FlatENN：训练平坦模型以增强量化深度神经网络的容错能力

Akul Malhotra普渡大学西拉法叶，印第安纳州
malhot23@purdue.edu

Sumeet Kumar Gupta普渡大学西拉法叶，印第安纳州
guptask@purdue.edu

arXiv:2301.00675v1 [cs.LG] 29 Dec 2022

摘要—通过量化和稀疏性增强进行模型压缩，已受到广泛关注，有望使深度神经网络（DNN）能够部署于资源受限的边缘环境。尽管这些技术在降低深度神经网络的能耗、延迟和内存需求方面已取得了令人鼓舞的成果，但它们在非理想的现实条件下（例如存在硬件故障时）的表现仍未得到充分了解。本文研究了位翻转和固定型故障对激活稀疏量化深度神经网络（QDNN）的影响。我们发现，较高的激活稀疏性会以更易受故障影响为代价。例如，激活稀疏量化深度神经网络的准确率比标准量化深度神经网络最多低 11.13%。我们还指出，导致准确率下降的一个主要原因是激活稀疏量化深度神经网络的损失景观具有更尖锐的极小值，使其对故障引起的权重值扰动更加敏感。基于这一观察，我们提出采用一种感知尖锐度的量化（SAQ）训练方案来减轻故障的影响。与采用传统方法训练的对应模型相比，采用 SAQ 训练的激活稀疏量化深度神经网络和标准量化深度神经网络，其推理准确率分别最高提升 19.50% 和 15.82%。此外，我们表明，与采用传统方法训练的标准量化深度神经网络相比，采用 SAQ 训练的激活稀疏量化深度神经网络在故障条件下表现出更高的准确率。因此，所提出的技术有望在不牺牲容错能力的前提下，实现与稀疏性相关的能耗／延迟优势。

索引词—**深度神经网络加速器、模型压缩、平坦极小值、容错能力。**

I. 引言

深度神经网络（DNN）在决策和感官处理任务中取得的显著成功[1] [2] ，推动了针对各种应用探索深度神经网络加速器设计 [3]。然而，最先进的深度神经网络所带来的性能优势，是以大量存储和计算需求为代价的，这也引发了各种设计挑战，尤其是在受能耗和存储限制的边缘应用中。降低深度神经网络的规模和计算复杂度的需求，促使量化深度神经网络（QDNN）的兴起。研究者提出了多种量化技术，通过降低深度神经网络权重和激活值的位精度，在缓解相较于全精度对应模型的准确率下降的同时，实现存储、计算和通信方面的能耗节省 [4][5]。

另一种广受欢迎的降低深度神经网络资源需求的方法，是对权重和激活值进行稀疏化。

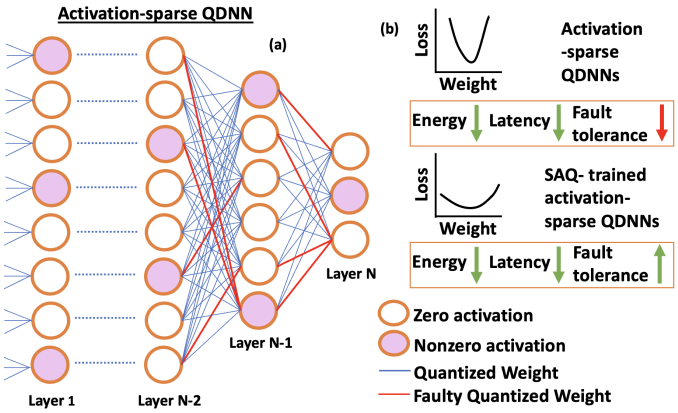


图 1：（a）展示了发生故障时的激活稀疏（AS）量化深度神经网络。我们的研究表明，与标准模型相比，激活稀疏量化深度神经网络更容易出现严重的准确率下降。（b）描述了增强激活稀疏性带来的能耗／延迟优势与降低的容错能力之间的权衡。我们提出了一种故障缓解策略，通过采用基于尖锐度感知量化（SAQ）的训练来展平其权重损失景观，使激活稀疏量化深度神经网络能够保持容错能力。

通过系统地移除网络中不太重要的部分，可以实现这一点。各种用于提高权重稀疏性的技术，也称为权重剪枝，已在准确率几乎不受影响或完全不受影响的情况下，显著压缩了深度神经网络 [6][7]。近年来，提高并利用激活值稀疏性的方法也逐渐受到关注，以减少内存需求并提升推理速度 [8] [9]。这类涉及权重与激活值稀疏化的技术可与量化结合使用，从而设计出针对边缘应用优化的量化深度神经网络。

尽管稀疏化量化深度神经网络为资源受限系统带来了诸多优势，但在现实环境中部署时，也需要考虑硬件的非理想行为。深度神经网络加速器的权重通常存储在片上存储器中，而存储器会遭遇各种故障，例如固定值故障和位翻转故障。这些故障会破坏权重值并降低系统准确率。尽管人们对传统深度神经网络加速器中故障的影响已有充分了解[10] [11], ，但对权重稀疏或激活稀疏深度神经网络的容错能力研究仍然有限。一些研究探讨了故障和非理想因素对剪枝后的

模型的影响，并表明与未剪枝模型相比，剪枝深度神经网络在存在故障和非理想因素时准确率下降更严重 [12][13]。然而，人们对故障对激活稀疏深度神经网络的影响缺乏了解。此外，还需要开发能够减轻故障对激活稀疏量化深度神经网络准确率造成的不利影响的技术，以提升其在非理想环境下的计算鲁棒性。

本文通过以下方式应对这些关键需求：(a) 全面分析激活稀疏量化深度神经网络在存在固定故障和位翻转故障时的性能；(b) 提出一种提升容错能力的训练技术。我们的主要贡献如下：

- 我们发现，激活稀疏性的提高会以降低量化深度神经网络的容错能力和推理准确率为代价。据我们所知，这是首项探索激活稀疏量化深度神经网络在存在故障时性能的研究。
- 借助[14], 中的权重损失景观可视化方法，我们证实，激活稀疏量化深度神经网络对故障更加敏感，是因为其损失景观中的极小值更加尖锐（相较于标准量化深度神经网络）。
- 基于上述发现，我们提出采用尖锐度感知量化（SAQ） [15] （这是为量化深度神经网络设计的锐度感知最小化（SAM） [16]的一种变体），以减轻故障对系统准确率的影响。
- 我们表明，通过降低权重损失景观的尖锐度，所提出的方法可将激活稀疏量化深度神经网络和标准量化深度神经网络的推理准确率分别提高最多 19.50% 和 15.82%。这一提升是在不牺牲基线软件准确率的情况下实现的。
- 我们还表明，与未采用 SAQ 训练的标准量化深度神经网络相比，经过 SAQ 训练的激活稀疏量化深度神经网络具有更高的推理准确率。这使得设计兼具激活稀疏性和容错能力的量化深度神经网络成为可能，非常适合边缘应用。

图 1 概述了我们的工作和贡献。

II. 背景与相关工作

A. 深度神经网络中的激活稀疏性

激活稀疏性是指深度神经网络激活值中存在大量零值。深度神经网络中最常见的激活函数是修正线性单元（ReLU），它对每个负输入都输出零，因此激活稀疏性较高。因此，可以使用稀疏存储方案存储激活值，从而降低内存需求[17]。此外，可以跳过涉及零值激活值的计算，从而降低深度神经网络的能耗和延迟 [18]。请注意，激活稀疏性具有动态特性，这意味着零值的数量和位置会因输入而异。利用激活稀疏性来降低计算和内存需求时，需要考虑这一特性。因此，若能恰当利用高激活稀疏性将有助于设计资源受限型深度神经网络加速器。

由于这一原因，近年来，用于增强激活稀疏性的算法技术逐渐受到关注。这些技术主要通过向损失函数中显式添加正则化项来实现，该正则化项会惩罚稠密激活。例如，文献 [9] 将激活值的 L_1 范数（$||x_{l,\ n}||_1$）添加到原始损失函数中，以促进稀疏性：

$$L_{reg}(x,w) = L_0(x,w) + \sum_{n=1}^{N}\sum_{l=1}^{L}\alpha_l||x_{l,n}||_1 \qquad (1)$$

此处，$L_{reg}(x,\ w)$ 是需要最小化的新损失函数，$L_0(x,\ w)$ 是原始损失函数值，L 是深度神经网络中的层数，N 是批次大小，α_l 是每层的正则化常数。通过利用 $L1$ 正则化 [19]，的稀疏化特性，文献 [9] 表明，在图像分类中，稀疏性最高可提升 60%，而推理准确率几乎不会损失。文献 [8] 使用基于 Hoyer 稀疏度指标的正则化器以及 ReLU 激活函数的变体来提升深度神经网络的激活稀疏性。需要注意的是，这些技术并不等同于对激活值进行“剪枝”，其原理类似于权重剪枝 [6] [7]。剪枝会针对所有输入将参数永久置零，而激活稀疏化技术不会永久移除任何激活值。相反，这些技术能够确保对于所有输入，非零激活值的比例都较低（平均而言）。

本文采用基于 L_1 正则化的方法训练激活稀疏量化深度神经网络，这是因为该方法直观且易于实现。我们将未使用 $L1$ 激活正则化训练的量化深度神经网络称为标准量化深度神经网络，将使用 L_1 激活正则化训练的量化深度神经网络称为激活稀疏（AS）量化深度神经网络。

B. 深度神经网络加速器中的内存故障

激进的缩放以及对新型存储器技术的探索，使得内存故障研究变得比以往任何时候都更加重要。用于深度神经网络加速器的所有主流存储器（静态随机存取存储器、阻变存储器、铁电场效应晶体管等）都会受到故障的影响。故障会破坏一部分存储值，导致乘加运算（深度神经网络中最常见的内核）计算不准确，进而降低推理准确率。困扰深度神经网络加速器的两种常见内存故障是位翻转故障和固定型故障。当存储单元的值不可更改地 固定为 ’1’ 和 ’0’ 时，分别会发生固定为 1 故障（SA1）和 SA0 故障。SA1 和 SA0 故障通常由制造缺陷 [20] 以及有限的耐久性引起。位翻转故障发生在存储单元的值发生翻转时（从 ’0’ 变为 ’1’，反之亦然）。这类故障可能是永久性的，也可能是瞬态的，其成因包括半选读取扰动和 α 粒子撞击等现象。

近年来，内存故障对深度神经网络性能的影响引起了广泛关注。已有研究

如 [21] 表明，单个位翻转就可能显著降低浮点深度神经网络的性能。研究也发现，量化深度神经网络很容易受到固定型故障和位翻转故障的影响 [10][11]。一些研究还分析了在存在故障 [13],电路非理想性和变化 [12],的情况下权重稀疏（剪枝后）网络的性能，结果表明，与未剪枝网络相比，剪枝深度神经网络更容易受到故障及其他非理想因素的影响。然而，据我们所知，尚无研究分析故障对激活稀疏量化深度神经网络的影响。

在故障缓解方面，人们分析了大量硬件和算法解决方案。基于硬件的技术通常通过识别并复制深度神经网络中的关键部分来增加冗余。 [22]利用穷举故障注入识别深度神经网络中易受影响的部分，并复制这些部分，以减轻故障的影响。 [23] 在深度神经网络中添加一个备用神经元，当其他神经元发生故障时，可将其配置为替代其中任意一个。从算法角度来看，故障缓解策略包括使用纠错码、故障感知再训练和容错训练。诸如[24]的研究采用纠错输出码，以降低对变化和固定型故障的敏感性。训练期间进行故障注入 [25] 也很有前景，但当系统容易遭受多种故障时，效果可能不佳。故障感知再训练基于在线故障检测，随后对未发生故障的参数进行再训练，以恢复深度神经网络的准确率[26]。该方法展现出良好的效果，但对于不支持在线故障检测的边缘应用而言，可能难以实现。

在本研究中，我们将深入探究固定型故障和位翻转故障对标准量化深度神经网络及激活稀疏量化深度神经网络性能的影响。我们还将提出一种故障缓解策略，利用 SAQ 展平量化深度神经网络的权重损失景观，使其参数对扰动不那么敏感，从而提升容错能力。我们提出的方法可以与基于硬件的容错技术结合使用，并能防护多种故障，后文将对此进行讨论。

C. 尖锐度感知量化（SAQ）

深度神经网络在训练过程中收敛到的损失景观极小值的尖锐度，对其泛化能力有着重要影响 [27][28]。理论和实证研究均已表明，收敛到平坦极小值有助于提升泛化能力。因此，文献 [16] 探索了锐度感知训练方案，例如锐度感知最小化（SAM）。SAM 同时最小化损失值和权重损失景观的尖锐度，从而学习在邻域内损失值均匀较低的最优解。研究表明，采用 SAM 训练的深度神经网络在多个基准数据集和模型上实现了最先进的准确率。然而，SAM 在量化深度神经网络上的效果不够理想，因为它没有考虑权重的量化。尖锐度感知量化（SAQ）是针对量化深度神经网络设计的 SAM 变体 [15]。SAQ 同时最小化损失值，并通过使用经过对抗性扰动的量化权重最小化损失函数，使极小值附近的损失曲率变平。这有助于提升量化深度神经网络的泛化能力和准确率。在这项工作中，我们基于这样的直觉，进一步将 SAQ 应用于训练容错量化深度神经网络：收敛到更平坦极小值的量化深度神经网络，对故障引起的权重扰动不那么敏感。

III. 故障对激活稀疏量化深度神经网络的影响

A. 实验框架

为研究故障对常规量化深度神经网络和激活稀疏（AS）量化深度神经网络的影响，我们采用以下方法。首先，使用 [5] 在 Pytorch 中的量化框架训练两个 4 位量化深度神经网络：采用 FashionMNIST（FMNIST）数据集 [30] 的 LeNet 5 [29]，以及采用 CIFAR-10 数据集 [32] 的 ResNet 18[31]。激活稀疏量化深度神经网络通过结合 [9] 中描述的 $L1$ 激活正则化与量化框架进行训练。为确保公平比较，标准量化深度神经网络和激活稀疏量化深度神经网络的推理准确率均训练至近似相等（$\leq 0.5\%$）。

训练完成 QDNNs 后，分别注入两类故障，即位翻转故障和固定型故障（SA0 和 SA1）。故障在各模型中随机且均匀地分布。我们分别针对 LeNet 5 和 ResNet 18 开展基于蒙特卡罗的故障注入实验。我们分析了不同故障率和故障类型下的推理准确率均值（见图4）。对于 LeNet 5 和 ResNet 18 QDNNs，我们分别在 1%–5% 和 0.5%–3% 的故障率范围内进行实验。由于更高的故障率会导致准确率严重下降，使 QDNNs 无法使用，因此我们在分析 ResNet 18 QDNNs 时将故障率上限设为 3%。

为了进一步了解二者在容错能力方面的差异，我们将 LeNet 5 标准模型和 AS QDNNs 的权重损失景观进行了可视化。许多研究利用损失景观可视化来加深对模型性能和行为的理解，并展示了深度神经网络的泛化能力与极小值形状之间的相关性。据我们所知，这是首项研究 QDNN 的容错能力与极小值形状之间关系的工作。我们采用 [14]中所述的基于归一化的可视化技术，生成 LeNet 5 标准模型和 AS QDNNs 的权重损失景观。

B. 结果

1) 激活稀疏性：为了理解故障对激活稀疏型量化深度神经网络的影响，我们先来讨论 $L1$ 激活正则化带来的激活稀疏性提升。图 2a 和图 2b 分别展示了 LeNet 5 和 ResNet 18 量化深度神经网络中零激活值的百分比（对所有层和测试输入取平均值）。对于 LeNet 5，使用$L1$正则化进行训练可使激活稀疏性提升 95.34%（从标准训练时的 43% 提升至使用 $L1$ 正则化时的 84%）。对于 ResNet-18，激活稀疏性提升了 41.51%（从 53% 提升至 75%）。LeNet 5 与 ResNet-18 的激活稀

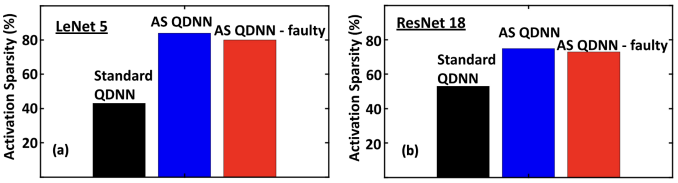


图 2：无故障和有故障条件下，(a) LeNet 5 和 (b) ResNet 18 标准量化深度神经网络与激活稀疏量化深度神经网络的激活稀疏性。LeNet 5 和 ResNet 18 激活稀疏量化深度神经网络的激活稀疏性分别比对应的标准量化深度神经网络高 95.34% 和 41.51%，且在有故障的环境中仍能保持。

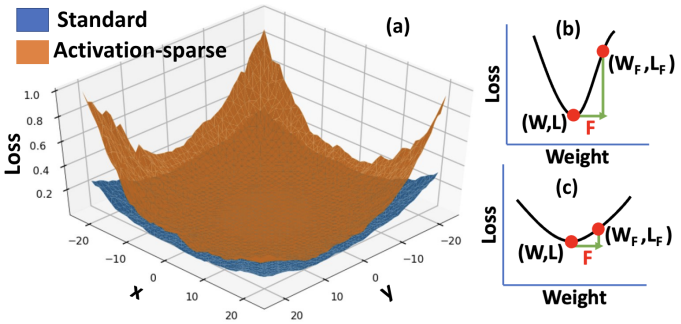


图3：(a) 使用 [14] 中的技术，将标准 LeNet 5 量化深度神经网络和激活稀疏（AS）LeNet 5 量化深度神经网络的权重损失景观可视化。x 和 y 为归一化随机方向。(b) 和 (c) 展示了故障（F）对损失景观中损失值的影响，其中 (b) 为尖锐极小值，(c) 为平坦极小值。在尖锐极小值的情况下，故障会导致损失值发生更大的变化。

疏性提升幅度存在差异，这是因为与 LeNet-5（FMNIST 数据集）相比，ResNet 18 的工作负载复杂度更高（在 CIFAR-10 数据集上进行图像分类）。因此，为了保持较高的准确率，ResNet-18 平均需要更多非零激活。

我们进一步分析了在有故障条件下激活稀疏量化深度神经网络的稀疏性受到的影响。根据故障注入分析，我们观察到激活稀疏性的变化可以忽略不计。在最坏情况下（位翻转故障率分别为 5% 和 3%），LeNet 5 和 ResNet 18 的稀疏性分别降低了 4% 和 2%。这表明，即使存在故障，激活稀疏性的提升仍能保持。

2）故障对推理准确率的影响：图 4a–4f 中的虚线展示了位翻转、SA0 和 SA1 故障对标准版及激活稀疏版 LeNet 5 和 ResNet 18 量化深度神经网络的影响。我们观察到，与标准量化深度神经网络相比，激活稀疏量化深度神经网络在存在故障时的推理准确率更低。对于位翻转故障，LeNet 5 和 ResNet 18 激活稀疏量化深度神经网络的准确率分别比对应的标准量化深度神经网络低 2.40% 至 11.13% 和 0.52% 至 8.00%（准确率的绝对差值）。SA0 和 SA1 故障也呈现这一趋势，但准确率下降幅度小于位翻转故障。这是因为固定型故障可能会被屏蔽，因此其严重程度低于位翻转故障。这些结果表明，LeNet 5 和 ResNet 18 激活稀疏量化深度神经网络在存在故障时的准确率下降幅度均大于对应的标准量化深度神经网络，这意味着激活稀疏性的提高是以降低故障容忍度为代价的。

需要指出的是，在相同故障率（最高为 3%）下，ResNet 18 激活稀疏量化深度神经网络的准确率下降（最高为 8.00%）大于 LeNet 5 激活稀疏量化深度神经网络（最高为 5.56%）。这是因为，相较于 FMNIST 上的 LeNet-5，CIFAR-10 上的 ResNet 18 工作负载更为复杂，因此更容易因故障而导致准确率下降。

3）权重损失景观可视化：图3a展示了标准 LeNet 5 量化深度神经网络和激活稀疏量化深度神经网络的权重损失景观（使用 [14] 创建）。直观来看，如图3b和图3c所示，与具有平坦极小值的量化深度神经网络相比，具有更尖锐极小值的量化深度神经网络在权重受到某种扰动时，其损失函数值会发生更大的变化。可以看出，激活稀疏量化深度神经网络收敛于更尖锐的极小值，这从经验上解释了第三节中观察到的容错能力下降。因此，激活程度的增加由于 $L1$ 激活正则化导致的稀疏性，与极小值尖锐度增加以及容错能力降低相关。

总而言之，我们得出结论：额外的激活稀疏性虽然能够降低优化后深度神经网络加速器的延迟和能耗，但代价是在发生故障时性能下降。鉴于提升激活稀疏量化深度神经网络容错能力的需求，我们接下来提出一种方法，通过基于尖锐度感知量化（SAQ）的训练来展平权重损失景观，从而减轻故障的影响。

IV. 缓解策略：SAQ

在本节中，我们提出并研究基于 SAQ 训练的故障缓解策略的性能，以降低存在位翻转和固定为故障时，标准量化深度神经网络和激活稀疏量化深度神经网络中的准确率下降。我们使用第三节所述的框架进行实验。

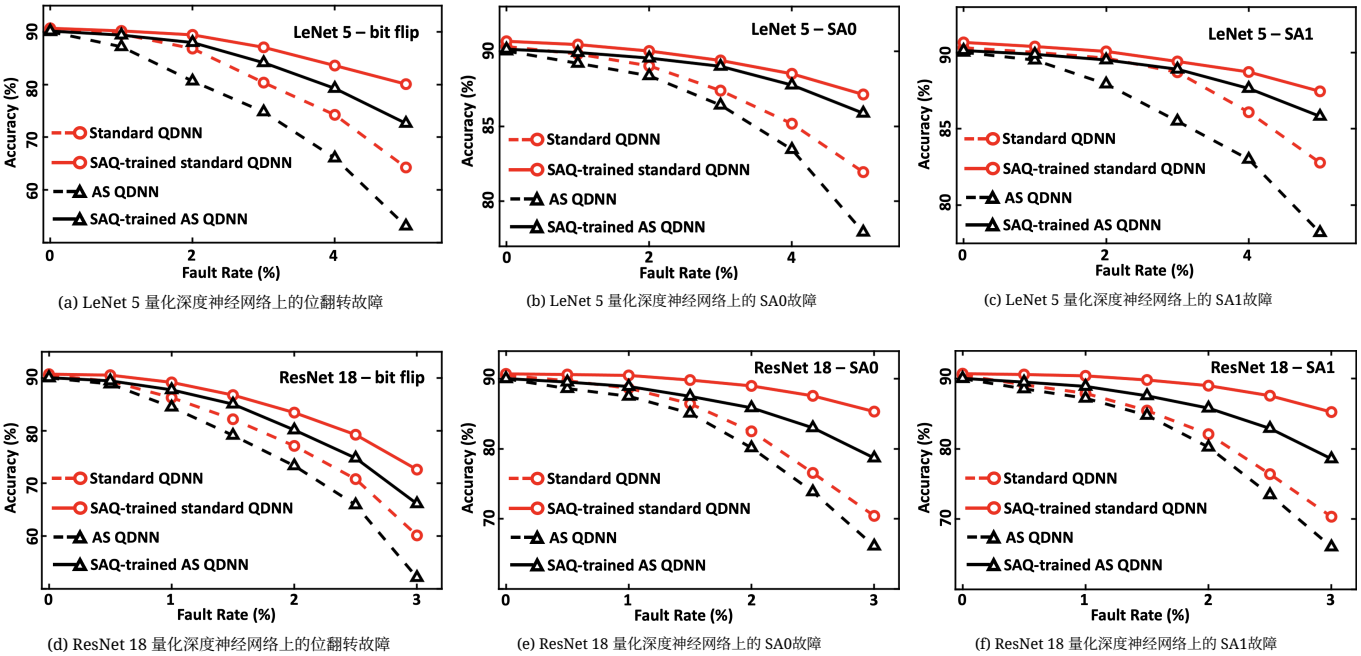
基于 *SAQ* 的故障缓解策略

我们的故障缓解技术基于展平量化深度神经网络的权重损失景观，使权重对故障不那么敏感。为实现并评估所提出的故障缓解策略，我们使用 SAQ 训练标准量化深度神经网络和激活稀疏量化深度神经网络。

SAQ 通过求解以下极小极大优化问题，同时最小化损失值和损失尖锐度：

$$\min_{w} \max_{||\epsilon||_2 \leq \rho} ((L_S(Q_w(w,b)) + \epsilon) - L_S(Q_w(w,b))) + L_S(Q_w(w,b)) + \frac{\lambda}{2}||w||_2^2 \quad (2)$$

公式中的第一项定义了尖锐度度量，即对于某个权重扰动 ϵ，损失值的最大变化量。第二项是损失函数本身，第三项是标准的 L_2 正则化项。此外，扰动 ϵ 是对抗性的，其选取方式为



(a) LeNet 5 量化深度神经网络上的位翻转故障 (b) LeNet 5 量化深度神经网络上的 SA0故障 (c) LeNet 5 量化深度神经网络上的 SA1故障

(d) ResNet 18 量化深度神经网络上的位翻转故障 (e) ResNet 18 量化深度神经网络上的 SA0故障 (f) ResNet 18 量化深度神经网络上的 SA1故障

图4：比较采用 SAQ 训练和传统训练的标准量化深度神经网络与激活稀疏（AS）量化深度神经网络在不同故障场景下分类准确率受到的影响。可以看出，激活稀疏量化深度神经网络的容错能力低于对应的标准量化深度神经网络。此外，采用 SAQ 训练的量化深度神经网络比采用传统训练的同类网络具有更高的容错能力。

使最大尖锐度最小化。 ϵ 由以下公式给出，其最大值可为 ρ。其值为：

$$\epsilon \simeq \hat{\epsilon} = \rho \frac{\nabla_{Q_w(w,b)} L_S(Q_w(w,b))}{||\nabla_{Q_w(w,b)} L_S(Q_w(w,b))||_2} \tag{3}$$

其中，$\nabla_{Q_w(w,b)} L_S(Q_w(w,b))$ 是损失函数相对于量化权重的梯度。 ϵ 通过对量化深度神经网络进行前向和反向传播来估算。

需要注意的是，SAQ 训练的一个训练轮次所需时间与传统训练的两个训练轮次相同。因此，为了公平比较，我们对基于 SAQ 的量化深度神经网络进行训练的训练轮次数设为传统训练的量化深度神经网络的一半。基于 SAQ 的量化深度神经网络的 ρ 超参数经过选择，使得该网络在无故障情况下（故障率为= 0 %）的准确率达到最大。借助这一框架，我们分析 SAQ 在提高标准量化深度神经网络和激活稀疏量化深度神经网络容错能力方面的有效性。

B. 结果

1）标准量化深度神经网络：图4展示了 LeNet 5 和 ResNet 18 标准量化深度神经网络在 SAQ 训练后与传统训练后在存在故障时的性能对比。可以看出，经过 SAQ 训练的标准量化深度神经网络比传统训练的同类模型具有更强的容错能力。对于故障率为 0%–5% 的 LeNet 5，SAQ 训练模型在位翻转、SA0 和 SA1 故障下的推理准确率分别高出 0.35%–15.82%、0.35%–5.18% 和 0.35%–4.68%。即使故障率为 0%（无故障环境），经过 SAQ 训练的标准量化深度神经网络的表现也优于传统训练的标准量化深度神经网络，这与 [16] 和 [15]中的结果一致。同样，对于 ResNet 18，在无故障和故障条件下，经过 SAQ 训练的标准量化深度神经网络均比传统训练的标准量化深度神经网络具有更高的推理准确率。位翻转、SA0 和 SA1 故障下的准确率分别提升了 0.21%–12.48%、0.21%–14.89% 和 0.21%–14.92%。

2）激活稀疏（*AS*）量化深度神经网络：图4展示了经 SAQ 训练和常规训练的 LeNet 5 与 ResNet 18 激活稀疏量化深度神经网络的推理准确率对比。经 SAQ 训练和常规训练的量化深度神经网络采用相同的 $L1$ 正则化常数，因此两者的激活稀疏性几乎相同（数值见图2a和图2b）。

与标准量化深度神经网络的趋势一致，经 SAQ 训练的激活稀疏量化深度神经网络，其推理准确率高于对应的常规训练模型。对于 LeNet 5，经 SAQ 训练的激活稀疏量化深度神经网络在位翻转、SA0 和 SA1 故障下的推理准确率分别高出 0.12%–19.50% 、 0.12%–7.94% 和 0.12%–7.63%。对于 ResNet 18，经 SAQ 训练的激活稀疏量化深度神经网络在无故障和故障条件下的推理准确率同样更高；在位翻转、SA0 和 SA1 故障下，其准确率分别比常规训练模型高出 0.04%–14.00%、0.04%–12.59% 和 0.04%–12.56%。

值得注意的是，在存在故障的情况下，使用 SAQ 训练的激活稀疏（AS）量化深度神经网络表现优于未经 SAQ 训练的标准量化深度神经网络。

例如，我们的实验表明，在故障率为= 5%时，使用 SAQ 训练的 LeNet 5 激活稀疏量化深度神经网络，其准确率比采用传统方式训练的 LeNet 5 标准量化深度神经网络高 8.37%。因此，使用 SAQ 训练的激活稀疏量化深度神经网络兼具更强的容错能力和更高的激活稀疏性（从而降低延迟和能耗），因此非常适用于边缘应用。

最后，我们发现，采用 SAQ 训练能够限制增强激活稀疏性所造成的准确率下降。对于 LeNet 5（故障率 = 5%）和 ResNet 18（故障率= 3%），分别采用 SAQ 训练的激活稀疏量化深度神经网络，其推理准确率比采用 SAQ 训练的标准量化深度神经网络低 7.45% 和 6.49%；而采用传统方式训练的激活稀疏量化深度神经网络，其准确率则低 11.13% 和 8.00%。因此，采用 SAQ 训练能够减轻激活稀疏性增强对容错能力的不利影响。

V. 结论

我们研究了存在故障时激活稀疏（AS）量化深度神经网络的性能，并提出了一种缓解策略，以降低故障对量化深度神经网络准确率的影响。通过在 LeNet 5 和 ResNet 18 量化深度神经网络中均匀注入不同类型故障进行实验，我们发现，激活稀疏性的提升是以故障存在时推理准确率下降为代价的；与标准量化深度神经网络相比，激活稀疏量化深度神经网络的准确率最多低 11.13%。为了解这一容错能力下降的原因，我们可视化了标准量化深度神经网络和激活稀疏量化深度神经网络的权重损失景观，并表明激活稀疏量化深度神经网络具有更尖锐的极小值，因而容错能力更低。基于这一观察，我们提出利用基于尖锐度感知量化（SAQ）的训练来展平损失景观，从而增强容错能力。结果表明，采用 SAQ 训练的激活稀疏量化深度神经网络和标准量化深度神经网络，准确率分别比采用传统方式训练的对应模型高出最多 19.50% 和 15.82%。我们还观察到，采用 SAQ 训练的激活稀疏量化深度神经网络，其推理准确率高于采用传统方式训练的标准量化深度神经网络，因此能够构建兼具激活稀疏性和容错能力的量化深度神经网络。因此，采用 SAQ 训练的量化深度神经网络具有更强的容错能力，适合部署于易发生故障的边缘场景。

参考文献

[1] J. Yu 等，“Coca：对比式图像描述模型是图文基础模型”，2022年。[2] D. Silver 等，“利用深度神经网络和树搜索掌握围棋”，2016年。[3] B. Reagen 等，“Minerva：实现低功耗、高精度的深度神经网络加速器”，载于*2016*年*ACM/IEEE*第*43*届年度国际计算机体系结构研讨会（*ISCA*），2016年，第267–278页。[4] D. Zhang 等，“Lq-nets：面向高精度、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*），2018年9月。[5] B. Jacob 等，“神经网络的量化与训练：实现高效的纯整数运算推理”，载于*2018*年*IEEE/CVF*计算机视觉与模式识别会议，2018年，第2704–2713页。[6] J. Frankle 和 M. Carbin，“彩票假说：寻找稀疏、可训练的神经网络”，载于国际学习表征会议（*ICLR*），2019年。

[7] X. Dong 等，“通过逐层最优脑外科剪枝深度神经网络”，载于第31届神经信息处理系统国际会议（NeurIPS）论文集，2017年，第4860–4874页。[8] M. Kurtz 等，“诱导并利用激活稀疏性，加速深度神经网络推理”，载于第37届国际机器学习会议（ICML）论文集，2020年，第5533–5543页。[9] G. Georgiadis，“通过激活图压缩加速卷积神经网络”，载于2019年IEEE/CVF计算机视觉与模式识别会议（CVPR），2019年，第7078–7088页。[10] U. Zahid 等，“Fat：训练神经网络，使其在硬件故障下实现可靠推理”，载于2020年IEEE国际测试会议（ITC），2020年。[11] J. Zhan 等，“利用边界感知激活提高深度神经网络的容错能力”，IEEE计算机辅助设计集成电路与系统汇刊，第1–1页，2021年。[12] A. Bhattacharjee 等，“研究并缓解交叉阵列非理想特性对稀疏深度神经网络精确实现的影响”，载于2022年欧洲设计、自动化与测试会议暨展览（DATE），2022年，第1119–1122页。[13] G. Yuan 等，“利用权重剪枝和差分交叉阵列映射提高基于阻变存储器的边缘人工智能的深度神经网络容错能力”，载于2021年第22届国际电子设计质量研讨会（ISQED），2021年，第135–141页。[14] H. Li、Z. Xu 等，“神经网络损失景观的可视化”，载于《神经信息处理系统进展》，2018年。[15] J. Liu 等，“面向深度神经网络的尖锐度感知量化”，2021年。[16] P. Foret 等，“锐度感知最小化：高效提升泛化能力”，载于《国际学习表征会议》，2021年。[17] S. Han 等，“深度压缩：通过剪枝、训练后量化和霍夫曼编码压缩深度神经网络”，《国际学习表征会议》（ICLR），2016年。[18] J. Albericio 等，“Cnvlutin：无需无效神经元的深度神经网络计算”，载于2016年ACM/IEEE第43届年度国际计算机体系结构研讨会（ISCA），2016年，第1–13页。[19] S. Sra 等，《机器学习优化》。麻省理工学院出版社，2011年。[20] C.-Y. Chen 等，“基于March测试和新型挤压搜索方案的阻变存储器缺陷建模与故障分析”，IEEE计算机汇刊，第180–190页，2015年。[21] S. Hong 等，“终末期脑损伤：揭示深度神经网络在硬件故障攻击下的急剧退化”，载于第28届USENIX安全研讨会（USENIX安全会议19），2019年，第497–514页。[22] F. Libano 等，“面向现场可编程门阵列中神经网络的选择性加固”，IEEE核科学汇刊，第216–222页，2019年。[23] C. Schorn 等，“一种高效的深度神经网络位翻转韧性优化方法”，载于2019年欧洲设计、自动化与测试会议暨展览（DATE），2019年，第1507–1512页。[24] T. Liu 等，“一种容错神经网络架构”，载于2019年第56届ACM/IEEE设计自动化会议（DAC），2019年，第1–6页。[25] T. Cho 等，“无需无噪声基线的节点扰动学习”，《神经网络》，第267–272页，2011年。[26] C. Liu 等，“通过高缺陷容忍度挽救基于忆阻器的神经形态设计”，载于2017年第54届ACM/EDAC/IEEE设计自动化会议（DAC），2017年，第1–6页。[27] Y. Jiang 等，“绝妙的泛化度量及其寻找方法”，载于《国际学习表征会议》，2020年。[28] G. K. Dziugaite 和 D. M. Roy，“为参数远多于训练数据的深度（随机）神经网络计算非平凡泛化界”，2017年。[29] Y. Lecun 等，“应用于文档识别的基于梯度的学习”，《IEEE会议录》，第2278–2324页，1998年。[30] H. Xiao 等，“Fashion-MNIST：用于机器学习算法基准测试的新型图像数据集”，arXiv，2017年。[31] K. He 等，“用于图像识别的深度残差学习”，CoRR，2015年。[32] A. Krizhevsky，“从微小图像中学习多层特征”，技术报告，2009年。