RescueSNN：在永久性故障下实现脉冲神经网络加速器的可靠执行

Rachmad Vidya Wicaksana Putra [1,*]、**Muhammad Abdullah Hanif** [2] **和 Muhammad Shafique** [2]

[1] 嵌入式计算系统，计算机工程研究所，维也纳工业大学（*TU Wien*），奥地利维也纳[2] *eBrain*实验室，工程系，纽约大学阿布扎比分校（*NYUAD*），阿拉伯联合酋长国阿布扎比

通讯作者*：Rachmad Vidya Wicaksana Putra rachmad.putra@tuwien.ac.at

arXiv:2304.04041v1 [cs.NE] 8 Apr 2023

摘要

为了最大限度地提升资源受限嵌入式系统上脉冲神经网络（脉冲神经网络）处理的性能和能耗效率，通常会采用专用硬件加速器/芯片。然而，这些SNN芯片可能遭受永久性故障，从而影响权重存储器的功能和神经元行为，进而导致准确率显著下降以及系统故障。这类永久性故障可能源于制造工艺期间的制造缺陷，和/或运行时操作期间的器件/晶体管损坏（例如由于磨损）。然而，SNN芯片中永久性故障的影响及相应的缓解技术尚未得到充分研究。为此，我们提出了RescueSNN，这是一种用于缓解SNN芯片计算引擎中永久性故障的新型方法，无需额外的重新训练，从而在保持吞吐量和质量的同时，显著减少设计阶段和重新训练成本。RescueSNN 方法的核心思想包括：（1）分析永久性故障下SNN的特征；（2）利用该分析，通过有效的故障感知映射（FAM）提升SNN的容错能力；以及（3）设计轻量级硬件增强，以支持FAM。我们的FAM技术利用SNN计算引擎的故障图来：（i）在将权重位映射到故障存储单元时，最大限度地减少权重损坏；以及（ii）在考虑脉冲神经网络操作和处理数据流的同时，有选择地使用不会导致准确率显著下降的故障神经元，以保持准确率和吞吐量。实验结果表明，与在未进行缓解的故障芯片上运行脉冲神经网络相比，在高故障率（例如，电势故障位置的0.5）下，我们的RescueSNN可将准确率提升最高达80%，同时将吞吐量下降保持在25%以下。通过这种方式，采用RescueSNN增强型芯片的嵌入式系统能够在其运行寿命期间，有效确保抵御永久性故障的可靠运行。

关键词：脉冲神经网络、加速器、容错能力、制造缺陷、可靠性、弹性、永久性故障。

1 引言

近年来，由于采用稀疏脉冲操作，脉冲神经网络有望以超低功耗/能耗实现高准确率（Putra 和 Shafique，2020年）。此外，脉冲神经网络可以利用脉冲时序依赖可塑性，在无标签数据上进行无监督学习；这在现实应用（如无人机和机器人技术中的自主智能体）中备受期待，原因有二：这类系统通常会遇到不可预见的场景（Putra 和 Shafique，2021b年，2022a年）；而且，收集无标签数据比收集有标签数据更便宜（Rathi 等人，2019年）。支持无监督学习的SNN架构如图 1(a)所示。为最大限度地提升脉冲神经网络处理的性能和能效，会采用专用的脉冲神经网络加速器/芯片（Painkras 等人，2013年；Akopyan等人，2015年；Davies等人，2018年；Frenkel等人，2019年）。然而，这些脉冲神经网络芯片可能会遭遇永久性故障，故障可能发生于以下阶段：(1) 芯片制造工艺 由于制造缺陷导致故障，因为制造由数百万至数十亿个纳米级晶体管构成且功能100%正确的脉冲神经网络芯片十分困难，而激进的技术尺度缩小又会进一步加剧这一问题（Zhang等人，2018年；Hanif等人，2018年；Hanif等人，2021年）；以及 (2) 运行阶段操作 由于器件/晶体管老化和损坏所致，这些问题由热载流子注入（HCI）、偏置温度不稳定性（BTI）、电迁移或时间相关介质击穿（TDDB）引起（Radetzki等人，2013年；Werner等人，2016年；Hanif等人，2018年；Baloch等人，2019年；Mercier等人，2020年；Hanif等人，2021年）。

永久性故障会影响脉冲神经网络加速器/芯片计算引擎的功能，包括本地权重存储器/寄存器和神经元，并通过破坏权重值和神经元行为（即膜电位动态和脉冲生成）降低准确率，如图 2 所示。例如，永久性故障会导致权重值固定为0或1，从而改变权重值，如图 1(b) 所示。在运行阶段简单地停止故障芯片上的执行会导致运行寿命缩短，而在设计阶段丢弃故障芯片则会降低良率，并增加无故障芯片的单位成本。因此，需要采用低成本的替代方案来缓解脉冲神经网络计算引擎中的永久性故障[1] 。这些方案将延长脉冲神经网络芯片的运行寿命。此外，这些方案还能提高晶圆级芯片在脉冲神经网络中的适用性，因为接纳永久性故障对于维持良率至关重要。

目标问题：如何有效缓解脉冲神经网络计算引擎（即本地权重寄存器和神经元）中的永久性故障对准确率的影响，从而提升脉冲神经网络的容错能力并维持吞吐量。针对这一问题的高效解决方案，将使能量受限嵌入式系统（例如物联网边缘设备和自主智能体）能够在故障芯片上可靠地运行脉冲神经网络，而无需重新训练。

1.1 现有技术及其局限性

除了舍弃发生故障的芯片外，还可以采用标准的超大规模集成电路容错技术，例如双模块冗余（DMR）（Vadlamani 等人，2010）、三模冗余（TMR）（Lyons 和 Vanderkulk，1962）以及纠错码（ECC）（Sze，2000），来缓解永久性故障。然而，这些技术需要额外的（冗余）硬件和/或执行操作，从而带来巨大的面积和能耗开销。现有最先进的研究从不同的设计角度出发，研究了如何理解脉冲神经网络中的故障并制定缓解技术，具体如下。

- **脉冲神经网络故障建模：** Vatajelu等人（2019年）已识别出可能影响脉冲神经网络的故障。在模拟域中，针对模拟神经元硬件及其容错能力的故障建模

[1] 为简洁起见，我们使用“脉冲神经网络计算引擎 引擎”或“计算引擎”可互换使用，以表示组件 ute 引擎的脉冲神经网络加速器/芯片。

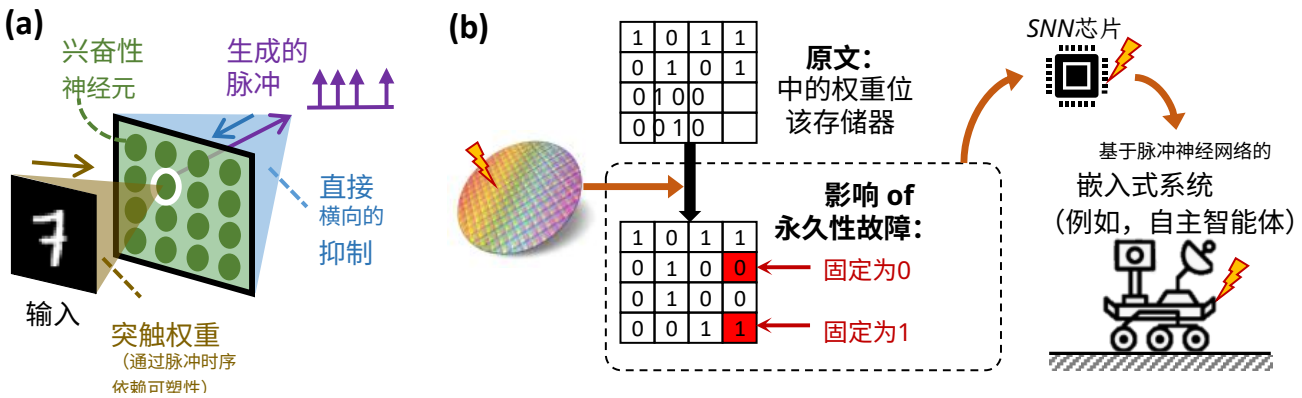


图1。 (a) 一种在无监督学习场景中实现高准确率的SNN架构，即单层全连接（FC）网络（Putra和Shafique，2020年）。(b) SNN计算引擎的权重存储器中可能存在固定为0和固定为 1 故障形式的永久性故障。

El-Sayed等人（2020年）和 Spyrou 等人（2021年）已经研究了相关策略，但由于我们针对的是数字域中的脉冲神经网络实现，这些策略不在本文的研究范围内。

- **脉冲神经网络容错：**以往研究探讨了故障对脉冲神经网络权重的影响，但未考虑底层硬件架构和处理数据流（Schuman 等人，2020年；Rastogi等人，2021年），且每项研究都针对特定故障，例如位翻转或突触移除。近期研究针对脉冲神经网络硬件的权重存储器中的故障设计了缓解技术（Putra 等人，2021年a、b、2022a年），而 Putra 等人（2022b年）的研究旨在应对瞬态故障。

总而言之，当前最先进的研究仍然仅关注对脉冲神经网络硬件中的权重存储器进行永久性故障建模和注入。因此，脉冲神经网络计算引擎（即突触和神经元）中的永久性故障的影响，以及相应的故障缓解技术，尤其是着眼于避免重新训练成本，仍未得到探索。为了研究缓解永久性故障所面临的挑战，我们在第1.2节开展了一项实验案例研究。

1.2 案例研究与研究挑战

在本案例研究中，我们考虑图2(a)所示的脉冲神经网络加速器架构。我们假设所有神经元均无故障，并以随机分布的方式在权重寄存器中注入永久性故障（即固定为0或固定为1），同时设置不同的故障存储单元比例，以考察故障寄存器对准确率的影响。实验设置的详细信息见第4节。根据图2(b)中的实验结果，我们得出以下关键观察结论。

- 在固定为0和固定为1两种情况下，随着故障存储单元比例的增加，分类准确率均会下降，这表明突触中的永久性故障会产生负面影响。
- 在固定为0的情况下，存储的权重值要么保持不变，要么从原始值减小。当权重值减小时，相应的神经元将需要更多刺激（输入脉冲）来提高其膜电位，并达到生成脉冲所需的阈值电位，而该脉冲表示对特定类别的识别。然而，在脉冲神经网络模型中，可能有多个神经元负责识别同一类别。因此，如果权重位发生故障的神经元无法识别输入类别，其他神经元仍可能识别该类别。因此，存储单元中固定为0所导致的准确率下降相对较小，在某些情况下甚至可以忽略不计。

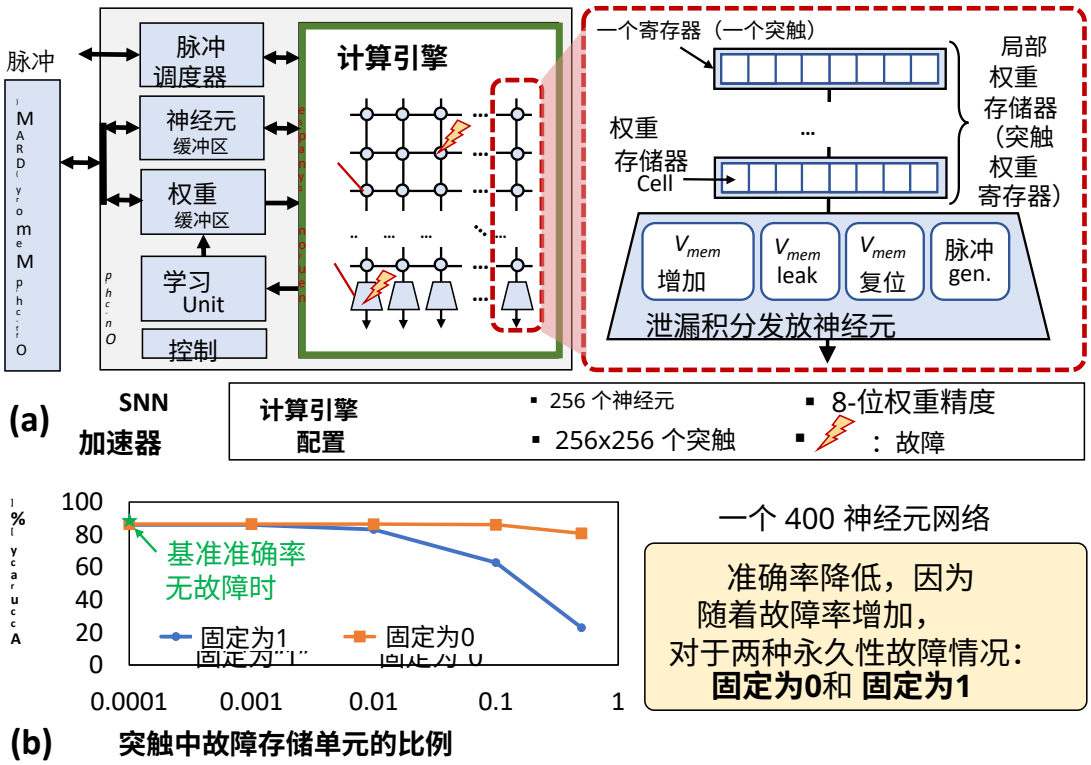


图 2. (a) 典型的脉冲神经网络加速器架构采用基于交叉开关的突触连接（Basu 等，2022年）。突触和神经元可能受到永久性故障的影响，具体讨论见第 2.2 节。(b) 本地权重存储器（突触权重寄存器）中的固定型故障可能降低准确率。

- 在固定为1的情况下，存储的权重值要么保持不变，要么从原始值增加。当权重值增加时，相应的神经元将需要更少的刺激（输入脉冲）来提高其膜电位，并达到生成脉冲所需的阈值电位，而该脉冲表示对特定类别的识别。因此，对于任何输入类别，该神经元都可能变得更加活跃并生成更多脉冲，从而导致更多分类错误。因此，存储单元中固定为1所导致的准确率下降比固定为0的情况更为显著/明显。

- 故障类型和故障率的不同组合会导致不同的准确率，这反映了实际芯片中不同的故障模式，因此在设计阶段难以预测。

基于这些观察，我们概述了以下研究挑战，以便为所针对的问题设计一种高效的解决方案。

- 缓解技术不应采用重新训练，因为重新训练需要巨大的计算/存储成本、处理时间，以及在某些情况下由于限制政策而可能无法获得的训练数据集。此外，考虑到已制造芯片的数量庞大，重新训练并不是一种可扩展的方法，因为它需要考虑每个芯片独有的故障图，从而针对每个芯片分别进行重新训练。需要注意的是，故障图信息可以在制造完成后通过标准的晶圆/芯片测试流程获得，因此该测试不会引入新的成本，仅产生芯片测试的典型成本（Xu 等人，2020年；Fan 等人，2022年）。

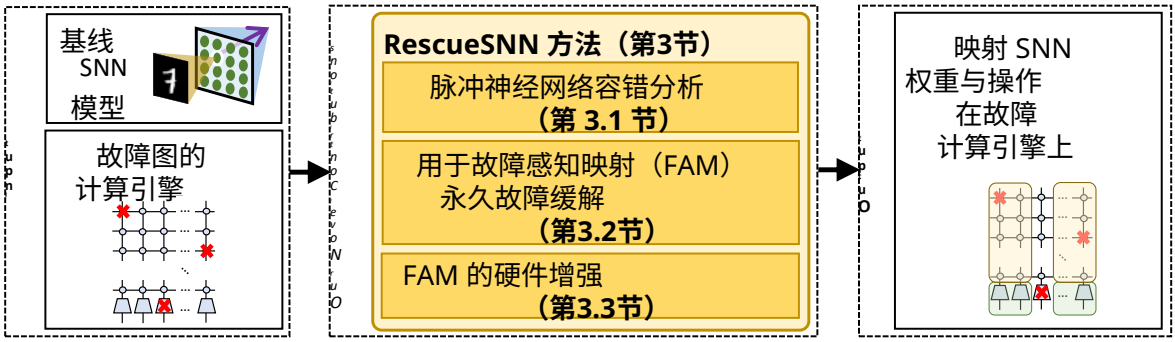


图 3. 我们新颖贡献的概述。

- 与不采用故障缓解技术的基线设计相比，该缓解技术在运行阶段应尽可能减少性能/能耗开销，从而使其适用于能耗受限的嵌入式系统。

- 该技术不应避开使用发生故障的脉冲神经网络组件（即突触和神经元），因为这意味着省略脉冲神经网络计算引擎中相应列的全部计算，从而导致吞吐量下降。

- 脉冲神经网络需要专门的故障缓解技术，相比其他神经网络计算模型（例如深度神经网络），因为脉冲神经网络具有不同的操作和数据流。

1.3 我们的创新贡献

为应对上述挑战，我们提出了 RescueSNN，一种能够在存在永久性故障的情况下实现脉冲神经网络加速器可靠执行的新型方法论。据我们所知，这是首个缓解脉冲神经网络加速器/芯片中永久性故障的工作。以下是 RescueSNN 方法的关键步骤（概述如图 3 所示）。

- **分析脉冲神经网络容错能力**，以了解故障组件（即突触和神经元）在给定故障率下对准确率的影响。
- **采用故障感知映射（FAM）技术** 将脉冲神经网络权重和神经元运算安全地映射到故障计算引擎上，从而保持准确率和吞吐量。我们的FAM技术利用计算引擎的故障图来执行以下关键机制。1. 将重要权重位映射到突触（权重寄存器）的无故障存储单元上，以尽可能减少对权重值的污染/更改。2. 根据神经元膜电位动态和脉冲生成的行为，有选择地采用在推理过程中不会导致显著准确率下降的故障神经元。

- **设计简单的硬件增强**，以实现高效的 FAM 技术。我们的增强方案采用多路复用器等简单组合逻辑单元，对突触的权重位进行随机打乱，使这些权重位能够用于脉冲神经网络计算。

关键结果： 我们在一台多 GPU 机器上使用基于 Python 的仿真（Hazan 等，2018）评估了 RescueSNN 方法。实验结果表明，与未进行故障缓解的脉冲神经网络处理相比，我们的 RescueSNN 无需重新训练，即可将 MNIST 和 Fashion MNIST 上的脉冲神经网络准确率分别提高至多 80% 和 47%。

2 背景

2.1 脉冲神经网络（SNNs）

概述： 脉冲神经网络是受大脑启发的计算模型，利用动作电位（脉冲）对信息进行编码（Maass，1997年；Putra 和 Shafique，2022b年）。在脉冲神经网络模型中，神经元和突触按照特定架构相互连接（Tavanaei 等，2019年；Mozafari 等，2019年）。本研究采用图 1(a) 所示的全连接（FC）网络架构，因为该架构在使用生物合理学习规则（例如脉冲时序依赖可塑性（STDP））时已展现出较高的准确率。该网络架构将输入中的每个数据项（例如图像中的一个像素）连接到所有兴奋性神经元，以产生脉冲。每个脉冲都会抑制除产生该脉冲的神经元以外的其他兴奋性神经元。这类生物合理学习规则通过利用脉冲事件在每个突触中执行局部学习，无需全局损失函数，从而提供高效的学习机制，并支持无监督学习和在线学习能力，这对自主智能体尤其有益（Putra 和 Shafique，2022a年；Pfeiffer 和 Pfeil，2018年）。本研究采用成对权重依赖型脉冲时序依赖可塑性，并将每个权重值 (wgh) 限定在规定范围 $wgh = [0, 1]$ 内，因为以往研究广泛采用了这一方法（Diehl 和 Cook，2015年；Srinivasan 等，2017；Hazan 等，2019年）。这里采用成对权重依赖型脉冲时序依赖可塑性学习规则，因为它定义了允许的最大权重，适用于定点格式（见公式 1）。

$$\Delta wgh = \begin{cases} -\eta_{pre} \cdot x_{post} \cdot wgh^{\mu} & \text{on presynaptic spike} \\ \eta_{post} \cdot x_{pre} \cdot (wgh_m - wgh)^{\mu} & \text{on postsynaptic spike} \end{cases} \tag{1}$$

Δwgh 表示权重更新，η_{pre} 和η_{post} 分别表示突触前脉冲和突触后脉冲的学习率，而x_{pre} 和x_{post} 分别表示突触前迹和突触后迹。 wgh_m 表示允许的最大权重， wgh 表示当前权重， μ 表示权重依赖因子。我们考虑采用积分发放漏电（LIF）神经元模型，因为该模型具有较低的计算复杂度，同时具备较高的生物合理性（Izhikevich，2004年）。此外，突触通过权重值（wgh）进行建模，该权重值表示相应神经元之间突触连接的强度。此外，脉冲神经网络模型通常采用特定的脉冲编码方案，将数据信息编码为脉冲，或从脉冲中解码数据信息。在本文中，我们考虑采用频率编码方案，该方案利用脉冲频率按比例表示数据，即较高的数据值由更多的脉冲表示。这是因为，频率编码方案结合具有生物合理性的基于 STDP 的学习规则，能够实现高准确率；这些学习规则可在无监督设置下执行高效的学习机制，从而实现节能且智能的计算系统（Diehl 和 Cook，2015年；Rathi 等人，2019年；Putra 和 Shafique，2020年）。

脉冲神经网络加速器架构： 我们考虑图2(a)和图4中分别展示的典型脉冲神经网络加速器和计算引擎架构，这些架构改编自 Frenkel 等人（2019）的设计。我们重点关注计算引擎，因为它负责生成决定准确率的脉冲。计算引擎包含一个突触交叉开关阵列，每个突触都有一个用于存储权重值的寄存器。我们采用 8 位权重精度，因为它在准确率和存储器之间取得了良好的权衡（Putra 和 Shafique，2020；Putra 和 Shafique，2021a）。每一位都存储在一个权重存储单元中。为了优化芯片面积，每个突触将其权重与同一列中前一个突触的累积值相加，并将累加值存储在一个 32 位寄存器中。这样，神经元与相连突触之间只需一条连接。此外，计算引擎的每一列都包含一个 LIF神经元。如果脉冲到达某个神经元，膜电位（V_{mem}）就会增加

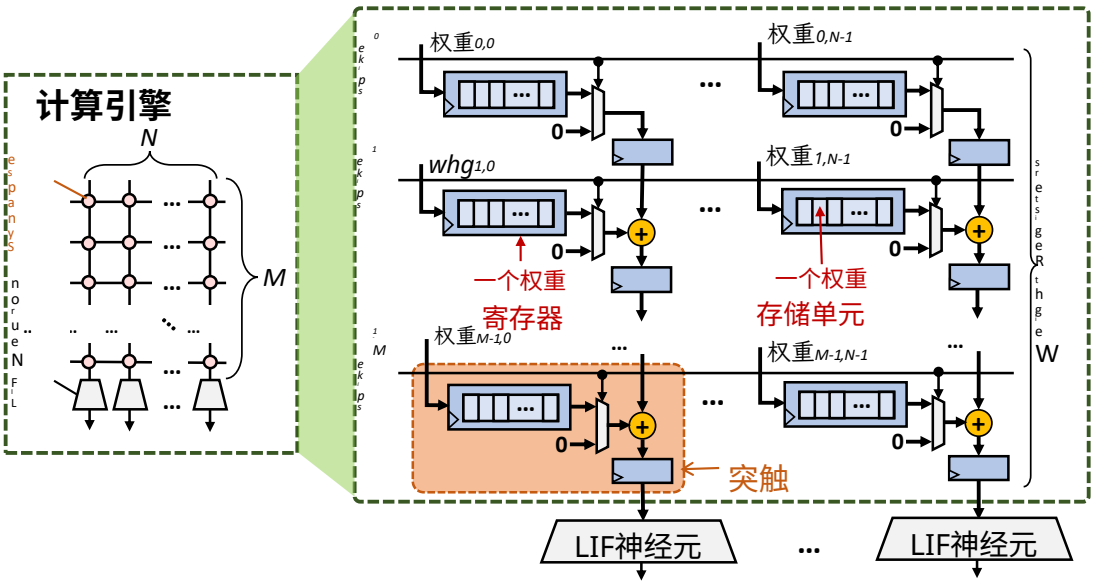


图4. 脉冲神经网络加速器的计算引擎架构。

相应的权重（wgh）；否则，V_{mem} 就会降低/泄漏。如果 V_{mem} 达到阈值电位（V_{th}），就会生成一个脉冲，随后 V_{mem} 返回复位电位（V_{reset}）。因此，LIF神经元有四种主要操作：（1）V_{mem} 增加，（2）V_{mem} 泄漏，（3）V_{mem} 复位，以及（4）脉冲生成，如图2(a)所示。神经元操作（1）至（3）定义了膜电位动态。

2.2 永久性故障建模

2.2.1 概述

脉冲神经网络计算引擎由两个主要组件组成，即突触和神经元，它们具有不同的硬件电路。因此，我们需要为每个组件定义故障模型，以实现快速的设计空间探索。

1. **突触：** 每个突触硬件使用一个寄存器来存储权重值。因此，突触中的每个永久故障都可能以固定为0或固定为 1 故障的形式影响单个位权重。
2. **神经元：** 每个神经元硬件都依赖神经元模型来执行其操作。因此，永久性故障可能表现为不同形式，具体取决于数字神经元硬件正在执行的操作类型，以下将对此进行讨论（概述见图 5）。

- “增加V_{mem}”操作中的故障会使神经元无法增加其膜电位。因此，该神经元无法产生任何脉冲（即休眠神经元）。
- “V_{mem} 泄漏”操作中的故障会使神经元无法降低其膜电位。因此，该神经元的行为类似于积分发放（IF）神经元模型。
- “V_{mem} 复位”操作中的故障会使神经元无法复位其膜电位。因此，该神经元将持续产生脉冲。

- “脉冲生成”中的故障会导致神经元无法生成脉冲（即休眠神经元）。

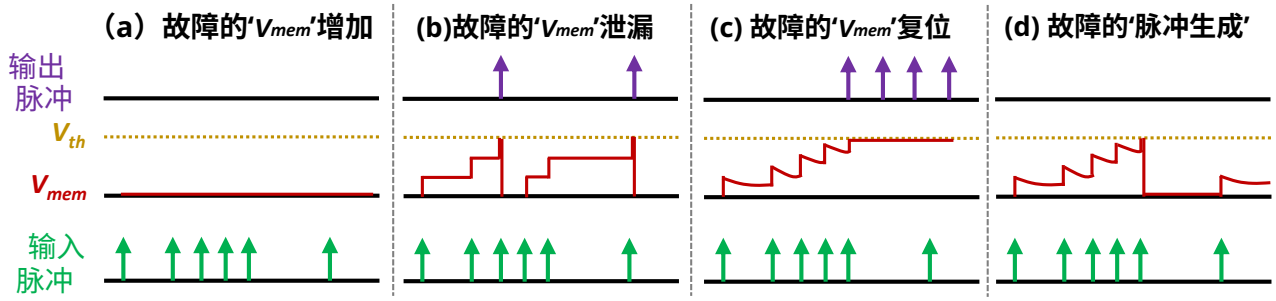


图5。不同故障 LIF 神经元运算概述：(a) 故障的"V_{mem} 增加"，(b) 故障的"V_{mem}泄漏"，(c) 故障的"V_{mem} 复位"，以及 (d) 故障的"脉冲生成"。

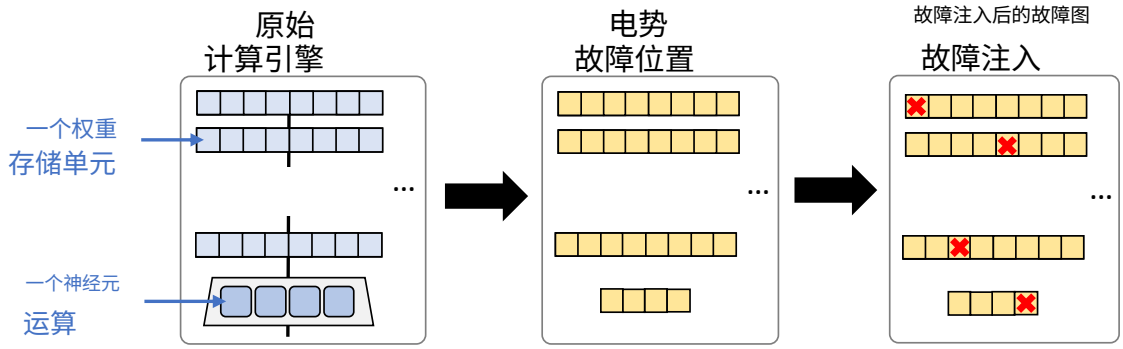


图6。永久　　性故障在脉冲神经网络计算引擎上的生成与分布步骤。

2.2.2 故障生成与分布

以往研究表明，永久性故障会出现在芯片的随机位置，从而形成特定的故障图（Zhang等人，2018年；Radetzki等人，2013年；Werner等人，2016年；Mercier等人，2020年；Stanisavljevi´c等人，2011）。如图6所示，以下是生成并分布脉冲神经网络计算引擎上永久性故障的关键步骤。

1. 我们将权重存储单元和神经元运算视为潜在故障位置。2. 我们根据给定的故障率生成永久性故障，并将其随机分布到各个潜在故障位置。故障率表示故障权重存储单元和神经元运算的总数与潜在故障位置总数（即权重存储单元和神经元运算的总数）之比。3. 如果本地权重存储单元发生故障，则随机选择固定型故障的类型（即固定为0或固定为1）。与此同时，如果神经元运算发生故障，则随机选择永久故障运算的类型。

3 RescueSNN 方法

我们的 RescueSNN 方法概述如图 7 所示，其关键步骤将在以下各小节中讨论。

3.1 永久故障下脉冲神经网络容错分析

脉冲神经网络容错分析对于了解给定脉冲神经网络模型在特定运行条件下（例如特定故障率、固定型故障类型和计算引擎架构等因素的组合）的表现至关重要

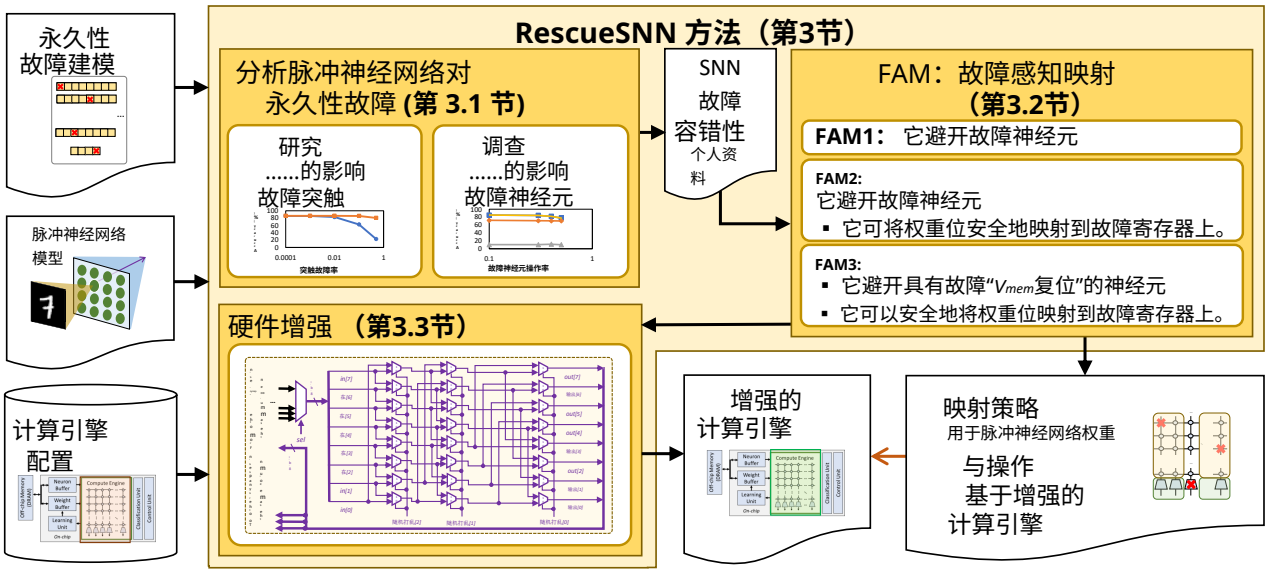


图7. 我们提出的 RescueSNN 方法概述及其在脉冲神经网络计算引擎中缓解永久性故障的关键步骤。

等）。此类分析可提供有助于制定高效故障缓解技术的信息。因此，我们的 *RescueSNN* 方法研究故障组件（即突触和神经元）与所得准确率之间的相互作用。为此，我们开展了以下实验案例研究，研究对象是一个采用全连接架构的 400 神经元网络，如图 1(a) 所示。

- **我们研究存在故障权重寄存器时的准确率**，方法是在权重寄存器中注入特定的固定型故障（即固定为0或固定为1），同时假设神经元无故障。实验结果如图2(b)所示。我们观察到，固定为0和固定为1故障都会降低准确率。因此，缓解技术应同时应对这两种固定型故障。

- **我们研究存在故障神经元操作时的准确率**，方法是在神经元运算中注入故障，同时假设权重寄存器无故障。实验结果如图8所示，我们据此得出以下观察结果。

  1. 故障的"脉冲生成"、"V_{mem} 增加"和"V_{mem} 泄漏"操作具有可容忍的准确率，这是因为其故障行为不会主导脉冲活动，和/或分类中相应故障神经元的功能可以由识别同一类别的其他神经元替代。因此，这些神经元仍可用于脉冲神经网络处理。

  2. 故障的"V_{mem} 复位"操作会导致显著的准确率下降，因为这些操作会使相应神经元主导分类。因此，不应将这些神经元用于脉冲神经网络处理。

请注意，具有多层和不同计算架构（例如卷积层和全连接层）的复杂脉冲神经网络模型，其观察结果可能与图 8 中的结果有所不同。不过，先前的研究观察到了与我们的研究相似的趋势，即，复位操作发生故障的神经元会持续产生脉冲（即饱和神经元），与其他类型的故障神经元操作相比，会导致最显著的准确率下降（Spyrou 等人，2021）。研究还指出，饱和神经元会影响脉冲神经网络模型任意层的分类准确率，因为这些故障神经元始终主导分类活动，从而导致准确率显著下降。这一发现

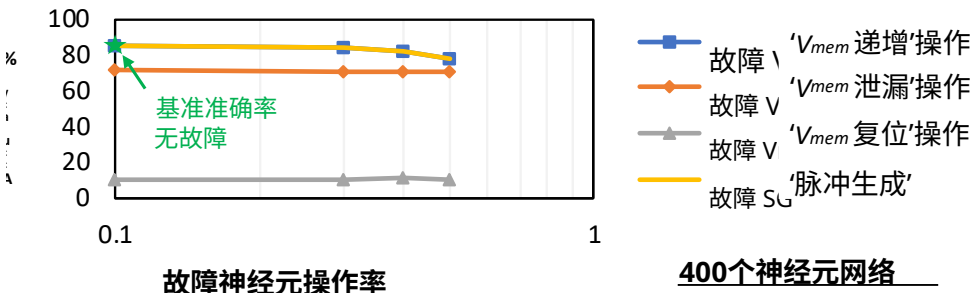


图 8. 故障神经元操作对准确率的影响。不同的故障神经元操作对准确率的影响不同。当采用故障"V_{mem} 复位"操作时，会出现显著的准确率下降。

与我们的研究所得见解一致。然而，在复杂脉冲神经网络模型上采用基于 STDP 的学习仍难以实现高准确率（Rathi 等人，2023），这阻碍了它们在多种应用中的适用性，例如需要在线训练和无监督学习的系统（如自主移动智能体）。因此，在本研究中，我们考虑采用图 1(a) 所示的基于全连接层的脉冲神经网络，以实现多项优势，例如高准确率、无监督学习能力和高效的在线训练。

3.2 提出的故障感知映射（FAM）

SNN芯片中的永久性故障可以在设计阶段和运行阶段被识别。制造后测试可用于在设计阶段找出SNN计算引擎中一组故障位置（由于制造缺陷导致）（即故障图）（Zhang等人，2018年；Putra等人，2021a年）。与此同时，诸如内建自测试（BIST）技术之类的在线测试策略可用于在运行阶段获取故障图（由于器件老化或物理损坏导致）（Baloch等人，2019年；Wang等人，2019年；Mercier等人，2022年）。我们的*RescueSNN* 方法利用该故障图，在计算引擎上安全地映射*SNN*权重和操作，从而最大限度地减小永久性故障的负面影响。为此，RescueSNN采用故障感知映射（FAM）技术，通过以下关键机制减轻突触和神经元中的故障。

1. **面向突触的 FAM：** 通过简单的位重排技术，将重要权重位放置在无故障的权重存储单元中，而将不重要位放置在故障存储单元中。可通过实验识别权重位的重要性：修改特定位后观察准确率（Putra 等人，2021a年）。通常，先前研究发现，权重位的重要性与其位位置成正比。例如，在 8 位定点精度下，位 7 比其他位更重要。此外，一个突触可能有一个或多个故障位。因此，我们提出了一种映射策略，可通过以下步骤处理这两种情况（概述见图9）。

- 我们识别故障权重位（例如，通过制造后测试），以获取每个突触硬件的故障图和故障率信息。
- 我们确定每个突触硬件中可安全存储权重的最大故障率。本文根据第 3.1 节的分析，考虑一个 8 位权重最多有 2 个故障位；该故障率仍可提供可接受的准确率。

- 我们找出每个突触中后续无故障存储单元数量最多的段。我们利用这一信息，尽可能将重要位存储在

无故障单元中。因此，我们也会检查边界情况（即最右侧和最左侧的存储单元），将其视为可能的后续无故障存储单元；见图9(b)中包含data-3的第三行。

- 我们对每个数据字执行循环移位技术，以高效实现位混洗。

2. **神经元FAM：** 如果神经元的‘V_{mem} 复位’操作存在故障，则应避免使用这些神经元，因为这些故障操作会导致准确率显著下降。同时，具有其他类型故障的神经元仍可用于脉冲神经网络处理，因为其故障行为不会主导脉冲活动。为维持吞吐量，会将用于识别同一输入类别的不同脉冲神经网络操作映射到故障神经元和无故障神经元上，同时弥补故障的‘V_{mem} 递增’、‘V_{mem} 泄漏’和‘脉冲生成’操作造成的损失。

随后，我们利用这些机制设计了三种映射策略，作为*FAM*技术的变体（即*FAM1*、*FAM2*和*FAM3*），在准确率与映射复杂度之间提供不同的权衡，具体讨论如下。

- **FAM1：** 如图10(a)所示，该策略避免将脉冲神经网络权重和操作映射到计算引擎中包含故障神经元的列上，因为故障神经元造成的准确率下降可能大于故障寄存器，尤其是在‘V_{mem} 复位’故障的情况下。不过，FAM1无法减轻寄存器故障带来的负面影响，因此准确率提升并非最优。FAM1的优势在于其机制简单，能够实现低复杂度的控制机制。

- **FAM2：** 它将脉冲神经网络权重和操作映射到具有无故障神经元的计算引擎列上（与FAM1一样），并采用位重排技术将重要权重位映射到无故障存储单元上，如图 10(b) 所示。因此，FAM2 可以提高脉冲神经网络容错能力，但其代价是需要比 FAM1 更复杂的控制机制。

- **FAM3：** 它有选择地将脉冲神经网络权重和操作映射到计算引擎中不具有故障“V_{mem} 复位”操作的列上，同时使用位重排技术将重要权重位映射到无故障存储单元上，如图10(c)所示。因此，与FAM1 相比，FAM3 以更复杂的控制机制为代价增强了脉冲神经网络容错能力，并且与 FAM1 和 FAM2 相比能够提高吞吐量。

有关如何在计算引擎上映射脉冲神经网络权重和操作的信息通过软件程序（例如固件）提供，从而使所提出的FAM技术（例如FAM1、FAM2或FAM3）能够适用于计算引擎上不同的可能故障映射。该信息的元数据存储在片上缓冲区中，可供计算引擎中的操作访问。

3.3 我们为 FAM 提供的硬件增强

我们的 FAM2 和 FAM3 策略可能会使存储的权重位以随机打乱的形式排列。因此，需要额外的机制将这些权重位转换回原始顺序，以便用于脉冲神经网络的执行。为此，我们提出了轻量级硬件增强，以支持重新随机打乱机制，从而撤销数据变换，即通过一个简单的 *8* 位桶形移位器实现。其核心思想是将每个突触的输出线顺序重新随机打乱为原始顺序，使相应的权重能够直接用于神经元运算。为优化开销（例如面积），我们让计算引擎同一列中的所有突触共享一个硬件增强模块 （HEB），不同突触将在不同时间访问该增强模块；见图 11。通过这种方式，HEB 的数量等于脉冲神经网络计算引擎中的列数。此外，为控制 HEB 的功能，我们采用一个增强控制单元 （ECU）。该 ECU 存储位移信息，并利用这些信息控制 HEB 中的桶形移位器。

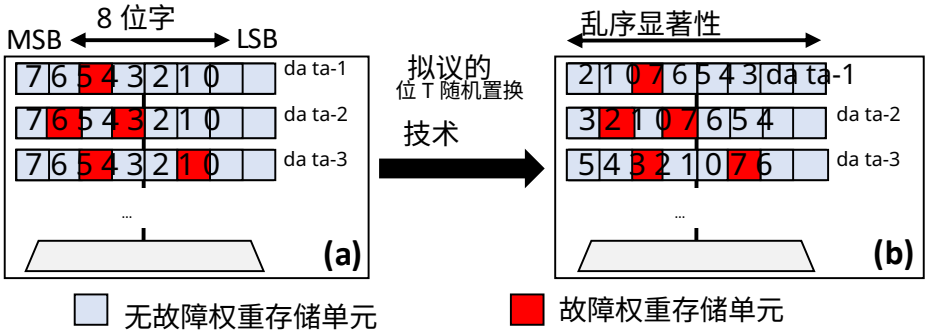


图 9. (a) 故障权重寄存器中可能故障位置（故障图）的示意图。(b) 针对相应故障图提出的循环移位位重排技术。

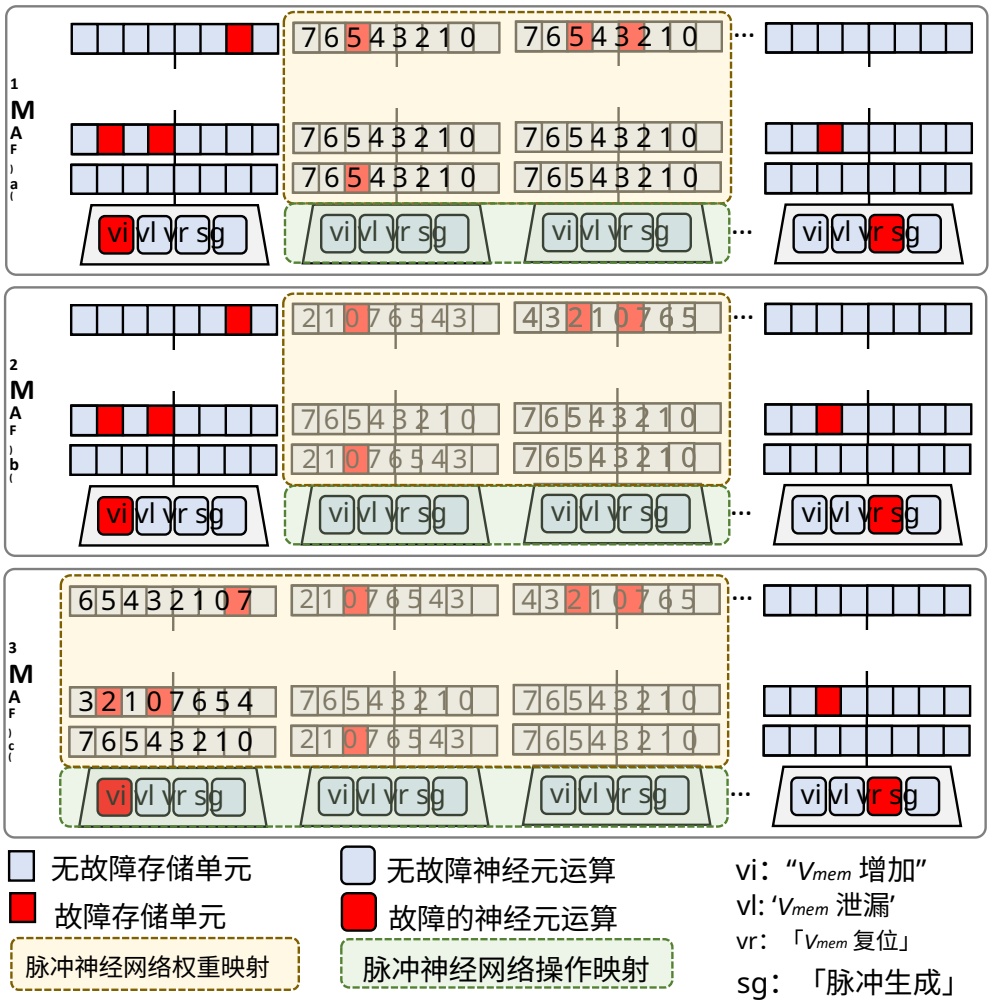


图 10. 我们的 FAM策略：(a) FAM1，(b) FAM2，以及 (c) FAM3。

信息，并利用这些信息控制 HEB 中的桶形移位器。对于计算引擎的每一列，ECU 使用：（1）一个专用的选择信号*sel*，用于确定每次应在 HEB 中处理哪个权重；以及（2）一组寄存器，用于存储同一列中所有权重的位移信息随机打乱*[2：0]*。

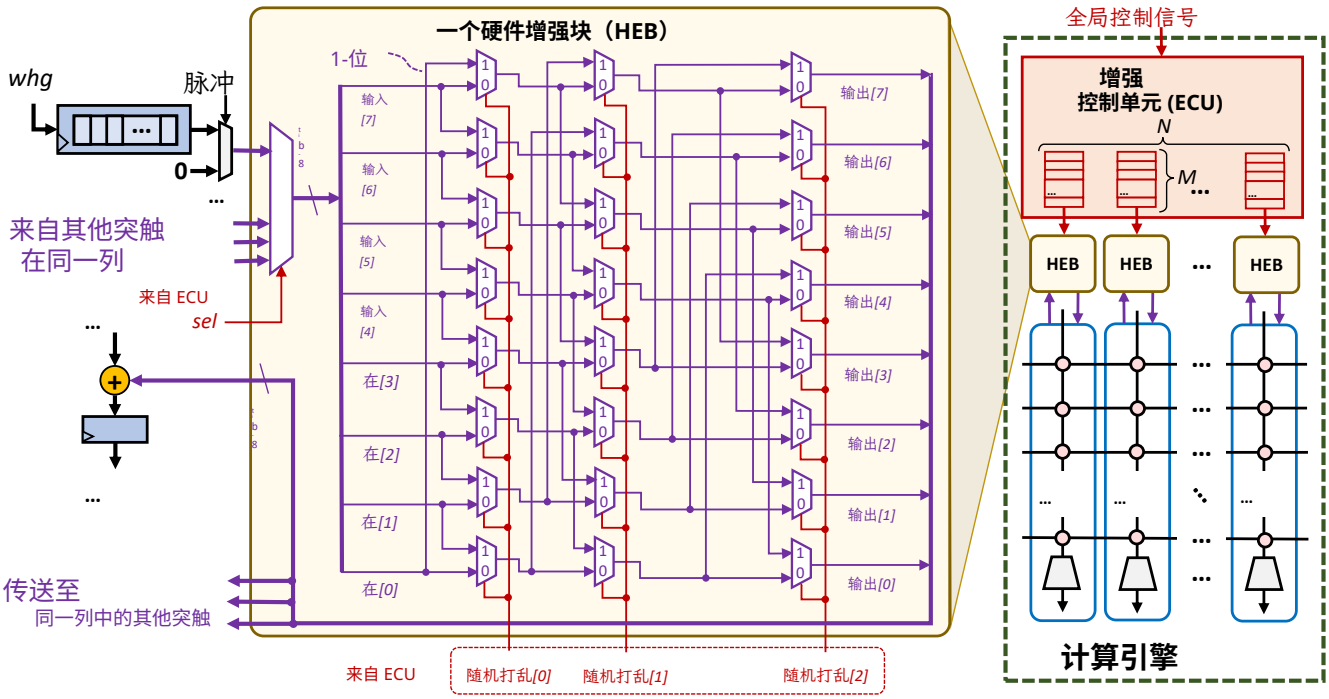


图11。 所提出增强机制的架构，包括用于适配FAM策略的硬件增强模块（HEBs）和增强控制单元（ECU）。

4 评估方法学

为评估 RescueSNN 方法，我们采用图 12 所示的实验设置。我们使用图 1(a)所示的全连接网络，并设置不同数量的神经元，以评估RescueSNN 方法的通用性。为简洁起见，我们将具有 i 个神经元的网络表示为Ni。我们使用MNIST 和FashionMNIST 作为工作负载，并采用脉冲神经网络社区广泛使用的相同测试条件（Diehl和 Cook，2015年）。作为比较，我们将未经故障缓解的脉冲神经网络视为基线。

故障生成与注入： 永久性故障根据修订稿第2.2节中的故障建模生成。为此，我们首先根据给定的故障率生成二进制值（即0和1），同时考虑潜在的故障位置（如图6所示）。其中，“0”表示突触中的非故障存储单元或神经元中的非故障操作；“1”则表示突触中的故障存储单元或神经元中的故障操作。随后，将这些二进制值随机分配到一个表示潜在故障位置的数组中，使每个值对应一个特定的权重存储单元或神经元运算。图13展示了可能受到永久性故障影响的潜在位置/组件，这些故障会导致存储单元故障，以及“V_{mem} 增加”、“V_{mem} 泄漏”、“V_{mem} 复位”和“脉冲生成”操作故障。对于权重存储单元（突触）中的每个故障，我们随机确定故障类型，即固定为0或固定为1。若固定为0，则向相应存储单元注入值0；若固定为1，则注入值1。同时，神经元中的每个故障 对应于“V_{mem} 增加”、“V_{mem} 泄漏”、“V_{mem} 复位”或“脉冲生成”操作故障（如图5所示）。相应神经元中的每种故障行为均通过不同方法实现，具体如下。

- 故障的‘V_{mem} 递增’操作：这主要是由‘V_{mem} 递增’部分的加法故障引起的，因此尽管有输入脉冲，V_{mem} 仍未增加。

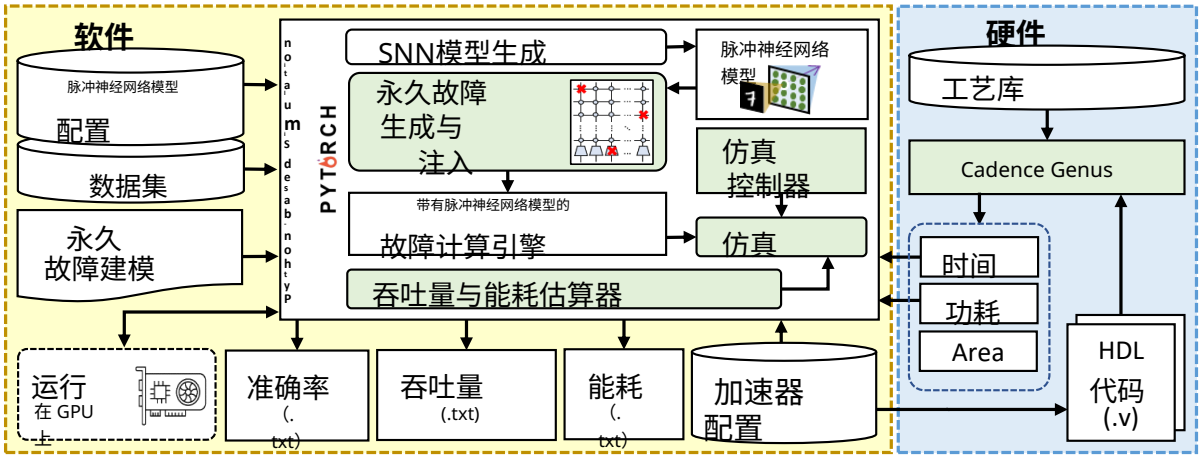


图 12. 实验设置概述及工具流程。

- 故障的'V_{mem} 泄漏'操作：这主要是由'V_{mem} 泄漏'部分的减法故障引起的，因此V_{mem} 尽管没有传入脉冲，却没有减少。
- 错误的'V_{mem} 复位'操作：这主要是由"V_{mem} 复位"部分的比较错误造成的，因此脉冲生成器会被激活并持续生成脉冲。
- 错误的'脉冲生成'：这主要是由"脉冲生成"部分的多路复用错误造成的，因此脉冲生成器始终处于停用状态，不会产生输出脉冲。

准确率评估：我们使用基于 Python 的仿真（Hazan 等人，2018年），在 Nvidia RTX 2080 Ti GPU 上运行，同时考虑图2(a)和图4所示的脉冲神经网络加速器架构。

硬件评估：我们评估采用 RescueSNN 方法的原始计算引擎（未经增强）和增强型计算引擎的面积、能耗和吞吐量。为此，我们分别为这两种计算引擎（原始和增强型）设计 RTL 代码，然后使用 Cadence Genus 工具，在采用 65nm 互补金属氧化物半导体工艺的条件下对其进行综合，以获取面积、功耗和时序（即计算引擎进行脉冲神经网络处理时的时钟周期延迟）。随后，结合综合得到的时序以及计算引擎中活跃突触和神经元的映射策略，计算处理一个输入样本所需的周期数和计算延迟（即每样本延迟）。然后，根据每样本延迟的信息，估算脉冲神经网络推理在一秒内可处理的样本数量，从而得出吞吐量。此外，我们还利用综合得到的功耗信息和每样本延迟，估算脉冲神经网络推理的能耗。吞吐量和能耗的估算也通过基于 Python 的仿真框架完成（Hazan 等，2018年）。

5 结果与讨论

我们评估了不同的设计方面，包括准确率、吞吐量、能耗和面积，具体讨论如下。

5.1 保持准确率

图 14 展示了不同故障缓解技术的准确率实验结果，包括基线和我们基于 FAM 的策略，即 FAM1、FAM2 和 FAM3。正如 所示，我们观察到基线的准确率显著下降，因为它没有进行缓解

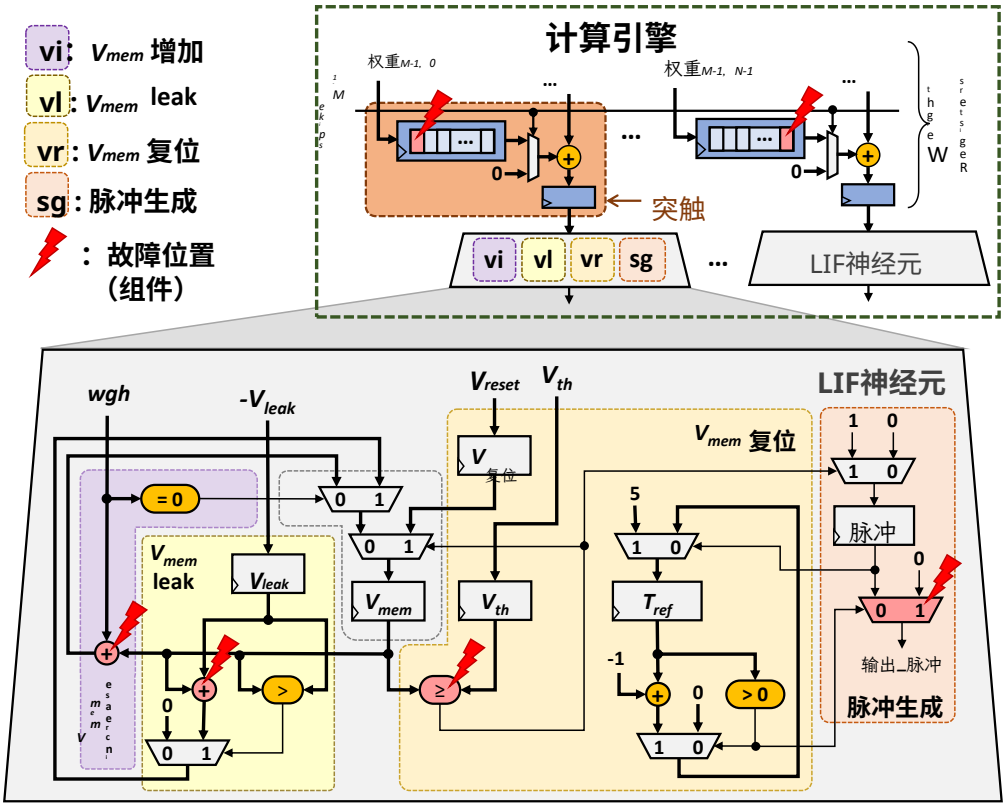


图 13。可能受永久性故障影响、导致故障存储单元以及故障的‘V_{mem} 增加’、‘V_{mem} 泄漏’、‘V_{mem} 复位’和‘脉冲生成’操作的电势位置/组件。

突触和神经元中的故障，从而导致不可靠的脉冲神经网络执行。显著的准确率下降主要是由于故障模型对故障‘V_{mem} 复位’操作的设定，该操作会使相应神经元的膜电位 V_{mem} 达到阈值电位 V_{th}后持续产生脉冲，从而主导分类活动并导致较高的分类错误率。我们还观察到，与基线相比，FAM1显著提高了脉冲神经网络容错能力，因为FAM1避免使用故障神经元，尤其是针对故障‘V_{mem} 复位’操作，如 所示。与FAM1相比，我们的FAM2进一步提高了脉冲神经网络容错能力，因为除了避免使用故障神经元之外，FAM2还缓解了权重寄存器中的故障，如 所示。同时，我们的FAM3也显著提高了相较于基线和FAM1的脉冲神经网络容错能力，并获得了与FAM2相当的准确率，因为FAM3缓解了权重寄存器中的故障，并选择性地使用故障神经元。在MNIST数据集上，与基线相比，其准确率最高提升了80%，如 所示。我们还观察到，相同的原因同样适用于不同的工作负载，因此Fashion MNIST的准确率曲线与MNIST的准确率曲线呈现出相似的趋势。这些结果表明，我们的*FAM*策略（*FAM1*、*FAM2*和*FAM3*）无需重新训练，即可在不同模型规模、故障率和工作负载下，有效缓解计算引擎中的永久性故障。

5.2 保持吞吐量

图 15(a) 展示了不同缓解技术（即基线和我们的 FAM 策略（FAM1、FAM2 和 FAM3））在吞吐量方面的实验结果。我们观察到，在不同模型规模和故障率下，基线的吞吐量最高，因为它使用所有突触和神经元来

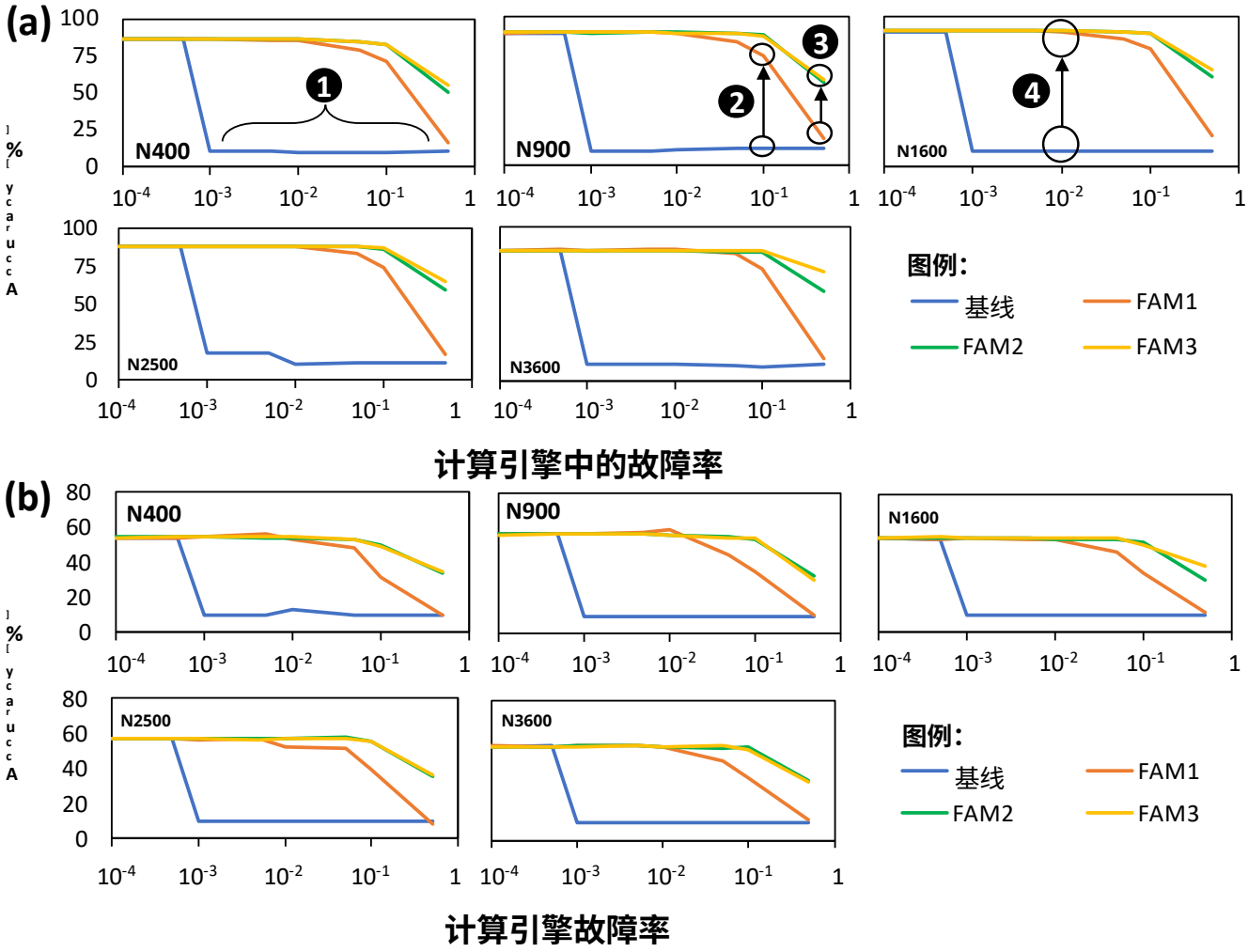


图 14。 不同缓解技术（即基线、FAM1、FAM2 和 FAM3）、不同模型大小（即 N400、N900、N1600、N2500 和 N3600）、不同故障率以及不同工作负载下的准确率曲线：(a) MNIST 和 (b) Fashion MNIST。

执行脉冲神经网络运算，如 1 所示。同时，FAM1 和 FAM2 由于避免使用故障神经元，因而会省略脉冲神经网络计算引擎中对应的列，可能导致吞吐量下降。例如，如 2 所示，当 N1600 的故障率为 0.1 时，FAM1 和 FAM2 的吞吐量可能下降 30%。相比之下，我们的 FAM3 能够将吞吐量维持在接近基线的水平（例如，在故障率为 0.5 时将吞吐量降幅控制在 25% 以下），因此与 FAM1 和 FAM2 相比，吞吐量显著提升。原因在于，FAM3 仅在对应神经元的“V_{mem} 复位”操作发生故障时，才会省略计算引擎中的相应列。例如，如 3 所示，对于 N1600，FAM3 的吞吐量降幅不到 15%。这些结果表明，我们的 *FAM3* 能够有效地在不同模型规模、故障率和工作负载下维持吞吐量。

5.3 能耗与面积开销

图 15(b) 展示了不同缓解技术（即基线和我们的 FAM策略（FAM1、FAM2 和 FAM3））的能耗实验结果。如标签所示，对于较低的故障率，不同技术的能耗相近-4。这是因为故障率较低时，出现故障神经元的概率较低，因此不同技术的资源利用率相似。

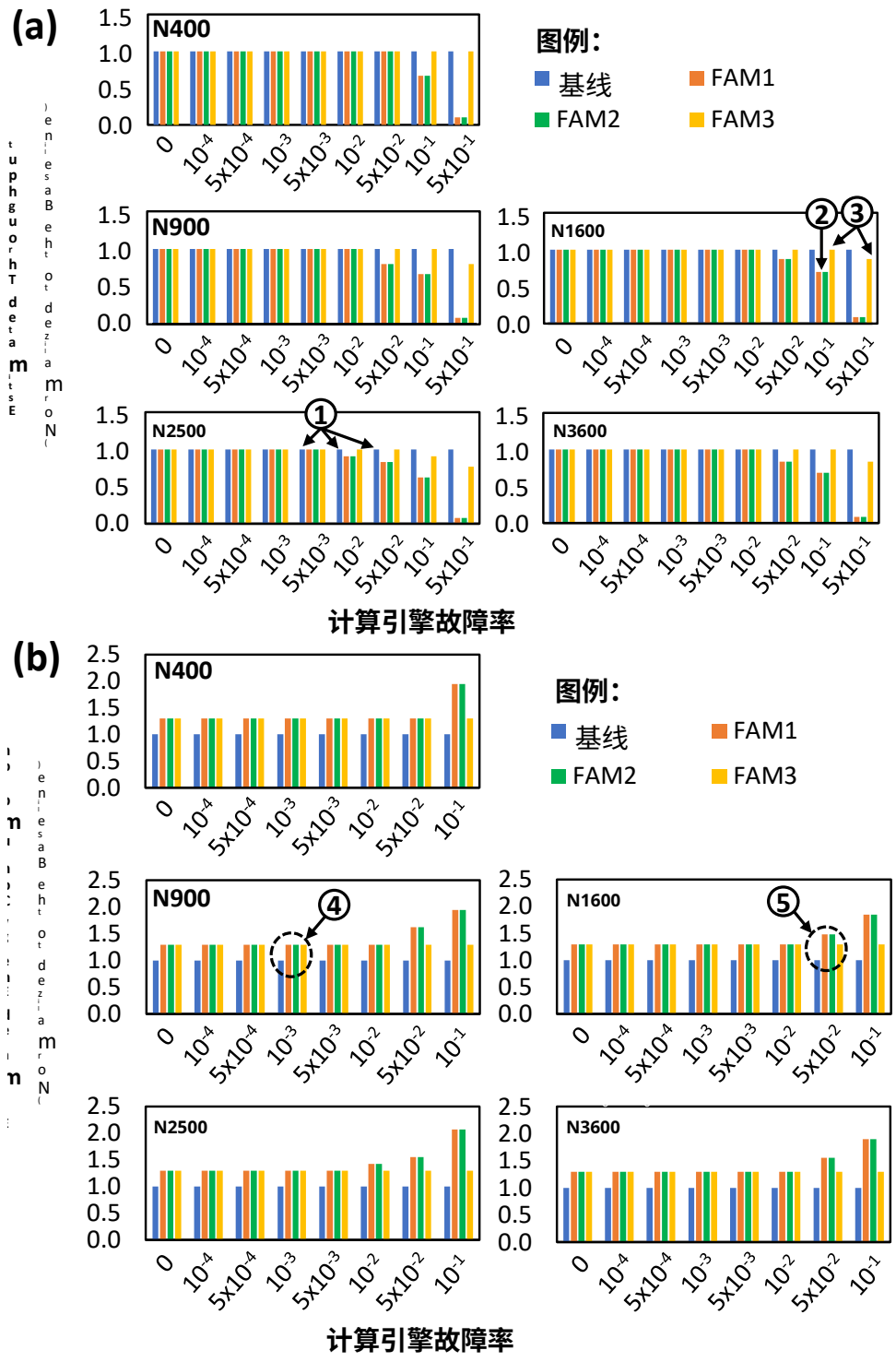


图 15. (a) MNIST 和 Fashion MNIST 在不同缓解技术、模型规模和故障率下的吞吐量。两者的脉冲神经网络权重和操作数量相近。(b) MNIST 和 Fashion MNIST 在不同缓解技术、模型规模和故障率下的能耗。两者的脉冲神经网络权重和操作数量相近。

对于较高的故障率，FAM1 和FAM2 的能耗高于基线和FAM3，如标签-5 所示。原因在于，较高的故障率意味着出现故障神经元的概率较高，因此不同技术的资源利用率不同，即FAM1和FAM2 避免使用故障神经元，从而带来更高的计算延迟和能耗。基线和 FAM3 则具有

能耗相近，这是因为 FAM3 采用了简单的硬件增强：（1）在每个 HEB 中复用操作，这些操作由计算引擎同一列中的所有突触共享；（2）在 ECU 中访问寄存器，从而将 FAM3 的能耗开销降至最低（即在 30% 以内）。就面积占用而言，原始计算引擎的面积约为 6.27 mm^2，而采用所提增强措施后的计算引擎面积约为 8.56 mm^2。因此，所提增强措施带来的面积开销约为 36.5%，其中 ECU 约占 36.2%，硬件增强模块约占 0.3%。ECU 的面积占增强措施总面积的主要部分，因为它主要采用一组 3 位寄存器（即 256x256 个寄存器），与硬件增强模块（即 256x25 个多路复用器）相比，这会占用更大的面积。这些结果表明，在不同模型规模、故障率和工作负载下，我们的 *FAM3* 在能耗和面积方面均实现了最低开销。

总之，上述讨论表明，我们的 *RescueSNN* 方法能够在无需重新训练的情况下有效缓解 *SNN* 芯片中的永久性故障。由于我们的 RescueSNN 在设计阶段和运行阶段都会处理永久性故障，因此能够提高 SNN 芯片的良率，并在其运行寿命内实现高效、可靠的 SNN 执行。此外，由于无需重新训练，我们的 RescueSNN 还能避免碳排放，从而提供一种环保的解决方案（Strubell 等人，2019年、2020年）。

5.4 进一步讨论

总的来说，我们观察到，一个有故障的“V_{mem} 复位”操作可能会导致显著的准确率下降，因为它使神经元偏离预期行为。其原因在于，生成的（有故障）脉冲会影响脉冲神经网络模型理解信息的方式，因为脉冲神经网络模型通常采用某种脉冲编码方案，即本文中的频率编码。因此，执行有故障“V_{mem} 复位”操作的神经元将生成大量脉冲，并主导分类活动，从而导致较高的分类错误率和显著的准确率下降。我们还观察到，有故障“V_{mem} 复位”操作生成的更多脉冲也表明，脉冲神经网络模型执行了更频繁的神经元运算，而这些运算对应于脉冲生成。这种情况会导致脉冲神经网络处理的功率/能耗更高，这一点已在先前的研究中得到观察和研究（Krithivasan 等，2019年；Park 等，2020年；Putra 和 Shafique，2023）。

与重新训练技术的比较：在标准芯片制造工艺中，制造出的芯片会经过晶圆/芯片测试流程（即晶圆验收测试和芯片探针测试）。该测试流程旨在评估每个芯片的质量，包括芯片中的任何故障（Xu 等人，2020年；Fan 等人，2022年）。在这一步中，会识别制造缺陷导致的永久性故障及相应的故障图信息。因此，这一步不会产生额外成本，只需承担标准晶圆/芯片测试流程通常所需的费用（Xu 等人，2020年；Fan 等人，2022年）。在重新训练技术中，随后会在重新训练过程中纳入故障图信息，并考虑权重和神经元运算如何映射到脉冲神经网络计算引擎上，即所谓的故障感知训练（FAT）。通过这种方式，脉冲神经网络模型有望适应故障的存在，从而保持较高的准确率。这表明，重新训练技术需要：（1）芯片测试流程提供的故障图信息；（2）完整的训练数据集，而由于政策限制，该数据集可能无法获取。此外，每个芯片都有独特的故障图，因此需要单独进行重新训练，进而产生巨大的时间和能耗成本。否则，重新训练技术将无法发挥作用。与此同时，我们提出的 RescueSNN 方法论中的 FAM 技术利用故障图信息，将权重和神经元运算安全地映射到脉冲神经网络计算引擎上。它能确保脉冲神经网络处理不受永久性故障的不利影响，从而保持较高的准确率。尽管每个芯片都有独特的故障图

，需要进行特定映射，但设计映射策略的成本远低于重新训练的成本。此外，我们的 FAM 技术无需任何训练数据集，因此适用于广泛的脉冲神经网络应用。

剪枝的优势与局限性： 如图 1(a) 所示，基于全连接层（FC）的脉冲神经网络架构中的神经元可以被剪枝，同时保持接近原始网络的准确率，因为较高比例的故障“V_{mem} 增加”操作并不会显著降低准确率。先前的研究（Rathi 等人，2019年）已经证明了基于全连接层的脉冲神经网络架构中剪枝的优势，包括减少内存占用和能耗。如果依赖离线训练，即使用训练数据集对脉冲神经网络模型进行离线训练，并在运行阶段推理时保持训练阶段所学知识不变，那么剪枝技术是适用的。然而，如果考虑的是需要在运行阶段定期更新知识以适应不同运行环境（即所谓的动态环境）的基于脉冲神经网络的系统，例如自主移动智能体（如无人地面车辆（UGV）），则剪枝技术并不适用。这是因为，基于脉冲神经网络的系统可能会在不同环境中遇到新的输入特征，而离线训练得到的知识可能无法代表相应类别的识别特征，从而导致运行阶段的准确率较低，并需要通过在线训练来更新知识（Putra 和 Shafique，2021b年，2022a年）。因此，具有未剪枝神经元和无监督学习能力的脉冲神经网络模型，有利于在在线训练期间学习和识别来自运行环境的（未标记）数据样本中的新特征。总之，用户可以根据设计要求选择所采用的脉冲神经网络模型。另一种方案是采用图 1(a) 所示的基于全连接层的脉冲神经网络，并选择是否进行剪枝，因为这类模型能够带来多种优势，例如在无监督设置下采用基于 STDP 的学习时具有较高准确率，以及具备高效的在线训练能力。

6 结论

我们提出了RescueSNN，这是一种用于缓解脉冲神经网络芯片中永久性故障的新型方法论。RescueSNN利用计算引擎的故障图，对脉冲神经网络权重和操作执行故障感知映射，并为所提出的映射技术采用高效的硬件增强。结果表明，RescueSNN无需重新训练即可提高脉冲神经网络容错能力。因此，发生故障的脉冲神经网络芯片可以在其运行寿命期间得到修复，并用于可靠的脉冲神经网络处理。

数据可用性声明

本研究使用了公开可用的数据集。这些数据集可在以下网站找到：http://yann.lecun.com/exdb/mnist/（MNIST），http://fashion-mnist.s3-website.eu-central-1.amazonaws.com/（Fashion MNIST）。

资助

本研究部分由人工智能与机器人中心（CAIR）和网络安全中心（CCS）提供支持；前者由 Tamkeen 根据纽约大学阿布扎比分校研究院 CG010 项目资助，后者由 Tamkeen 根据纽约大学阿布扎比分校研究院 G1104 项目资助。本研究还部分获得项目“eDLAuto：面向自主系统中节能嵌入式深度学习的自动化框架”的支持，该项目由纽约大学阿布扎比分校研究提升基金（REF）资助。作者感谢维也纳工业大学图书馆通过其开放获取资助计划提供出版费用支持。

REFERENCES

Akopyan, F., Sawada, J., Cassidy, A., Alvarez-Icaza,R.,Arthur, J., Merolla, P., 等人（2015年）。TrueNorth：一种65毫瓦、拥有100万个神经元的可编程神经突触芯片的设计与工具流程。*IEEE*集成电路与系统计算机辅助设计汇刊34。doi:10.1109/TCAD.2015.2474396 Baloch, N. K., Baig, M. I. 和Daneshtalab,M.（2019年）。Defender：一种用于可靠片上路由器的低开销高效容错机制。*IEEE Access*7，142843–142854 Basu, A., Deng,L.,Frenkel,C. 和 Zhang,X.（2022年）。脉冲神经网络集成电路：趋势与未来方向综述。载于*2022*年 *IEEE*定制集成电路会议（*CICC*）。1–8。doi:10.1109/CICC53496.2022.9772783 Davies, M., Srinivasa, N., Lin,T., Chinya, G.,Cao, Y., Choday, S.H.,等人（2018年）。Loihi：一种具有片上学习功能的神经形态众核处理器。*IEEEMicro* 38，82–99。doi:10.1109/MM.2018.112130359Diehl,P. 和Cook, M.（2015年）。利用脉冲时序依赖可塑性进行数字识别的无监督学习。《计算神经科学前沿》9，99。doi:10.3389/fncom.2015.00099El-Sayed, S. A. 等人（2020年）。脉冲神经元硬件级故障建模。载于*2020*年*IEEE*第*26*届在线测试与鲁棒系统设计国际研讨会。1–4。doi:10.1109/IOLTS50870。2020。9159745 Fan, S.-K. S.,Cheng,C.-W. 和Tsai,D.-M.（2022年）。前道工艺与后道工艺之间晶圆验收测试和芯片探测的故障诊断。*IEEE*《自动化科学与工程汇刊》19，3068–3082。doi:10.1109/TASE.2021.3106011 Frenkel, C., Lefebvre, M.,Legat,J.和Bol, D.（2019年）。采用28纳米互补金属氧化物半导体工艺的0.086毫米^2 12.7皮焦耳/突触、64K突触、256神经元在线学习数字脉冲神经形态处理器。*IEEE*《生物医学电路与系统汇刊》13，145–158。doi:10.1109/TBCAS.2018.2880425Hanif, M. A., Khalid, F., Putra,R.V. W., Rehman, S.和Shafique, M.（2018年）。鲁棒机器学习系统：深度神经网络的可靠性与安全性。载于*2018*年*IEEE*第*24*届在线测试与鲁棒系统设计国际研讨会。257–260。doi:10.1109/IOLTS.2018.8474192Hanif, M. A., Khalid, F., Putra, R.V. W., Teimoori, M. T., Kriebel,F., Zhang, J.J., 等人（2021年）。面向基于机器学习的系统的鲁棒计算。载于《可靠嵌入式系统》（Springer，Cham）。479–503 Hazan, H., Saunders, D. J., Khan, H., Patel,D., Sanghavi, D. T., Siegelmann, H. T., 等人（2018年）。BindsNET：面向机器学习的Python脉冲神经网络库。《神经信息学前沿》doi:10.3389/fninf.2018.00089 Hazan, H., Saunders, D.J., Sanghavi, D. T.,Siegelmann, H.和Kozma,R.（2019年）。用于图像数据聚类和分类的格映射脉冲神经网络（lm-snns）。《数学与人工智能年刊》doi:10.1007/s10472-019-09665-3Izhikevich, E. M.（2004年）。皮层脉冲神经元应使用哪种模型？*IEEE*《神经网络汇刊》15，1063–1070 Krithivasan, S., Sen, S.,Venkataramani,S.和 Raghunathan, A.（2019年）。用于高能效脉冲神经网络的动态脉冲捆绑。载于*2019*年*IEEE/ACM*低功耗电子与设计国际研讨会。1–6。doi:10.1109/ISLPED.2019.8824897 Lyons, R.E. 和 Vanderkulk, W.（1962年）。利用三模冗余提高计算机可靠性。《*IBM*研究与开发杂志》6，200–209。doi:10.1147/rd.62.0200 Maass, W.（1997年）。脉冲神经元网络：第三代神经网络模型。*NeuralNetworks*10，1659–1671。doi:https://doi.org/10.1016/S0893-6080(97)00011-7

Mercier, R., Killian, C., Kritikakou,A., Helen,Y. 和Chillet,D.（2020年）。通过位混洗缓解片上网络架构中的多重永久性故障。载于*2020*年*IEEE*第*38*届计算机设计国际会议。205–212。doi:10.1109/ICCD50377.2020.00046 Mercier, R., Killian, C., Kritikakou,A., Helen, Y.和 Chillet, D.（2022年）。BiSuT：一种基于片上网络的位混洗技术，用于缓解多重永久性故障。*IEEE*集成电路与系统计算机辅助设计汇刊41，2276–2289。doi:10.1109/TCAD.2021.3101406 Mozafari, M., Ganjtabesh, M.,Nowzari-Dalini, A. 和 Masquelier, T.（2019年）。SpykeTorch：每个神经元最多产生一个脉冲的卷积脉冲神经网络高效仿真。《神经科学前沿》13，625。doi:10.3389/fnins.2019.00625 Painkras, E., Plana, L. A., Garside, J., Temple, S., Galluppi, F., Patterson, C. 等（2013年）。SpiNNaker：用于大规模并行神经网络仿真的1 W、18核片上系统。*IEEE*固态电路杂志48，1943–1953 Park, S.,Kim, S.,Na, B. 和 Yoon,S.（2020年）。T2FSNN：采用首次脉冲时间编码的深度脉冲神经网络。载于第*57*届*ACM/IEEE*设计自动化会议。1–6。doi:10.1109/DAC18072.2020。9218689 Pfeiffer,M.和 Pfeil, T.（2018年）。脉冲神经元深度学习：机遇与挑战。《神经科学前沿》12。doi:10.3389/fnins.2018.00774 Putra, R. V.W.,Hanif, M. A. 和Shafique,M.（2021a年）。ReSpawn：考虑不可靠存储器的脉冲神经网络高能效容错。载于*2021*年*IEEE/ACM*计算机辅助设计国际会议。1–9。doi:10.1109/ICCAD51958.2021.9643524 Putra, R. V.W.,Hanif, M. A. 和 Shafique,M.（2021b年）。SparkXD：一种使用近似动态随机存取存储器进行脉冲神经网络推理的弹性高能效框架。载于*2021*年第*58*届*ACM/IEEE*设计自动化会议。379–384。doi:10.1109/DAC18074.2021.9586332 Putra, R. V.W.,Hanif, M. A. 和Shafique, M.（2022a年）。EnforceSNN：面向嵌入式系统、考虑近似动态随机存取存储器的弹性高能效脉冲神经网络推理。《神经科学前沿》，1242doi:10.3389/fnins.2022.937782 Putra, R. V.W.,Hanif, M. A.和 Shafique,M.（2022b年）。SoftSNN：软错误下脉冲神经网络加速器的低成本容错。载于第*59*届*ACM/IEEE*设计自动化会议论文集。151–156。doi:10.1145/3489517.3530657 Putra, R. V.W.和Shafique,M.（2020年）。FSpiNN：面向存储器高效和能耗高效脉冲神经网络的优化框架。*IEEE*集成电路与系统计算机辅助设计汇刊39，3601–3613。doi:10.1109/TCAD.2020.3013049 Putra, R. V. W.和Shafique,M.（2021a年）。Q-SpiNN：脉冲神经网络量化框架。载于*2021*年*International Joint Conference onNeuralNetworks*。1–8。doi:10.1109/IJCNN52387.2021。9534087 Putra, R. V. W.和 Shafique, M.（2021b年）。SpikeDyn：一种具备动态环境中持续学习和无监督学习能力的高能效脉冲神经网络框架。载于*2021*年第*58*届*ACM/IEEE*设计自动化会议。1057–1062。doi:10.1109/DAC18074.2021.9586281 Putra, R. V. W. 和Shafique,M.（2022a年）。lpSpikeCon：为自主智能体实现低精度脉冲神经网络处理，以进行高效的无监督持续学习。载于*2022*年*International Joint Conference onNeuralNetworks*。1–8。doi:10.1109/IJCNN55064.2022.9892948 Putra, R. V. W. 和Shafique, M.（2022b年）。Mantis：利用脉冲神经网络实现高能效自主移动智能体。*arXiv*预印本 *arXiv:2212.12620*Putra, R. V. W.和 Shafique,M.（2023）。TopSpark：面向自主移动智能体上高能效脉冲神经网络的时间步优化方法论。*arXiv*预印本 *arXiv:2303.01826*。

Radetzki, M., Feng, C., Zhao, X., and Jantsch, A. (2013). 片上网络中的容错方法。*ACM Computing Surveys*46. doi:10.1145/2522968.2522976 Rastogi, M., Lu, S., Islam, N., and Sengupta, A. (2021). 脉冲时序依赖可塑性支持的无监督脉冲神经网络中星形胶质细胞的自修复作用。《神经科学前沿》14, 1351. doi:10.3389/fnins.2020.603796 Rathi, N., Chakraborty, I., Kosta, A., Sengupta, A., Ankit, A., Panda, P., et al. (2023). 探索基于脉冲神经网络的神经形态计算：从算法到硬件。*ACMComputingSurvey*55. doi:10.1145/3571155 Rathi, N., Panda, P., and Roy, K. (2019). 基于脉冲时序依赖可塑性的脉冲神经网络连接剪枝与权重量化，实现节能识别。*IEEE*集成电路与系统计算机辅助设计汇刊38, 668–677. doi:10.1109/TCAD.2018.2819366 Schuman, C. D., Parker Mitchell, J., Johnston, J. T., Parsa, M., Kay, B., Date, P., et al. (2020). 脉冲神经网络在神经形态系统中的弹性与鲁棒性。载于*2020 International JointConferenceon Neural Networks*。1–10. doi:10.1109/IJCNN48605.2020.9207560 Spyrou, T., El-Sayed, S., Afacan, E., Camu˜nas-Mesa, L., Linares-Barranco, B., and Stratigo poulos, H.-G. (2021). 脉冲神经网络中的神经元容错。载于*2021 Design, Automation and Test inEuropeConference and Exhibition*。743–748. doi:10.23919/DATE51398.2021.9474081 Srinivasan, G., Roy, S., Raghunathan, V., and Roy, K. (2017). 基于脉冲时序依赖可塑性的增强型自学习，实现脉冲神经网络中的高效模式识别。载于 *IJCNN*。1847–1854. doi:10.1109/IJCNN.2017.7966075 Stanisavljevi´c, M., Schmid, A., and Leblebici, Y. (2011). 可靠性、故障与容错。载于*Reliability of Nanoscale Circuits and Systems*（Springer）。7–18 Strubell, E., Ganesh, A., and McCallum, A. (2019). 自然语言处理深度学习的能耗与政策考量。载于 *Proc. 57th Annual Meetingof the Association for ComputationalLinguistics*。3645–3650. doi:10.18653/v1/P19-1355 Strubell, E., Ganesh, A., and McCallum, A. (2020). 现代深度学习研究的能耗与政策考量。*Proc. AAAI Conference on Artificial Intelligence*34, 13693–13696. doi:10.1609/aaai. v34i09.7123 Sze, H. Y. (2000). 使用循环冗余校验快速检查纠错码的电路与方法 美国专利 6,092,231 Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. (2019). 脉冲神经网络中的深度学习。*Neural Networks*111, 47–63. doi:10.1016/j.neunet.2018.12.002 Vadlamani, R., Zhao, J., Burleson, W., and Tessier, R. (2010). 使用自适应双模冗余稳定多核软错误率。载于*2010 Design, Automation and Test in Europe Conference andExhibition*。27–32. doi:10.1109/DATE.2010.5457242 Vatajelu, E.-I., Di Natale, G., and Anghel, L. (2019). 专题会议：硬件实现的脉冲神经网络（脉冲神经网络）的可靠性。载于 *2019 IEEE* 第*37*届超大规模集成电路测试研讨会。1–8. doi:10.1109/VTS.2019. 8758653 Wang, J., Ebrahimi, M., Huang, L., Xie, X., Li, Q., Li, G., et al. (2019). 片上网络的高效可测试性设计方法。*IEEE Transactions on Computers*68, 198–213. doi:10.1109/TC.2018.2865948 Werner, S., Navaridas, J., and Luj´an, M. (2016). 规避片上网络永久故障的设计方法综述。*ACM Computing Surveys*48. doi:10.1145/2886781 Xu, H., Zhang, J., Lv, Y., and Zheng, P. (2020). 半导体制造中晶圆验收测试参数的混合特征选择。*IEEE Access* 8, 17320–17330. doi:10.1109/ACCESS.2020. 2966520

Zhang, J.J.,Gu, T.,Basu,K.和Garg, S.（2018年）。分析并缓解永久性故障对基于脉动阵列的神经网络加速器的影响。载于*2018*年第*36*届*IEEE*超大规模集成电路测试研讨会。1–6。doi:10.1109/VTS.2018.8368656