通过深度量化神经网络提高对对抗性攻击的鲁棒性

Ferheen Ayaz‡、Idris Zakariyya*、Jos´e Cano*、Sye Loong Keoh*、
Jeremy Singer*、Danilo Pau†、Mounia Kharbouche-Harrari†
*英国格拉斯哥大学 †意法半导体 ‡英国萨塞克斯大学

arXiv:2304.12829v1 [cs.LG] 25 Apr 2023

摘要——**缩减机器学习（ML）模型，尤其是深度神经网络（DNN）的内存占用，对于将其部署到资源受限的微型设备上至关重要。然而，深度神经网络模型的一项缺点是容易受到对抗攻击：只需向输入添加轻微扰动，就可能使其产生错误判断。因此，如何构建准确、鲁棒且体积小巧的深度神经网络模型，并将其部署到资源受限的嵌入式设备上，是一项挑战。本文报告了一种微型深度神经网络模型的设计成果，该模型能够抵御对抗性黑盒和白盒攻击，并通过自动化的量化感知训练框架 QKeras 进行训练，同时在学习循环中纳入深度量化损失，从而提高所设计的深度神经网络在微型设备上的部署准确率。我们研究了 QKeras 与一种对抗鲁棒性技术——雅可比正则化（JR）——如何通过利用深度神经网络拓扑结构和逐层 JR 方法，形成协同优化策略，以构建鲁棒且体积小巧的深度量化深度神经网络模型。最终，我们构思、开发并测试了一种采用该协同优化策略的新型深度神经网络模型；测试使用了三个包含图像和音频输入的数据集，并将其性能与现有基准方法在多种白盒和黑盒攻击下进行了比较。实验结果表明，在白盒和黑盒攻击下，我们提出的深度神经网络模型在 CIFAR-10 图像数据集和 Google Speech Commands 音频数据集的一个子集上，平均准确率分别比 MLCommons/Tiny 基准高出 8.3% 和 79.5%。此外，在 SVHN 图像数据集上，该模型对黑盒攻击的准确率也高出 6.5%。**

索引词—**深度神经网络（DNN）、QKeras、雅可比正则化（JR）、对抗攻击。**

I. 引言

深度神经网络（DNN）在自然语言处理、网络安全、计算机视觉、智能应用等诸多任务中展现出卓越的性能 [1]。然而，深度神经网络模型资源消耗较大，具有较高的内存占用和计算需求。此外，边缘智能日益增长的需求催生了机器学习（ML）领域的新型优化策略，这些策略力求在同时缩减DNN模型架构的情况下达到最佳准确率 [2]。机器学习中专门生成受约束机器学习工作负载、以部署到目标边缘设备（如微控制器单元（MCU）和传感器）上的子领域，被称为深度量化机器学习（DQML）。然而，深度神经网络模型通常容易受到对抗攻击的影响，导致输入发生人眼无法察觉的变化 [3]。这些漏洞至关重要，因为这限制了深度神经网络模型作为现实世界应用的有效解决方案的部署能力，例如自动驾驶汽车、智慧城市、智能应用和响应式人工智能（AI） [4]。两类广泛使用的对抗攻击是白盒攻击和黑盒攻击。在白盒攻击中，攻击者完全了解深度神经网络模型及其结构和参数；而在黑盒范式中，攻击者不了解所使用的深度神经网络模型。在这两种场景下，这些攻击都旨在故意造成错误分类或扰乱模型性能。因此，亟需在减少微型嵌入式设备上深度神经网络模型的内存占用与增强其抵御对抗攻击的能力之间取得平衡。

DQML 有望在资源受限的微控制器上运行极低位深的机器学习模型，这些微控制器的功耗为μW，内存仅有几 MB [5]。各种框架被用于优化机器学习模型的资源占用，例如用于后训练优化的 TensorFlow Lite（TFLite） [6] 、用于深度量化感知训练的 Larq [7], Brevitas [8] 和 QKeras [9]。其中大多数框架利用量化，根据数据精度优化所使用的机器学习模型。

QKeras 旨在提供低至单比特的量化，同时通过将量化误差以随机噪声的形式引入，并在训练期间学习克服该误差，从而保持模型准确率 [10]。它基于 Keras 的即插即用替换函数，因此可以灵活地添加量化器，并针对每一层的激活值、偏置和权重分别选择量化位深。这对于高效训练量化的深度神经网络模型非常有用。在 QKeras 提供的各种深度量化策略中，有一种随机量化 [11], : 它并非对深度神经网络模型的所有元素（参数）进行量化，而是以与量化误差成反比的随机概率，对部分元素进行量化，同时将其余部分保持为全精度（FP）。每次迭代都会逐步增加量化部分，直到最终可能将整个深度神经网络量化。该程序能够大幅度、渐进式地补偿量化误差，从而使极低位深的深度神经网络获得更高的准确率。

与此同时，探索深度神经网络模型的鲁棒性对于将其集成到DQML中至关重要。事实上，在支持模型部署于资源受限的微控制器上的同时提升安全性，是一项关键挑战。文献中已有各种针对深度神经网络的防御技术和模型

这项工作是在 Ferheen Ayaz 任职于格拉斯哥大学期间完成的。

用于提供抵御对抗攻击的鲁棒性 [12]。然而，这些防御机制可能导致模型规模增大，或使其在干净数据集上的准确率下降。对于DQML模型而言，为达到最优的模型规模和准确率，通常需要大量学习计算，同时还可能容易受到对抗攻击的影响，因此，模型规模的任何增加或准确率的下降都可能影响部署性能。鉴于此，近期研究 [13], [14] 表明，量化可以降低计算需求，同时在一定程度上增强对抗白盒对抗攻击的鲁棒性。

受这一观察结果的启发，本文研究以下假设：逐层混合量化方案能否在保持干净数据集准确率性能与微型嵌入式设备有限资源需求之间权衡的同时，继承抵御白盒攻击和黑盒攻击的鲁棒性？我们提出了一种可部署于微型设备且对对抗攻击具有高度鲁棒性的 DQML 模型，该模型使用 QKeras框架进行训练。我们的理论研究表明，QKeras 将雅可比正则化（JR）用作对抗攻击防御机制。基于此，我们使用 QKeras 提出了一种随机三值量化（STQ）深度神经网络模型，其准确率性能适合部署在微型微控制器上，并有望在同一传感器封装中应用于图像和音频。该模型抵御各种对抗攻击的能力已得到证明。本文的贡献如下：

- 开发了一种基于 STQ 的模型，该模型复杂度更低，可部署于微控制器上，仅需极少的内存占用，并且在干净数据集上的准确率性能得到提升（第三节）。
- 根据其与雅可比正则化（JR）技术的相似性分析 STQ 模型的鲁棒性，并展示其抵御各种白盒和黑盒对抗攻击的能力（第 III 节和第 IV 节）。
- 全面比较 STQ 模型针对业界认可的 MLCommons/Tiny 基准在图像和音频输入方面的抵抗能力，同时研究其在 K 折交叉验证下的性能（第 IV 节）。

II. 背景与相关工作

A. 针对机器学习模型的对抗攻击

安全性考量对于智能交通系统、医疗保健和金融系统等敏感应用至关重要，这些系统会使用基于 AI 的图像和语音识别模型 [15]。为这类任务开发的机器学习和 AI 模型存在一个缺点：在现实世界应用中，它们容易受到攻击，从而危及其完整性、机密性和隐私。具体而言，对抗攻击会在输入中添加微小扰动，以在受限范数约束下最大化模型的损失函数 [16]。方程 (1) 描述了向输入数据引入扰动的这一程序，其中：$\mathcal{L}(\theta,\ x',\ y)$ 是损失函数，θ 表示模型参数，x' 是扰动输入，y 是模型输出，δ 表示扰动，p 是扰动范数 [17]。本研究考虑第 I 节定义的两种攻击，即白盒攻击和黑盒攻击。

一种常见的对抗性攻击技术是快速梯度符号法（FGSM） [18]。这是一种白盒攻击，通过单步迭代，根据输入估计模型训练损失函数的梯度。FGSM [18]攻击过程如式（2）所示，其中 Δ 表示梯度，ϵ 表示用于限制扰动的小常数。FGSM 的一种变体是投影梯度下降（PGD） [18]。这是一种计算开销更高的多步威胁模型，会运行多次迭代，以找到具有尽可能低 δ 的对抗性输入，如式（3）所示，其中 i 是迭代索引，α 表示梯度步长，S 表示扰动集合。

$$\max_{||\delta||_p} \mathcal{L}(\theta, x', y) \tag{1}$$

$$x' = x + \epsilon \cdot sign(\Delta_x \mathcal{L}(\theta, x, y)) \tag{2}$$

$$x'_{i+1} = \prod_{x+S} (x + \alpha \cdot sign(\Delta_x \mathcal{L}(\theta, x, y))) \tag{3}$$

一种高效的黑盒攻击是 Square 攻击 [19]，它基于一种经过多次迭代的随机搜索优化技术。在每次迭代中，它会随机选择位置，将输入中一小部分改变为正方形。与基于梯度的优化类似，它也依赖于步长缩减，其中尺寸指的是正方形的维度 [19]。黑盒攻击的另一种形式是 Boundary 攻击 [20], ，其中利用查询来估计输出类别的决策边界。从一幅干净图像开始，通过查询执行梯度估计，在每次迭代中沿估计方向移动，并投影出新的扰动，直到模型决策发生改变 [20]。

克服白盒和黑盒攻击方法需要适当的防御机制。因此，增强模型在不同攻击变体下的性能对于促进其部署非常重要。

B. 对抗攻击鲁棒性

人们提出了各种防御方法，以提高深度神经网络抵御对抗攻击的鲁棒性 [21]。一些策略旨在检测对抗性输入 [22] ，或通过附加网络或模块执行变换，以移除扰动 [23] [24] 。对抗训练会在模型学习过程中引入对抗性输入，使其学会不对这些输入产生误判 [25]。然而，这些方法无法保证模型抵御训练期间未引入的攻击，因为模型并未学习过这些攻击。其他方法，例如 [26] ，通过在损失函数中添加正则化项来降低模型对微小扰动的敏感性。公式（4）表示这一联合损失函数，其中$\mathcal{L}_{reg}$ 表示正则化项，λ是一个超参数，用于调节正则化项与实际损失之间的权重。

$$\mathcal{L}_{joint} = \mathcal{L}(\theta, x, y) + \lambda L_{reg}, \tag{4}$$

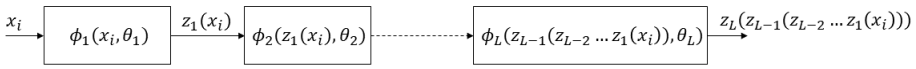


图 1：DNN 将输入 x_i 转换为输出 z_L 。

C. 雅可比正则化

雅可比正则化（JR）是一种用于提升对抗鲁棒性的技术，其中输入梯度正则化会对交叉熵损失的梯度进行归一化，使得 $\mathcal{L}_{reg} = \sum_{k=1}^{K} ||\Delta_x z_k^L(x_i)||_2^2 = ||J(\mathbf{x}_i)||_F^2$，而 $||J(\mathbf{x}_i)||_F^2$ 表示在输入数据 [27]上计算得到的模型雅可比矩阵的弗罗贝尼乌斯范数。为降低计算复杂度，文献 [28]提出了逐层雅可比正则化。

为展示逐层雅可比正则化的鲁棒性，我们考虑一个由 N 个训练样本组成的 D 维网络输入 X 。如图 1 所示，该网络包含 l =1、2、..... L 层。 z_l 表示第 l 层的输出，z_l^k 表示第 l^{th} 层中第 k^{th} 个神经元的输出。考虑网络输出端的 softmax 运算；对于输入 x_i ，用于计算 Top-1 准确率的最终预测输出为 $f(x_i) = argmax\{z_L^1, z_L^2, .., z_L^K\}$，其中 K 为输出向量的维度。DNN 的雅可比矩阵在 L^{th} 层进行计算，即 $J_L(x_i) = \nabla_x z_L(x_i)$，其定义如下：

$$J_L(x_i) = \begin{bmatrix} \frac{\partial z_L^1 x_1}{\partial x_1} & \frac{\partial z_L^1 x_1}{\partial x_2} & \cdots & \frac{\partial z_L^1 x_1}{\partial x_D} \\ \frac{\partial z_L^2 x_2}{\partial x_1} & \frac{\partial z_L^2 x_2}{\partial x_2} & \cdots & \frac{\partial z_L^2 x_2}{\partial x_D} \\ . & . & . & . \\ . & . & . & . \\ \frac{\partial z_L^K x_K}{\partial x_1} & \frac{\partial z_L^K x_K}{\partial x_2} & \cdots & \frac{\partial z_L^K x_K}{\partial x_D} \end{bmatrix} \epsilon \mathbb{R}^{K \times D}. \quad (5)$$

I对于输入 JR，其定义为 i i d fi d

$$||J(x_i)||_F^2 = \sum_{d=1}^{D} \sum_{k=1}^{K} \left(\frac{\partial}{\partial x_d} z_L^k(x_i) \right)^2 = \sum_{k=1}^{K} ||\nabla_x z_L^k(x_i)||_2^2, \quad (6)$$

在训练的标准损失中加入式（6）所示的正则化项，以提高 DNN 的鲁棒性。该方法作为训练后处理提出，即使用新的损失函数 [27]对网络进行更少迭代次数的重新训练。如 $i\epsilon\{1, 2, ..., N\}$所示，雅可比正则化需要计算 N 个梯度，而在逐层雅可比正则化中，每一层仅对一个随机i 计算雅可比矩阵 [28]。其基本思想是减小雅可比矩阵的弗罗贝尼乌斯范数，从而扩大分类间隔，即输入与网络分类器所诱导的决策边界之间的距离。

D. 量化模型的对抗鲁棒性

模型量化技术广泛应用于各个领域 [29]。在 [30], 中，它被用于物联网（IoT）网络的异常检测和网络攻击防御。然而，量化模型的局限性在于 FP 模型的分类边界发生偏移，这可能会影响模型对抗对抗扰动的脆弱程度 [31]。因此， [31] 中的作者研究了采用基于边界的再训练方法，通过使用非线性映射作为防御白盒对抗攻击的机制，来减少对抗损失和量化损失。

表 I：对抗量化深度神经网络模型的比较。

研究工作		[32]	[33]	[34]	[35]	Ours
压缩技术	方法	N 位宽量化				SQ*
	权重上的	✓	✓	✓	✓	✓
	激活值上的	✓	✓	✓	✗	✓
输入	图像	✓	✓	✓	✓	✓
	音频	✗	✗	✗	✗	✓
数据集		MNIST C-10*	MNIST SVHN	MNIST C-10* TinyI*	MNIST 螺旋	C-10* SVHN GSC
攻击	白色 box	FGSM PGD C&W	FGSM BIM C&W	✗	FGSM	FGSM PGD C&W
	黑色 box	✗	SPSA ZOO	✗	✗	Square Boundary ZOO
	随机	✗	✗	高斯噪声 噪声	✗	✗
内存占用		✗	✗	✗	✗	410 千字节

*C-10 表示 CIFAR-10 数据集，*TinyI 表示 TinyImageNet，*SQ 表示随机量化

其他先前研究探讨了扰动对不同模型的影响。表 I 对比了若干研究量化深度神经网络模型鲁棒性的现有工作。这些研究探讨了量化深度神经网络模型的韧性，但没有考察其在微控制器或传感器等资源受限设备上的部署可行性，而且大多数研究仅考虑了白盒攻击。

III. 提出的深度神经网络模型

本节介绍一种基于随机三值量化（STQ）方法的新型、复杂度更低且鲁棒性更强的深度神经网络模型，该模型旨在部署于微型微控制器上。

A. QKeras 中的对抗鲁棒性

QKeras [9] 是一个面向量化的 DNN 框架，以 Keras [36],为基础，为构建和训练量化神经网络提供了高效的方法，支持从 1 位到 32 位的分数位宽或整数位宽。QKeras 执行量化感知训练 [37]。使用 q 位宽权重、激活值和梯度参数训练神经网络的基础算法在 [38]中进行了概念化。具体而言，QKeras 中的网络训练包括反向传播，在此过程中，参数梯度会被随机量化为低位宽数值。图 2 展示了 QKeras 中每一层的训练流程。每次训练迭代都包含一个前向传播步骤，用于量化权重、计算输出，并将量化误差加入每一层的输出中。随后执行反向传播步骤，将梯度随机量化为低位宽数值。最后，更新未量化的权重和梯度，以供下一次迭代使用。通过 QKeras 训练深度量化网络，可以使其具备抵御对抗攻击的鲁棒性，原因在于

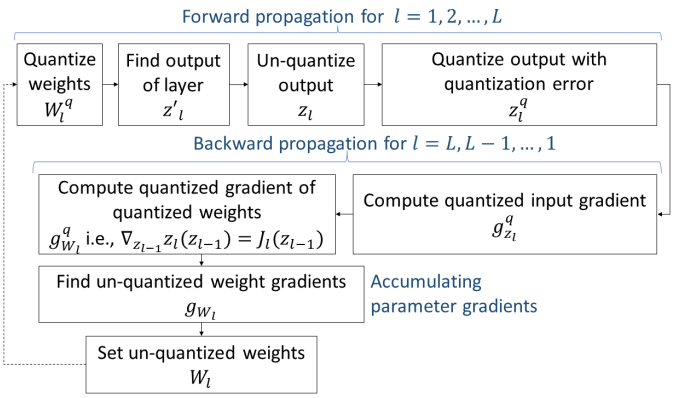


图 2：QKeras量化感知训练流程。

以下两点：i）在梯度量化过程中引入噪声函数，以克服训练期间的量化误差，从而使 DNN 对噪声和扰动效应具备鲁棒性；ii）QKeras 包含 JR特征，具体说明如下。

定理： *QKeras*引入逐层雅可比正则化，因此提高了深度神经网络抵御对抗攻击的鲁棒性。

证明： 考虑图 1 所示的深度神经网络逐层结构，最后一层的输出 z_L L 为

$$z_L = \phi_L((\phi_{L-1}(...\phi_1(x_i, \theta_1), ...\theta_{L-1}), \theta_L)), \tag{7}$$

其中，ϕ_l（即，θ_l）表示 l^{th} 层的函数，θ_l表示第 l层的模型参数，且 $z_0 = x_i$ [28]。l^{th} 层的雅可比矩阵定义为

$$J_l(z_{l-1}) = \frac{dz_l}{dz_{l-1}}, \tag{8}$$

该矩阵会在QKeras训练期间进行反向传播（ [38]中的算法 1 的步骤 10 至 16 ）。式（ 8 ）所表示的推导结果是 l^{th} 块层的雅可比矩阵 [28]。因此，具有此类模式的QKeras在训练期间融入了逐层雅可比正则化。

为证明逐层雅可比正则化能够增强对抗鲁棒性，考虑一个干净输入 x_c 和一个对抗性输入 x_p，二者都接近输入 x_i ，并且都属于同一类别 k。由于 $f(x_i) = f(x_c) \neq f(x_p)$ ，因此，按照 [27],中的定义，网络输入与输出的 ℓ_2距离度量为

$$\frac{||x_p - x_i||_2}{||x_c - x_i||_2} \approx 1, \tag{9}$$

$$\frac{||z_L(x_p) - z_L(x_i)||_2}{||z_L(x_c) - z_L(x_i)||_2} > 1, \tag{10}$$

结合（9）和（10），并利用均值定理 [39], 可以证明，较小的弗罗贝尼乌斯范数会使网络对扰动不那么敏感，即:

$$\frac{||z_L(x_p) - z_L(x_i)||_2^2}{||x_p - x_i||_2^2} \leq ||J(x')||_F^2, \tag{11}$$

其中 $x' \epsilon [x_i, x_p]$。与（10）类似，对于每一层 l，我们有

$$\frac{||z_l(x_p) - z_l(x_i)||_2}{||z_l(x_c) - z_l(x_i)||_2} > 1. \tag{12}$$

误分类误差在每一层传播，因此

$$\frac{||z_l(z_{l-1}(x_p)) - z_l(z_{l-1}(x_i))||_2}{||z_l(z_{l-1}(x_c)) - z_l(z_{l-1}(x_i))||_2} > 1. \tag{13}$$

该误差会经过反向传播，然后在每一层进行调整。因此，学习优化过程提高了通过 QKeras 训练的深度神经网络模型的鲁棒性，因为它区分了由干净输入与扰动输入所导致的误差。由此可以得出结论：如图 2 以及 [38] 中算法 1 的步骤 11 至 15 所示，QKeras 深度量化感知过程会针对前一层反向传播的输入引入逐层雅可比正则化。尽管反向传播旨在对参数进行量化，但最终会生成更加鲁棒的网络。其学习过程会增加训练计算时间，但具有双重优势，即深度量化以及对对抗扰动的鲁棒性。

B. 随机三值量化*QKeras*架构

由于 QKeras [37] 应用程序接口是 Keras 的扩展，最初使用 [36] 构建了一个 32位浮点深度神经网络模型。该 FP 模型包含六个卷积层，包括深度卷积层、可分离卷积层，以及三个全连接层。前几层用于捕获通道相关性，并能以更少的参数提供更多特征，尤其适用于图像输入 [40]。模型采用带有残差连接的多分支拓扑结构，以优化特征图。为提高准确率，隐藏层采用批归一化 [41]，ReLU [42] 和 sigmoid [43]激活函数，而输出层使用 softmax。根据模型摘要()的结果，该 FP 模型包含 998,824 个参数，其中 997,460 个可训练，1,364 个不可训练。默认情况下，此深度神经网络模型未集成雅可比正则化，因此容易受到对抗攻击。然而，由于模型体积过大（4 MB），无法将其部署在微型微控制器上。由于我们的目标设备是 STM32H7，其最大 RAM内存为 1 MB，因此我们使用 QKeras 对该模型应用了 STQ方法。

图3展示了使用 Netron [44] 渲染深度神经网络图及其超参数后可视化的 STQ 模型架构。可以看到，卷积层、深度卷积层、可分离层和激活函数后都附有 Q，表示浮点 Keras [36] 模型的量化版本，而 量化*ReLU* 则表示 Keras 中量化激活函数的版本。所提出的 STQ 模型采用了 QKeras 的异构量化特性，支持对深度神经网络 [10]中的每一层进行独立量化。这有助于降低模型内存占用和复杂度，同时提高性能准确率。

IV. 评估

本节介绍所提出的 STQ 模型的评估程序，使用图像和音频数据集来处理干净样本和对抗样本。此外，本文将我们的 STQ 模型的 Top-1 准确率与其他基准方法在三种黑盒对抗攻击和三种白盒对抗攻击下的结果进行比较。

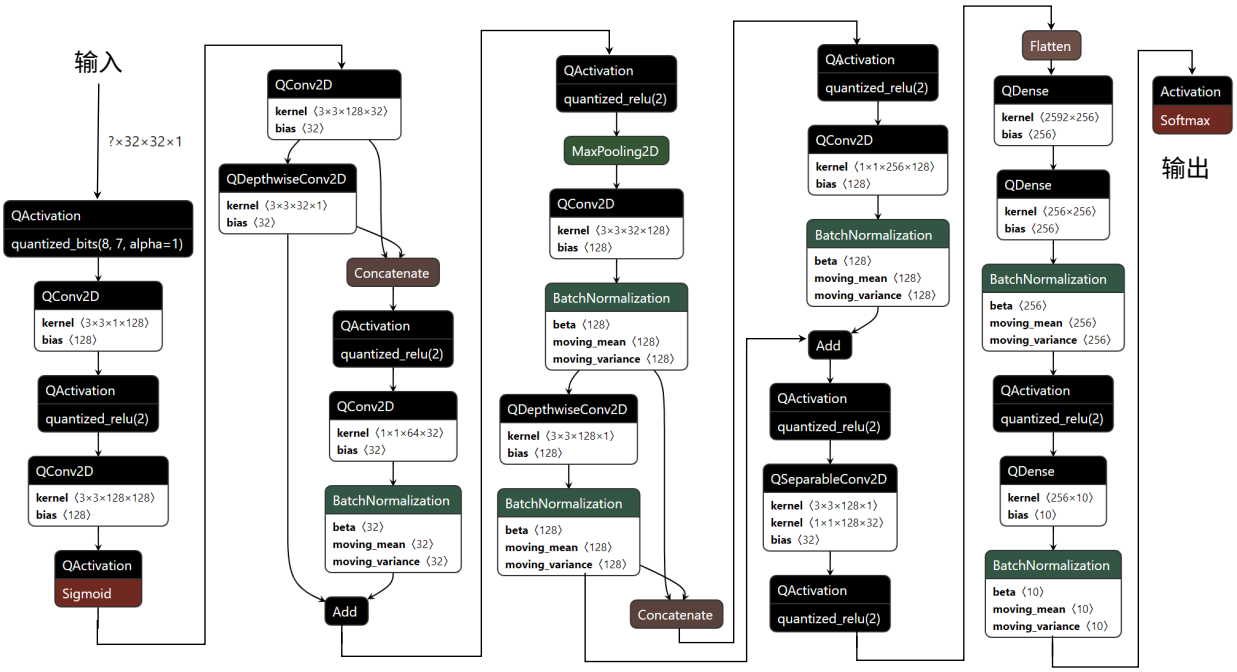


图 3：使用 QKeras 开发的基于 STQ 的深度神经网络模型的架构。

A. 实验设置

1) 数据集和预处理： 为评估所设计 STQ 模型的有效性，选取了以下音频和图像基准数据集：

- **CIFAR-10** 包含 60,000 张图像，分属 10 个不同类别。每个类别由 50,000 张训练图像和 10,000 张测试图像组成 [45]。
- **街景门牌号数据集 (SVHN)** 是一个真实世界图像数据集，取自 Google 街景图像中的门牌号。该数据集包含 10 个类别，每个类别对应一个数字。训练集和测试集分别包含 73,257 和 26,032 个数字 [46]。
- **Google Speech Commands**（GSC）由数千名不同人士录制的 65,000 条时长为一秒、包含 30 个短词的语音组成。其 12 个类别包括实验中使用的“yes”、“no”以及从“zero”到“nine”的数字。训练样本和测试样本的数量分别为 31,257 和 15,636 [47]。

请注意，在将输入图像样本提供给 DNN 输入之前，先将其像素值归一化到 0 和 1 之间，然后转换为灰度图像。这在避免使用已知具有欺骗性的颜色的同时，有效降低了处理图像样本的计算成本。彩色图像处理不在本文的研究范围内，未来可在此基础上进一步研究。

2) 模型训练流程：表 II 列出了用于评估全精度（FP）、所提出的 STQ 以及其他经过测试的量化模型的训练参数。采用余弦退火学习率（LR）函数 [48]，并结合 Adamax 优化器，以加快收敛。这些训练参数经过选择，既能对每个模型进行微调，又能降低其计算复杂度，同时保持更好的或业界领先的性能。

表 II：QKeras DNN 的训练参数

参数	值
训练轮数	1000
批量大小	64
学习率	$[1 \times 10^{-6}, 1 \times 10^{-3}]$
学习率调度器	余弦退火
Loss	分类交叉熵
优化器	Adamax

由于所提出的 STQ 模型面向图像和音频数据集，因此还采用并测试了 ResNetv1 和 DS-CNN TFLite 模型，以进行比较。这是为了研究该模型相对于业界采用的 MLCommons/Tiny 模型的鲁棒性，从而深入了解所提出模型在各种数据集和其他基准方法上的性能表现。

3) 对抗攻击程序：通过使用多种对抗攻击评估所提出的 STQ 模型，以展示其韧性和鲁棒性。评估考虑了三种白盒攻击：FGSM、PGD [18]和 Carlini 和 Wagner (C&W-L2) [49]；以及三种黑盒攻击：Square [19]，Boundary 攻击 [20]和零阶优化 (Zoo) [50]。针对受测数据集，使用对抗鲁棒性工具箱 (ART) [51] 生成所有攻击对应的受扰动数据样本。对于 FGSM 和 PGD 样本，使用 L1 范数，ϵ 值设为 0.6。对于 Square 攻击，使用无穷范数，ϵ 值设为 0.6；对于 Boundary 攻击，使用无穷 范数，ϵ 值设为 0.01。最大

表 III：CIFAR-10、SVHN 和 GSC 干净数据集上的性能（Top-1 准确率）对比。

模型	闪存 (KB)	CIFAR-10 准确率 (%)	SVHN 准确率 (%)	GSC 准确率（%）
FP	4496	52.73	67.94	89.50
8 位	1124	52.46	70.04	85.60
4 位	527	50.88	64.72	82.06
三元	410	77.35	93.11	88.30
STQ	410	**80.57**	**95.53**	**94.76**
2 位	281	46.35	89.74	81.56
二元	140	39.03	56.91	77.10
S-Binary	140	52.25	63.72	82.51

C&W-L2 和 Zoo 使用的最大迭代次数为 10，Zoo 使用的二分搜索树大小为 10。C&W-L2 和 Zoo 使用 CIFAR-10 数据样本。用于生成扰动的样本数分别为 CIFAR-10、SVHN 和 GSC 数据集的 100、240 和 2400。

为进一步考察我们的 STQ 模型对 FGSM、PGD [18] 和 Square [19] 攻击的抵抗能力，我们使用 CIFAR-10 的 10,000 个测试样本，在不同的攻击强度下构造了这些攻击。FGSM 攻击针对 ϵ 进行构造，其范围为：$\{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$，如 [14]所示。对于 PGD 攻击，使用了 $t = 7$ 次迭代，步长为 $\alpha = 2/255$，其 ϵ 范围为：$\{8/255, 16/255, 32/255\}$。对于 Square [19] 攻击，第一种变体采用无穷范数，ϵ 值为 0.05，最大迭代次数为 10,000。第二种 Square 攻击变体采用 ℓ_2 范数，最大迭代次数为 10,000，如 [52]所示。

4) K折交叉验证：为了估计每个测试模型上干净数据样本和对抗性数据样本的方差，采用了 K 折交叉验证。该技术将整个数据集划分为 K （折）个相等的子集，其中 $K - 1$ 个子集用于训练，剩余的一个子集用于验证 [53]。对于每一折，模型在训练集上进行拟合，并在验证集上进行预测，以估计平均性能。出于实现方面的考虑，取 $K = 10$，因为较大的折数可能会提高预测性能 [54]。此外，还使用了 scikit-learn [55] 机器学习 Python API。

B. 量化方案

表 III 展示了使用每个数据集中的干净输入时，不同量化模型在所测试数据集上的性能（Top-1 测试准确率）对比。此外，表中还包括每个量化模型的 FLASH MCU 存储器大小，该大小根据使用 qstats QKeras 库获得的权重配置计算得出。可以看出，与浮点、随机

二元（S-Binary）模型以及其他量化模型。此外，该模型比浮点、8 位和 4 位模型更轻量，并且可以部署在微控制器设备上。这些结果促使我们进一步研究所提出的 STQ 模型对抗对抗性攻击的鲁棒性，因为浮点模型没有集成任何防御机制，在干净样本分类方面的准确率性能也没有更好。

表 IV：CIFAR-10、SVHN 和 GSC 数据集上的模型鲁棒性（Top-1 测试准确率）比较。

数据集	模型	清洁 准确率（%）	FGSM 准确率（%）	PGD 准确率（%）	Square 准确率（%）	Boundary 准确率（%）
CIFAR-10	ResNetv1	85.0	65.6	64.1	65.5	82.9
	STQ	80.6	**80.8**	**80.8**	**73.6**	**83.8**
SVHN	ResNetv1	94.8	**94.0**	**94.0**	82.1	94.0
	STQ	**95.5**	93.1	93.1	**91.7**	**97.3**
GSC	DS-CNN	89.1	15.0	15.4	15.1	14.7
	STQ	**94.8**	**94.3**	**94.2**	**94.3**	**95.4**

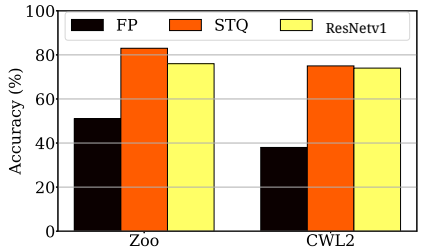


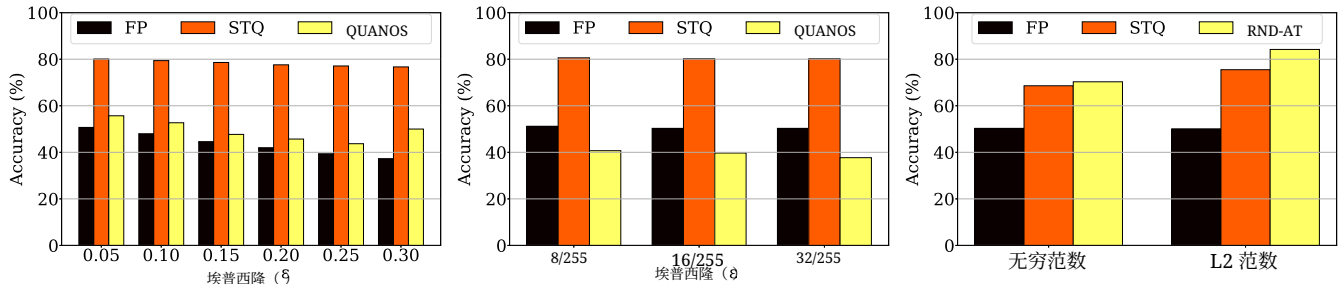
图 4：针对 CIFAR-10 数据集，模型在 Zoo 和 C&W-L2 攻击下的鲁棒性（Top-1 测试准确率）对比。

C. 针对对抗攻击的鲁棒性评估

表 IV 展示了我们的 STQ 模型针对两种白盒攻击（FGSM、PGD）和两种黑盒攻击（Square、Boundary）的鲁棒性（Top-1 测试准确率）。研究结果与两个 MLCommons/Tiny 基准进行了比较：用于图像分类、在 CIFAR-10 和 SVHN 数据集上训练的 ResNetv1；以及用于关键词识别、在 GSC 数据集上训练的 DS-CNN TFLite [56]。此外，图 4 展示了模型针对另一种白盒攻击（C&W-L2）和一种黑盒攻击（Zoo）的鲁棒性。研究结果与用于图像分类、在 CIFAR-10 数据集上训练的 ResNetv1 进行了比较。

如表 IV 所示，面对所有对抗攻击，两个 MLCommons/Tiny 基准的准确率都有所下降，因为其中没有采用任何防御机制，因此集成雅可比正则化的 STQ 是一种值得关注的方案。平均而言，对抗攻击分别导致 CIFAR-10、SVHN 和 GSC 数据集的准确率下降 15.5%、3.8% 和高达 74.1%。关于 C&W-L2 和 Zoo 攻击，从图 4 可以看出，ResNetv1 在 CIFAR-10 数据集上的平均准确率下降为 10.5 %。上述准确率下降都凸显了鲁棒微型模型的重要性。请注意，图 4 还展示了我们的 FP 模型的鲁棒性，其鲁棒性显然低于 STQ。

然而，在攻击存在的情况下，所提出的 STQ 模型的准确率要么有所提升，要么略有下降。对于 Square 和 C& W-L2 攻击，以及 CIFAR-10 数据集，观察到的最大准确率下降分别为 7.0% 和 5.6%，尽管这两种情况下模型都优于 ResNetv1 基准模型。值得注意的是，对于 SVHN 数据集上的 FGSM 和 PGD 扰动，ResNetv1 的表现仅比我们的 STQ 模型好 0.9%。最后，在所有实验中，对于 GSC 数据集的干净样本和对抗样本，我们的模型准确率均高于 DS-CNN 模型。



(a) FGSM (b) PGD (c) Square

图5：在不同 FGSM 和 PGD 扰动强度以及不同 Square 攻击变体下，我们的浮点模型和 STQ 模型与 Quanos [14] 及 RND-AT [52] 之间的鲁棒性比较。

这些结果表明，所提出的 STQ 模型在抵御多种对抗攻击方面的韧性和鲁棒性优于经过测试的 MLCommons/Tiny 基准模型。因此，将逐层JR集成到我们的 STQ 模型中，既能降低模型复杂度，又能增强模型的鲁棒性；与传统基准模型相比，该模型也展现出有效的性能。对于微控制器以及其他受内存和计算资源限制的微型设备而言，这一发现颇具意义。

表V展示了 K折交叉验证的方差报告，其中 $K = 10$。MLC/T 列表示 CIFAR-10 和 SVHN 数据集上的 ResNetv1 模型，以及 GSC 数据集上的 DS-CNN 模型。在 CIFAR-10 的许多样本和 GSC 数据集的所有实例中，STQ 模型基准测试的方差都较低。对于 SVHN，ResnetV1 模型的方差低于 STQ 模型。因此， K折交叉验证结果会随数据集和攻击类型而变化，尽管总体平均结果表明，采用类别交叉熵损失的 STQ 具有良好的一致性。尤其是在 GSC 数据集上，无论是干净样本还是受到攻击的样本，STQ 往往都优于 MLC/T。这些结果证明了 STQ 作为一种鲁棒且有效的模型所具备的性能能力，适合部署在资源受限的环境中。

最后，图5比较了我们的浮点模型和 STQ 模型在以下攻击下的表现：i) 在不同 FGSM 和 PGD 扰动强度下，与 QUANOS [14] 基线模型进行比较；ii) 在不同 Square [19] 攻击下，与 RND-AT [52] 进行比较。所有攻击均使用 CIFAR-10 数据集中的 10,000 个测试样本。如图所示，在检测不同强度的 FGSM 和 PGD 攻击时，STQ 明显优于 QUANOS；而在应对 Square 攻击时，其表现略逊于 RND-AT。我们没有找到可供公平比较的 Boundary 攻击相关先前研究，因此将 Zoo 和 C&W-L2 留待未来研究。

表V： K折交叉验证所得的方差。

数据集	程序	STQ	MLC/T
CIFAR-10	无攻击	**3.04**	26.25
	FGSM	**5.81**	6.01
	PGD	3.7	3.35
	Square	**11.78**	15.58
	Boundary	**1.4**	5.69
SVHN	无攻击	0.19	0.1
	FGSM	4.27	1.99
	PGD	1.64	1.40
	Square	**4.2**	5.36
	Boundary	**0.39**	0.45
GSC	无攻击	**0.37**	0.80
	FGSM	**1.57**	4.45
	PGD	**3.17**	5.45
	Square	**2.96**	3.64
	Boundary	**0.24**	0.34
	平均值	**2.98**	5.39

V. 结论

本文研究了深度量化机器学习（DQML）模型针对各种白盒和黑盒对抗攻击的鲁棒性。研究采用了 QKeras 的深度量化功能，以构建一种节省内存、准确且具备对抗鲁棒性的量化模型。这是因为该模型融合了与一种防御技术雅可比正则化（JR）相似的机制。如研究所示，所提出的随机三值量化（STQ）模型通过引入逐层 JR 的量化感知训练程序，在面对多种对抗攻击时，比业界采用的 MLCommons/Tiny 基准更具鲁棒性。事实上，该随机量化方案能够在支持对抗攻击鲁棒性的同时有效实现模型压缩。通过利用两个图像数据集（CIFAR-10 和 SVHN）以及一个音频数据集（GSC），在各种对抗攻击下观察模型的准确率，实验验证了这种鲁棒性。这对于在能力有限的资源受限环境中部署高效且有效的模型具有重要意义。我们的初步结果表明，有必要进一步探索具有不同攻击强度的其他复杂白盒和黑盒攻击。未来的工作将进一步评估所提出的 STQ QKeras模型应对新型及最新攻击的有效性。

致谢

本研究得到PETRAS物联网系统网络安全国家卓越中心的支持，该中心由英国工程与自然科学研究委员会根据资助编号 EP/S035362/1提供资助。

参考文献

[1] W. Liu、Z. Wang、X. Liu等，《深度神经网络架构及其应用综述》，Neurocomputing，第234卷，第11–26页，2017年。[2] J. Turner、J. Cano、V. Radu等，《深度卷积神经网络跨堆栈优化特性》，载于2018年IEEE国际工作负载特征研讨会（IISWC），2018年。[3] I. Rosenberg、A. Shabtai、Y. Elovici和L. Rokach，《网络安全领域中的对抗性机器学习攻击与防御方法》，ACM Comput. Surv.，第54卷，第5期，第108:1–108:36页，2021年。[4] Y. Lin、H. Zhao、X. Ma等，《卷积神经网络调制识别中的对抗攻击》，IEEE Trans. Reliab.，第70卷，第1期，第389–401页，2020年。[5] B. Sudharsan、J. G. Breslin、M. Tahir等，《OTA-TinyML：Tiny ML模型的空中部署及其在物联网设备上的执行》，IEEE Internet Comput.，第26卷，第3期，第69–78页，2022年。[6] R. David、J. Duke、A. Jain等，《TensorFlow Lite Micro：面向TinyML系统的嵌入式机器学习》，载于Machine Learning and Systems会议论文集，第3卷，2021年。[7] L. Geiger和P. Team，《Larq：用于训练二值化神经网络的开源库》，J. Open Source Softw.，第5卷，第45期，第1746页，2020年。[8] A. Pappalardo，《Xilinx/brevitas》，2022年。 [在线]。可访问：https://doi.org/10.5281/zenodo.3333552。[9] F. Loro、D. Pau和V. Tomaselli，《用于深度量化成像的QKeras神经网络Zoo》，载于RTSI国际论坛论文集，2021年。[10] L. N. C. Jr、A. Kuusela、S. Li等，《面向粒子探测器边缘端低延迟推理的深度神经网络自动异构量化》，Internet of Things，2021年。[11] Y. Dong、R. Ni、J. Li等，《通过随机量化学习精确的低比特深度神经网络》，arXiv，2017年。[12] S. H. Silva和P. Najafirad，《深度学习对抗鲁棒性中的机遇与挑战：综述》，arXiv，2020年。[13] J. Lin、C. Gan和S. Han，《防御性量化：效率与鲁棒性相遇之时》，arXiv，2019年。[14] P. Panda，《Quanos：由对抗噪声敏感性驱动的神经网络混合量化》，载于ACM/IEEE ISPLED论文集，2020年。[15] A. K. Sikder、G. Petracca、H. Aksu等，《智能设备及应用中的基于传感器的威胁与攻击综述》，IEEE Commun. Surv. Tutor.，第23卷，第2期，第1125–1159页，2021年。[16] J. Su、D. V. Vargas和K. Sakurai，《用于欺骗深度神经网络的单像素攻击》，IEEE Trans. Evol. Comput.，第23卷，第5期，2019年。[17] Y. Jiang、G. Yin、Y. Yuan和Q. Da，《针对多源遥感图像场景分类的投影梯度下降对抗攻击》，Security Commun. Netw.，第2021卷，2021年。[18] T. Huang、V. Menkovski、Y. Pei和M. Pechenizkiy，《弥合FGSM与PGD对抗训练之间的性能差距》，arXiv预印本arXiv:2011.05157，2020年。[19] M. Andriushchenko、F. Croce、N. Flammarion和M. Hein，《Square攻击：一种通过随机搜索实现查询高效的黑盒对抗攻击》，载于ECCV论文集，2020年。[20] H. Li、X. Xu、X. Zhang等，《QEBA：查询高效的基于边界的黑盒攻击》，载于IEEE/CVF CVPR论文集，2020年6月。[21] H. Xu、Y. Ma、H.-C. Liu等，《图像、图和文本中的对抗攻击与防御：综述》，Int. J. Autom. Comput.，第17卷，第151–178页，2020年。[22] G. Cohen、G. Sapiro和R. Giryes，《使用影响函数和最近邻检测对抗样本》，载于IEEE/CVF CVPR论文集，2020年6月，第14 453–14 462页。[23] S. Hussain、P. Neekhara、S. Dubnov等，《WaveGuard：理解和缓解音频对抗样本》，载于第30届USENIX安全研讨会论文集，2021年。[24] R. Bernhard，P.-A. Mo¨ellic和J.-M. Dutertre，《诱导深度神经网络中的可迁移对抗扰动》，载于2021年国际联合神经网络会议（IJCNN），2021年，第1–8页。[25] R. A. Khamis和A. Matrawy，《评估对抗训练对深度学习入侵检测系统中不同类型神经网络的影响》，载于ISNCC论文集。IEEE，2020年，第1–6页。[26] M. Picot、F. Messina、M. Boudiaf等，《通过Fisher-Rao正则化提升对抗鲁棒性》，IEEE Trans. Pattern Anal. Mach. Intell.，2022年。[27] D. Jakubovitz和R. Giryes，《使用雅可比正则化提升DNN对对抗攻击的鲁棒性》，载于ECCV论文集，2018年9月，第1–16页。

[28] J. Sokoli´c、R. Giryes、G. Sapiro 和 M. R. D. Rodrigues，《鲁棒大间隔深度神经网络》，IEEE信号处理汇刊，第65卷第16期，第4265–4280页，2017年。[29] A. Gholami、S. Kim、Z. Dong 等，《高效神经网络推理的量化方法综述》，arXiv:2103.13630，2021年。[30] D. Pau、A. Khiari 和 D. Denaro，《供水系统异常检测中的微型控制器在线学习》，载于ICCE会议论文集，第11卷。IEEE，2021年，第1–6页。[31] C. Song、E. Fallon 和 H. Li，《提升权重量化神经网络的对抗鲁棒性》，arXiv电子预印本，2020年。[32] A. Galloway、G. W. Taylor 和 M. Moussa，《攻击二值化神经网络》，载于国际学习表示会议（ICLR），2018年。[33] R. Bernhard、P.-A. Moellic 和 J.-M. Dutertre，《低位宽量化对嵌入式神经网络对抗鲁棒性的影响》，载于国际网络世界会议论文集，2019年。[34] Y. Kim、J. Lee、Y. Kim 和 J. Seo，《深度神经网络的鲁棒量化》，载于第29届国际编译器构造会议论文集，2020年，第74–84页。[35] M. Gorsline、J. Smith 和 C. Merkel，《量化神经网络的对抗鲁棒性》，载于GLSV论文集，2021年，第189–194页。[36] Google，《Keras框架：Python深度学习API》，2017年。 [在线]。可访问：https://keras.io/ [37] ——，《QKeras：Keras量化扩展》，2017年。 [在线]。可访问：https://github.com/google/qkeras [38] S. Zhou、Y. Wu、Z. Ni 等，《DoReFa-Net：使用低位宽梯度训练低位宽卷积神经网络》，arXiv预印本，第arXiv:1606.06160卷，2018年。[39] E. Shishkina，《B-调和函数的平均值定理》，Lobachevskii数学杂志，第43卷第6期，第1401–1407页，2022年。[40] Y. Guo、Y. Li、L. Wang 和 T. Rosing，《深度卷积足以实现多个视觉域的学习》，载于AAAI人工智能会议论文集，2019年。[41] E. Wang、J. J. Davis、D. Moro 等，《在边缘端实现二值神经网络训练》，载于第5届国际EMDL研讨会论文集，2021年。[42] K. Eckle 和 J. Schmidt-Hieber，《采用ReLU激活函数的深度网络与线性样条型方法的比较》，神经网络，第110卷，第232–242页，2019年。[43] N. Papernot、A. Thakurta、S. Song 等，《采用温度调节Sigmoid激活实现具有差分隐私的深度学习》，载于AAAI人工智能会议论文集，2021年。[44] L. Roeder，《Netron》，2022年。 [在线]。可访问：https://netron.app/ [45] A. Krizhevsky，《从微型图像中学习多层特征》，多伦多大学，加拿大安大略省多伦多，技术报告，2009年。[46] Y. Netzer、T. Wang、A. Coates 等，《通过无监督特征学习读取自然图像中的数字》，载于NIPS深度学习与无监督特征学习研讨会，2011年。[47] G. S. Commands，《Google语音命令数据集》，2016年。 [在线]。可访问：https://pyroomacoustics.readthedocs.io/en/pypi-release/pyroomacoustics.datasets.google speech commands.html [48] I. Loshchilov 和 F. Hutter，《SGDR：采用热重启的随机梯度下降》，载于ICLR，2017年。[49] N. Carlini 和 D. Wagner，《迈向神经网络鲁棒性的评估》，载于2017年IEEE安全与隐私研讨会（SP），2017年。[50] P.-Y. Chen、H. Zhang、Y. Sharma 等，《Zoo：无需训练替代模型的基于零阶优化的深度神经网络黑盒攻击》，载于第10届ACM人工智能与安全研讨会论文集，2017年，第15–26页。[51] M.-I. Nicolae、M. Sinn、M. N. Tran 等，《对抗鲁棒性工具箱v1.0.0》，arXiv预印本arXiv:1807.01069，2018年。[52] Z. Qin、Y. Fan、H. Zha 和 B. Wu，《针对基于查询的黑盒攻击的随机噪声防御》，神经信息处理系统进展，第34卷，第7650–7663页，2021年。[53] D. Berrar，《交叉验证》，载于《生物信息学与计算生物学百科全书》，2019年，第542–545页。[54] T.-T. Wong 和 P.-Y. Yeh，《通过K折交叉验证获得可靠的准确率估计》，IEEE知识汇刊，第32卷，1594，2019 8，1586，第期，第页。[55] 《Scikit-learn：机器学习Python API》，2019年。 [在线]。可访问：https://scikit-learn.org/stable/ [56] C. Banbury、V. J. Reddi、P. Torelli 等，《MLPerf Tiny基准测试》，arXiv预印本，第arXiv:2106.07597卷，2021年。