieu Guillaumin 和 Luc Van Gool。Food-101——利用随机森林挖掘判别性成分。载于 David J. Fleet、Tom´as Pajdla、Bernt Schiele 和 Tinne Tuytelaars（编），《计算机视觉——ECCV 2014——第13届欧洲计算机视觉会议，瑞士苏黎世，2014年9月6日至12日，论文集》第VI部分，计算机科学讲义第8694卷，第446–461页。施普林格，2014年。4[4] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习器。《神经信息处理系统进展》，第33卷：1877–1901，2020年。1[5] Minsik Cho、Keivan A Vahid、Saurabh Adya 和 Mohammad Rastegari。DKM：用于神经网络压缩的可微 k 均值聚类层。arXiv预印本 arXiv:2108.12659，2021年。1[6] Franc¸ois Chollet。Xception：采用深度可分离卷积的深度学习。载于《IEEE计算机视觉与模式识别会议论文集》，第1251–1258页，2017年。7[7] Mostafa Dehghani、Josip Djolonga、Basil Mustafa、Piotr Padlewski、Jonathan Heek、Justin Gilmer、Andreas Steiner、Mathilde Caron、Robert Geirhos、Ibrahim Alabdulmohsin 等。将视觉变换器扩展至220亿个参数。arXiv预印本 arXiv:2302.05442，2023年。1[8] Tim Dettmers、Mike Lewis、Younes Belkada 和 Luke Zettlemoyer。LLM.int8()：面向大规模 Transformer 的8位矩阵乘法。arXiv预印本 arXiv:2208.07339，2022年。2[9] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。一张图像胜过16×16个词：用于大规模图像识别的 Transformer。arXiv预印本 arXiv:2010.11929，2020年。1、4、8[10] Jiawei Du、Zhou Daquan、Jiashi Feng、Vincent Tan 和 Joey Tianyi Zhou。无需额外开销的锐度感知训练。载于《神经信息处理系统进展》。3、8[11] Jiawei Du、Hanshu Yan、Jiashi Feng、Joey Tianyi Zhou、Liangli Zhen、Rick Siow Mong Goh 和 Vincent Tan。用于改进神经网络训练的高效尖锐度感知最小化。载于《国际学习表征会议》。3、4、8[12] Jun Fang、Ali Shafiee、Hamzah Abdel-Aziz、David Thorsley、Georgios Georgiadis 和 Joseph H Hassoun。深度神经网络的训练后分段线性量化。载于《计算机视觉–ECCV 2020：第16届欧洲会议》，英国格拉斯哥，2020年8月23–28日，会议论文集，第II部分16，第69–86页。施普林格，2020年。7[13] Pierre Foret、Ariel Kleiner、Hossein Mobahi和Behnam Neyshabur。利用尖锐度感知最小化高效提升泛化能力。arXiv预印本 arXiv:2010 01412，2020年 4，8。[14] Jonathan Frankle和Michael Carbin。彩票假设：寻找稀疏且可训练的神经网络。arXiv预印本 arXiv:1803.03635，2018年。1，6，7[15] Elias Frantar和Dan Alistarh。大规模语言模型可以通过一次性剪枝实现高精度。arXiv预印本 arXiv:2301.00774，2023年。1[16] Amir Gholami、Sehoon Kim、Zhen Dong、Zhewei Yao、Michael W. Mahoney和Kurt Keutzer。高效神经网络推理的量化方法综述。CoRR，abs/2103.13630，2021年。7[17] Song Han、Huizi Mao和William J Dally。深度压缩：通过剪枝、训练量化和霍夫曼编码压缩深度神经网络。arXiv预印本 arXiv:1510.00149，2015年。1，4，6，7[18] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar和Ross Girshick。掩码自编码器是可扩展的视觉学习器。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第16000–16009页，2022年。1，4[19] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun。用于图像识别的深度残差学习。载于《IEEE计算机视觉与模式识别会议论文集》，第770–778页，2016年。1，7[20] Yihui He、Ji Lin、Zhijian Liu、Hanrui Wang、Li-Jia Li和Song Han。AMC：面向移动设备的模型压缩与加速自动机器学习。载于《欧洲计算机视觉会议（ECCV）论文集》，第784–800页，2018年。7[21] Yihui He、Xiangyu Zhang和Jian Sun。通过通道剪枝加速超深度神经网络。载于《IEEE国际计算机视觉会议论文集》，第1389–1397页，2017年。1，7[22] Jung Hwan Heo、Arash Fayyazi、Amirhossein Esmaili和Massoud Pedram。用于降低卷积神经网络加速器延迟和功耗的稀疏周期性脉动数据流。载于《ACM/IEEE低功耗电子学与设计国际研讨会论文集》，第1–6页，2022年。7[23] Jung Hwan Heo、Arash Fayyazi、Mahdi Nazemi和Massoud Pedram。用于视觉变换器的快速免训练压缩框架。arXiv预印本 arXiv:2303.02331，2023年。2，3[24] Geoffrey Hinton、Oriol Vinyals和Jeff Dean。在神经网络中提炼知识。arXiv预印本 arXiv:1503.02531，2015年。1[25] Sepp Hochreiter和J¨urgen Schmidhuber。长短期记忆。Neural computation，9(8)：1735–1780，1997年。4，8[26] Andrew Howard、Mark Sandler、Grace Chu、Liang-Chieh Chen、Bo Chen、Mingxing Tan、Weijun Wang、Yukun Zhu，

Ruoming Pang、Vijay Vasudevan等。搜索MobileNetV3。载于《IEEE/CVF国际计算机视觉会议论文集》，第1314–1324页，2019年。1[27] Andrew G Howard、Menglong Zhu、Bo Chen、Dmitry Kalenichenko、Weijun Wang、Tobias Weyand、Marco An-dreetto和Hartwig Adam。MobileNets：面向移动视觉应用的高效卷积神经网络。arXiv预印本 arXiv:1704.04861，2017年。7[28] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang和Weizhu Chen。LoRA：大语言模型的低秩适配。arXiv预印本 arXiv:2106.09685，2021年。2，8[29] Binh-Son Hua、Minh-Khoi Tran和Sai-Kit Yeung。逐点卷积神经网络。载于《IEEE计算机视觉与模式识别会议论文集》，第984–993页，2018年。7[30] Simon Kornblith、Jonathon Shlens和Quoc V Le。更好的ImageNet模型是否具有更好的迁移能力？载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 2661 2671, 2019 4 –页。Jonathan [31] Krause、Michael Stark、Jia Deng和Li Fei-Fei。用于细粒度分类的三维对象表示。载于《2013年IEEE国际计算机视觉会议研讨会》，ICCV Workshops 2013，澳大利亚悉尼，2013年12月1–8日，第554–561页。IEEE计算机学会，2013年。4[32] A. Krizhevsky。从微小图像中学习多层特征。技术报告，多伦多大学，2009年。4[33] Alex Krizhevsky、Ilya Sutskever和Geoffrey E Hinton。使用深度卷积神经网络进行ImageNet分类。《Communications of the ACM》，60(6)：84–90，2017年。1[34] Jungmin Kwon、Jeongseop Kim、Hyunseo Park和In Kwon Choi。ASAM：用于深度神经网络尺度不变学习的自适应尖锐度感知最小化。载于Marina Meila和Tong Zhang（编），《第38届国际机器学习会议论文集》，《机器学习研究论文集》第139卷，第5905–5914页。PMLR，2021年7月18–24日。8[35] Yann LeCun、John Denker和Sara Solla。最优脑损伤。《神经信息处理系统进展》，第2卷，1989年。7[36] Ji Lin、Yongming Rao、Jiwen Lu和Jie Zhou。运行时神经网络剪枝。《神经信息处理系统进展》，第30卷，2017年。7[37] Ji Lin、Ligeng Zhu、Wei-Ming Chen、Wei-Chen Wang、Chuang Gan和Song Han。在256KB内存下进行设备端训练。arXiv预印本 arXiv:2206.15472，2022年。8[38] Ji ng Liu、Jianfei Cai和Bohan Zhuang。面向深度神经网络的尖锐度感知量化。arXiv预印本 arXiv:2111.12273，2021年。1，8[39] Ren Liu、Fengmiao Bian和Xiaoqun Zhang。利用尖锐度感知最小化进行二值量化网络训练。《Journal of Scientific Computing》，94(1)：16，2023年。1[40] Yong Liu、Siqi Mai、Xiangning Chen、Cho-Jui Hsieh和Yang You。迈向高效且可扩展的尖锐度感知最小化。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12360–12370页，2022年。8[41] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：使用移位窗口的分层视觉Transformer。载于 IEEE/CVF国际计算机视觉会议论文集，第 10012–10022 页，2021年。1, 8[42] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feicht-enhofer、Trevor Darrell 和 Saining Xie。面向 2020 年代的卷积网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 11976 11986 页，2022年 1 – 。[43] Zhuang Liu、Mingjie Sun、Tinghui Zhou、Gao Huang 和 Trevor Darrell。重新思考网络剪枝的价值。arXiv 预印本 arXiv:1810.05270，2018年。1[44] Raphael Gontijo Lopes、Stefano Fenu 和 Thad Starner。面向深度神经网络的无数据知识蒸馏。arXiv 预印本 arXiv:1710.07535，2017年。1[45] Ilya Loshchilov 和 Frank Hutter。SGDR：采用热重启的随机梯度下降。载于第 5 届国际学习表征会议（ICLR 2017），法国土伦，2017年4月24–26日，会议论文集。OpenReview.net，2017年。4[46] Subhransu Maji、Esa Rahtu、Juho Kannala、Matthew B. Blaschko 和 Andrea Vedaldi。飞机的细粒度视觉分类。CoRR，abs/1306.5151，2013年。4[47] Lu Miao、Xiaolong Luo、Tianlong Chen、Wuyang Chen、Dong Liu 和 Zhangyang Wang。通过 Frank-Wolfe 学习易于剪枝的网络：一次性、任意稀疏度且无需重新训练。载于国际学习表征会议，2021年。1, 8[48] Asit Mishra、Jorge Albericio Latorre、Jeff Pool、Darko Stosic、Dusan Stosic、Ganesh Venkatesh、Chong Yu 和 Paulius Mi- ci kevicius。加速稀疏深度神经网络。arXiv 预印本 arXiv:2104.08378，2021年。7[49] Clara Na、Sanket Vaibhav Mehta 和 Emma Strubell。先训练平坦模型，再进行压缩：尖锐度感知最小化能够学习到更易压缩的模型。arXiv 预印本 arXiv:2205 12694，2022年 2, 3, 8 。 。[50] Markus Nagel、Mart van Baalen、Tijmen Blankevoort 和 Max Welling。通过权重均衡和偏置校正实现无数据量化。载于 IEEE/CVF国际计算机视觉会议论文集，第 1325–1334 页，2019年。1, 2[51] Maria-Elena Nilsback 和 Andrew Zisserman。大规模类别的自动花卉分类。载于第六届印度计算机视觉、图形学与图像处理会议（ICVGIP 2008），印度布巴内斯瓦尔，2008年12月16–19日，第 722–729 页。IEEE计算机学会，2008年。4[52] Eunhyeok Park、Junwhan Ahn 和 Sungjoo Yoo。面向深度神经网络的加权熵量化。载于 IEEE计算机视觉与模式识别会议论文集，第 5456–5464 页，2017年。7[53] Omkar M. Parkhi、Andrea Vedaldi、Andrew Zisserman 和 C. V. Jawahar。猫与狗。载于 2012年 IEEE计算机视觉与模式识别会议，美国罗德岛州普罗维登斯，2012年6月16–21日，第 3498–3505 页。IEEE计算机学会，2012年。4

[54]Alexandra Peste、Adrian Vladu、Dan Alistarh 和 Christoph H Lampert。Cram：一种压缩感知的极小化器。arXiv预印本 arXiv:2207.14200，2022年。1，8[55] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein 等。ImageNet大规模视觉识别挑战赛。《国际计算机视觉杂志》，115：211–252，2015年。4[56] Andreas Steiner、Alexander Kolesnikov、Xiaohua Zhai、Ross Wightman、Jakob Uszkoreit 和 Lucas Beyer。如何训练视觉变换器？视觉变换器中的数据、增强与正则化。arXiv预印本 arXiv:2106.10270，2021年。2[57] Mingjie Sun、Zhuang Liu、Anna Bair 和 J Zico Kolter。一种简单有效的大语言模型剪枝方法。arXiv预印本 arXiv:2306.11695，2023年。1，4，6[58] Mingxing Tan 和 Quoc Le。EfficientNet：重新思考卷积神经网络的模型缩放。在国际机器学习会议论文集，第6105–6114页。PMLR，2019年。1[59] Mingxing Tan、Ruoming Pang 和 Quoc V Le。EfficientDet：可扩展且高效的目标检测。在IEEE/CVF计算机视觉与模式识别会议论文集，第10781–10790页，2020年。1[60] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Herv´e J´egou。通过注意力训练数据高效的图像变换器并进行蒸馏。在国际机器学习会议论文集，第10347–10357页。PMLR，2021年。1[61] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar 等。LLaMa：开放且高效的基础语言模型。arXiv预印本 arXiv:2302.13971，2023年。1，4[62] Tongzhou Wang、Jun-Yan Zhu、Antonio Torralba 和 Alexei A Efros。数据集蒸馏。arXiv预印本 arXiv:1811.10959，2018年。1[63] Ross Wightman。PyTorch图像模型。https://github.com/rwightman/pytorch-image-models，2019年。2，4[64] Mitchell Wortsman、Gabriel Ilharco、Samir Ya Gadre、Rebecca Roelofs、Raphael Gontijo-Lopes、Ari S Morcos、Hongseok Namkoong、Ali Farhadi、Yair Carmon、Simon Kornblith 等。模型汤：对多个微调模型的权重求平均，可在不增加推理时间的情况下提高准确率。在国际机器学习会议论文集，第23965–23998页。PMLR，2022年。1[65] Lei Wu、Zhanxing Zhu 和 Weinan E。迈向理解深度学习的泛化：损失景观视角。CoRR，abs/1706.10239，2017年。8[66] Guangxuan Xiao、Ji Lin、Mickael Seznec、Julien Demouth 和 Song Han。SmoothQuant：面向大语言模型的准确高效训练后量化。arXiv预印本 arXiv:2211.10438，2022年。2，7[67] Huanrui Yang、Hongxu Yin、Pavlo Molchanov、Hai Li 和 Jan Kautz。NViT：视觉变换器压缩与参数重分配。arXiv预印本 arXiv:2110.04869，2021年。1

[68] Jiwei Yang、Xu Shen、Jun Xing、Xinmei Tian、Houqiang Li、Bing Deng、Jianqiang Huang 和 Xian-sheng Hua。量化网络。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 7308–7316 页，2019年。 1[69] Zhihang Yuan、Chenhao Xue、Yiqi Chen、Qiang Wu 和 Guangyu Sun。Ptq4vit：面向视觉变换器的训练后量化框架。 *arXiv* 预印本 *arXiv:2111 12293*，2022年4，5。。[70] Shuchang Zhou、Zekun Ni、Xinyu Zhou、He Wen、Yuxin Wu 和 Yuheng Zou。DoReFa-Net：使用低位宽梯度训练低位宽卷积神经网络。 *CoRR*，abs/1606.06160，2016年。 4，5