# 安全感知近似脉冲神经网络

Syed Tihaam Ahmad，Ayesha Siddique，Khaza Anuarul Hoque密苏里大学
电气工程与计算机科学系，美国密苏里州哥伦比亚市tawm9@umsystem.edu，
ayesha.siddique@mail.missouri.edu，hoquek@missouri.edu

arXiv:2301.05264v1 [cs.CR] 12 Jan 2023

摘要——**深度神经网络（DNNs）和脉冲神经网络（SNNs）都容易受到对抗攻击。因此，近年来研究人员广泛研究了对抗攻击下深度神经网络和脉冲神经网络的鲁棒性与防御方法。与精确脉冲神经网络（AccSNN）相比，近似脉冲神经网络（AxSNN）在超低功耗应用中的能效最高可提升 4 倍。然而，近似脉冲神经网络在对抗攻击下的鲁棒性尚未得到探索。本文首先全面分析了在两种基于梯度的攻击和两种神经形态攻击下，具有不同结构参数和近似级别的近似脉冲神经网络的鲁棒性。随后，我们提出两种新颖的防御方法，即精度缩放和近似量化感知滤波（AQF），以保护近似脉冲神经网络。我们使用静态数据集和神经形态数据集评估了这两种防御方法的有效性。结果表明，与精确脉冲神经网络相比，近似脉冲神经网络更容易受到对抗攻击，但精度缩放和 AQF 能显著提升其鲁棒性。例如，在静态 MNIST 数据集中，PGD攻击使近似脉冲神经网络的准确率相比未受攻击的精确脉冲神经网络降低 72%；而同一攻击对经过精度缩放的近似脉冲神经网络仅造成 17% 的准确率损失（鲁棒性提升 4 倍）。类似地，在神经形态 DVS128手势数据集中，稀疏攻击使近似脉冲神经网络的准确率相比未受攻击的精确脉冲神经网络降低 77%；而同一攻击对采用 AQF 的近似脉冲神经网络仅造成 2% 的准确率损失（鲁棒性提升 38 倍）。**

索引词—**脉冲神经网络、近似脉冲神经网络、对抗鲁棒性、近似防御。**

## I. 引言

脉冲神经网络（SNNs）是第三代神经网络，具备事件驱动计算能力 [1]。近年来，人们针对数据分析开发了许多不同规模的 SNN 模型，例如手势识别、目标检测和图像分类。然而，众所周知，大规模 SNN 模型比小规模模型能够识别更多特征。因此，最先进的大规模 SNN 模型包含大量参数，在训练和推理阶段都需要对其加以考虑。这限制了 SNN 在超低功耗、资源受限的边缘设备上的部署。为解决这一问题，SNN 中的近似计算近年来逐渐成为一种高能效解决方案。基于近似计算的近似脉冲神经网络（AxSNNs）在容错应用中放宽了近乎完美准确率的要求，从而实现较低的能耗。例如，通过近似权重得到的 AxSNN 可将能耗降低 4 倍 [2]，相比精确脉冲神经网络（AccSNN）而言。

与传统深度神经网络（DNNs）类似，近似脉冲神经网络也容易受到对抗攻击。对抗攻击以隐蔽性极强而著称：它们只向输入添加极小的扰动噪声，这些噪声对人眼不可察觉，却能成功欺骗脉冲神经网络分类器。[3]。受此启发，近期多项研究对 AccSNN 中的鲁棒性分析与对抗防御进行了深入探讨 [4]–[7]。最近， [8]，中的作者表明，与精确 DNN（AccDNN）相比，近似深度神经网络（AxDNNs）更容易受到对抗攻击。然而，令人意外的是，近似脉冲神经网络的鲁棒性尚未得到探索。因此，有必要开展更全面的研究，以了解近似脉冲神经网络与 AccSNN 的内在行为，尤其是在遭受对抗攻击时。探索这类行为有助于设计专门针对近似脉冲神经网络的防御技术。

### A. 动机案例研究与关键观察

作为激发我们研究动机的初步研究，我们开展了一项动机案例研究，重点分析对抗攻击对近似脉冲神经网络与 AccSNN 的影响。为此，我们首先训练了一个 5层精确脉冲神经网络，该网络包含 3 个卷积层和 2 个全连接层，用于对 MNIST[9] 数据集进行分类。随后，我们构建了一个 AxSNN（采用近似级别 0.1），作为 AccSNN 的近似对应模型。最后，我们通过改变扰动预算 $\epsilon$ （范围为 0 到 1.0），比较了 AccSNN 和 AxSNN 在基于 $l_\infty$ 范数的投影梯度下降（PGD）攻击下的性能。结果如图 1 所示。我们观察到，与 AccSNN 相比，AxSNN 在受到攻击时的鲁棒性显著较低。例如，在没有攻击（$\epsilon$=0）时，AccSNN 和 AxSNN 的准确率分别为 97% 和 52%。然而，对于($\epsilon$=0.5），其准确率为

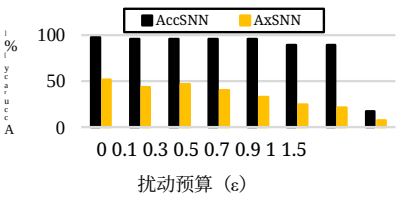



图1：在不同扰动预算下，AccSNN和AxSNN对PGD攻击的鲁棒性比较。

AccSNN 和 AxSNN 的准确率分别为 95% 和 40%，两者相差 55%。此外，当扰动预算调整为 1.0（$\epsilon$=1.0）时，比较 AccSNN 和 AxSNN 的准确率可以发现，两者相差 68%。这些结果促使我们深入研究近似脉冲神经网络的鲁棒性，并探索潜在的防御技术，以设计鲁棒的近似脉冲神经网络。

*B.* 新颖贡献

本文提出了一种安全感知型 AxSNN 设计方法，具体包含以下新颖贡献：

1）通过精度缩放识别其鲁棒性感知旋钮，提出了一种设计对抗鲁棒型 AxSNN 的新方法，即寻找近似级别、结构参数（阈值电压、时间步）和量化的适当组合。我们还提出了一种通过近似量化感知滤波（AQF）实现的防御方法，专门用于抵御神经形态攻击。

**[第 IV 节]**

2）在不同阈值电压、时间步、精度尺度和近似级别下，针对两种基于梯度的攻击和两种神经形态攻击，对 AxSNN 与 AccSNN 展开了广泛的脆弱性分析。具体而言，我们评估了投影梯度下降（PGD）和基本迭代方法（BIM）攻击对静态数据集 MNIST[9]上的 AccSNN 和 AxSNN 的影响。另一方面，我们评估了稀疏攻击和帧攻击对神经形态数据集 DVS128手势上的 AccSNN 和 AxSNN 的影响 [10]）。 **[第 V 节]**

我们的结果表明，AxSNN 比 AccSNN 更容易受到对抗攻击。然而，精度缩放和近似量化感知滤波显著提升了其鲁棒性。例如，在 MNIST 分类中，对 AxSNN 施加扰动预算为 1.0 的 PGD攻击会导致 72% 的准确率损失，而对 AccSNN 施加相同攻击仅导致 9% 的准确率损失。有趣的是，采用精度缩放的相同 AxSNN 仅出现 17% 的准确率损失，表明其鲁棒性提升了 4 倍。同样，在 DVS128 手势分类中，对 AxSNN 进行稀疏攻击会导致 77% 的准确率损失，而 AccSNN 在未受到任何攻击时则不会出现这种损失。然而，使用我们设计的近似量化感知滤波器后，准确率损失仅为 2%，表明鲁棒性提升了 38 倍。

## II. 预备知识

本节简要介绍脉冲神经网络和对抗攻击，以帮助读者理解本文。

**近似脉冲神经网络：** AxSNN 采用近似计算，在超低功耗应用中以分类准确率换取能效。AxSNN 通常为每个脉冲神经元指定一个近似级别 $a_{th}$。 $a_{th}$ 根据神经元对误差和脉冲活动 [11]的敏感性，决定是否激活或停用相应的神经元。与 AccSNN 类似，AxSNN 使用标准的漏积分发放（LIF）模型：当膜电位超过阈值电压时，神经元会发出输出脉冲，并重置其膜电位。具体而言，它们处理脉冲编码输入，这些输入通常采用频率编码；其中，激活活动对应于一定时间步内的平均脉冲发放率。时间步指的是脉冲神经网络接收相同输入的观测周期。

**对抗攻击：** 对抗攻击是指通过微小扰动使分类器预测出错误标签。此类攻击包括基于梯度的投影梯度下降（PGD）和 BIM，它们是对抗机器学习领域中威力强大的攻击。基于梯度的攻击利用反向传播的概念；然而，它们并非计算模型权重的梯度，而是通过扰动输入图像来构造对抗样本。近期研究表明，由于神经形态数据集具有事件驱动特性，这些攻击无法用于扰动神经形态数据集。因此，人们会采用专门的神经形态攻击，例如稀疏攻击和帧攻击 [6]。稀疏攻击是一种隐蔽攻击，它基于输出标签概率的损失函数，迭代地扰动神经形态图像以生成扰动事件。另一方面，帧攻击是一种简单而有效的神经形态攻击，它通过攻击所有事件边界上的每个像素来生成扰动事件。

## III. 威胁模型

本节提出一种威胁模型，用于探索近似脉冲神经网络的对抗鲁棒性。

**对手知识：** 我们假设对手使用准确的分类器模型来制作对抗样本。此外，对手对 AxSNN 只有部分了解，即对手知道分类器模型的内部架构，但不知道其近似误差，也不知道阈值电压、时间步长、精度尺度和近似级别等模型参数。

**攻击生成：** 假设对手能够在预测阶段篡改输入图像以逃避分类器模型，同时不影响训练数据。对手通过寻找能够最大化模型在输入上的损失、同时使扰动幅度低于扰动预算 $\epsilon$ 的扰动，来制作对抗样本。如前所述，我们具体采用迭代式基于梯度的攻击，即基于 $l_\infty$ 范数的 *BIM* 和 *PGD*，这些被认为是针对静态数据集的高强度攻击。我们还采用神经形态攻击，具体包括稀疏攻击和帧攻击；这些攻击隐蔽性强，却能有效扰动高分辨率的神经形态图像。本文使用 $\epsilon$=1.0，因为超过这一值后，AccSNN 和 AxSNN 的准确率都会显著下降，且无法恢复。例如，当 $\epsilon$=1.5 时，准确率会降至 10%。

## IV. 安全感知型 AxSNN 设计方法

OACH

本节将详细讨论我们提出的鲁棒、安全 AxSNN 设计方法。

*A.* 精度缩放

传统上，近似级别是通过确定神经网络中的最大可容忍扰动（MTP）来确定的 [12]。然而，由于阈值电压、时间步和精度尺度的变化会导致准确率波动，因此在精度缩放的近似脉冲神经网络中，这一过程颇具挑战性。直观而言，增加时间步数和阈值电压可能会使某些神经元产生更多无关紧要的脉冲，进而影响分类准确率。跳过此类神经元有望提高近似脉冲神经网络的准确率；但在攻击下，其鲁棒性可能会下降。由于精度缩放有望提升精确深度神经网络的鲁棒性 [12]，因此依据近似脉冲神经网络中的精度尺度确定近似级别，有助于提升对抗鲁棒性。然而，除阈值电压和时间步外，还要探索精度尺度以实现鲁棒近似，这似乎颇具挑战。本文通过以下方程确定近似脉冲神经网络中的近似级别 $a_{th}$：

$$a_{th} = (cN_s/T) \cdot min(1, V_m/V_{th}) \cdot \sum_{i=1}^{c} w_i^p, \qquad (1)$$

其中，$c$、$N_s$、$T$、$V_m$、$V_{th}$ 和 $w^p$ 分别表示神经元的输出连接数、脉冲数、时间步、膜电位、阈值电压和精度缩放权重。此外，$min(1, V_m/V_{th})$ 表示脉冲概率。当 $V_m$ 超过 $V_{th}$时，最大脉冲概率为 1；否则为 $V_m/V_{th}$。脉冲概率由与连接对应的所有权重的均值加权 $c$ 即 $\sum_{i=1}^{c} w_i^p$，其中包含权重的精度缩放。

算法1概述了在 AxSNN 中实现该鲁棒性方程所涉及的步骤。首先，我们初始化一个用于成功生成攻击的计数器 $adv$ （第 1 行）。然后，使用给定的阈值电压 $v$ 和时间步 $t$ 训练一个近似脉冲神经网络模型，并保存训练后的模型以及每一层 l 对应的权重（第 3 行）。通过质量约束 $Q$ 对这一学习阶段进行定量验证，即低于该最低基准准确率时，我们认为脉冲神经网络的学习效率低下（第 4 行）。$Q$ 的取值取决于给定的脉冲神经网络架构、数据集和应用。接着，我们针对对抗攻击和扰动预算 $\epsilon$ 构造对抗样本（第 5 行）。通过精度缩放进行对抗防御始于第 6 行：首先按升序排列训练后的权重。这种排序有助于之后根据权重的重要性对其进行近似处理。随后，我们依据每个精度尺度 $s$ 执行精度缩放，并计算层 $l$中所有连接的均值（第 8—9 行）。利用该均值，我们为每一层 $l$ 确定 $a_{th}$ （如前所述），并通过移除权重低于$a_{th}$ 的连接来近似精度缩放后的模型（第 10 行）。如果输入的是神经形态数据集，且相应标志 $F_d$ 为高，则我们还会使用算法 2 中的特殊近似量化感知掩码滤波器（AQF），如第 IV-B 节所述（第 12—14 行）。接下来，算法会检查构造的对抗样本是否能够欺骗 AxSNN，即攻击是否成功迫使模型输出错误标签；如果成功，则相应地递增对抗成功计数器（第 15—18 行）。最后，我们将扰动预算 $\epsilon$ 下的鲁棒性评估为攻击未能生成可欺骗受攻击的脉冲神经网络的对抗样本的比例（第 21 行）。本文使用该算法，在阈值电压、时间步和精度尺度构成的鲁棒集合中，寻找近似级别 $a_{th}$ 。因此，我们比较精度缩放后的 AxSNN 模型在所有这些参数下的准确率，并返回满足质量约束的参数值（第 22—24 行）。

**算法1：** 近似脉冲神经网络中的精度缩放

**输入：** 对抗攻击类型：$attack$;
扰动预算：$\epsilon$;
时间步：$T = [t_1, t_2, ..., t_n]$;
阈值电压：$V_{th} = [v_1, v_2, ...v_n]$;
训练数据集：$\mathcal{D}_{tr}$ = ($X$, $L$);
测试数据集：$\mathcal{D}_{ts}$ = ($x$, $l$);
扰动预算：$\epsilon$;
精度缩放级别：$s_l = [s_1, s_2, ..., s_n]$;
质量约束：$Q$;
神经形态数据集标志：$F_d$
**输出：** 鲁棒性级别 $R$，最佳 $V_{th}$，时间步 $ts$，
近似级别 $a_{th}$ 和精度缩放级别 $s$

1: $adv$ = 0;
2: **对于每个** ($v$, $ts$) in ($V_{th}$, $T$) **执行**
3: ($model$, $w_l$) = trainAccurateSNN ($v$, $ts$, $\mathcal{D}_{tr}$)
4: **如果**准确率 ($model$) $\geq Q$ **则**
5: ($x_k^*$, $l_k^*$) = AdvExGen ($model$, $\epsilon$, $attack$, $X_k^t$)
6: $w$ =SortInAscendingOrder ($w_l$);
7: **对每个** $s$在 $s_l$ **中执行**
8: $w^p$ =精度缩放 ($w$, s);
9: $m_l^c$ = $\sum_{j=1}^{c}(w_j^p)$;
10: $a_{th}$ = ($cN_s/T$) $m_l^c$ · 分钟 (1, $V_m$ / $v$)
11: $model$ = 近似脉冲神经网络($model$, $w^p$, $m_l^c$ $a_{th}$);
12: **如果** $T_d$ 为 TRUE **则**
13: $\mathcal{D}_{ts}$ = ApproximateQuantizedFilter($q_t$, $\mathcal{D}_{ts}$)
14: **结束如果**
15: ($x'_k$, $l'_k$) = 对抗攻击 ($model$, $\epsilon$, $attack$, $x_k^*$, $l_k^*$)
16: **如果** $l'_k \neq l_k$ **则**
17: adv++;
18: **else**
19: NOP;
20: **结束如果**
21: $R$ ($\epsilon_i$) = (1 - $adv$/ size($\mathcal{D}_{ts}$)) * 100;
22: **如果** $R \geq Q$ **则**
23: **返回** ($R$, $V_{th}$, $T$, $s$, $a_{th}$)
24: **结束如果**
25: **结束循环**
26: **结束如果**
27: **结束循环**

*B.* 近似量化感知滤波

对于由动态视觉传感器（DVS）提供输入的脉冲神经网络，由于神经形态图像具有事件驱动特性，上文讨论的针对基于帧的传感器的防御技术无法直接应用。因此，我们提出一种额外的近似量化感知滤波器（AQF），用于从神经形态图像中去除不相关事件。我们提出的算法 2 从事件驱动神经形态数据集 $E$ 中去除对抗扰动噪声。神经形态数据集 $E$ 表示为 ($x$, $y$, $p$, $t$) 的形式，其中 $x$、$y$、$p$ 和 $t$ 分别表示事件的 x 坐标、y 坐标、极性

和时间戳 $E$。事件 $e$ 与时空域相关联，因此彼此相关。我们的算法计算事件 $e$之间的相关性。如果事件 $e$ 的相关性低于某些时空阈值（$s$、$T1$、$T2$），则将其移除，因为相关性极低的这些事件更可能是由对抗扰动导致的噪声。

**算法 2：** ApproximateQuantizedFilter

**输入：** 事件列表：$D_{ts}(x, y, p, t) = Events$
量化步长：$qt$;
**输出：** 滤波后的量化数据集 $D_q$

```
1:  M = 0
2:  activity = 0, s = 2, T1 = 5, T2 = 50
3:  对于 e 在 Events 执行
4:      e =四舍五入(e/q_t) · q_t
5:      对于 i 在 (x_e − s, x_e + s)) 执行
6:          对于 j 中的 (y_e − s, y_e + s) ) 执行
7:              如果 非(i == x_e 且 j == y_e) 则
8:                  M[i][j] = t_e
9:              结束如果
10:             如果 非(i == x_e 且 j == y_e) 则
11:                 activity[i][j]+ = 1
12:             结束如果
13:         结束循环
14:     结束循环
15:     如果 activity[i][j] >T1 则
16:         M[i][j] = 1
17:     结束条件判断
18:     如果 t_e − M[x_e][y_e] > T2 或 M[x_e][y_e] == 1) 则
19:         从 Events中移除 e
20:     结束 if
21: 结束 for
22: D_q = Events
23: 返回 D_q
```

算法 2 描述了对神经形态数据集中的扰动事件进行近似量化感知滤波的步骤。首先，使用固定的量化步长 $q_t$ 对数据集进行量化（第 4 行），然后检查数据集每个事件 $e$ 中的不相关值（第 5–9 行）。每当相关性较低时，就会增加计数器变量 $activity$ 并将其标记（第 10–16 行）。最后，从数据集中移除已标记的不相关值事件 $e$ （第 18–20 行），以获得量化和滤波后的数据集 $D_q$ （第 23 行）。

## V. 结果与讨论

本节介绍我们对近似脉冲神经网络的对抗脆弱性分析与对抗防御的结果。

### A. 数据集与架构

我们使用静态数据集和神经形态数据集，即 MNIST[9] 和 DVS128手势 [10]。这两个数据集都常用于评估脉冲神经网络 [13] 在嵌入式平台上的性能。对于 MNIST 数据集，我们采用了一个 7 层脉冲神经网络作为分类器，其中包含三个卷积层、两个池化层和两个全连接层。该架构在干净输入（未受到任何对抗攻击）上的测试准确率为 97%。另一方面，对于 DVS128手势数据集，我们采用了一个 8 层脉冲神经网络作为分类器，其中包含两个卷积层、两个全连接层、三个池化层和一个丢弃层。该架构在干净输入上的测试准确率为 92%。

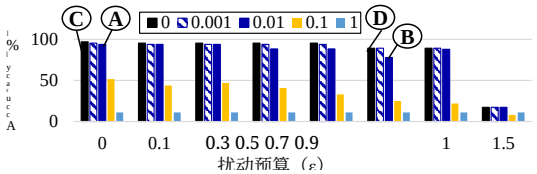



图2：AccSNN（近似级别为 0）和 AxSNN MNIST分类器在针对近似级别为 0.001、0.01、0.1 和 1 的 PGD攻击下的鲁棒性分析。

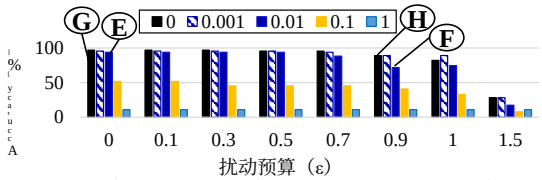



图3：AccSNN（近似级别为 0）和 AxSNN MNIST分类器在针对近似级别为 0.001、0.01、0.1 和 1 的 BIM攻击下的鲁棒性分析。

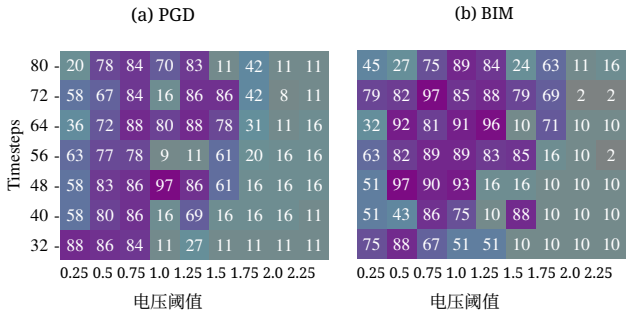



图4：MNIST 上 AxSNN（近似级别 0.01，精度尺度 FP32）遭受攻击（$\epsilon$ = 1）时的准确率。

*1)* 近似脉冲神经网络的脆弱性分析：我们首先通过改变近似级别，并在静态 MNIST数据集上针对不同扰动预算将近似脉冲神经网络与 AccSNN 进行比较，评估近似脉冲神经网络的对抗鲁棒性。在本实验中，我们将阈值电压和时间步分别固定为 0.25 和 32。图2和图3中的结果表明，近似脉冲神经网络的鲁棒性会随着扰动预算的增加而降低。例如，PGD攻击（$\epsilon$ = 0.9）会使近似级别为 0.01 的近似脉冲神经网络的准确率从 93%（$\epsilon$ = 0）降至 77%（见图2中的标签 A 和 B），而相同的攻击会使 AccSNN 的准确率从 96% 降至 89%（见图2中的标签 C 和 D）。这表明，AccSNN 在 PGD攻击下的准确率损失为 7%，而相同的攻击会使近似脉冲神经网络的准确率损失达到 16%。在这种情况下，近似脉冲神经网络比 AccSNN 脆弱 2 倍。同样，BIM攻击（$\epsilon$ = 0.9）会使近似级别为 0.01 的近似脉冲神经网络的准确率从 93%（$\epsilon$ = 0）降至 71%（见图3中的标签 E 和 F），而相同的攻击会使 AccSNN 的准确率从 96% 降至 82%（见图3中的标签 G 和 H）。这表明，AccSNN 的准确率下降了 14%，而近似脉冲神经网络在 PGD攻击下的准确率下降约 22%。在这种情况下，近似脉冲神经网络比 AccSNN 脆弱 1.5 倍。

此外，我们观察到，近似脉冲神经网络的鲁棒性会随着近似级别的变化而变化。例如，

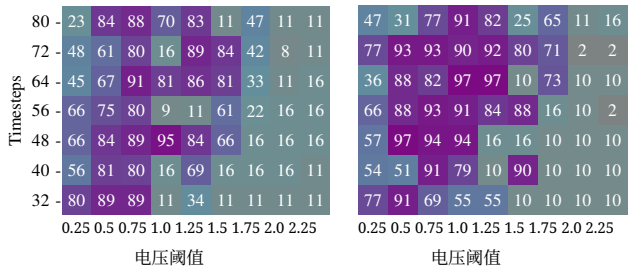



图5：攻击下（$\epsilon$ = 1）MNIST 上 AxSNN 的准确率（近似等级 0.01，精度范围 FP16）。（a）PGD

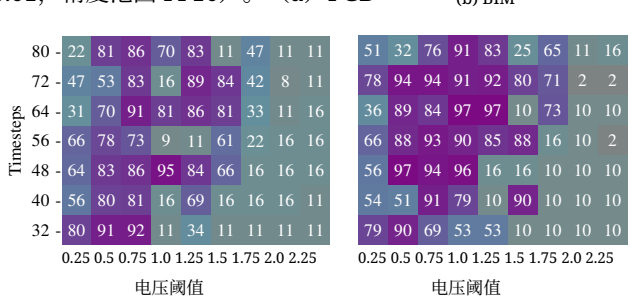



图6：MNIST 上 AxSNN（近似等级 0.01，精度缩放 INT8）在攻击（$\epsilon$ = 1）下的准确率。

在无攻击（$\epsilon$=0）的情况下，当我们将近似级别从0调整为0.01、0.1和1.0时，近似脉冲神经网络的准确率分别从96%下降至93%、51%和10%。相反，在PGD攻击（$\epsilon$=0.9）下（见图2），当我们将近似级别从0调整为0.01、0.1和1.0时，近似脉冲神经网络的准确率分别从96%（$\epsilon$=0）下降至77%、25%和10%。对于BIM攻击，也观察到了类似趋势（见图3）。

接下来，我们使用神经形态 DVS128手势数据集，探索 AxSNN 在稀疏攻击和帧攻击下的脆弱性。在本实验中，我们将阈值电压和时间步分别固定为 1.0 和 80。结果表明，AxSNN 容易受到神经形态攻击。例如，AxSNN 的准确率在稀疏攻击和帧攻击下分别从 92% 降至 12% 和 10%；如图7所示，AccSNN 也出现了几乎相同的情况。

*2）基于精度缩放的对抗防御*：为了评估针对静态 MNIST 数据集的 AxSNN 基于精度缩放的对抗防御，我们首先从近似级别为 0.01，并将精度缩放尺度分别设为 INT8、FP16 和 FP32。具体而言，我们通过比较图 4、图 5 和图 6 中各精度尺度对应的准确率与图 7a 中 MNIST 分类的基础模型准确率（AccSNN），分析各精度缩放后的 AxSNN 对 PGD 和 BIM 攻击的鲁棒性，扰动预算为（$\epsilon$=1.0）。我们使用 $\epsilon$=1.0，因为在超过该值后，AccSNN 和 AxSNN 的准确率都会显著下降，且无法恢复。例如，当 $\epsilon$=1.5 时，准确率会降至 10%。

此外，我们观察到，精度缩放近似脉冲神经网络的鲁棒性会随着阈值电压、时间步、精度缩放

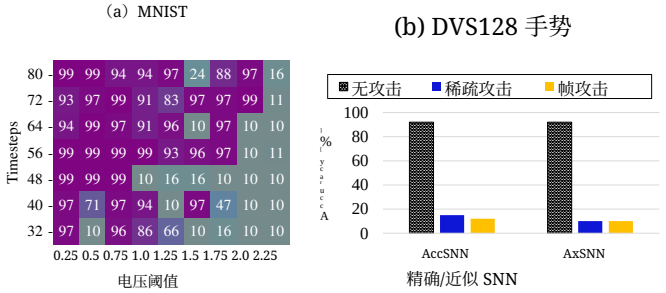



图 7：MNIST 上 AccSNN 在无攻击（$\epsilon$=0）条件下的准确率，以及 DVS128 手势上 AccSNN 和 AxSNN 在有攻击和无攻击条件下的准确率。

尺度和近似级别等参数的变化而变化。例如，与 AccSNN 相比（见图 7a），在近似级别为 0.01、精度尺度为 FP32、阈值电压为 0.75、时间步为 32 的情况下，PGD攻击（$\epsilon$ = 1.0）会使精度缩放 AxSNN 的准确率损失 12%。然而，将精度尺度改为 FP16 和 INT8 后，在相同的 AxSNN 中，与 AccSNN 相比，准确率损失分别为 7%（见图 5a）和 4%（见图 6a）。同样地，将阈值电压和时间步分别改为 1.0 和 48 后，AxSNN 的准确率可恢复至 95%。类似地，与 AccSNN 相比（见图 7a ），在近似级别为 0.01、精度缩放为 FP32、阈值电压为 0.5、时间步为 72 的情况下，BIM攻击（$\epsilon$ = 1.0）会使精度缩放 AxSNN 的准确率损失 15%。然而，将精度尺度改为 FP16 和 INT8 后，在相同的 AxSNN 中，与 AccSNN 相比，准确率损失分别为 4%（见图 5a）和 3%（见图 6a）。同样地，将阈值电压和时间步分别改为 1.25 和 64 后，AxSNN 的准确率可恢复至 97%。这充分凸显了使用我们提出的算法1的重要性，即在攻击条件下，为最合适的阈值电压、时间步、近似级别和精度尺度寻找稳健的最佳平衡点。需要注意的是，在阈值电压非常高（即大于 1.75）且近似级别较高（即大于 0.1）时，通常会观察到一些偏离上述鲁棒性趋势的行为。这并不令人意外，因为对于非常高的阈值电压和较高的近似级别，脉冲神经网络通常缺乏性能效率：触发漏积分发放神经元的脉冲可能无法越过阈值电压，从而导致整个网络的性能受损。

我们使用阈值电压 0.25 至 2.25（间隔为 0.25）、时间步 32 至 80（间隔为 8），以及 Int8、FP16 和 FP32 精度尺度，测试所提出的算法1。如表I所示，即使在 PGD 和 BIM 攻击下，我们的算法1也能找出最佳参数配置。例如，即使在 PGD攻击（$\epsilon$ = 1.0）下，我们的算法仍能确定：对于精度缩放的 AxSNN，当精度尺度为 FP32、阈值电压为 1.0、时间步为 48 时，近似级别为 0.01，准确率可达 97%。然而，对于精度缩放的 AxSNN，当精度尺度为 INT8、阈值电压为 0.75、时间步为 32 时，它确定的近似级别为 0.011，准确率为

表I：所提出的精度缩放 AxSNN 分类器在 MNIST 数据集上的最佳鲁棒性设置。

| $(V_{th},\ T)$ | 攻击 | $(q_{th},\ a_{th})$ | 准确率[%] |
|---|---|---|---|
| (0.25,32) | PGD | (FP32, 0.01) | 88 |
| | BIM | (INT8, 0.009) | 80 |
| (0.75,32) | PGD | (INT8, 0.011) | 92 |
| | BIM | (FP16, 0.013) | 91 |
| (1.0,48) | PGD | (FP32, 0.01) | **97** |
| | BIM | (INT8, 0.0125) | **96** |

表 II：在 AxSNN ($V_{th}$, $T$) = (1.0, 80) 中使用 DVS128手势和 AQF滤波后，与基准准确率相比的恢复准确率 $A_r$ 及准确率损失 $A_l$ 。

| 攻击 | $(q_t,\ a_{th})$ | $A_r$[%] | $A_l$[%] |
|---|---|---|---|
| 稀疏攻击 | (0.015, 0.1) | **90.01** | **2.0** |
| | (0.01, 0.15) | 88.4 | 3.6 |
| | (0.0, 0.001) | 84.3 | 7.7 |
| 帧攻击 | (0.015, 0.1) | **91.1** | **1.0** |
| | (0.01, 0.15) | 89.9 | 2.1 |
| | (0.0, 0.001) | 88.2 | 3.8 |

在相同攻击下为 88%。值得一提的是，AxSNN 的鲁棒性会随精度尺度而变化，因此，对于阈值电压、时间步和精度尺度的不同组合，我们的算法会给出不同的近似级别。例如，在静态MNIST数据集上，对近似级别为 0.1、时间步为 32、阈值电压为 0.25 的 AxSNN 发起 PGD攻击（$\epsilon = 1.0$），会导致准确率损失 72%；而对近似级别为 0.01、精度尺度为 Int8 的精度缩放 AxSNN 发起相同攻击，准确率损失仅为 17%（鲁棒性提升 4 倍）。BIM攻击也呈现出类似趋势。

*3) 基于 AQF 的对抗防御*：为评估精度缩放 AxSNN 中所提出的基于 AQF 的对抗防御的有效性，我们使用算法 2，在稀疏攻击和帧攻击下进行 DVS128 手势分类。不幸的是，训练近似脉冲神经网络需要很长时间；因此，我们仅使用阈值电压为 1.0、时间步为 80 的设置，测试所提出的算法 2。需要注意的是，这些参数设置是脉冲神经网络研究领域针对神经形态数据集最常用的设置 [14]。表 II 列出了我们的算法针对精度尺度 0.015 和 0.01 所确定的近似级别。我们观察到，与未受到任何攻击的 AccSNN 相比，稀疏攻击会使 AxSNN 的准确率损失 77%。然而，精度缩放 AxSNN 中的基于 AQF 的对抗防御能够将准确率恢复到接近基准准确率的水平。例如，在精度缩放 AxSNN 中结合使用精度尺度 0.015、近似级别 0.1 的基于 AQF 的对抗防御，并在稀疏攻击下进行测试，可将准确率恢复到近 90%，仅比基准准确率低 2%。有趣的是，与 AccSNN 相比，帧攻击会使 AxSNN 的准确率损失 82%。然而，在精度缩放 AxSNN 中采用精度尺度 0.015、近似级别 0.1 的基于 AQF 的对抗防御，并在帧攻击下进行测试，可将准确率恢复到近 91%，仅比基准准确率低 1%。鲁棒性之所以得到如此显著的提升，是因为 AQF 会屏蔽受扰神经形态数据集中彼此相关性较低的噪声事件。

## VI. 结论

本文深入探讨了近似脉冲神经网络的对抗鲁棒性，并提出了两种用于设计对抗鲁棒型 AxSNN 的新型防御方法：精度缩放和近似量化感知滤波（AQF）。为展示这些防御方法的有效性，我们采用了静态 MNIST 和神经形态 DVS128手势数据集。结果表明，与 AccSNN 相比，近似脉冲神经网络更容易受到对抗攻击，但精度缩放和 AQF 能够显著提升近似脉冲神经网络的鲁棒性。例如，与未受到任何攻击的 AccSNN 相比，AxSNN 在 PGD攻击下会导致 72% 的准确率损失；而在静态 MNIST 数据集中，同样的攻击对精度缩放 AxSNN 仅造成 17% 的准确率损失（鲁棒性提升 4 倍）。同样地，与未受到任何攻击的 AccSNN 相比，AxSNN 在稀疏攻击下会导致 77% 的准确率损失；而在神经形态 DVS128手势数据集中，同样的攻击作用于采用 AQF 的 AxSNN 时，仅造成 2% 的准确率损失（鲁棒性提升 38 倍）。

## VII. 致谢

本研究部分获得美国海军研究实验室通过项目编号 N0017321C2016 提供的资助。本出版物中表达的任何观点、研究结果、结论或建议均为作者个人意见，不一定反映美国政府或其相关机构的立场。

## 参考文献

[1] A. Viale 等人，“Carsnn：一种在 Loihi 神经形态研究处理器上用于基于事件的自动驾驶汽车的高效脉冲神经网络”，载于国际神经网络联合会议。电气电子工程师学会，2021 年，第 1–10 页。[2] S. Sen 等人，“脉冲神经网络的近似计算”，载于设计、自动化与测试欧洲会议。电气电子工程师学会，2017 年，第 193–198 页。[3] A. Marchisio 等人，“R-snn：一种分析与设计方法，通过动态视觉传感器的噪声滤波器增强脉冲神经网络抵御对抗攻击的鲁棒性”，载于智能机器人与系统国际会议。电气电子工程师学会，2021 年，第 6315–6321 页。[4] A. Bhattacharjee 等人，“研究脉冲神经网络在非理想忆阻器交叉阵列上的鲁棒性”，载于低功耗电子与设计国际研讨会，2022 年，第 1–6 页。[5] O. Nomura 等人，“基于首次脉冲时间编码的脉冲神经网络抵御对抗攻击的鲁棒性”，*IEEE*电路与系统汇刊第二辑：简报，第 69 卷，第 9 期，第 3640–3644 页，2022 年。[6] A. Marchisio 等人，“Dvs-attacks：针对脉冲神经网络的动态视觉传感器对抗攻击”，载于国际神经网络联合会议。电气电子工程师学会，2021 年，第 1–9 页。[7] R. El-Allami 等人，“通过固有结构参数保护深度脉冲神经网络免受对抗攻击”，载于设计、自动化与测试欧洲会议。电气电子工程师学会，2021 年，第 774–779 页。[8] A. Siddique 等人，“近似计算是否能普遍防御深度神经网络中的对抗攻击？”，载于设计、自动化与测试欧洲会议。电气电子工程师学会，2022 年，第 364–369 页。[9] Y. LeCun 等人，“MNIST手写数字数据库”，*ATT*实验室 *[在线]*。网址：*http://yann.lecun.com/exdb/mnist*，第 2 卷，2010 年。[10] A. Amir 等人，“一种低功耗、完全基于事件的手势识别系统”，载于计算机视觉与模式识别会议，2017 年，第 7243–7252 页。[11] S. Venkataramani 等人，“Axnn：采用近似计算的高能效神经形态系统”，载于低功耗电子与设计国际研讨会。电气电子工程师学会，2014 年，第 27–32 页。[12] F. Khalid 等人，“Qusecnets：基于量化的防御机制，用于保护深度神经网络免受对抗攻击”，载于测试技术与系统国际研讨会。电气电子工程师学会，2019 年，第 182–187 页。[13] R. V. W. Putra 和 M. A. Shafique，“Q-spinn：脉冲神经网络量化框架”，国际神经网络联合会议，第 1–8 页，2021 年。[14] E. P. Frady 等人，“利用英特尔 Pohoiki Springs 进行神经形态最近邻搜索”，载于神经形态与类脑计算国际会议，2020 年，第 1–10 页。