arXiv:2302.02210v3 [cs.CV] 2 Jun 2023

低比特视觉 Transformer 的无振荡量化

Shih-Yang Liu [*1] **Zechun Liu** [*2] **Kwang-Ting Cheng** [1]

摘要

权重振荡是量化感知训练中一种不良副作用，量化权重会频繁地在两个量化等级之间跳变，导致训练不稳定，并使最终模型陷入次优状态。我们发现，可学习缩放因子这一量化中的广泛采用的事实标准设置会加剧权重振荡。本研究探讨了可学习缩放因子与量化权重振荡之间的联系，并以 ViT 为案例进行分析，阐释相关发现及解决方法。此外，我们还发现，自注意力层中查询与键 的量化权重之间存在相互依赖，使 ViT 容易发生振荡。因此，我们相应地提出了三种技术：统计权重量化（StatsQ），相比普遍采用的基于可学习缩放因子的方法，可提高量化鲁棒性；置信度引导退火（CGA），冻结置信度高 的权重并抑制振荡权重；以及查询-键 重参数化（QKR），以解决查询-键交织振荡问题，并缓解由此产生的梯度误估。大量实验表明，这些技术成功抑制了权重振荡，并在 ImageNet 上持续显著提升准确率。具体而言，我们的 2位 DeiT-T/DeiT-S/Swin-T 算法分别比此前的当前最优水平高出 9.8%/7.7%/4.64%。代码和模型见：https://github.com/nbasyl/OFQ。

[*]贡献相同 [1]香港科技大学 [2]Meta 公司旗下 Reality Labs。通讯作者：Shih-Yang Liu <sliuau@connect.ust.hk>、Zechun Liu <zechun- liu@fb.com>。

论文集 40^{th} 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年，作者所有。

1. 引言

深度神经网络在众多应用中取得了巨大成功，应用范围涵盖计算机视觉（He 等，2016；Dosovitskiy 等，2021）和自然语言处理（Vaswani 等，2017；Kenton 与 Toutanova，2019）。然而，模型规模庞大且计算耗费资源，使得在资源受限的设备上部署大型模型难以实现。在众多研究神经网络压缩与加速的方法中（Liu 等，2019；Wu 等，2019；Zhou 等人，2016），基于量化的方法凭借较高的压缩率，以及采用高效位运算带来的显著吞吐时间缩减而脱颖而出（Zhu 等，2020；Zhang 等，2018a）。

尽管量化效果显著，量化模型与其全精度对应模型之间仍存在不可忽视的精度差距。准确率下降源于多个方面，包括但不限于量化的离散性及其有限的表征能力（Liu 等人，2022；Zhang 等，2018a；Li 等人，2020；Miyashita 等，2016）、难以对不可微量化函数进行梯度近似（Gong 等，2019；Liu 等，2018），以及妨碍优化的量化振荡（Nagel 等人，2022）。最后一点相对鲜少受到探讨，也是本研究的主要关注点。

在本研究中，我们选择量化视觉变换器作为研究振荡成因的对象。先前的工作（Nagel 等人，2022）假设量化过程中缩放因子是固定的，但事实并非如此，因为大多数主流量化方法都采用可学习缩放因子，其中也包括（Nagel 等人，2022）的方法。令人意外的是，我们发现可学习缩放因子会加剧振荡，导致量化感知训练不稳定，并且往往产生次优模型。可学习缩放因子由噪声梯度更新，并决定量化阈值。因此，噪声可学习阈值与阈值附近权重之间相互交织的振荡会使训练变得不稳定。

我们对振荡权重的组成进行了可视化，并观察到它们与位置无关，这意味着在优化过程中，不存在持续停留在振荡区域内的权重。但当前优化器无法在清除权重的同时阻止新的权重进入振荡区域，

因此振荡无法消退。此外，我们发现，在朴素量化的自注意力层中，量化查询的振荡会直接影响键中权重的梯度估计，反之亦然，从而导致查询-键权重协同振荡。

为此，我们基于分析提出了三种旨在减轻量化振荡的新技术：统计权重量化（StatsQ）和置信度引导退火（CGA），用于稳定训练并最终消除权重振荡；以及查询-键重新参数化（QKR），用于解耦量化后的查询与键之间的负面相互影响。

我们证明，所提出的技术彼此互补，能够协同工作，实现无振荡的量化视觉变换器，并在 ImageNet 上持续优于此前的当前最优水平。具体而言，与此前的当前最优方法相比，我们的量化方法使 2位 DeiT-T 和 DeiT-S 的 ImageNet Top-1准确率分别提高了 9.88% 和 7.72%；我们还首次表明，3位 DeiT-T 和 DeiT-S 的准确率可与全精度模型相当。

2. 相关工作

模型量化可分为训练后量化（PTQ）（Liu 等，2021c；Nagel 等，2020；Banner 等，2019；Nagel 等，2019）和量化感知训练（QAT）（Zhou 等人，2016；Choi 等人，2018；Gong 等，2019）。一般而言，PTQ 无需重新训练，只依赖校准数据，因此量化流程比 QAT 更快；另一方面，QAT 训练时间较长，但更适合对精度敏感的场景。本文聚焦于 QAT。

以往研究尝试从不同方面改进量化感知训练。一些研究提出以非均匀量化取代均匀量化，从而提升量化表示能力，例如（Liu 等人，2022；Li 等人，2020；Zhang 等，2018a）。另一类研究则建议通过可微函数近似离散量化，以减少非可微舍入函数导致的梯度误估（Gong 等，2019；Zhou 等人，2016）。此外，还有多项研究致力于寻找最优缩放因子。例如，（Zhou 等人，2016）引入非线性函数来缩放权重，以限制其取值范围。（Choi 等人，2018）引入可训练的裁剪参数，以自动确定合适的量化尺度。最近，（Esser 等人，2020）提出了一种简单、直观且有效的方法，其核心是可学习缩放因子；该方法已广泛成为量化感知训练的事实标准方法。

最近，（Nagel 等人，2022）指出，权重振荡会严重影响量化感知训练，这主要源于深度卷积（DW-Conv）和批量归一化（BN）层。该研究以卷积神经网络（CNNs）为案例进行分析。在我们的研究中，我们发现，视觉变换器（ViT）中即使没有深度卷积和批量归一化，振荡仍然存在；这促使我们选择 ViT 作为研究案例，探究量化权重振荡。此外，以往工作（Nagel 等人，2022）假设量化阈值是静态的，却忽略了可学习缩放因子与权重振荡之间相互交织的关系。为此，我们深入研究 ViT 结构与量化权重振荡之间错综复杂的关系，并探究可学习缩放因子对加剧此类振荡的影响。

3. 预备知识

3.1. 量化感知训练

在量化感知训练（QAT）中，缩放因子对于在表示值范围和量化步长之间取得良好平衡至关重要，尤其是在低比特量化中。以往研究（Zhou 等人，2016；Choi 等人，2018；Esser 等人，2020）广泛探讨了如何获得合适的缩放因子，以弥合量化模型与其全精度对应模型之间的精度差距。在这些方法中，LSQ（Esser 等人，2020）凭借简洁性和学习缩放因子的有效性，已成为最常用的方法。因此，本文采用 Esser 等人（2020）的方法，研究可学习缩放因子对 QAT 振荡的影响。

在 LSQ（Esser 等人，2020）中，量化函数定义如下：

$$\mathbf{W}_q^{(i)} = \alpha \hat{\mathbf{W}}_q^{(i)} = \alpha \cdot \left\lfloor \mathrm{Clip}\left(\frac{\mathbf{W}^{(i)}}{\alpha}, Q_n, Q_p\right)\right\rceil \quad (1)$$

其中$\mathbf{W_q}$ 和 $\mathbf{W}$ 分别表示量化权重和实值权重，上标 i 表示第 i^{th} 个元素。 α 是可学习缩放因子，而 Q_n、Q_p表示量化范围。

为近似梯度通过不可微的舍入函数时的传播，采用直通估计器（STE）（Bengio 等人，2013）：

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(i)}} \overset{STE}{\approx} \frac{\partial \mathcal{L}}{\partial \mathbf{W}_q^{(i)}} \cdot \mathbf{1}_{Q_n \le \mathbf{W}^{(\mathrm{i})}/\alpha \le Q_p} \quad (2)$$

其中$\mathbf{1}$ 表示指示函数：当$\mathbf{W}^{(\mathrm{i})}/\alpha$ 位于裁剪范围内时输出 1，否则输出 0。

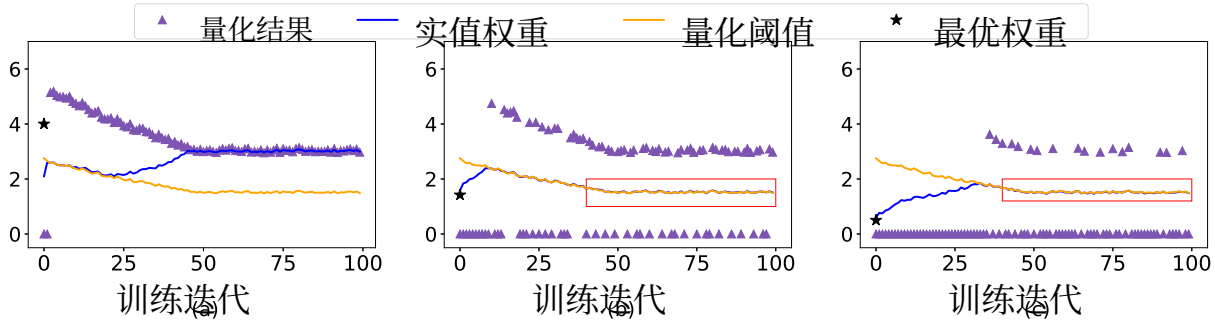


图 1：三维玩具回归示例中每个权重的振荡（即，(a) $\mathbf{W}^{(1)}$ (b) $\mathbf{W}^{(2)}$，以及 (c) $\mathbf{W}^{(3)}$）。红色方框显示，权重与可学习缩放因子相互交织，加剧了双方的振荡。

3.2. 量化 ViT 架构

视觉变换器（ViT）与卷积神经网络（CNN）的主要区别在于其变换器层结构，该结构由两个子模块组成：多头自注意力层（MHSA）和前馈网络（FFN），并大量使用全连接（FC）层。对于 FC 层中的权重和激活值量化，以及注意力层中量化后的激活值之间的乘法，我们采用按行粒度：

$$\mathbf{X}_q \cdot \mathbf{Y}_q^{\mathrm{T}} = \alpha_{\mathbf{x}_q} \alpha_{\mathbf{Y}_q} \odot (\hat{\mathbf{X}}_q \otimes \hat{\mathbf{Y}}_q^{\mathrm{T}}) \tag{3}$$

其中，$\mathbf{X}_q$ 和 $\mathbf{Y}_q$ 表示量化张量，$\otimes$ 表示整数矩阵乘法，$\odot$ 表示高精度标量-张量乘法。请注意，缩放因子应与乘法方向对齐，以使公式 3 成立，并便于实现低成本整数矩阵乘法加速（Esser 等人，2020；Xiao 等，2022）。例如，当输入为$\mathbf{X}_q \in \mathbb{R}^{N \times D_1}$和$\mathbf{Y}_q \in \mathbb{R}^{D_2 \times D_1}$时，我们有缩放因子向量$\alpha_{\mathbf{x}_q} \in \mathbb{R}^N$ 和 $\alpha_{\mathbf{Y}_q} \in \mathbb{R}^{D_2}$。我们将此称为沿最后一个维度量化，即缩放因子沿最后一个维度共享。具体而言，在视觉变换器中，权重矩阵和激活矩阵均沿最后一个维度量化，唯独与注意力矩阵$\mathbf{Attn} \in \mathbb{R}^{N \times N}$相乘的值 矩阵$\mathbf{V} \in \mathbb{R}^{N \times D}$ 除外；其中$\mathbf{V}$ 沿序列长度维度N 量化，并具有缩放因子向量 $\alpha_{\mathbf{V}_q} \in \mathbb{R}^D$。

4. QAT 中的振荡

4.1. 振荡与可学习缩放因子

尽管可学习缩放因子在量化感知训练中十分常见，但其对训练稳定性的负面影响却鲜少得到研究。我们通过提出一个包含三个权重 $\mathbf{W} = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}\}$ 的简化示例来说明这些问题，并根据公式 1，使用一个缩放因子α 来确定这三个权重的量化阈值。优化目标是一个三维回归问题，其目标是最小化目标最优浮点权重 $\mathbf{W}_* \in \mathbb{R}^3$ 与量化权重 $\mathbf{W}_q$在对数据向量 $\mathbf{X} \in \mathbb{R}^3$加权时的 $l2$-损失：

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = \mathbb{E}_{\mathbf{X} \sim U} \left[\frac{1}{2} (\mathbf{X}\mathbf{W}_* - \mathbf{X}\mathbf{W}_q)^2 \right]. \tag{4}$$

这里$\mathbf{X}$ 是从区间 $[0, 1)$ 上的均匀分布中采样得到的 U。我们按照公式 2 来优化权重。

我们发现，由量化的离散特性引起的权重振荡，会因可学习缩放因子的存在而大幅加剧。在没有可学习缩放因子的情况下，距离最优值较远的权重往往会朝目标方向更新更多，而接近最优值的权重则保持稳定。然而，我们发现存在可学习缩放因子时，情况并非如此。

如图 1（b）和（c）所示，初始化时接近目标值的权重（$\mathbf{W}^{(2)}$ 和$\mathbf{W}^{(3)}$）会受到可学习缩放因子的影响，最终在量化阈值附近振荡。请注意，缩放因子根据公式 1 决定阈值。这一现象的原因是存在一个离群权重（$\mathbf{W}^{(1)}$），其初始化值远离最优值（$\mathbf{W}_*^{(1)}$）。该离群权重对可学习缩放因子的梯度贡献更大，促使缩放因子朝着适合该离群权重的最优值移动，进而导致另外两个初始值接近各自最优值的权重发生振荡。这一观察结果与 LSQ 的梯度推导（Esser 等人，2020）一致：位于 $[Q_n, Q_p]$ 范围内的权重对缩放因子梯度的影响有限，其取值被限制在 -0.5 到 0.5 之间。另一方面，超出该范围的权重对缩放因子梯度的贡献要大得多，例如，大于Q_p 的缩放后权重会对缩放因子的梯度贡献 Q_p。

更为严重的是，一旦权重与可学习阈值交织在一起（如图中红色框所示）

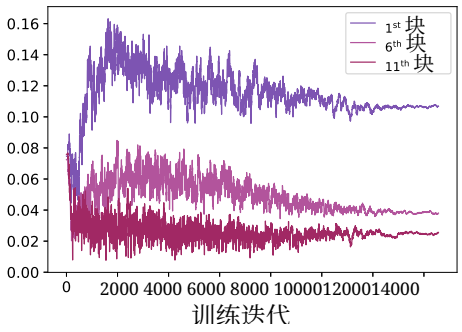


图 2：2位 DeiT-T 中来自 $1^{st}/6^{th}/11^{th}$ Transformer 块的三个可学习缩放因子 α的变化轨迹。纵轴表示 α 的值。可学习缩放因子的波动在整个训练过程中持续存在。

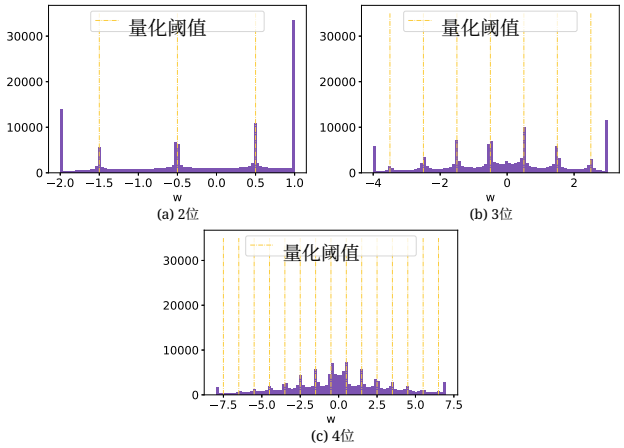


图 3：8 位 Transformer块中前馈网络（FFN）的第 1 个全连接层内权重分布的直方图：(a) 2位、(b) 3位和 (c) 4位量化 DeiT-T。

图 1 中，权重振荡会给可学习缩放因子的梯度引入噪声，导致其发生振荡。反过来，后者会改变量化阈值，加剧权重振荡。这种相互作用形成的恶性循环使振荡难以平息。

4.2. 量化视觉变换器中的振荡

可学习缩放因子导致的权重振荡加剧并非只存在于简化示例中，而是现实场景中的常见问题。本节选取视觉变换器（ViT）作为量化振荡的案例进行研究。我们分析量化视觉变换器中的权重振荡，并探究可学习缩放因子如何加剧这种振荡。

4.2.1. 可学习量化对量化视觉变换器的影响

在整个训练过程中，我们可视化了量化 DeiT-T（Touvron 等，2021）中三个可学习缩放因子的变化轨迹。图 2 显示，缩放因子发生了剧烈波动；即使在训练后期学习率已经很低、模型理应收敛时，这种波动仍然持续。根据我们在第 4.1 节中的分析，可学习缩放因子的振荡会导致权重振荡。图 3 展示了大量权重聚集在量化阈值附近的明显模式，证实了这一点。这一现象还表明，训练结束时仍有不可忽略的一部分权重未能收敛，并继续在阈值附近振荡。

我们进一步考察接近量化阈值的权重的平均梯度方向变化。我们将“边界范围（BR_x）”定义为一个$\widetilde{\ }$范围，其中包含所有与最近量化阈值的距离不超过 x 的权重$\mathbf{W}_q$，

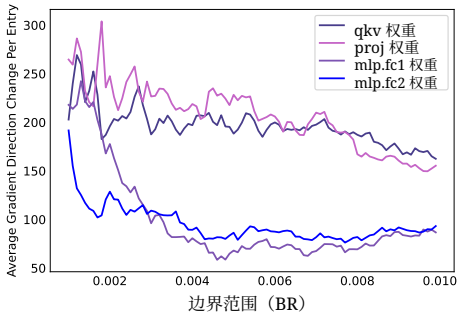


图 4：训练接近结束时，2位 DeiT-T 在 2000 次迭代中的梯度方向变化平均次数。梯度变化频率与权重到量化阈值（边界区域）的距离呈负相关。

$\widetilde{\ }$其中$\mathbf{W}_q = \mathbf{W}/\alpha$ 是舍入前重新缩放后的$\mathbf{W}$ 。图4清楚地表明，权重条目越接近量化阈值，其梯度方向翻转得越频繁，而且这一现象在 ViT 的所有模块中都一致存在。这一现象证实，在振荡过程中，阈值附近权重的梯度会不断改变方向，推动底层实值权重跨越阈值，使量化权重在两个量化值之间跳变，从而阻碍网络良好收敛。

4.2.2. 噪声注入分析

为进一步验证我们的假设——最终的量化网络会因阈值附近一定边界范围内的权重振荡而落入次优解——我们对量化

表1：对量化 DeiT-T 进行的噪声注入分析。1^{st}、 6^{th} 和 11^{th} 行表示模型在噪声注入前收敛后的准确率。“随机”指随机位置噪声注入。“边界区域内”指仅对边界范围内的权重注入噪声。“权重占比”指注入噪声的权重所占比例。 μ和 σ 分别表示十次实验试验的均值和方差。

方法	验证集 Top-1	权重占比
W2A2	54.20%	–
W2A2 随机（$\mu+\sigma$）	51.85%±0.0873	12.15%
W2A2 随机（最佳）	52%	12.15%
W2A2（边界区域内）（$\mu+\sigma$）	54.69%± 0.0803	12.15%
W2A2（边界区域内）（最佳）	54.86%	12.15%
W3A3	67.56%	–
W3A3 随机（$\mu+\sigma$）	66.72%±0.0441	7.94%
W3A3 随机（最佳）	66.78%	7.94%
W3A3（在边界区域内）（$\mu+\sigma$）	67.67%±0.0923	7.94%
W3A3（边界区域内）（最佳）	67.83%	7.94%
W4A4	72.58%	–
W4A4 随机 ($\mu+\sigma$)	72.35%±0.0617	3.61%
W4A4 随机 (最佳)	72.422%	3.61%
W4A4（在边界区域内）（$\mu+\sigma$）	72.49%±0.0499	3.61%
W4A4（在边界区域内）（最佳）	72.546%	3.61%

DeiT-T。以 $\mathrm{BR}_{0.005}$ 为例。我们首先向 $\mathrm{BR}_{0.005}$ 内的权重注入随机噪声，并训练模型一个训练轮次。表1中的结果表明，令人惊讶的是，向 $\mathrm{BR}_{0.005}$ 内的权重随机注入噪声，不会导致2位和3位模型的准确率下降，甚至在最佳情况下还能提高模型准确率。这表明最终模型只是振荡中模型的一个随机快照，而且在权重振荡区域附近存在大量更优解。

此外，为了检验向所有权重任意注入噪声是否会产生相同效果，我们均匀地向相同比例的总权重注入随机噪声，而不是只向 BR 内的权重注入。正如预期，这种设置会导致严重的准确率下降，尤其是在2位和3位设置下。我们还观察到，位宽更高的模型在振荡边界中的权重更少；向 $\mathrm{BR}_{0.005}$ 内的权重或任意位置的权重随机注入噪声，都会使4位 DeiT-T的精度略有下降，这表明振荡严重程度与位宽负相关。基于以上讨论，我们可以有把握地得出结论：这些振荡权重是阻碍模型收敛到良好局部最优解的主要因素。

4.2.3. 振荡权重的构成

出于好奇，我们想了解振荡权重的构成，即，权重张量中的某些特定位置是否容易发生振动或振荡

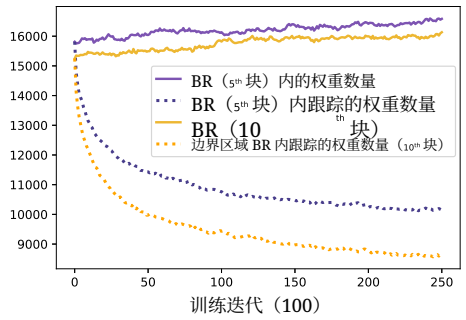


图5：在训练接近结束时，2位DeiT-T 中 $5^{th}/10^{th}$ Transformer块内初始位于边界范围 $\mathrm{BR}_{0.005}$ 且仍留在边界范围内的权重数量，与边界范围 $\mathrm{BR}_{0.005}$ 内权重总数的比较。

权重与位置无关。按照先前设置的思路，我们在 DeiT-T 训练的最后几个轮次中，追踪最初位于边界区域$_{0.005}$ 内的权重。如图 5 所示，我们注意到，持续停留在边界区域$_{0.005}$ 内的被追踪权重数量逐渐减少，而位于边界区域$_{0.005}$ 内的权重总数并未减少。这表明，尽管最初处于振荡区域的权重在优化过程中逐渐逃离，但也有新的权重进入边界区域，使网络无法摆脱振荡。这一观察还表明，振荡权重与位置无关；如果我们能够阻止新的权重进入边界范围，那么一旦所有权重都经过优化并移至振荡区域之外，振荡就会消退。这自然引出了第 5.2 节中的置信度引导退火方案。

4.2.4. 量化视觉变换器的瓶颈

先前研究发现，与其他模块相比，量化自注意力操作会导致最严重的准确率下降（Li和Gu，2022；Li 等，2022a；b）。然而，这些研究都未对此给出具体解释。在研究中，我们发现自注意力中对量化最敏感的部分，是量化查询 ($\mathbf{X}^{\mathcal{Q}_q}$)与键($\mathbf{X}^{\mathcal{K}_q}$)之间的乘法：

$$\mathbf{X}_{\mathcal{Q}_q}\cdot\mathbf{X}_{\mathcal{K}_q}^{\mathrm{T}}=F_q(F_q(\mathbf{X})\cdot F_q(\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}}))\cdot F_q(F_q(\mathbf{W}_{\mathcal{K}})\cdot F_q(\mathbf{X}^{\mathrm{T}})) \quad (5)$$

这里， F_q 表示量化函数，$\mathbf{X}$ 表示注意力块的输入。相应的反向传播公式如下：

$$\frac{\partial\mathcal{L}}{\partial\mathbf{W}_{\mathcal{K}}}\overset{STE}{\approx}\frac{\partial\mathcal{L}}{\partial\mathbf{X}_{out}}\cdot F_q(F_q(\mathbf{X})\cdot F_q(\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}}))\cdot F_q(\mathbf{X}^{\mathrm{T}}) \quad (6)$$

其中 $\mathbf{X}_{out}=\mathbf{X}^{\mathcal{Q}_q}\cdot\mathbf{X}^{\mathcal{K}_q}$。为简洁起见，术语$\mathbf{1}_{Q_n\le F_q(\mathbf{W}_{\mathcal{K}})\cdot F_q(\mathbf{X}^{\mathrm{T}})/\alpha\le Q_p}$ 以及 $\mathbf{1}_{Q_n\le\mathbf{W}_{\mathcal{K}}/\alpha\le Q_p}$ 均省略

于此处。梯度相对于 $\mathbf{W}_{\mathcal{K}}$ 取决于 $F_q(F_q(\mathbf{X})\cdot F_q(\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}}))$ 和 $F_q(\mathbf{X}^{\mathrm{T}})$。因此，$\mathbf{W}_{\mathcal{K}}$ 优化会受到 $F_q(\mathbf{W}_{\mathcal{Q}})$ 中振荡导致的不准确估计的影响，而由此产生的 $F_q(\mathbf{W}_{\mathcal{K}})$波动又会反过来加剧 $F_q(\mathbf{W}_{\mathcal{Q}})$ 振荡。为打破这一振荡加剧循环，我们在第 5.3 节提出对量化的查询-键 乘法进行重新参数化。

5. 克服量化视觉变换器中的振荡

基于上述观察，我们提出了三种新颖技术来减轻量化振荡。

5.1. 统计权重量化

为消除离群值权重对可学习缩放因子的不成比例影响，并减少权重与缩放因子之间交织的振荡，我们提出了基于统计的权重量化方法 StatsQ：

$$\mathbf{W}_q = \alpha_s \cdot ((\left\lfloor \mathrm{Clip}(\frac{\mathbf{W}}{\alpha_s}, -1, 1)\cdot n - 0.5 \right\rceil + 0.5)\cdot \frac{1}{n}) \quad \text{where } \alpha_s = 2\cdot \frac{||\mathbf{W}||_{l1}}{|\mathbf{W}|},\ n = 2^{b-1} \tag{7}$$

受（Liu 等人，2022）的启发，α_s 依据最大信息熵理论计算：当量化权重均匀分布于量化级别时，能够保留最高熵。此外，截断函数会忽略缩放后权重中的离群值，而缩放因子 α_s 的统计计算则在更新时对每个权重的贡献一视同仁。

我们进一步在图 6 中展示了统计缩放因子 α_s 在整个训练过程中的变化。与图 2 相比，可以清楚地看到，α_s 的变化更加平滑，并且在训练接近结束时几乎保持稳定；与可学习缩放因子相比，其波动更少，从而提升了整体量化感知训练稳定性。

5.2. 置信度引导退火

要抑制振荡，就需要阻止振荡区域之外的权重继续进入该区域，同时优化最初发生振荡的权重。因此，我们提出置信度引导退火（CGA），通过冻结振荡区域之外的权重、仅更新区域内的权重来微调最终模型。每次迭代都会冻结权重；一旦权重脱离振荡区域，就会被冻结，即不再更新。常规训练阶段结束后，CGA 将用于 n 个微调轮次。CGA 的详细实现总结于算法 1。

CGA 的直觉也与（Helwegen 等人，2019；Liu 等人，2021b）提出的权重置信度命题一致：

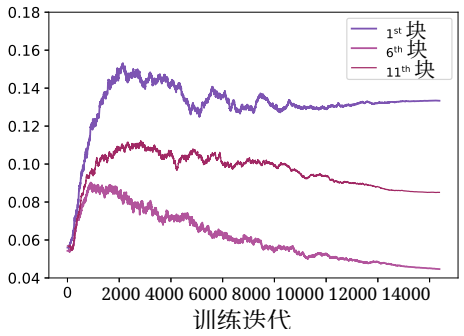


图6：训练过程中，2位DeiT-T中来自 $1^{\mathrm{st}}/6^{\mathrm{th}}/11^{\mathrm{th}}$ Transformer块的三个统计缩放因子 α_s的变化轨迹。纵轴表示 α_s的值。

量化权重的置信度与其实值潜在权重到最近量化阈值的距离成正比。距离量化阈值较远的权重被认为具有“高置信度”，而那些靠近阈值、处于振荡状态的权重则“置信度较低”。从这个意义上说，CGA 会冻结“置信度较高”的权重，使“置信度较低”的权重能够以这些固定锚点为前提进行优化，并逐渐脱离振荡边界。相比之下，（Nagel 等人，2022）提出的迭代冻结方法冻结的是振荡权重，而非非振荡权重。我们尝试在量化 DeiT-T 上使用迭代冻结方法，但没有发现明显的准确率提升。因此，我们认为应先冻结“置信度较高”的权重，再冻结“置信度较低”的权重，而不是反过来，这样才能更好地估计振荡权重的更新方向。

为进一步验证这一观点，我们展示了置信度引导退火期间，2位DeiT-T中来自 $1^{\mathrm{st}}/8^{\mathrm{th}}$ Transformer块的两个缩放因子 α_{s} 的变化轨迹。图7显示，缩放因子 α_{s} 的波动在经过一定次数的迭代后停止。由于 α_{s} 是根据整个权重张量的统计数据计算得出的，α_{s} 的冷却过程反映出所有权重都已成功脱离边界范围并被冻结，从而得到一个无振荡模型。此外，在 ImageNet 1K 上应用 CGA 后，2位量化 DeiT-T 的 Top-1准确率从 62.17% 提升至 64.33%，这证明了所提出方法的有效性。

5.3. 查询-键重参数化

最后，我们提出查询-键 重参数化（QKR），以消除量化查询 和键权重（$\mathbf{W}^{\mathcal{Q}_q}$，$\mathbf{W}^{\mathcal{K}_q}$）振荡之间的负向相互影响。方法简单，但并非显而易见：我们重新排列$\mathbf{X}_{\mathcal{Q}}$ 与

算法1置信度引导退火

给定： 权重矩阵 $\mathbf{W}$，边界范围 BR_{x}，退火迭代次数 n

对于 $j=1$ **到** n，**执行**

计算梯度 $\mathbf{G}_j = \frac{\partial \mathcal{L}_j}{\partial \mathbf{W}}$

$\mathbf{G}'_j = \mathbf{G}_j \cdot \mathbf{1}_{\widetilde{\mathbf{W}}_q \in \mathrm{BR}_{\mathrm{x}}}$

其中 $\widetilde{\mathbf{W}}_q =$ 裁剪$(\frac{\mathbf{W}}{\alpha_s} - 1,\ 1) \cdot n - 0.5$

使用 $\mathbf{G}'_j$更新$\mathbf{W}$

结束循环

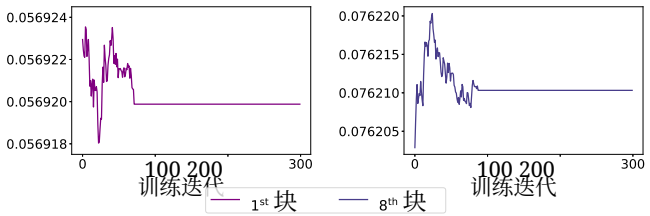


图7：在置信度引导退火训练周期内，2位 DeiT-T 中 $1^{\mathrm{st}}/8^{\mathrm{th}}$ Transformer块的统计缩放因子的变化轨迹。纵轴表示 α_s的值。经过一定次数的迭代后，统计缩放因子不再波动，这表明所有权重都已脱离边界范围并被冻结。

自注意力模块中$\mathbf{X}_{\mathcal{K}}$ 的乘法顺序，如下所示：

$$\mathbf{X}_{out} = \mathbf{X}_{\mathcal{Q}} \cdot \mathbf{X}_{\mathcal{K}}^{\mathrm{T}} = \mathbf{X} \cdot (\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}} \cdot \mathbf{W}_{\mathcal{K}}) \cdot \mathbf{X}^{\mathrm{T}} \tag{8}$$

然后，通过重新排列乘法，量化可表述如下：

$$\mathbf{X}_{out} = F_q(\mathbf{X}) \cdot F_q(F_q(\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}} \cdot \mathbf{W}_{\mathcal{K}}) \cdot F_q(\mathbf{X}^{\mathrm{T}})) \tag{9}$$

此处 F_q 表示量化函数。相应的反向传播表述为：

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{\mathcal{K}}} \overset{STE}{\approx} \frac{\partial \mathcal{L}}{\partial \mathbf{X}_{out}} \cdot F_q(\mathbf{X}) \cdot \mathbf{W}_{\mathcal{Q}}^{\mathrm{T}} \cdot F_q(\mathbf{X}^{\mathrm{T}}) \tag{10}$$

为简洁起见，此处省略了项$\mathbf{1}_{Q_n \le F_q(\mathbf{W}_{\mathcal{Q}}^{\mathrm{T}} \cdot \mathbf{W}_{\mathcal{K}}) \cdot F_q(\mathbf{X}^{\mathrm{T}}) / \alpha \le Q_p}$ 和$\mathbf{1}_{Q_n \le \mathbf{W}_{\mathcal{Q}}^{\mathrm{T}} \cdot \mathbf{W}_{\mathcal{K}} / \alpha \le Q_p}$。与式 6 相比，在这个关于关于 $\mathbf{W}_{\mathcal{K}}$的新梯度推导中，量化查询权重 $F_q(\mathbf{W}_{\mathcal{Q}})$ 不再存在，因此量化查询-键乘法层中的权重协同振荡得以解耦。此外，所提出的查询-键 重新参数化将量化操作次数从式 5 中的六次减少到式 9 中的四次，从而减轻前向传播中离散化造成的信息损失，并在反向传播中估计出更准确的梯度。

6. 实验

在本节中，我们在 ILSVRC12 ImageNet 分类数据集上评估所提出的方法，使用的架构包括 DeiT-tiny、DeiT-small（Touvron 等，2021）和 Swin-tiny（Liu 等，2021a）（Krizhevsky 等，2017）。

6.1. 实现细节

我们采用 DeiT† 和 Swin§的官方实现。我们的量化模型使用相应的全精度模型作为教师模型和初始化参数，并通过知识蒸馏训练 300 轮。对于量化 DeiT-T、2 位 DeiT-S 和 2 位 Swin-T，训练设置遵循 DeiT（Touvron 等，2021）的设置，但不使用 mixup/CutMix（Zhang 等，2018b；Yun 等，2019）数据增强。对于 3 位/4 位量化 DeiT-S 和 Swin-T，我们遵循（Li 等，2022a；b）中的训练方案。使用 CGA 对优化后的模型进行微调时，退火轮数设为 25。遵循（Esser 等人，2020；Li 等，2022a；b），我们对第一层（patch embedding）和最后几层（分类与蒸馏）采用 8 位量化。

6.2. 主要结果

我们将所提方法的组合命名为振荡无偏移量化 (OFQ)，并展示其在 DeiT-T、DeiT-S 和 Swin-T 上的整体性能，同时将结果与基线 LSQ（Esser 等人，2020）、提出混合精度量化视觉变换器的 Mix-QViT（Li 等人，2022b），以及 QViT‡（Li 等，2022a）在 2/3/4 位量化下进行比较。

从表 2 中，我们首先可以观察到，在更低位的量化中，性能退化更为严重。也就是说，使用基线 LSQ 量化时，从实值到 2位量化，DeiT-T、DeiT-S 和 Swin-T 的准确率分别下降了 17.57%、11.9% 和 10.8%。此外，与量化 DeiT-S 和 Swin-T 相比，量化 DeiT-T 的性能下降更大。上述两项观察结果与 Nagel 等人（2022）的发现一致：由于振荡，尺寸更小且位宽更低的模型会表现出更严重的准确率下降。相比之下，所提出的 OFQ 消除了最终模型中的振荡，因而显著提升了以往的当前最优水平，并缩小了量化模型与其全精度对应模型之间的精度差距。具体而言，采用 OFQ 量化的 2 位 DeiT-S/Swin-T 的准确率比先前当前最优水平 QViT（Li 等，2022a）高出 7.05%/4.64%，将与实值模型之间的差距缩小至仅 4.18%/2.68%。同样，3位 OFQ DeiT-T/DeiT-

† https://github.com/facebookresearch/deit§https://pytorch.org/vision/0.13/models/swin transformer.html_

‡我们已向 (Li 等，2022a) 的作者确认，由于第 3.2 节讨论的原因，他们的实现无法验证公式 3。因此，我们修正了他们的实现，按照其设置重新进行了实验，并在表 2 中报告结果为 QViT* 。

表 2：在 ImageNet-1K 数据集上，将所提出的 OFQ 与以往 ViT 量化方法进行比较。

网络	方法	位宽	Top-1 %
DeiT-T	全精度	W32A32	72.02
	LSQ（Esser 等人，2020）	W2A2	54.45
	QViT* （Li 等，2022a）	W2A2	50.37
	OFQ（我们的方法）	**W2A2**	**64.33**
	LSQ（Esser 等人，2020）	W3A3	68.09
	Mix-QViT（Li 等人，2022b）	~W3A3	69.62
	QViT* （Li 等，2022a）	W3A3	67.12
	OFQ（我们的方法）	**W3A3**	**72.72**
	LSQ（Esser 等人，2020）	W4A4	72.46
	Mix-QViT（Li 等人，2022b）	~W4A4	72.79
	QViT* （Li 等，2022a）	W4A4	71.63
	OFQ（我们的方法）	**W4A4**	**75.46**
DeiT-S	全精度	W32A32	79.9
	LSQ（Esser 等人，2020）	W2A2	68
	QViT* （Li 等，2022a）	W2A2	68.67
	OFQ（我们的方法）	**W2A2**	**75.72**
	LSQ（Esser 等人，2020）	W3A3	77.76
	Mix-QViT（Li 等人，2022b）	~W3A3	78.08
	QViT* (Li 等，2022a)	W3A3	78.45
	OFQ（我们的方法）	**W3A3**	**79.57**
	LSQ（Esser 等人，2020）	W4A4	79.66
	Mix-QViT（Li 等人，2022b）	~W4A4	80.11
	QViT* (Li 等，2022a)	W4A4	80.33
	OFQ（我们的方法）	**W4A4**	**81.10**
Swin-T	全精度	W32A32	81.2
	LSQ（Esser 等人，2020）	W2A2	70.40
	QViT* (Li 等，2022a)	W2A2	73.88
	OFQ（我们的方法）	**W2A2**	**78.52**
	LSQ（Esser 等人，2020）	W3A3	78.96
	Mix-QViT（Li 等人，2022b）	~W3A3	79.45
	QViT* （Li 等，2022a）	W3A3	80.06
	OFQ（我们的方法）	**W3A3**	**81.09**
	LSQ（Esser 等人，2020）	W4A4	80.47
	Mix-QViT（Li 等人，2022b）	~W4A4	80.59
	QViT* （Li 等人，2022a）	W4A4	81.29
	OFQ（我们的方法）	**W4A4**	**81.88**

S/Swin-T 显著优于 QViT（Li 等，2022a）和 Mix-QViT（Li 等人，2022b）中的 3 位模型。

在 4 位设置下，我们观察到 LSQ 的准确率可与全精度模型相当，甚至更高；大多数先前方法的表现也已超过全精度模型。这表明，由于模型分辨率更高，振荡在 4 位设置下造成的负面影响较小。尽管提升空间变小，OFQ 仍始终优于所有先前方法，在 DeiT-T、DeiT-S 和 Swin-T 上的准确率分别比全精度模型高出 3.24%、1.2% 和 0.68%。

表3：在量化 DeiT-S 上，对所提出的统计权重量化（StatsQ）、置信度引导退火（CGA）和 QK 重参数化（QKR）的单项效果进行消融研究。

网络	方法	位宽	Top-1 %
DeiT-S	基线（LSQ）	W2A2	68
	StatsQ	W2A2	74.3
	StatsQ +QKR	W2A2	75.00
	StatsQ +CGA	W2A2	75.07
	StatsQ + QKR + CGA	W2A2	**75.72**
	基线（LSQ）	W3A3	77.76
	StatsQ	W3A3	78.56
	StatsQ +QKR	W3A3	79.15
	StatsQ +CGA	W3A3	79
	StatsQ + QKR + CGA	W3A3	**79.57**
	基线 (LSQ)	W4A4	79.66
	StatsQ	W4A4	80.68
	StatsQ +QKR	W4A4	81.00
	StatsQ +CGA	W4A4	80.73
	StatsQ + QKR + CGA	W4A4	**81.10**

表4：应用 CGA 时边界范围（BR_i）选择的消融研究。实验采用 2位量化的DeiT-S 进行。

网络	方法	位宽	Top-1 %
DeiT-S	不使用 CGA	W2A2	75
	CGA（边界区域$_{0.003}$）	W2A2	75.57
	CGA（边界区域$_{0.005}$）	W2A2	**75.72**
	CGA（边界区域$_{0.007}$）	W2A2	75.63
	CGA（边界区域$_{0.01}$）	W2A2	75.67

6.3. 消融实验

本节首先考察所提出的三种方法（即 StatsQ、CGA 和 QKR ）分别在 DeiT-S 量化中的效果。如表 3 所示，在 2/3/4 位设置下，仅将 Esser 等人（2020）的方法替换为 StatsQ，准确率就分别提升了 6.3%、0.8% 和 1.0%。此外，将 QKR 和 CGA 与 StatsQ 配合使用，可在不同位宽下带来不可忽视的提升。例如，在 2 位/3 位设置下，加入 QKR 可使准确率提升 0.7%/0.59%，加入 CGA 则可提升 0.77%/0.44%。这三种方法协同发挥作用，与基线 LSQ 相比，结合这三种方法的最终模型在 2 位 DeiT-S 上将准确率提升了 7.72%。

随后，我们研究了在具有4种不同设置 [$BR_{0.003}$, $BR_{0.005}$, $BR_{0.007}$, $BR_{0.01}$]的2位量化DeiT-S上，CGA中边界范围（BR_x）选择的鲁棒性。表4显示，$BR_{0.005}$ 带来了最大的准确率提升，达0.72%；而$BR_{0.003}$、$BR_{0.007}$ 和 $BR_{0.01}$使准确率提升了 ~ 0.6%。我们进一步绘制了不同BR设置下统计缩放因子的变化轨迹，图8表明，边界

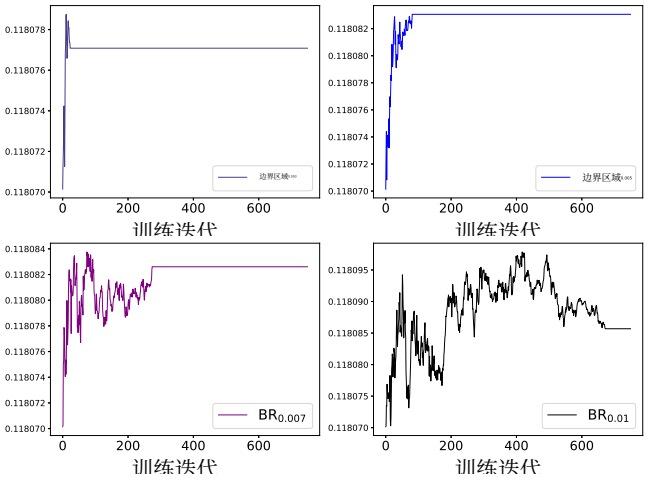


图 8：在置信度引导退火（CGA）期间，2 位 DeiT-S 中 10^{th} Transformer 块的统计缩放因子的变化轨迹，采用 4 种不同的边界范围设置（[$BR_{0.003}$、$BR_{0.005}$、$BR_{0.007}$、$BR_{0.01}$]）。纵轴表示 α_s 的值。3/4 位 DeiT-S 的可视化结果请参见附录 9.1。

范围越大，振荡权重完全退出边界范围所需的训练迭代次数就越多。结果表明，如果BR_x 设置得过宽松，就会导致优化时间更长，并增加影响“高置信度”权重的风险；反之，如果BR_x过窄，部分“低置信度”权重将被冻结，从而限制CGA更新振荡权重的效果。我们根据优化后模型中的平均梯度，经验性地设置边界范围；表4表明，只要BR_x 的设置不太激进，CGA对BR_x 的选择就相当鲁棒。

7. 结论

在这项工作中，我们以视觉变换器（ViT）为研究案例，探究量化振荡。我们揭示了可学习缩放因子对加剧权重振荡的负面影响。此外，我们发现，ViT 自注意力中的量化查询 和键 会相互产生负面影响，共同加剧权重振荡。基于这些观察，我们提出统计权重量化（StatsQ）、置信度引导退火（CGA）以及查询-键重参数化（QKR），以减轻量化权重的振荡。StatsQ 和 QKR 让量化感知训练更加稳定，并收敛到更好的局部极小值；CGA 则使振荡在训练后期进一步减弱，最终得到无振荡模型。未来，我们希望研究更多种类的深度神经网络在多种应用中的振荡现象，以考察所提方法的有效性和泛化能力。

8. 致谢

本研究由 ACCESS - 新兴智能系统人工智能芯片中心（由中国香港特别行政区 InnoHK 资助）以及香港特别行政区研究资助局优配研究金（GRF）项目编号 16203319 资助。

参考文献

Banner, R.、Nahshan, Y. 和 Soudry, D. 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，32，2019。

Bengio, Y.、L´eonard, N. 和 Courville, A. 通过随机神经元估计或传播梯度，以实现条件计算。*arXiv* 预印本 *arXiv:1308.3432*，2013。

Choi, J.、Wang, Z.、Venkataramani, S.、Chuang, P. I.-J.、Srinivasan, V. 和 Gopalakrishnan, K. PACT：用于量化神经网络的参数化裁剪激活。*arXiv* 预印本 *arXiv:1805.06085*，2018。

Dosovitskiy, A.、Beyer, L.、Kolesnikov, A.、Weissenborn, D.、Zhai, X.、Unterthiner, T.、Dehghani, M.、Minderer, M.、Heigold, G.、Gelly, S.、Uszkoreit, J. 和 Houlsby, N. 一张图像胜过 16×16 个词：用于大规模图像识别的变换器。载于国际学习表征会议，2021。

Esser, S. K.、McKinstry, J. L.、Bablani, D.、Appuswamy, R. 和 Modha, D. S. 学习步长量化。载于国际学习表征会议，2020 年。

Gong, R.、Liu, X.、Jiang, S.、Li, T.、Hu, P.、Lin, J.、Yu, F. 和 Yan, J. 可微软量化：连接全精度神经网络与低比特神经网络。载于*IEEE/CVF*国际计算机视觉会议论文集，第 4852–4861 页，2019 年。

He, K.、Zhang, X.、Ren, S. 和 Sun, J. 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。

Helwegen, K.、Widdicombe, J.、Geiger, L.、Liu,Z.、Cheng, K.-T. 和 Nusselder, R. 潜在权重并不存在：重新思考二值化神经网络优化。神经信息处理系统进展，32，2019 年。

Kenton, J. D. M.-W. C. 和 Toutanova, L. K.《Bert：用于语言理解的深度双向变换器预训练》

。载于*naacL-HLT*会议论文集，第4171–4186页，2019年。Krizhevsky, A.、Sutskever, I. 和 Hinton, G. E.《使用深度卷积神经网络进行ImageNet分类》。《*ACM*通讯》，60(6):84–90，2017年5月。ISSN 0001-0782。doi: 10.1145/3065386。网址：https://doi.org/10.1145/3065386。Li, Y.、Dong, X. 和 Wang, W.《加法二次幂量化：一种面向神经网络的高效非均匀离散化》。载于国际学习表征会议，2020年。Li, Y.、Xu, S.、Zhang, B.、Cao, X.、Gao, P. 和 Guo, G.《Q-ViT：精确且完全量化的低比特视觉变换器》。载于神经信息处理系统进展，2022a。Li, Z. 和 Gu, Q.《I-ViT：用于高效视觉变换器推理的纯整数数量化》。*arXiv*预印本 *arXiv:2207.01405*，2022年。Li, Z.、Yang, T.、Wang, P. 和 Cheng, J.《Q-ViT：面向视觉变换器的完全可微量化》。*arXiv*预印本 *arXiv:2201.07703*，2022b。Liu, Z.、Wu, B.、Luo, W.、Yang, X.、Liu, W. 和 Cheng, K.-T.《Bi-Real Net：通过增强表征能力和先进训练算法提升1比特卷积神经网络的性能》。载于欧洲计算机视觉会议（*ECCV*）论文集，第722–737页，2018年。Liu, Z.、Mu, H.、Zhang, X.、Guo, Z.、Yang, X.、Cheng, K.-T. 和 Sun, J.《MetaPruning：用于自动神经网络通道剪枝的元学习》。载于*IEEE/CVF*国际计算机视觉会议论文集，第3296–3305页，2019年。Liu, Z.、Lin, Y.、Cao, Y.、Hu, H.、Wei, Y.、Zhang, Z.、Lin, S. 和 Guo, B.《Swin Transformer：使用移位窗口的层次化视觉变换器》。载于*IEEE/CVF*国际计算机视觉会议论文集，第10012–10022页，2021a。Liu, Z.、Shen, Z.、Li, S.、Helwegen, K.、Huang, D. 和 Cheng, K.-T.《Adam和训练策略如何帮助二值神经网络优化》。载于国际机器学习会议，第6936–6946页。PMLR，2021b。Liu, Z.、Wang, Y.、Han, K.、Zhang, W.、Ma, S. 和 Gao, W.《视觉变换器的训练后量化》。神经信息处理系统进展，34：28092–28103，2021c。

Liu, Z.,Cheng, K.-T., Huang,D., Xing, E.P., and Shen, Z. 非均匀到均匀量化：通过广义直通估计实现准确量化。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第4942–4952页，2022年。Miyashita, D., Lee, E. H. 和 Murmann, B. 使用对数数据表示的卷积神经网络。*arXiv*预印本 *arXiv:1603.01025*，2016年。Nagel, M., Baalen, M. v., Blankevoort, T. 和 Welling, M. 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，第1325–1334页，2019年。Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C. 和 Blankevoort, T. 向上还是向下？训练后量化的自适应舍入。载于国际机器学习会议，第7197–7206页。PMLR，2020年。Nagel, M., Fournarakis, M., Bondarenko, Y. 和 Blankevoort, T. 克服量化感知训练中的振荡。载于Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G. 和 Sabato, S.（编），第*39*届国际机器学习会议论文集，第162卷《机器学习研究论文集》，第16318–16330页。PMLR，2022年7月17日至23日。Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. 和 J´egou, H. 通过注意力实现训练数据高效的图像变换器与蒸馏。载于国际机器学习会议，第10347–10357页。PMLR，2021年。Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. 和 Polosukhin, I. 注意力就是你所需要的一切。神经信息处理系统进展，第30卷，2017年。Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y. 和 Keutzer, K. FBNet：通过可微神经架构搜索进行硬件感知的高效卷积网络设计。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第10734–10742页，2019年。Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J. 和 Han, S. SmoothQuant：大型语言模型的精准高效训练后量化。*arXiv*，2022年。Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J. 和 Yoo, Y. CutMix：利用可定位特征训练强分类器的正则化策略。载于*IEEE/CVF*国际计算机视觉会议论文集，第6023–6032页，2019年。Zhang, D., Yang, J., Ye, D. 和 Hua, G. LQ-Nets：用于高精度紧凑型深度神经网络的学习式量化。载于欧洲计算机视觉会议论文集（*ECCV*），第365–382页，2018a。

Zhang, H., Cisse, M., Dauphin, Y. N. 和 Lopez-Paz, D. mixup：超越经验风险最小化。载于国际学习表征会议，2018b。

Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., and Zou, Y. Dorefa-net：使用低位宽梯度训练低位宽卷积神经网络。*arXiv* 预印本 *arXiv:1606.06160*，2016。

Zhu, S.、Duong, L. H. 和 Liu, W. Xor-net：一种用于在边缘设备上进行二值神经网络推理的高效计算流水线。发表于*2020 IEEE* 第 *26* 届并行与分布式系统国际会议（*ICPADS*），第 124–131 页。IEEE，2020。

9. 附录

9.1. 不同边界区域下 α_s CGA中的进展边界区域x

本节展示了在置信度引导退火（CGA）期间， α_s 来自 10^{th} Transformer 块的统计缩放因子轨迹可视化，涵盖不同边界范围 [边界区域$_{0.003}$、边界区域$_{0.005}$、边界区域$_{0.007}$、边界区域$_{0.01}$]，并比较它们在 2/3/4 位 DeiT-S 中的不同表现。从图 9–11 可以看出，无论模型位宽和边界区域$_x$如何，CGA 都能有效引导模型消除振荡。此外，显而易见的是，对于所有位宽，要使所有权重脱离更大的边界范围，都需要更多训练迭代次数。另外，我们可以观察到，与 2 位/3 位模型相比，4 位模型只需较少的训练迭代次数，就能使所有权重停止振荡，例如，4 位 DeiT-S 需要例如 ∼250 次迭代，3 位 DeiT-S 需要 ∼1500 次迭代，2 位 DeiT-S 需要 ∼700 次迭代。这一观察结果与我们的发现一致：由于模型具有更高的分辨率，振荡在 4 位设置下的危害较小。

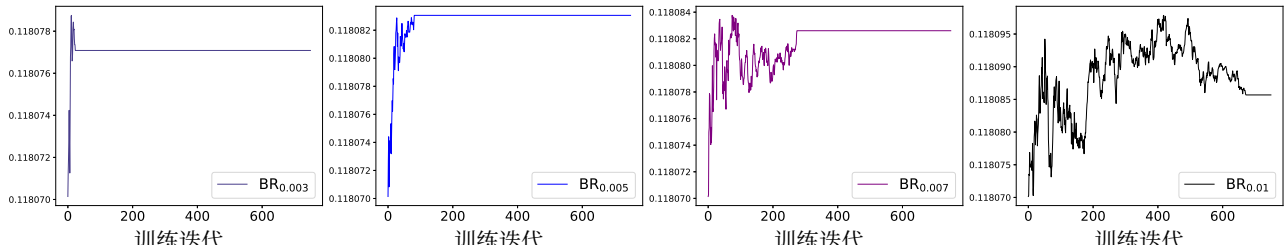


图 9：在 CGA 的整个过程中，**2位** DeiT-S 中 10^{th} Transformer块的统计缩放因子 α_s 轨迹，其中采用了 4 种不同的边界范围 [BR$_{0.003}$，BR$_{0.005}$、BR$_{0.007}$、BR$_{0.01}$]。纵轴表示 α_s 的值。

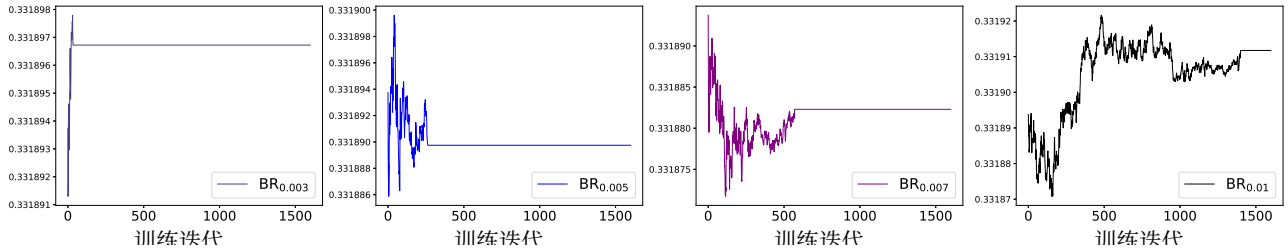


图 10：在 CGA 的整个过程中，**3位** DeiT-S 中 10^{th} Transformer块的统计缩放因子 α_s 轨迹，其中采用了 4 种不同的边界范围 [BR$_{0.003}$，BR$_{0.005}$、BR$_{0.007}$、BR$_{0.01}$]。纵轴表示 α_s 的值。

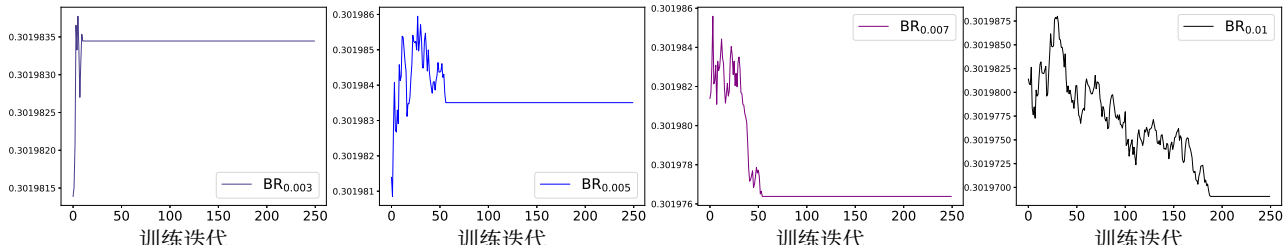


图 11：在 CGA 的整个过程中，**4位** DeiT-S 中 10^{th} Transformer块的统计缩放因子 α_s 轨迹，其中采用了 4 种不同的边界范围 [BR$_{0.003}$，BR$_{0.005}$、BR$_{0.007}$、BR$_{0.01}$]。纵轴表示 α_s 的值。