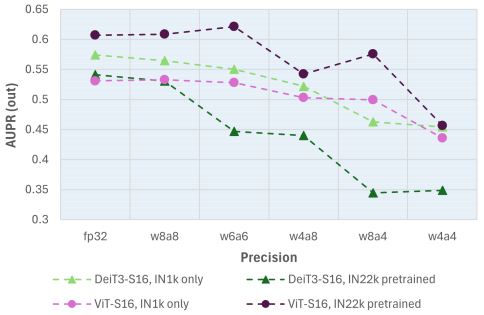
arXiv:2602.01459v1 [cs.CV] 1 Feb 2026

通过分布外检测理解视觉 Transformer 量化鲁棒性

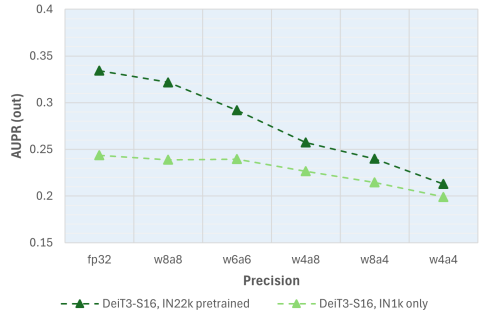
JoeyKuang，Alexander Wong 视觉与图像处理研究组，系统设计工程系，滑铁卢大学，加拿大安大略省{jjykuang，a28wong}@uwaterloo.ca

摘要

视觉 *Transformer* 在视觉任务中展现出了卓越的性能，但要实现易于使用的实时应用仍然具有挑战性。量化能够降低内存和推理成本，但可能带来性能损失。近年来，人们主要通过理解分布内（*ID*）任务行为来缓解低精度问题；然而，通过探索分布外（*OOD*）情境，注意力机制或许能为量化特性提供新的洞见。我们研究了常见分布外数据集上经过量化的热门小型视觉 *Transformer* 变体（*DeiT*、*DeiT3* 和 *ViT*）的行为。分布内分析显示，*4* 位模型最初存在不稳定性，尤其是那些在更大规模的 *ImageNet-22k* 上训练的模型。作为最强的 *FP32* 模型，*DeiT3* 因量化误差导致性能骤降 *17%*，从而成为最弱的 *4* 位模型之一。尽管 *ViT* 在分布内校准方面表现出合理的量化鲁棒性，但分布外检测揭示了更多信息：在从全精度到 *4* 位的变化中，在 *ImageNet-22k* 上预训练的 *ViT* 和 *DeiT3*，其 *AUPR-out* 的平均量化差异分别为 *15.0%* 和 *19.2%*；而仅使用 *ImageNet-1k* 的对应模型，其差异分别为 *9.5%* 和 *12.0%*。总体而言，我们的结果表明，在大规模数据集上进行预训练可能会削弱分布外检测中的低比特量化鲁棒性，而数据增强或许是更有益的选择。



（a）五个分布外数据集上的分布外精确率-召回率曲线下面积的平均值（第 2节列出了所有数据集，但不包括 ImageNet-O）。随机猜测的平均水平为 0.175。



（b）ImageNet-O 上的分布外精确率-召回率曲线下面积。随机猜测水平为 0.167。

图 1. 性能更高的模型会为分布外样本赋予更高的异常分数。关于将分布外精确率-召回率曲线下面积作为分布外检测指标的说明，请参见第 2 节。在较大数据集上预训练的模型，在分布外检测中对量化误差更为敏感。

1. 引言

近年来，视觉 Transformer（ViT）已在大规模视觉任务中表现出色，并日益被视为深度计算机视觉领域的新标准。它们依赖必然具有二次复杂度的自注意力机制 [8]，，在一个模块中采用多个注意力头，并在其大型骨干网络中堆叠多个模块。因此，ViT 并不适合边缘部署。不过，量化等模型压缩技术或许能让 ViT 应用于更强大的实时场景。

量化通过降低表示参数所需的精度来压缩模型，从而减少内存开销和计算复杂度。降低精度会引入噪声，可能导致性能变化。量化通常有两种框架：量化感知训练（QAT）和训练后量化（PTQ）。QAT 需要使用能够最小化量化模型误差的目标函数对模型进行完整的重新训练。这通常会消耗大量资源，且难以

	Top-1 准确率 (%)						NLL					
	FP32	W8A8	W6A6	W4A8	W8A4	W4A4	FP32	W8A8	W6A6	W4A8	W8A4	W4A4
DeiT3-S16，基于 IN22k 预训练	82.700	82.594	81.380	73.491	76.287	65.288	0.733	0.819	1.002	1.171	1.325	1.662
DeiT3-S16，仅使用 IN1k	81.302	81.212	80.144	74.768	74.052	66.873	0.894	0.897	0.974	1.334	1.440	1.927
ViT-S16，基于 IN22k 预训练	81.337	81.151	80.265	75.484	73.173	64.812	0.673	0.682	0.726	0.926	1.057	1.486
DeiT-S16，仅使用 IN1k	79.786	79.714	78.829	74.831	73.367	67.933	0.883	0.888	0.945	1.148	1.242	1.512

表 1. ImageNet-1k（分布内）上的量化性能。每个数据点均为三个随机种子的平均值。所有模型具有相同的架构，仅训练方法和数据集不同，但对较低精度的鲁棒性各不相同。

	ViT-S16，IN22k [25]	DeiT-S16，IN1k [26]	DeiT3-S16，IN1k [27]	Deit3-S16，IN22k [27]
批次大小	4096	1024	2048	2048
训练轮数（预训练/微调）	300/30	300	400	90/50
优化器	AdamW	AdamW	LAMB	LAMB
权重衰减	0.1	0.05	0.02	0.02
标签平滑 ϵ	0.1	0.1	✗	0.1
梯度裁剪	归一化，1.0	✗	归一化，1.0	归一化，1.0
RRC	✓	✓	✓	✗
RandAugment（$N/M/M_{std}$）	2/15/0.0	2/9/0.5	✗	✗
3 增强	✗	✗	✓	✓
LayerScale	✗	✗	1e-4	1e-4
Mixup 混合系数	0.2	0.8	0.8	✗
CutMix 混合系数	✗	1.0	1.0	1.0
随机擦除概率	✗	0.25	✗	✗
颜色抖动	✗	✗	0.3	0.3

表 2. 对所评估模型的训练流程进行总结。对于 ViT， [25]. 请注意，仅在 ImageNet-1k 上训练的模型也训练了 300 个周期，且配置取自特定的模型检查点[1]。我们展示此表，以突出训练方法之间的差异，这些差异可能有助于解释我们在量化后的分布外性能方面的观察结果。

配置 [14]。PTQ 的精度较低，但更实用，只需一个小型校准数据集即可确定量化参数 [20]。卷积神经网络（CNNs）已成功应用该领域开发的技术，但 ViT 却未能从中受益。早期探索 ViT 量化的研究发现，层归一化、Softmax 和 GeLU 激活值的分布各不相同且较为复杂（前两者属于注意力块），因此需要提出新的 PTQ 方案 [7, 16, 17, 32]。

尽管在缩小准确率方面的量化差距上已取得值得称道的进展，我们也不禁想问，对于分布外（OOD）任务是否同样如此。如果我们考虑将 ViT 用于部署，那么实时使用将使模型面临未见过的噪声输入的挑战。在全精度形式下，注意力块所产生的特征已被证明对应对分布偏移的总体鲁棒性很有价值，其原因在于缺乏归纳偏置，从而允许更丰富的表示学习 [4, 9, 19, 24]。然而，Pinto 等人 [22]讨论了在将 ViT 与卷积神经网络进行比较时，预训练的作用实际上比归纳偏置更为显著。无论如何，注意力块中的量化完整性是表面上看，这对其在任何分布上的性能都至关重要。

我们研究与分布外相关的任务能否为理解视觉Transformer在量化方面的常见鲁棒性提供新的见解。我们在六个常见的分布外数据集上测试量化模型，并评估它们区分分布外样本与分布内（ID）样本的能力。图1中的结果表明，大规模预训练会削弱低比特量化的鲁棒性，这一点在分布外检测中尤为明显。因此，我们对注意力进行可视化，以观察定量结果是否符合我们对量化后注意力完整性的预期。我们的研究结果表明，即使在较低精度下，数据增强也有益于分布外任务。

2. 方法

2.1. 研究动机

视觉Transformer架构将图像块嵌入后输入各个模块，每个模块由一个多头

[1]https : / / storage . googleapis . com / vit _ models / augreg/ [S _16-i21k-300ep-lr_0.001-aug_medium2-wd_0.1-do_0.0-sd_0.0--imagenet2012- steps_20k- lr_0.03- res_224.npz, S_16- i1k-300ep- lr_0.001- aug_medium2-wd_0.1-do_0.0-sd_0.0.npz]

自注意力（MHSA）模块和一个多层感知机模块组成。MHSA 将这些分词图像块投影为三种表征（查询、键、值），并应用 Soft-max 来表示 $[0, 1]$ 中的注意力分数。这些注意力分数旨在让模型了解相关的图像块间相关性；因此，在视觉任务中，量化时保留注意力特征至关重要。我们认为，这不仅意味着要保持对分布内样本进行分类的性能，还要保留有助于模型区分 OOD 样本与分布内样本的已学习特征。

我们希望考察量化后的视觉 Transformer 在分布内（ID）和分布外（OOD）场景中的表现。如果模型在分布偏移前后的表现一致，那么当前用于基准测试量化性能的范式很可能可以照常沿用。否则，我们或许能对如何开展视觉 Transformer 量化获得新的认识。为了缩小对模型在 OOD 场景下行为的考察范围，我们特别关注仅训练方案不同的模型；以往研究曾探讨某些训练技术对于模型鲁棒性的价值 [10, 12]。

我们首先使用分类准确率和负对数似然（NLL）对一个 ID 任务进行简要分析。这些是从业者常用于评估量化鲁棒性的指标。后者是一种适当评分，也是模型校准的有力指标，而模型校准与模型对不同分布的鲁棒性相关。鉴于我们假设注意力机制的量化可能导致学得的相关性退化，我们应当考察：当模型正确识别样本为 OOD 时，与错误地将其判定为 OOD 时相比，它区分 OOD 样本的能力如何。我们采用Hendrycks等人提出的最大Softmax概率 [11], , 计算 OOD 样本的精确率 - 召回率曲线下面积（ AUPR ）分数（将 OOD 作为“正”类，ID 作为“负”类）。下一节将详细介绍实验设置。

2.2. 设置

我们研究了采用相同架构的小型变体视觉 Transformer，这些模型要么基于 ImageNet-22k (IN22k) [6] 预训练，要么基于 ImageNet-1k (IN1k) [23]训练：ViT [25], DeiT [26], 和 DeiT3 [27]。其他量化研究也常考察这些规模的变体，而且它们在边缘设备场景中更具现实意义（尽管对边缘设备而言仍然十分庞大）。模型权重来自 Timm库[2] 和 Google Research 的云存储。

我们使用开源的 RepQ-ViT [15, 16], 对所有网络进行量化。RepQ-ViT 是一种适用于 ViT 的强大训练后量化方法。我们遵循原论文中的量化配置。使用随机种子 0、13 和 37 来生成一组稳健的结果。评估分布内性能时，我们使用 ImageNet-1k。评估分布外性能时，我们使用六个常用数据集：ImageNet-O [13], 纹理 [3], Places365 [33], SUN [30], iNaturalist[28], 和 OpenImage-O [29]。基线分布内分析报告 Top-1 准确率和负对数似然，以深入了解与任务性能和校准相关的量化误差。我们报告 AUPR-out [11]分数来评估分布外鲁棒性，并使用 ImageNet-1k 验证集作为分布内对照集。进行注意力可视化时，我们采用注意力传播，并遵循 Gildenblat 的建议，丢弃最低像素（比例为 0.9），并取各注意力头中的最大值，以获得最佳可视化效果 [1, 2]。

[2]https://github.com/huggingface/pytorch-image-models/tree/main/timm

3. 主要发现

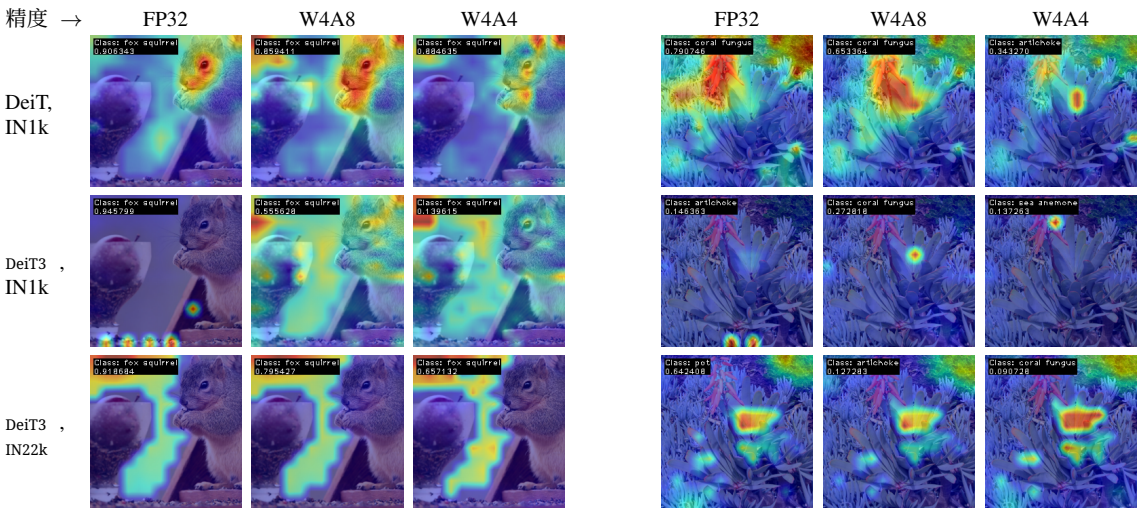
3.1. 视觉 Transformer 对 4 位量化的鲁棒性难以预测

类 ViT 模型在 RepQ-ViT 框架下对 8 位和 6 位量化具有较强的鲁棒性。一旦模型面临 4 位量化，性能最高的模型便开始出现不稳定现象。表 1 显示，FP32 下性能最强的模型 DeiT3-S16（基于 IN22k 预训练）在完全采用 4 位量化后，准确率大幅下降超过 17%。ViT-S16（基于 IN22k 预训练）的准确率下降 16.5%，而两个基于 IN1k 训练的模型在 4 位量化下更能保持分类性能（DeiT3 下降 14.4%，DeiT 下降 11.9%）。预训练规模的差异并不能解释所有结果，因为两个 DeiT3 模型的负对数似然均大幅上升。与此同时，ViT-S16 在所有精度下始终是校准效果最好的模型，而 DeiT-S 再次表现出极小的量化误差。

由于所有模型都采用相同的架构，我们可以将重点放在训练流程上。表2列出了四种模型所采用的训练方法。值得注意的是，DeiT 没有使用梯度裁剪，却大量应用了数据增强。梯度裁剪能够稳定训练中的优化过程，几乎可以确保收敛 [18, 21]。如果这种技术会导致 4 位量化稳定性较差，那就意味着模型在训练数据范围（FP32）内的最优模型参数具有极高的方差。更合理的解释是，DeiT 对数据增强的激进运用使其损失景观变得平坦； [10], 对此进行了探讨，而这一现象似乎也延续到了量化中。接下来的分析将重点考察 [25] 和 [27] 的两组模型：一组在 IN1k 上训练，另一组在 IN22k 上预训练，以进一步审视预训练规模和其他训练决策的影响。

3.2. 在 ImageNet-22k 上预训练的视觉 Transformer 在分布外场景下对量化的鲁棒性较差

现在，我们分析 ViT 和 DeiT3 在分布外场景中的表现。尽管预训练并非 ID 情况下的性能指标，但我们在分布外测试中观察到了更一致的量化稳定性模式。无论 FP32 下的能力如何，使用 IN22k 预训练的模型区分分布外样本与 ID 样本的能力都会大幅下降。见图。



（a）ID 样本（ImageNet-1k）的注意力传播图。　　（b）分布外样本（iNaturalist）的注意力传播图。

图 2. DeiT [26] 即使在较低精度下也能生成可解释的注意力图。我们观察到，与 [5] 中 DeiT3特征的发现类似的离群伪影。尽管 AUPR 对 4 位量化的鲁棒性较差，但在 IN22k 上预训练的 DeiT3-S 所生成的注意力图似乎很稳定。在分布外数据下，注意力图似乎不会发生明显变化。每张图中均标注了预测类别标签和概率。

1（a）中，两个 FP32-DeiT3 模型之间 3.3% 的差距在 4 位量化时扩大到 10.5%，且 IN1k 变体始终优于使用 IN22k 预训练的变体。我们在 ImageNet-O 上也发现了类似的稳定性表现。这一点意义重大，因为使用 IN22k 预训练的模型实际上见过这些数据（ImageNet-O 是从 IN22k 的一个子集中构建的），因此我们预期它能更好地区分这些分布外数据。表2 展示了在 IN1k 训练模型中额外采用 RRC 和 Mixup 数据增强的效果。这进一步印证了数据增强对损失景观的益处在量化设置下同样成立。

我们还发现，两种 FP32-ViT 模型之间 7.6% 的分布外精确率-召回率曲线下面积差距，在 4 位时缩小至 2.1%。在这种情况下，训练中的唯一区别是所见数据和训练总时长，这再次表明，大规模预训练可能会在低比特情形下对分布外数据产生不利影响。

3.3. 注意力图中的高范数离群令牌在较低精度下可能过于占主导地位

为了说明量化对视觉 Transformer 的影响，我们在图中展示了 DeiT 和 DeiT3 模型在不同精度下的注意力归因。2。DeiT3特征包含许多伪影斑块（出现在无关区域的热点），类似于 [5] 所讨论的高范数离群词元现象。Darcet 等人 [5] 研究了这一现象，并指出，大规模预训练、延长训练时长以及增大模型尺寸，可能会导致离群词元的形成。他们还发现，这些离群词元包含全局信息。鉴于我们使用的是 Small变体，DeiT3 的训练时长可能是模型学习到这些词元的一个因素。

有趣的是，DeiT3-S-IN22k 的注意力图在 4 位精度下似乎具有鲁棒性；而在使用 IN1k 训练的对应模型中，离群词元会在图像中“移动”，表明其对 4 位精度较为敏感。虽然这与分布内校准中的量化误差相符，但离群属性的变化并未妨碍分布外检测能力。Yellapragada等人 [31]表明，蒸馏到寄存器中的高范数词元的学习到的表征，在分布外场景中泛化能力较差。使用 IN22k 预训练的模型似乎在较低位宽下保留了这些离群词元的全局信息，但其他局部表征可能已经饱和。与此同时，IN1k 模型中的离群范数可能并不显著，因此在较低位宽下，全局信息会产生量化误差（将范围调整为较低的幅值），但其他有助于异常检测的信息得以更好地保留。

4. 结论

为提升视觉Transformer在量化情境下的性能所做的努力，并不一定能转化为分布外场景中的性能提升。我们的研究结果表明，预训练和数据增强在决定低比特量化下的模型稳定性方面发挥着不可忽视的作用。我们还初步探讨了携带全局信息的高范数词元在量化下的行为，以及这对异常检测的影响。未来的工作包括进一步研究高范数词元及其他图块词元所学习的表征在较低精度下的表现，并探讨数据增强和大规模预训练对分布外量化的价值。

参考文献

[1] 探索视觉Transformer的可解释性。3[2] Samira Abnar 和 Willem Zuidema。量化 Transformer 中的注意力流。载于计算语言学协会第58届年会论文集，第4190–4197页。计算语言学协会。3[3] Mircea Cimpoi、Subhransu Maji、Iasonas Kokkinos、Sammy Mohamed 和 Andrea Vedaldi。描述自然场景中的纹理。第3606–3613页。3[4] Luca Cultrera、Lorenzo Seidenari 和 Alberto Del Bimbo。利用视觉注意力进行分布外检测。载于2023年IEEE/CVF国际计算机视觉会议研讨会（ICCVW），第4449–4458页。IEEE。2[5] Timoth´ee Darcet、Maxime Oquab、Julien Mairal 和 Piotr Bojanowski。视觉Transformer需要寄存器。4[6] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型层次化图像数据库。载于2009年IEEE计算机视觉与模式识别会议，第248–255页。ISSN：1063-6919。3[7] Yifu Ding、Haotong Qin、Qinghua Yan、Zhenhua Chai、Junjie Liu、Xiaolin Wei 和 Xianglong Liu。迈向精准的视觉Transformer训练后量化。载于第30届ACM国际多媒体会议论文集，第5380–5388页。计算机协会。2[8] Feyza Duman Keles、Pruthuvi Mahesakya Wijewardena 和 Chinmay Hegde。论自注意力的计算复杂度。载于第34届算法学习理论国际会议论文集，第597–619页。PMLR，2023年。1[9] Stanislav Fort、Jie Ren 和 Balaji Lakshminarayanan。探索分布外检测的极限。载于神经信息处理系统进展，第7068–7081页。Curran Associates, Inc. 2[10]Jonas Geiping、Micah Goldblum、Gowthami Somepalli、Ravid Shwartz-Ziv、Tom Goldstein 和 Andrew Gordon Wilson。数据增强究竟有多大价值？关于缩放定律、不变性和隐式正则化的研究。3[11] Dan Hendrycks 和 Kevin Gimpel。用于检测神经网络中误分类样本和分布外样本的基线方法。3[12] Dan Hendrycks、Kimin Lee 和 Mantas Mazeika。使用预训练可以提升模型鲁棒性和不确定性。载于第36届国际机器学习会议论文集，第2712–2721页。PMLR。ISSN：2640-3498。3[13] Dan Hendrycks、Kevin Zhao、Steven Basart、Jacob Steinhardt 和 Dawn Song。自然对抗样本。3[14] Benoit Jacob、Skirmantas Kligys、Bo Chen、Menglong Zhu、Matthew Tang、Andrew Howard、Hartwig Adam 和 Dmitry Kalenichenko。用于高效仅整数算术推理的神经网络量化与训练，2017年。2[15]Zhikai Li、Xuewen Liu、Jing Zhang 和 Qingyi Gu。RepQuant：通过尺度重参数化实现大型Transformer模型的精准训练后量化。3[16] Zhikai Li、Junrui Xiao、Lianwei Yang 和 Qingyi Gu。RepQ-ViT：用于视觉 Transformer 训练后量化的尺度重参数化。载于《2023 IEEE/CVF 国际计算机视觉会议（ICCV）》，第 17181–17190 页。IEEE，。2, 3[17] Yang Lin、Tianyu Zhang、Peiqin Sun、Zheng Li 和 Shuchang Zhou。FQ-ViT：全量化视觉变换器的训练后量化。载于《第三十一届国际人工智能联合会议论文集》，第 1173–1179 页。国际人工智能联合会议组织。2[18] Vien V. Mai 和 Mikael Johansson。随机梯度裁剪的稳定性与收敛性：超越 Lipschitz 连续性与光滑性。3[19]Matthias Minderer、Josip Djolonga、Rob Romijnders、Frances Hubis、Xiaohua Zhai、Neil Houlsby、Dustin Tran 和 Mario Lucic。重新审视现代神经网络的校准。载于《神经信息处理系统进展》，第 15682–15694 页。Curran Associates, Inc.，2021。2[20] Markus Nagel、Marios Fournarakis、Rana Ali Amjad、Yelysei Bondarenko、Mart van Baalen 和 Tijmen Blankevoort。神经网络量化白皮书，2021。2[21] Razvan Pascanu、Tomas Mikolov 和 Yoshua Bengio。循环神经网络训练的难点。载于《第 30 届国际机器学习会议论文集》，第 1310–1318 页。PMLR。ISSN：1938-7228。3[22] Francesco Pinto、Philip Torr 和 Puneet K. Dokania。视觉 Transformer 是否总是比卷积神经网络更鲁棒？2[23] Olga Russakovsky、Jia Deng、Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein、Alexander C. Berg 和 Li Fei-Fei。ImageNet 大规模视觉识别挑战赛。3[24] Minho Sim、Jongwhoa Lee 和 Ho-Jin Choi。通过注意力掩码改进近分布外图像检测。载于《2023 IEEE 国际大数据与智能计算会议（BigComp）》，第 195–202 页。ISSN：2375-9356。2[25] Andreas Peter Steiner、Alexander Kolesnikov、Xiaohua Zhai、Ross Wightman、Jakob Uszkoreit 和 Lucas Beyer。如何训练 ViT？视觉 Transformer 中的数据、增强与正则化。2, 3[26] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Herve Jegou。通过注意力训练数据高效的图像变换器并进行蒸馏。载于《第 38 届国际机器学习会议论文集》，第 10347–10357 页。PMLR，。ISSN：2640-3498。2, 3, 4[27] Hugo Touvron、Matthieu Cord 和 Herv´e J´egou。DeiT III：ViT 的复仇，。2, 3[28] Grant Van Horn、Oisin Mac Aodha、Yang Song、Yin Cui、Chen Sun、Alex Shepard、Hartwig Adam、Pietro Perona 和 Serge Belongie。iNaturalist 物种分类与检测数据集。第 8769–8778 页。3

[29] Haoqi Wang、Zhizhong Li、Litong Feng 和 Wayne Zhang。ViM：通过虚拟 logit 匹配实现分布外检测。第 4921–4930 页。3[30] Jianxiong Xiao、James Hays、Krista A. Ehinger、Aude Oliva 和 Antonio Torralba。SUN 数据库：从修道院到动物园的大规模场景识别。载于*2010 年 IEEE 计算机学会计算机视觉与模式识别会议*，第 3485–3492 页。ISSN：1063-6919。3[31]Srikar Yellapragada、Kowshik Thopalli、Vivek Narayanaswamy、Wesam Sakla、Yang Liu、Yamen Mubarka、Dimitris Samaras 和 Jayaraman J. Thiagarajan。在视觉 Transformer 中利用寄存器实现鲁棒适配。4[32] Zhihang Yuan、Chenhao Xue、Yiqi Chen、Qiang Wu 和 Guangyu Sun。PTQ4vit：采用双均匀量化的视觉 Transformer 训练后量化。载于计算机视觉 — *ECCV 2022*，第 191–207 页。Springer Nature Switzerland。丛书名称：计算机科学讲义。2[33] Bolei Zhou、Agata Lapedriza、Aditya Khosla、Aude Oliva 和 Antonio Torralba。Places：用于场景识别的 1,000 万张图像数据库。40(6):1452–1464。3