脉冲神经网络对抗攻击与防御研究进展

丁健豪[1),2)] 刘俣伽[1),2)] 卜 通[1),3)] 郝泽成[1),2)] 余肇飞[1),3)] 黄铁军[1),2)]

[1)](北京大学视频与视觉技术国家工程研究中心 北京 100871)

[2)](北京大学计算机学院 北京 100871)

[3)](北京大学人工智能研究院 北京 100871)

摘 要 随着神经形态计算的发展,脉冲神经网络有望广泛部署于实时场景和安全关键型应用。现有研究表明,脉冲神经网络内部信息表示的离散特点导致其相比于人工神经网络,可能对微小扰动不敏感,因此被认为相比于传统人工神经网络具有更强的抗干扰能力。近年来的研究表明,脉冲神经网络同样面临与人工神经网络类似的对抗攻击威胁。目前针对脉冲神经网络的现有综述主要关注脉冲神经网络的结构设计、训练方法及硬件实现。本文聚焦于脉冲神经网络的鲁棒性,首次对脉冲神经网络对抗攻击和防御方法进行了系统综述。针对攻击方法,本文归纳了基于数据模态的攻击方法、基于可微近似的梯度攻击方法以及梯度无关的攻击方法;在防御上,当前已有多项研究围绕脉冲神经网络鲁棒性展开探索,显示出其在对抗防御中的潜力。本文归纳了当前常用的防御方法,包括输入防御方法、网络防御方法、输出防御方法等,旨在提升模型应对扰动时的安全性与稳定性。最后,本文总结了当前研究面临的挑战,并展望了脉冲神经网络在对抗攻击与防御研究中的未来发展方向。

关键词 脉冲神经网络;脉冲编码;对抗攻击;对抗鲁棒性;对抗训练

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2026.00211

Research Progress on Adversarial Attack and Defense of Spiking Neural Networks

DING Jian-Hao[1),2)] LIU Yu-Jia[1),2)] BU Tong[1),3)] HAO Ze-Cheng[1),2)]
YU Zhao-Fei[1),3)] HUANG Tie-Jun[1),2)]

[1)](National Engineering Research Center for Visual Technology, Peking University, Beijing 100871)

[2)](School of Computer Science, Peking University, Beijing 100871)

[3)](Institute for Artificial Intelligence, Peking University, Beijing 100871)

Abstract With the advancement of neuromorphic computing, spiking neural networks (SNNs) are expected to be widely deployed in real-time scenarios and safety-critical applications. SNNs simulate the behavior of neurons in biological brains through spatiotemporal neuronal dynamics. Neurons in SNNs update their membrane potentials over time and output spike signals of 0 and 1. Existing studies have shown that the discrete characteristics of internal information representation of SNNs make them less sensitive to small perturbations than traditional artificial neural networks (ANNs), and therefore they are considered to be more robust to adversarial attacks than traditional ANNs. Research in recent years has shown that SNNs also face similar threats from adversarial attack as ANNs. Existing reviews of SNNs primarily focus on the structural design,

收稿日期:2025-05-08;在线发布日期:2025-10-29。本课题得到国家自然科学基金(Nos. 62422601,U24B20140,62176003)、北京市科技新星(Nos. 20230484362,20240484703)资助。**丁健豪**,博士,博士后,中国计算机学会(CCF)会员,主要研究领域为类脑计算、脉冲神经网络。E-mail:djh01998@alumni.pku.edu.cn。**刘俣伽**,博士,博士后,主要研究领域为脉冲神经网络、对抗攻击与鲁棒性。**卜 通**,博士研究生,主要研究领域为机器学习、类脑计算。**郝泽成**,博士研究生,主要研究领域为视觉信息处理与神经形态计算。**余肇飞(通信作者)**,博士,助理教授,主要研究领域为类脑计算、神经网络。E-mail:yuzf12@pku.edu.cn。**黄铁军**,博士,教授,主要研究领域为视觉信息处理与神经形态计算。

training methods, and hardware implementation. Given the fundamental differences between SNNs and ANNs, the robustness mechanisms as well as attack and defense strategies of SNNs are unique, and thus cannot be directly derived from the well-established ANN framework. This paper focuses on the robustness and systematically reviews adversarial attacks and defense methods for SNNs for the first time. Regarding attack methods, this paper summarizes attack methods based on data modality, gradient attack methods based on differentiable approximation, and gradient-independent attack methods. The first category targets input data modalities, including attacks on static images, attacks on neuromorphic event datasets, and general attack methods applicable across data modalities. The second category of attacks leverages the differentiable approximation training methods commonly used in deep SNNs, generating input perturbations through techniques such as conversion approximation, backpropagation through time approximation, and spike rate approximation. The third category comprises gradient-independent attack methods, designed to overcome the issues associated with differentiable approximations. In terms of defense mechanisms, many studies have explored their robustness, showing their potential in adversarial defense. This paper summarizes several common defense methods, including input defense methods, network defense methods, output defense methods, etc., aiming to improve the security and stability of the model when dealing with perturbations. The defense mechanisms cover the entire network processing pipeline. At the input stage, spike encoding and filtering methods are employed. During the network processing stage, robustness is enhanced through improved neuron design, interval bound propagation, adversarial training, regularization training, and network lightweighting. At the output stage, spike decoding methods are utilized. Existing research indicates that SNNs are inherently more robust than traditional ANNs. These defense methods aim to further exploit their intrinsic mechanisms or draw insights from traditional defense methods. Finally, this paper summarizes several challenges in current research and looks forward to the future development direction of SNNs in adversarial attack and defense research. On the attack side, researchers should explore novel methods that target defense vulnerabilities, exploit spiking encoding mechanisms, and leverage event-based data properties, while developing lightweight and highly generalizable gradient-based or gradient-free approaches. On the defense side, it is essential to enhance robustness by integrating spiking encoding principles, temporal processing, and bio-inspired mechanisms, design specialized defense solutions for emerging SNN architectures, achieve effective protection under neuromorphic hardware constraints, and validate effectiveness in real-world scenarios. The secure deployment of SNNs depends on the continuous evolution of attack and defense technologies alongside synergistic advances in neuromorphic computing.

Keywords spiking neural network; spike coding; adversarial attack; adversarial robustness; adversarial training

1 引 言

人工智能(Artificial Intelligence, AI)的终极目标是构建具有与人类智慧水平相当的强人工智能系统,以完成人类各种智能行为。自 2012 年以来,深度学习作为人工智能的重要分支迅猛发展,成为第三次人工智能浪潮的推动力量[1-3]。深度学习利用多层神经网络从数据中学习到有用的特征和模式。然而,对抗攻击的出现对计算机视觉、自然语言处理、医疗诊断等各个领域构成严重挑战[4-7]。脉冲神经网络(Spiking Neural Network, SNN)被誉为继基于非线性激活函数人工神经网络(Artificial Neural Network, ANN)的第三代人工神经网络[8-10]。

传统 ANN 使用浮点数表示信息。而 SNN 通过时空动力学和脉冲表示模拟生物大脑中的神经元行为[11-13]。SNN 的神经元随着时间更新其膜电位并输出 0 和 1 的脉冲信号[14-17]。

对抗攻击通过向原始数据添加微小扰动产生使深度学习模型误判的对抗样本[4-5,18-19],带来安全和隐私方面的风险。例如,一张猫的图片,经过添加人类难以察觉的扰动后,可能就会被深度学习模型错误地识别为狗。这说明深度学习模型对于输入数据的特定变化非常敏感。在自动驾驶系统中,对抗样本可能误导模型无法正确地识别道路标志或行人,造成严重的交通事故[20]。或者在人脸识别系统中,对抗样本可能使用户无法正确地验证身份或授权,导致身份盗窃或信息泄露[21]。

SNN 被认为在鲁棒性上具有相比于传统 ANN 的优势[4,22]。国内外知名科研机构正在积极开展如何利用 SNN 改善鲁棒性的研究工作[23-25]。研究认为 SNN 离散的信息表示和天然时域滤波机制有助于解决深度学习中模型易受对抗扰动误导的问题。脉冲的离散特点使攻击者更难通过微小的输入连续变化改变 SNN 输出[26]。而 SNN 时序编码使得扰动可以在不同时间被处理,减小了网络处理扰动的压力[27],有助于在面临对抗攻击时保持稳定性和高准确性[28]。

即使 SNN 具有一定抗扰动特质,其仍会受到对抗攻击的威胁[29-30]。构造对抗扰动需要让网络测试时损失函数增大,近年发展的端到端反向传播训练 SNN 方法为攻击提供了有效梯度。传统 ANN 中梯度的计算方式通常是固定的。而 SNN 得到梯度的方式非常多样,每一种梯度计算方法都能用于对抗样本构造[25]。例如,SNN 可以通过从传统 ANN 转换得到,也可以通过端到端随时间反向传播训练得到。因此,SNN 的对抗攻击挑战实际比传统 ANN 更复杂。

目前关于 SNN 的综述大多聚焦于网络结构设计、训练方法以及神经形态硬件实现等方面[31-34],对鲁棒性研究的整理与分析较为有限。相比之下,传统 ANN 在对抗攻击与防御方面已形成较为成熟的研究体系[35-36]。由于 SNN 与 ANN 存在本质差异,SNN 鲁棒性表现及攻击防御机制具有独特性,不能简单套用 ANN 领域的研究成果。在此背景下,有必要系统综述 SNN 的对抗攻击与防御方法,梳理研究进展,分析挑战,为后续研究提供理论和技术参考。图 1 基于 SNN 推理过程,对现有国内外研究进行了整理与归纳。本文进一步汇总相关攻击与防御方法,便于读者理解与应用。

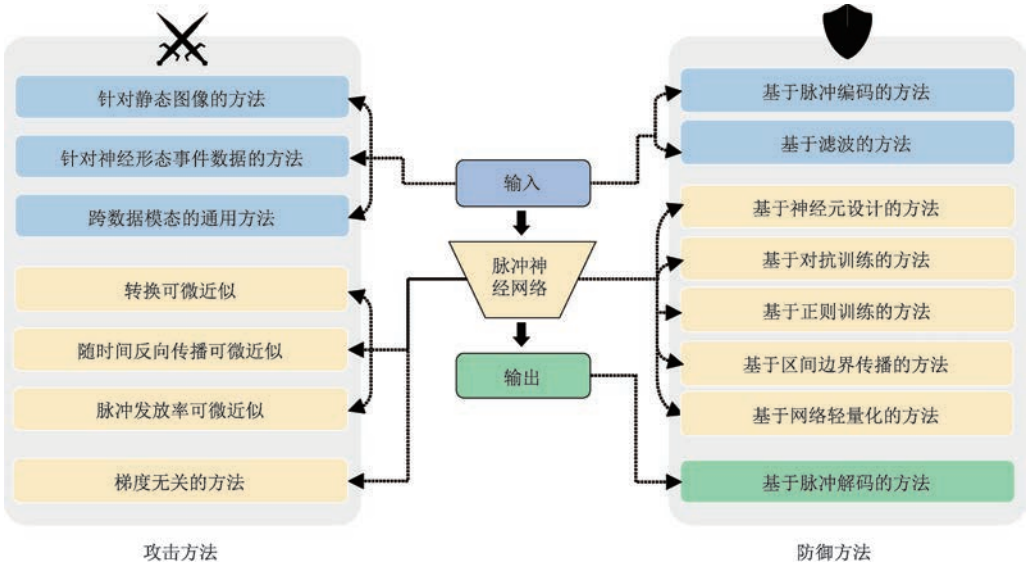

图 1　脉冲神经网络对抗攻击及防御方法研究进展总结

攻击方法通过影响输入模式,使得网络得到错误输出。攻击方法可以被归纳为:(1)在输入数据攻击上,研究者提出了针对不同输入数据模态的攻击方法。在视觉任务中,输入数据可以是包含时间长度、数据高度和数据宽度的神经形态数据,也可以是传统的图像数据。因此,攻击方法从输入类型上可分为:①针对静态图像的攻击方法;②针对神经形态事件数据集的攻击方法;③跨数据模态的通用攻击方法。其中,针对传统静态图像的攻击方法主要结合 ANN 中的攻击方法。针对神经形态事件数据集的攻击方法

考虑了数据的多维性和时间性质[37-39]。跨数据模态的通用攻击方法可攻击的数据包括神经形态数据和图片数据等模态，旨在融合考虑数据模态[40-41]。(2)深度 SNN 训练目前最常用可微近似。结合 SNN 可微近似，研究者提出了结合不同 SNN 可微近似的攻击方法。这些方法主要包括以下几种：①转换可微近似的攻击方法[25-26]；②随时间反向传播可微近似的攻击方法[26,29]；③脉冲发放率可微近似的攻击方法[25,30]。这些方法借助 SNN 训练方法的有益经验反过来为 SNN 构造输入扰动。(3)除了基于可微近似的方法之外，为了绕开可微近似带来的模型误差和对抗攻击适应性问题，研究者还提出了梯度无关的攻击方法。

防御方法利用 SNN 在输入、网络处理和输出上的时序特点和离散特点设计相关方法提高 SNN 的鲁棒性。研究者提出方法旨在挖掘 SNN 的内在鲁棒性机理，或旨在从传统 ANN 防御方法中借鉴经验。现有研究已明确指出，SNN 天然具有比传统 ANN 更好的鲁棒性[22,25]。根据处理阶段，防御方法可归纳为：(1)针对网络输入，研究者提出了①基于脉冲编码的方法[26-28,42-43]以及②基于滤波的方法[44-45]。(2)针对 SNN，研究者提出了①基于神经元设计的防御方法[46]；②基于区间边界传播的防御方法[25]；③基于对抗训练的防御方法[24]；④基于正则训练的防御方法[25]；⑤基于网络轻量化的防御方法[47-48]。(3)针对网络输出，研究者提出了基于脉冲解码的方法以提高 SNN 的鲁棒性。

研究 SNN 对抗鲁棒性具有两方面意义。其一，理解 SNN 的安全可靠性有助于开发高可信的生物启发的深度学习模型，提高其在工业应用中的稳定性。SNN 更加适合部署在易成为黑客攻击对象的终端边缘设备的神经形态系统中，因此研究 SNN 的鲁棒性具有极高的现实意义；其二，SNN 是类脑计算的重要研究方向，探索 SNN 对抗鲁棒性相比于传统方案有助于理解人脑鲁棒工作机制。综上，SNN 对抗攻击与防御研究具有应用与理论双重意义。

本文将面向 SNN 对抗攻击与防御学术研究前沿，探讨网络可微性质、神经形态数据特点等因素对鲁棒性的影响。本文后续内容组织结构安排如下：第 2 节简要介绍 SNN 相关背景，包括传统 ANN 常用的攻击防御方法以及 SNN 研究进展；随后，第 3 节介绍了 SNN 对抗攻击方法；在第 4 节，本文介绍了 SNN 的防御方法；第 5 节本文提出了未来可能的研究方向；第 6 节对全文进行了总结。

2 脉冲神经网络及攻防背景

SNN 的信息处理模式相比于传统 ANN 有所不同，在对抗攻击上挑战也与 ANN 不完全相同。本节首先对 ANN 对抗攻击与防御方法进行概述；接着介绍 SNN 及其学习方法；最后，介绍 SNN 应用场景与挑战。

2.1 深度学习对抗攻击与防御

对抗攻击旨在改变模型输入导致模型预测与标签有较大偏差。Szegedy 等人通过攻击研究指出了深度神经网络关于输入中扰动的脆弱性，只需添加微小的扰动构造对抗样本即可误导网络产生错误输出[49]。对抗攻击还被发现具有可迁移性，相同扰动图像可以影响多个模型。图 2 展示了一个经典案例。一个图像分类模型在没有扰动的情况下以 57.7%的置信度正确识别一张熊猫图像。在对抗攻击后，该图像分类模型却以 99.3%的置信度错误地将其分类为长臂猿[4]。

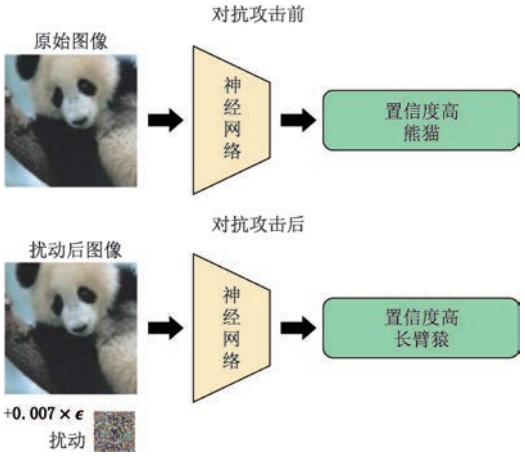

图 2 深度学习对抗攻击

对抗攻击可以被形式化为优化问题，求解在未受扰动数据 $\boldsymbol{x}$ 的 l_p 邻域空间如何设计扰动 δ 满足式(1)：

$$\underset{\delta}{\operatorname{argmax}} L(f(x+\delta;W),y)\ \text{s.t.}\ \|\delta\|_p \leqslant \epsilon \quad (1)$$

其中，L 为损失函数，f 为深度网络，其为参数 W。y 为真实标签，ϵ 为扰动强度。攻击方法的代表为基于梯度的方法，如快速梯度符号法(Fast Gradient Sign Method, FGSM)[4]。该类方法可以沿着梯度方向对数据实施有效扰动。从目标模型参数是否对攻击者的可见性上看，攻击方法可依照是否了解目标模型信息被大体分为白盒攻击[50]和黑盒攻击[18]。

为了应对攻击方法，研究者提出了多种神经网

络防御方法,旨在提升模型的鲁棒性:

其一是预处理方法。该类方法在将图像输入神经网络前对其进行处理,以减小对抗样本的影响。预处理通常包括去噪、随机分块、重构、缩放与像素量化等操作。该类方法通常不需要对模型进行修改,可以直接应用于已经训练好的模型,计算开销相对较低[51-53]。然而,其鲁棒性提升受限,难以防御针对预处理机制本身设计的攻击。

其二是对抗训练,被广泛认为是提升深度模型鲁棒性的最有效策略之一。其核心思想是对抗样本参与训练,使模型在学习过程中可以在潜在攻击影响下获得泛化能力,从而具备在攻击下维持性能的能力。然而,该方法需在每个训练步骤内确定数据的对抗样本,计算开销大大增加;此外,对抗训练后模型对某种范数下生成的对抗样本有效性强,但对其他类型扰动鲁棒性有限[54-56]。

其三是正则化方法。该类方法通过正则化约束梯度变化,提高模型对扰动的抵抗力。例如,Parseval 正则化约束网络中的权重矩阵保持近似正交,从而限制扰动在每一层的放大。流形正则化网络在模型损失中加入样本在数据流形上的邻近性假设,通过对同一类样本在嵌入空间中的相似性进行约束,限制扰动带来的决策边界偏移。此外,梯度惩罚类正则通过约束输入梯度范数,引导模型学习平滑的判别函数,提高模型对输入微小扰动的鲁棒性[57-59]。

其四是特征去噪。该类方法在卷积模块中加入处理特征图的去噪模块。去噪模块如基于注意力机制的噪声选择、鲁棒池化操作或非局部均值去噪可以显式地抑制中间特征的噪声。虽然本身可能不会提高模型在原始数据集上的分类准确率,但当其与对抗训练结合使用时可提高模型的鲁棒性[60]。另外,一些研究还提出特征压缩方法,通过删除冗余通道,减少攻击者可操控空间,从而降低对抗攻击成功率。该类方法在计算效率方面相对较高,易于与现有模型结构兼容[61]。

总体而言,深度学习的防御方法主要的思路是训练时进行网络全局的鲁棒性优化,而在推理时进行扰动抑制。这一思路同样适用于 SNN。前者可为鲁棒训练方法借鉴,如 SNN 适合的对抗训练;而后者则需结合 SNN 输入格式设计相应的预处理模块,以发挥其时序优势并增强其抵抗输入扰动的能力。

2.2　脉冲神经网络及其学习方法

SNN 在信息处理模式上类似生物神经元,需要接收一段时间的离散脉冲并产生输出[14,62]。通常处理过程包含输入编码、脉冲表征、输出解码等流程,如图 3 所示。与最常用的深度 ANN 不同,SNN 类似循环神经网络具有天然时间维度,当输入图像等静态数据需要额外编码方法参与[16]。最常用的脉冲神经元是漏电积分-发放模型(Leaky Integrate-and-fire model, LIF)。积分发放神经元的动态过程可以被形式化为式(2)~(4)。

$$v_i^l[t]=\lambda u_i^l[t-1]+\sum_j w_{ij}^l s_j^{l-1}[t] \tag{2}$$

$$s_i^l[t]=H(v_i^l[t]-\theta) \tag{3}$$

$$u_i^l[t]=v_i^l[t](1-s_i^l[t]) \tag{4}$$

式中,来自 j 个突触前神经元的脉冲输入 s_j^{l-1} 加权构成第 l 层第 i 个神经元的输入电流。w_{ij}^l 表示权重。λ 表示膜电位泄漏参数。当充电后的膜电位 v_i^l 超过阈值 θ 时,在该时刻产生脉冲。在脉冲产生后,神经元的膜电位将被重置为静息电位,默认为 0,如式(4)所示。这种“重置到静息电位 0”的方式被称为硬重置。相比之下,另一种常用方式为软重置。硬重置将膜电位归零,软重置则是在现有膜电位基础上减去阈值。软重置下,神经元状态会在脉冲发放后减少与阈值相等的电压但并不保证回到静息电位。其过程可表示为式(5):

$$u_i^l[t]=v_i^l[t]-\theta s_i^l[t] \tag{5}$$

当表示积分发放模型(Integrate-and-fire model, IF)[9]时,$\lambda=1$。$H(\cdot)$ 为单位阶跃函数,只有当输入大于 0 时输出为 1,否则为 0。

SNN 接收序列输入,因此图片需要在经过编码后才能作为 SNN 首层输入,如图 3 所示。编码将静态图片表示为输入序列。目前深度 SNN 最常用直接编码,将每一个像素映射为像素值不变的时间序列。它能够有效地保留原始输入的信息,使得 SNN 可以更好地理解和学习复杂的数据特征。之后被广泛用于基于代替梯度训练的 SNN 中[63-65]。频率编码通常考虑脉冲计数[66-68]、时间相关频率[69]、群体活动频率[70]。SNN 解码通常采用脉冲计数频率对信息。

SNN 学习方法主要关注如何构建能够在神经形态芯片上执行计算功能 SNN。不同于深度学习的反向传播,目前 SNN 学习领域并没有出现公认的最佳方法[71-72]。目前常用的深度 SNN 训练方法更多考虑如何在 SNN 上传递监督信号,进而有监督地训练网络。与传统神经网络不同,SNN 中的脉冲产生函数是不可导的单位阶跃函数。依照学习监

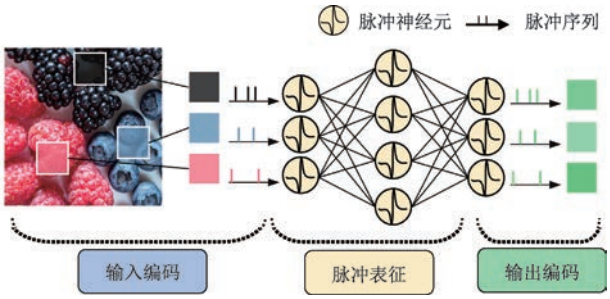

图 3 脉冲神经网络示意图

督方式可以被划分为早期有监督方法、间接监督方法、近期有监督方法。

早期有监督方法通过调整脉冲时间回避了脉冲产生函数不可导的问题，它们被视为早期的时序学习方法[73]，例如 SpikeProp[74] 和 ReSuMe[75]。SpikeProp 通过链式法则基于脉冲时间进行反向传播，但其具有一定的局限性：仅适用于 SRM 模型，并且神经元脉冲次数限制为 1。为了突破局限性，Ghosh-Dastidar 等人提出了 Multi-SpikeProp 学习方法，允许神经元多次触发[76]。相对于 SpikeProp，Multi-SpikeProp 在 EEG 和 Iris 分类任务上显著提高了分类准确率。然而，上述工作主要集中在浅层神经网络上，无法应用在深层 SNN 中。

间接监督方法间接利用 ANN 训练好的权重，将其转换到深度 SNN 上[77-78]。因此也被称为 ANN-SNN 转换方法。Perez 等人开展最早的 SNN 转换研究，实现了对神经形态视觉传感器数据的实时处理[79]。随后，Cao 等人提出了 ANN 和 SNN 的精确转换关系，首次将 ReLU 激活的 ANN 转换为 IF 神经元[80]。其后，Diehl 等人提出了一种权重标准化方法，解决神经元激活过高或过低导致的准确性损失问题。该方法首次为 ReLU 激活的转换工作提供了原则性的指导[81]，并设计了数据驱动的标准化方法自动调整神经元参数。Rueckauer 等人之后进行了对 ANN-SNN 转换的详细理论解释，发现 SNN 神经元的输出与 ReLU 激活输出存在误差，并提出软重置方式降低误差。这项工作还拓展了转换方法的网络模块适用范围，真正使转换方法变成实用技术[82]。进一步的工作集中在频率编码方法上，采用脉冲频率对应 ReLU 的方法，并将其应用于不同应用场景，如物体检测与跟踪[83]。考虑到神经形态芯片的约束和能耗要求，除频率编码之外的神经元编码方式也被提出用于转换，包括时序编码和相位编码[84-86]。上述工作推动了 ANN 到 SNN 转换的发展，使其适应了更广泛的应用需求。

近期有监督方法主要研究基于代替梯度的端到端反向传播方法，它通过代理梯度函数等工具近似梯度并沿时间展开进行反向传播训练[15,66]。代替梯度用于替代不可导的阶跃函数[15,87-89]克服了脉冲的不可导性。SNN 在前向传播中记录每个时间步上的膜电位和脉冲，而在反向传播中将每个时间步展开以进行梯度传递，类似于循环神经网络的时间展开梯度传播方式，因此也称为时空反向传播（Spatio-temporal backpropagation，STBP）或随时间反向传播（Back propagation through time，BPTT）[17]。图 4 中展示了反向传播过程。原本绿色箭头处是单位阶跃函数梯度无法传递。代替梯度函数通过平滑阶跃函数使得梯度可以通过绿色箭头传递。确定某时刻膜电位的损失函数需要来自下个时刻膜电位的梯度信息（红色箭头）以及来自本时刻的脉冲梯度信息（绿色箭头），而本时刻的脉冲梯度信息需要下个时刻的膜电位的梯度信息（蓝色箭头）。近期研究工作使用各种技术提高 SNN 分类任务的性能，例如批标准化等[90-92]。与转换方法相比，这些经过直接有监督训练的 SNN 具有更低的推理延迟[14,93]，但训练时需要更多的计算资源和内存[94]。因为反向传播需要在时间上对梯度进行积分，增加了计算和内存管理的复杂性。近年兴起了另一类端到端学习时序学习方法，通过前后层的脉冲时间传递梯度，可充分利用脉冲时间编码的优势[95-96]。时序学习方法也采用类似代替梯度的思想。Kim 等人提出了脉冲时间和膜电位的偏导数由前后两个时间步的膜电位的负倒数来替代。近年的工作极大地提高了时序学习方法的可用性。Zhang 等人提出了神经元间和神

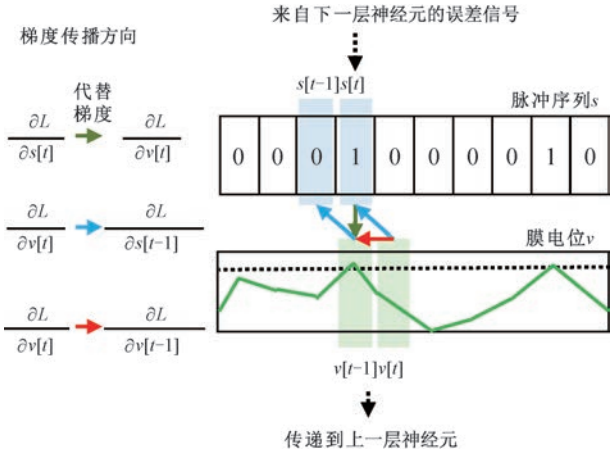

图 4 随时间反向传播

经元内反向传播机制,用于学习时序信息[95]。另一项由 Zhu 等人的工作发现了时序学习方法的梯度不变性性质,即前一层脉冲时间的偏导数之和等于后一层脉冲时间的偏导数之和。该发现极大地提高了时序学习的理论有效性,并提供了相应的改进方法[96-97]。

上述学习算法在为深度 SNN 带来更好性能的同时也为 SNN 带来了潜在的威胁。其梯度信息可以用来设计有效的攻击方法。

2.3　脉冲神经网络应用潜力与鲁棒性挑战

SNN 相对于传统 ANN 在硬件计算方面具有高并行性和低能耗的优势,因此在某些对响应速度和能效比要求较高的实际应用中具有巨大潜力。图 5 展示了 SNN 相关应用,如机器人控制[98-99]、智能飞控系统[100-101]、自动驾驶[102-103]、脑机接口[104-105]、神经形态相机处理[79,90]等方面。这些系统往往需在复杂动态环境中快速做出决策,或以较低能耗持续处理大量的传感数据[106]。SNN 的事件驱动计算机制与这些场景中对异步稀疏数据的处理需求天然契合。然而,鲁棒性不足将可能带来严重的安全隐患,导致系统在关键任务中做出错误响应。

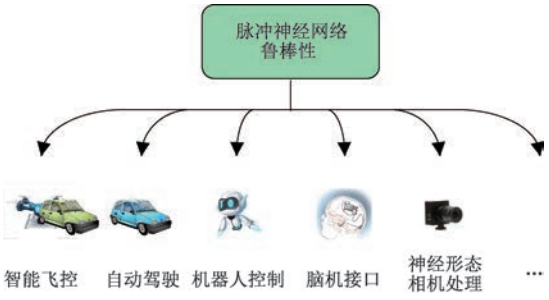

图 5　脉冲神经网络典型应用

SNN 适合处理高速、异步的传感器数据流,如视觉、听觉和触觉输入[107-108]。结合神经形态传感器,SNN 能充分发挥其在处理稀疏时序脉冲信息方面的能力。例如,动态视觉传感器(Dynamic Visual Sensor,DVS)采用地址事件编码格式生成稀疏脉冲流。当像素检测到亮度变化超过对数阈值时,传感器会产生包含像素位置、时间和极性的事件(+1 或 −1)。此类传感器通常具有较高的时间分辨率和动态范围[109],可以在毫秒或更短的时间内响应。由于其事件特性,适合与神经形态芯片结合用于高速识别、跟踪。该数据形式与 SNN 输入特征高度匹配,在快速移动目标跟踪和复杂光照条件下的场景感知中具有优势[110-112]。这些任务往往部署在资源受限的边缘设备中,需要神经形态计算芯片保持低功耗的同时具备高吞吐计算能力。

鲁棒性的提升不仅依赖于算法与训练策略的改进,还与底层硬件平台的协同设计密切相关。神经形态计算的发展推动了一系列支持 SNN 计算的芯片诞生如 Neurogrid[113]、TrueNorth[114]、SpiNNaker[115]以及 Darwin[116]。它们通过模拟神经元和突触之间的事件驱动通信,提供了低功耗、高并发的计算支持。与此同时,混合架构的 Tianjic 芯片,将 SNN 的生物拟态与传统 ANN 计算性能相融合,为 SNN 的应用提供了更具灵活性和计算性能的硬件支撑[11]。在传感端,神经形态视觉传感器根据视觉功能的生物启发原则设计,分为差分型[109,117]和积分型[110,112,118],前者更适合运动检测,后者则提升了纹理捕捉能力。上述高采样率、低冗余的传感器生成的事件流数据能够有效与 SNN 及其芯片平台协同工作,减少因处理高密度事件流的拥塞而带来的输入扰动影响。

随着 SNN 在各类关键场景中的广泛部署,其安全性问题亦愈发重要。研究证明,SNN 同 ANN 一样面临传统深度网络的对抗攻击威胁[22-24]。如图 6 所示,SNN 受到来自深度学习攻击方法和 SNN 可微近似的双重挑战。某些基于代替梯度的训练方法,可能使攻击者更易获取网络的近似梯度方向,从而生成对抗样本实施攻击[25-26]。其次,梯度攻击方法需要与可用的近似梯度方法叠加使用,提高了网络的鲁棒性挑战[30,40]。

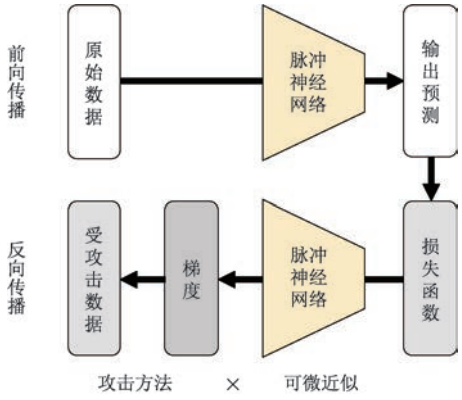

图 6　脉冲神经网络的鲁棒性挑战

因此,在推动 SNN 落地应用的同时需同步加强对抗攻击等防御方法的研究,提升其在现实应用中的鲁棒性与安全性。

3　脉冲神经网络对抗攻击方法

SNN 的对抗攻击与传统神经网络对抗攻击有

很大不同。传统 ANN 攻击方法输入和输出都是连续的实数值,通过梯度生成对抗样本,如 FGSM 与 PGD。这些方法需要 ANN 具有可微和反向传播能力。而 SNN 无法直接反向传播。因而发展出两类方法:可微近似梯度攻击和梯度无关攻击。本节将介绍相关设计。本节首先分析 SNN 的脆弱性成因。接着介绍数据模态、可微近似以及梯度无关的攻击方法。

3.1 脆弱性分析

从深度学习视角来看,SNN 可以被攻击并不意外。本节首先分析 SNN 的脆弱性。从神经元、编码以及网络层面分别介绍 SNN 的脆弱性成因。

(1)脉冲神经元

为实现训练,SNN 常采用可微近似方法,但可微化处理也暴露了其脆弱性。攻击者一旦知道网络结构和参数就能够借助可微近似构造精确扰动。攻击者可以通过扰动操控膜电位越过阈值或抑制其发放,导致输出错误。

(2)网络编解码

SNN 主流的输入输出编码方式多为频率编码。因此,SNN 在处理上也与传统 ANN 接近。频率编码降低了时序动态特性,信息表示为一种统计平均形式,忽略了时间结构也能够表示信息的事实。在频率编解码中,层间通过脉冲累积传递信息,类似 ANN 的加权求和加激活函数。SNN 前向传播在计算图上与 ANN 几乎等价,仅将连续值替换为脉冲频率。频率编码中的直接编码被认为给 SNN 带来了较差的鲁棒性。Kundu 等人认为直接编码的 SNN 鲁棒性较差[24]。Sharmin 等人认为基于直接编码转换的 SNN 包含较少的动态信息,导致网络鲁棒性较差[26]。

(3)网络编解码

SNN 中信息在层间通过脉冲逐步积累并影响最终决策。输入扰动一旦进入网络,便可能在膜电位与脉冲中逐层累积并传递至输出,在深层 SNN 中尤为明显。因此,即使单个时间步的扰动幅度很小,也可能在推理过程中导致误分类的行为。

3.2 基于数据模态的攻击方法

在计算机视觉领域,SNN 既可以用于静态图像相关的任务,又可以用于神经形态数据相关的任务。事件数据每个序列都是异步的,非常适合神经形态芯片处理。常用的神经形态数据又可分为事件数据和脉冲数据[119]。因此为了验证 SNN 鲁棒性需要为静态图像数据集合事件数据集分别设计攻击方法。对于静态图像数据的攻击可以适当结合深度学习方法完成,然而对事件数据则不同。并不是所有的可微近似都适合这种序列数据。例如有些可微近似要求将神经元序列响应看作浮点值,需要额外设计才能得到序列扰动方案。下面分别介绍针对静态图像的对抗攻击方法、针对事件数据的对抗攻击方法,以及跨数据模态的攻击方法。图 7 中给出了三种方法的示意图。

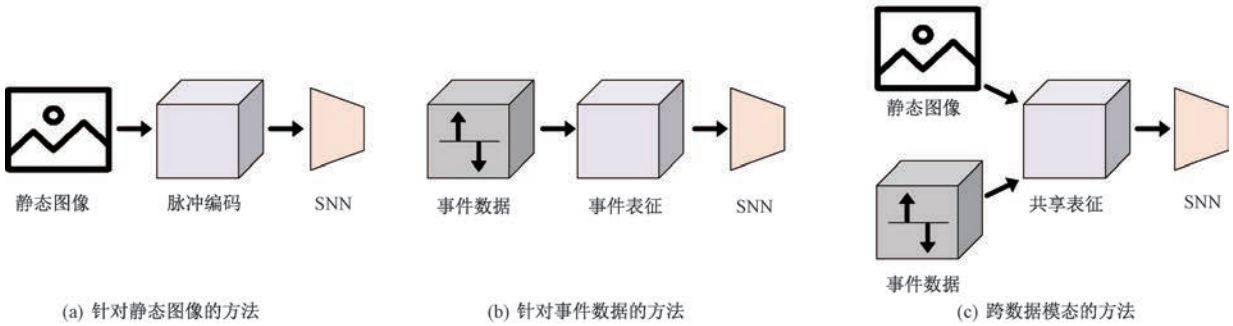

图 7 基于数据模态的攻击方法

(1)针对静态图像的攻击方法

图像输入 SNN 需要首先经过编码。攻击方法目前已经覆盖常用的频率编码方案,包括泊松编码和直接编码。

Sharmin 等人首先提出了针对泊松编码的静态图像的对抗攻击方案[22]。作者首先对图像进行泊松编码,即根据输入的像素强度以一定的概率在每个时间步触发脉冲。转换后的泊松序列输入到第一个卷积层并得到第一层卷积层的输出。对第一层卷积层的输出关于损失函数求导可得其梯度。将其梯度输出序列的梯度求平均并反卷积到和输入维度一样可得估计的泊松编码前图像的梯度,并据此设计 FGSM、PGD 等攻击。该方法建立在白盒设定下,需获取完整模型结构与梯度信息。但是作者也进行

了黑盒攻击测试,即攻击者不知道目标模型的参数结构,验证了攻击方法的迁移效果。该方法本质上是通过梯度近似。除了上述梯度近似方法,Sharmin 等人[26]还设计了一种基于 ANN 的间接 SNN 攻击方法,该方法规避了在脉冲域求导的挑战。下文将其命名为转换可微近似的攻击方法。Mukhoty 等人的工作中同样涉及针对泊松编码的白盒攻击[120]。作者采用的做法是攻击过程中利用直通估计器解决泊松编码中非可导节点的梯度传播问题,从而可在反向传播中计算扰动方向。

Kundu 等人针对直接编码的静态图像提出攻击方法[24]。直接编码相比于泊松编码更容易得到对抗扰动,因为直接编码并不涉及将图片编码到某个脉冲域上,梯度扰动可以直接施加在数据上。作者使用了 BPTT 方法计算损失函数对每个输入像素的梯度再在时间轴上累加起来。其余步骤与 ANN 攻击相同。作者验证了应用 FGSM 和 PGD 攻击的效果。然而复杂攻击方法的效果不明。这一缺点在后续论文中得到改进。Ding 等人将直接编码的图像攻击方法拓展到 RFGSM、BIM 等攻击方法中[25]。Liu 等人将这种攻击方法拓展到 AutoPGD 多步攻击,通过自动调节步长实现更有效的攻击[121]。此外,作者还采用了集成攻击的框架,在每个测试样本上尝试所有攻击组合,只要有一种成功使模型输出错误,就判定为攻击成功。该方式极大提高了模型鲁棒性评估的可靠性。Ding 等人将 l_2 约束的攻击加在了直接编码攻击上,拓展了该类攻击的范数约束[42]。Jiang 等人设计了基于 MIM 的直接编码梯度攻击[122]。

Bagheri 等人设计了适用于概率 SNN 的白盒攻击方法[41]。该方法特点在于 Bagheri 等人考虑了直接编码、泊松编码以及 TTFS 编码的攻击。添加脉冲、删除脉冲以及翻转脉冲。攻击受到 Hamming 距离控制。Marchisio 等人提出也提出了类似的工作,通过梯度反向传播在图像中添加特定模式输入噪声,在推理过程中导致 SNN 位翻转[123]。该方法高度隐蔽,适配了频率编码之外的方法。然而为了设计扰动,方法需在梯度空间中反复迭代优化,计算成本较高。

上述所有攻击方法都需要攻击者了解模型的结构和参数,也就是需要白盒信息。这对于部分攻击场景是不可行的。另外还有针对图像数据的黑盒攻击方法。Marchisio 等人提出黑盒对抗攻击算法攻击脉冲深度信念网络[124]。该方法不依赖模型结构、参数或训练数据,仅通过输入和输出信息进行攻击。作者通过定义感知距离度量图像扰动的可感知度和分类间隔描述目标类别概率与其余类别中最大概率之间的差值,实施迭代扰动。每轮迭代选择扰动优先级最高的像素,直至感知距离超过预设阈值。该方法优势是可以迁移到其他深度结构上,但缺点是攻击效率受限于启发式搜索,需逐像素评估扰动方向与幅度,难以扩展至高分辨率图像。

综上,针对静态图像的对抗攻击方法主要通过对白盒设定下频率编码设计梯度扰动策略,目前已经扩展到更复杂的多步攻击和范数约束攻击,同时也出现了无需模型结构信息的黑盒攻击,但在攻击效率方面仍存在挑战。

(2)针对事件数据的攻击方法

与传统的帧采集相比,DVS 采集异步事件数据[125]。事件数据集为事件相机采集或相机原理模拟出的数据集合,包括 DVS-Gesture 和 NMNIST 等,可用于训练和测试 SNN 模型。现有的对抗攻击方法应用到事件数据上存在挑战[37],数据形式不同。事件数据中事件时间是连续值,处理时可能需要转换为帧。因此,攻击工作主要在事件的表征上进行。

首先,Marchisio 等人提出 DVS-Attacks 扰动,包含了很多基于事件数据的扰动方法[126]。这些方法包括稀疏攻击方法、帧攻击方法等等。为了攻击 DVS 数据,作者首先将 DVS 数据变换为具有时空维度的数据格式。稀疏攻击方法根据数据每个序列的损失函数逐步更新扰动量。该方法利用掩码来选择应添加扰动的特定事件帧。之后,输出预测概率以及扰动下获得的损失函数值。最后,根据损失函数关于输入的梯度更新扰动值。之后,作者提出了若干非对抗扰动的方法。作者提出简单的帧攻击方法,在数据周围添加一个事件框。该方法将导致事件数量增加,影响 SNN 推理延迟。角点攻击依次修改图像角落像素,根据分类结果自适应调整扰动区域与强度。该方法的缺点是并非所有样本都受到相同强度的扰动。难攻击的数据上扰动更容易被发现。连字攻击在空间上集中于图像边缘的相邻像素点,扰动不会导致事件数量大量增加。上述方法中只有稀疏攻击方法是涉及白盒设定的梯度,其余均为黑盒方法。

受到 Marchisio 等人的工作启发,Büchel 等人提出了更具可拓展性以及更高效的事件攻击方案[37]。作者提出基于深度学习攻击方法 DeepFool

的 SpikeFool 方法。该方法同时具备可拓展性和稀疏性，通过迭代寻找网络决策边界，对扰动进行稀疏约束。为了保持事件数据的离散性，SpikeFool 在每次迭代时将输出四舍五入到最接近的整数。最后，SpikeFool 利用上下限约束来施加二进制约束或限制发放率，从而使其适应于不同类型的数据。该方法的优势在于基于 DeepFool 的方法攻击性能较好，劣势在于计算复杂度较高。

在上述工作后，Lee 和 Myung 提出了基于事件脉冲张量表征(Event Spike Tensor，EST)的事件数据攻击方法[39]。该方法首先通过一个多层感知器卷积核卷积，将事件数据转换成多通道网格表示构建 EST 表征[38]。作者为了攻击原始事件的时间，设计了基于梯度的攻击方法，其中攻击扰动被认为是数据中事件时间的偏移。此外，作者还随机生成了额外的对抗事件，在没有事件的位置添加空事件，并利用空事件的损失梯度迭代地来确定额外对抗事件的位置。该方法的优势在于，扰动后的数据格式和事件数据的原格式相同。且攻击方法对于 SNN 和 ANN 具有通用性。缺点在于扰动的范数仅仅建立在数据上，并没有验证扰动量是否可以在真实场景上被人类观察到。

Marchisio 等人设计了一种针对事件帧的时空对抗攻击算法[127]。该攻击采用掩码选定哪些帧需要扰动，从而控制扰动的稀疏性。作者分别针对无防御 SNN、有防御但是攻击者知道或者不知道防御的情况三种条件进行攻击。使用交叉熵损失对扰动进行迭代优化，基于目标类别的概率变化计算损失，并通过反向传播逐步调整扰动。该方法的优势是在不同事件数据集都具有较好性能，劣势是攻击仍基于帧数据，而不是原始事件数据。

为了直接攻击在事件数据上，克服上述工作的不足，Yao 等人设计了一种能够直接对事件相机输出的数据格式进行扰动的对抗攻击框架[128]。作者提出 Gumbel-Softmax 采样策略和直通估计器，将离散优化问题转化为可微的连续优化问题，从而使得梯度优化可行。该方法集合了其他事件工作的优势，通过梯度攻击提高了攻击性能，同时直接攻击原始事件数据。同时，攻击样本在稀疏性和扰动幅度上较小。但是其限制是攻击方法需要模型梯度信息，难以直接在黑盒条件下应用。

综上，针对事件数据的攻击方法通过扰动事件的时间、位置或数量实现攻击，包括帧转换、稀疏优化、EST 表征以及可微采样等多种策略，部分方法兼顾稀疏性与攻击性能，但仍存在计算复杂度高等限制。

(3)跨数据模态的攻击方法

除了上述特别针对图像或者事件的攻击方法，一些方法并没有针对任何数据类型，而是可以灵活地应用在图像或事件数据上。

Liang 等人提出了一种既适用于图像数据，也适用于事件数据的通用对抗攻击方法[29]。其核心在于将图像输入转换为泊松输入，而将事件输入转换为三值的离散值，从而使得两种数据得到统一，实现跨模态对抗攻击。扰动的形式为基于梯度计算出的 0/1 翻转。对于图像数据，作者通过反向传播获得时空梯度信息，再将这些梯度在时间维度上聚合得到像素扰动。作者提出将连续梯度映射为三值(−1，0，1)梯度，解决梯度与脉冲输入格式不兼容的问题。同时，为应对 SNN 中常见的梯度消失问题，作者还在全零梯度时引入随机翻转。该方法的优点是扰动梯度经过特别设计较为可靠。缺点是只能进行 01 翻转，对图像来说扰动粒度较粗。

Lin 等人提出了一种基于脉冲概率建模的对抗攻击方法 SPA，可兼容图像数据和事件数据[129]。其主要原理和 Liang 等人意一样，也是对图片利用泊松编码，使之转化为和事件相同的脉冲序列格式，并基于概率生成扰动，从而在脉冲空间中构造对抗样本。该方法优势是有较高的攻击效率，迭代次数少、收敛快。然而，该方法应用在黑盒攻击上时，计算开销较大。针对上述劣势，Lin 等人提出 SFTA，该方法为跨模态的黑盒攻击方法[130]。作者提出利用训练好的替代模型提取其中间层的脉冲特征并结合输出层梯度构造损失函数。在统一数据模态上，该方法也与前述方法类似。该方法的攻击效果依赖于构建的替代模型，好处是无需目标模型信息，可以实施黑盒攻击。然而若替代模型与目标模型分布偏差过大，攻击效果会下降。

上述方法输入数据为离散脉冲数据。然而还有一类方法，图片输入采用直接编码，而事件数据被处理成帧数据。Bu 等人提出的 RGA 攻击[30]和 Hao 等人提出的 HART 攻击[40]就属于这个类型。这两个方法相当于默认模型采用频率编码，好处是扰动在事件数据上时不是扰动脉冲，而是扰动事件一段时间平均而成的帧，如此可以直接使用连续梯度攻击，而不需要特殊处理成离散信号。

综上，跨数据模态的攻击方法通过统一编码或特征表征策略使图像和事件数据在同一域中，以此

支持同一攻击框架,但仍面临替代模型依赖性强和攻击效率较低等挑战。

3.3　基于可微近似的梯度攻击方法

之前介绍了数据模态视角的攻击方法,从 SNN 本身来看,其不可导特点带来了基于可微近似的梯度攻击方法这一分类。依照采用的可微近似类型不同,可分为转换可微近似的攻击方法、随时间反向传播可微近似攻击方法以及脉冲发放率可微近似攻击方法。图 8 中给出了三种方法的示意图。下面介绍相关方法:

(1)转换可微近似攻击方法

转换可微近似攻击(Conversion-based Approximation,CBA)方法的基本思想是将 SNN 转换为具有相同权重和偏置的 ANN,并用 ANN 生成 SNN 的对抗样本。Sharmin 等人提出了这种方法的初始版本,即利用一个随机初始化的 ANN 模型,将其权重替换为训练好的 SNN 的权重,然后使用 FGSM 方法对输入进行扰动,得到对抗样本[26]。这种方法利用了 SNN 和 ANN 之间的转换关系,可以直接使用 ANN 的梯度信息来构造对抗样本,而不需要对 SNN 进行可微近似或修改。在 Ding 等人的研究中,这种攻击方法正式被称作为转换可微近似的攻击方法[25]。转换可微近似的方法简单易实现,攻击者不需要知道 SNN 的实现细节(例如神经元模型、推理时间步长等),只需要知道其结构,就能依据 SNN 发放率与 ANN 激活值近似等价[87]构建近似,直接使用 ANN 的梯度信息来构造对抗样本,而不需要对 SNN 进行可微近似。然而,这种方法的不足也很明显:因为 ANN 和 SNN 存在时序动力学和脉冲编码上的差异,这些差异在发放率等价中无法表达,因此该方法生成的对抗样本不够精确,攻击性能有限。

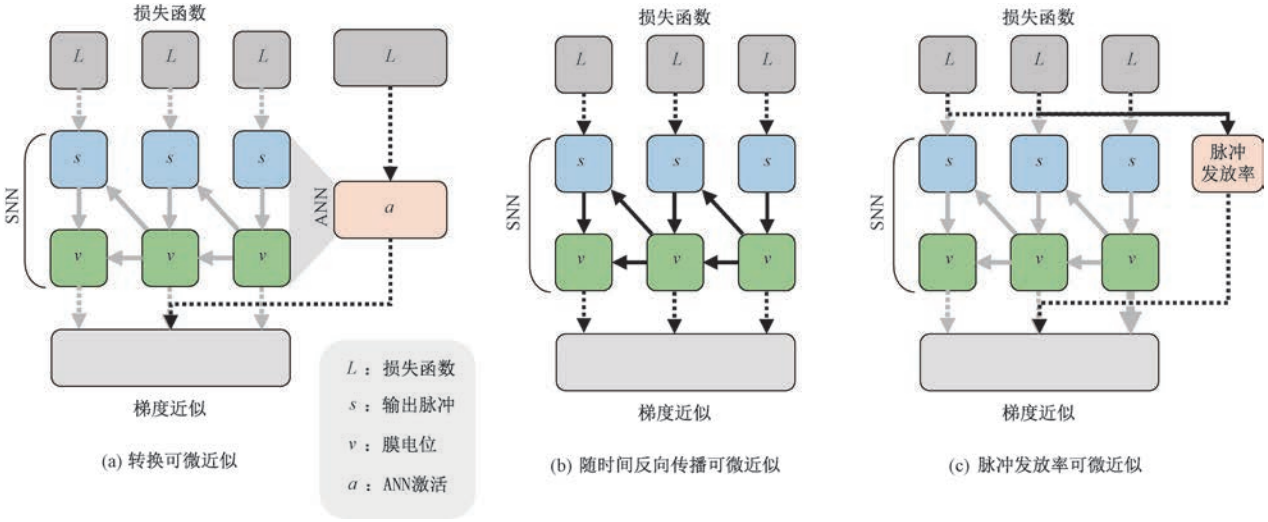

图 8　基于可微近似的梯度攻击方法

综上,转换可微近似攻击方法通过将 SNN 转换为结构相同的 ANN 并借助 ANN 的梯度生成对抗样本,具备实现简便和不依赖 SNN 细节的优点,但由于忽略了 SNN 的时序动态与脉冲特性,导致攻击精度和效果受限。

(2)随时间反向传播可微近似攻击方法

随时间反向传播可微近似攻击(Backward Pass Through Time,BPTT)方法利用了代替梯度来克服 SNN 的不可微性质[26]。基于随时间反向传播的可微近似攻击的设计思路是在反向传播时,用可微函数替代 SNN 中的阶跃函数,从而得到损失函数关于样本的梯度,并用它来生成对抗样本。随时间反向传播的过程可表示为式(6):

$$\frac{\partial L}{\partial s^l[t]}=\frac{\partial L}{\partial s^{l+1}[t]}\frac{\partial s^{l+1}[t]}{\partial v^{l+1}[t]}\frac{\partial v^{l+1}[t]}{\partial s^l[t]}+\frac{\partial L}{\partial s^l[t+1]}\frac{\partial s^l[t+1]}{\partial v^l[t+1]}\frac{\partial v^l[t+1]}{\partial u^l[t]}\frac{\partial u^l[t]}{\partial s^l[t]} \tag{6}$$

其中,t 代表第 t 时间步,$u^l[t]$ 指的是神经元充电前的膜电位,$v^l[t]$ 代表的是神经元充电后的膜电位,$s^l[t]$ 代表的是神经元充电后产生的脉冲,L 是损失函数。该方法的优势是可以利用 SNN 的时空动态信息生成与端到端训练梯度不正交的有效梯度,从而构造出有效的对抗样本。缺点是它需要在每一层每一时刻进行梯度计算,这会增加计算的复杂度和开销。为了改善攻击的计算开销,Krithivasan 等人提出减少攻击时 SNN 的推理时间生成梯度攻击[131]。另外,攻击效果会依赖于代替梯度函

数的选择,不同的代替梯度函数可能会导致不同的攻击效果。

目前,随时间反向传播可微近似攻击方法广泛用于 SNN 的鲁棒性验证[23,26,120],是当前众多攻击方法中的基线方法。

(3)脉冲发放率可微近似攻击方法

脉冲发放率可微近似攻击(Backward Pass Through Rate,BPTR)[25,62]同基于随时间反向传播的攻击一样,也是一种针对 SNN 的梯度攻击方法。区别在于,这类方法不是使用时间维度上的梯度近似,而是计算 SNN 的平均发放率用于反向传播。脉冲发放率可微近似公式可表示为式(7):

$$\frac{\partial L}{\partial s^l[t]}=\frac{\partial L}{\partial r^{l+1}[t]}\frac{\partial r^{l+1}}{\partial r^l} \tag{7}$$

其中,r^l 代表第 l 层的发放率。由于 SNN 的非线性激活函数不可导,因此该方法使用了直通估计器的思想近似 $\partial r^{l+1}/\partial r^l$。对于 IF 神经元,该偏导可以近似为常数 1。如此,攻击时就不需要考虑神经元的动力学和时序信息,有效降低了计算资源需求。它可以直接利用 SNN 的发放率信息来构造有效的梯度攻击,而不改变 SNN 的前向传播过程。然而,相比于基于随时间反向传播的攻击,它忽略了 SNN 的时序信息,可能无法捕捉到 SNN 的动态特征,因此会导致梯度估计不够准确。该方法首先由 Ding 等人提出,用于攻击 SNN[25]。

Bu 等人拓展了脉冲发放率可微近似攻击,针对 SNN 提出了更具有普适意义的速率梯度近似攻击 RGA[30]。RGA 使用基于发放频率的代理梯度进行反向传播进行攻击。此外,他们还提出了一种延长攻击仿真时间的增强版本,以减小攻击中随机性带来的影响,从而能够生成更有效的对抗样本。该方法在白盒攻击和黑盒攻击场景下都比 BPTT 更为有效。此外,该方法对神经元参数的敏感性较低,显示出较强攻击能力。然而,该方法目前局限于频率编码的神经元。

Hao 等人在 RGA 基础上提出了一种新的攻击框架 HART[40]。HART 方法强调仅基于频率并不能高效攻击 SNN。因此提出基于频率和时间信息的混合对抗攻击。HART 保持标准的前向传播模式,但修改了反向传播链以增强对频率和时间信息的捕捉,允许动态调整代替梯度函数中时序梯度信息的保留程度。同时,其在时间维度上降低了不同时间步之间的相关性,能够更加精准地捕获频率信息。HART 相较于其他脉冲发放率攻击方法提升了攻击成功率,并适用于多种攻击环境和数据类型。

综上,脉冲发放率可微近似攻击方法利用 SNN 的平均发放率进行梯度估计,降低了对时序信息的依赖,具备计算效率高和适用性强的优势,但也因忽略时序结构导致梯度精度有限。

针对可微近似的攻击方法具有鲜明特点,三者都旨在解决 SNN 梯度问题,其中 CBA 采用间接方案,而 BPTT 采用精确方案。相比于 CBA 和 BPTT,BPTR 类的方法在梯度的准确性和效率中维持了平衡。Ding 等人对这三种方法的攻击性能进行了比较,表 1 摘录了不同可微近似攻击下 SNN 的分类准确率和计算效率。详细参数设计以及模型设置参见文献[25]。作者发现 BPTT 和 BPTR 攻击效果更好,而 CBA 则攻击能力较差。作者认为,这是因为 CBA 改变了 SNN 网络前向传播过程,使用了 ReLU 激活函数来替代 SNN 的脉冲发放。这一过程与深度学习能够产生有效攻击的可微近似反向传播(Backward Pass Differentiable Approximation,BPDA)技术的原则相悖。BPDA 要求在前向传播过程中保持原始的非可微函数,而只在反向传播过程中使用可微近似函数。这样可以避免产生不准确或无效的梯度信息。BPTR 方法则正好能够匹配 BPDA 的需求。在计算效率上,BPTR 只需要在每一层计算一次梯度,而 BPTT 需要计算梯度次数与时间步成正比,因此 BPTR 的计算开销更低。作者实测计算时间为 BPTT 的三分之一。因此,BPTT 和 BPTR 同为性能较好的攻击方法,而 BPTR 的攻击计算代价较低。

表 1 脉冲可微近似梯度对抗攻击方法比较

	CBA	BPTT	BPTR
FGSM	54.34%	12.78%	10.59%
PGD	37.30%	0.04%	0.10%
计算时间	1×	3×	1×

3.4 梯度无关的攻击方法

可微近似的攻击方法依赖白盒设定,同时需要可微近似方法的参与。针对不了解目标模型、参数的黑盒设定,研究者提出了梯度无关的攻击方法用于设计对抗扰动。下面介绍相关方法:

Bagheri 等人提出针对概率 SNN 的几种梯度无关的攻击方法[41]。这些方法均采用直接修改输入脉冲的方法实现,不依赖梯度信息。该方法一边去除、翻转、添加脉冲,一边衡量哪个操作能够有效对数据进行攻击。作者汉明距离计算扰动大小,并设计了相应贪心算法,选择能够最大化目标类别的

修改动作。为了降低计算复杂度,攻击只在输入脉冲序列的前几个时间步内进行搜索。该攻击方式完全基于对输入脉冲模式的修改操作实现了对 SNN 的有效扰动。然而,该方法攻击搜索空间有限,难以应对高复杂度或深层网络结构。

Venceslai 等人提出了梯度无关攻击方法 NeuroAttack[123]。该方法通过精心设计的输入噪声影响 SNN 的安全性。与上述工作一样,该方法也不依赖网络的梯度信息,主要是通过扰动使得网络中间神经元脉冲模式发生变化,从而影响 SNN 鲁棒性。攻击者通过输入扰动使得特定神经元超过脉冲阈值,从而完成对目标权重位的翻转。攻击者首先需要选择攻击目标层和目标神经元,再确定输入触发区域,并采用优化算法迭代更新输入扰动以最大化目标神经元的激活程度。该方法存在一定不足,例如方法对攻击场景依赖性较强,需要攻击者能够在神经形态硬件中植入木马,实际中实现难度较大;其次,输入扰动位置选择需要进行模拟测试,缺乏自动化优化流程设计。

Marchisio 等人设计了在黑盒场景下脉冲深度置信网络的梯度无关攻击方法[124]。该方法通过贪心启发式算法生成对抗样本。方法首先在图像中选择一个区间,衡量每个像素对分类结果的敏感度,再根据优先级排序,选择对分类影响最大的像素进行调整。通过迭代优化,在设定扰动范围内持续调整输入像素。该方法由于需要频繁评估像素敏感度,计算代价较高,难以处理大规模复杂输入场景或实时攻击应用。

上述方法目标模型是可访问的,且攻击构建过程直接与目标模型相关。下面方法通过构建代理模型来攻击目标模型:

Lin 等人提出一种高效的无梯度攻击方法 SPA[129]。该方法基于泊松编码,将脉冲交叉熵损失和扰动的 l_2 范数作为扰动的优化目标。该方法更新扰动向量,直到判断攻击成功并且扰动足够小时退出优化过程。这一早停设计加快了优化的收敛速度。在黑盒攻击场景下,该方法为目标模型构建了代理模型,利用与目标模型输入输出之间的关系估计梯度方向,并优化代理模型,使其生成的对抗样本能有效迁移并误导目标模型。该方法的优势在于考虑了代理模型攻击的迁移性,具备迁移到更多模型上的可能。然而该方法的劣势在于攻击效果与代理模型的拟合质量有关,若代理模型与目标模型存在较大差异,迁移攻击效果可能大打折扣。其次,对于大规模脉冲输入,该方法仍然存在计算开销较大的问题。

之后,Lin 等人提出了 SNN 黑盒攻击方法 SFTA[130]。该方法通过修改代替模型的特征表示实施攻击。该方法为某一隐藏层的脉冲表示设计扰动,通过抑制和增强特定特征提升对抗样本的迁移性与攻击效果。SFTA 首先估计代替模型隐藏层的特征梯度,设计扰动的优化目标函数。在优化过程中,SFTA 抑制正向梯度对应的特征以及增强负向梯度对应的特征,逐步生成具有强迁移性的对抗样本。然而该方法存在以下不足:首先,该方法需要提取目标模型中间层特征,若代替模型与目标模型差异过大,则转移攻击效果受限。其次,该方法存在较多参数调优,缺少合适的调参方法。此外,该方法攻击时迭代次数和仿真步长,计算效率有提升空间。

综上,梯度无关的攻击方法通过直接修改脉冲输入、设计扰动策略或构建代理模型,在无需访问梯度信息的黑盒设定下就可实现对 SNN 的有效攻击,但仍面临搜索空间受限、代理模型影响及计算成本较高等问题。

3.5　攻击方法对比与总结

本文将攻击方法分为基于数据模态的方法、可微近似的梯度攻击方法以及梯度无关的方法三个大类。

基于数据模态的方法主要是与 SNN 常用的输入数据模态有关。如果采用针对特定数据模态的攻击将有较好的性能。跨模态方法通过统一编码策略提升输入数据的适配性。攻击的复杂度与采用的数据形式有关,若采用帧形式则计算量较小;而若需逐对事件数据进行逐事件调整,则可能计算量较大。单模态的攻击方法迁移性较差,而跨模态攻击方法在通用性和迁移性上较好。

可微近似的梯度攻击方法可以借鉴深度学习的梯度攻击方法。其主要用于白盒场景,高度依赖模型梯度信息、神经元类型与网络参数。这类方法由于梯度信息可用拥有精准的扰动控制能力,能够基于模型梯度计算构造强攻击样本。然而,可微近似的梯度攻击方法中性能最好的随时间反向传播攻击复杂度高,原因是需构建完整的反向传播链。而脉冲发放率攻击在计算复杂度上介于随时间反向传播攻击和转换攻击之间,而攻击能力较好。在白盒设定下通用性较强,适配各类 SNN 架构。

梯度无关的方法适合黑盒场景,无需模型梯度或结构信息即可实施攻击。这类方法通常需要设计启发式策略提高其攻击能力;且需频繁评估扰动效果,或根据目标模型构建代理模型,导致整体效率下降。这些方法迁移性较好,适合难以访问模型信息

的场景。

表 2 总结了文中提到的攻击方法在具体算法、数据模态、可微近似、范数约束、输出格式、计算效率与攻击类型多个角度的特征。可以观察到以下趋势：

表 2 脉冲神经网络攻击方法总结

参考文献	数据模态	可微近似	具体算法	约束	输出	计算效率	类型
[41]	图像	—	—	l_0	离散张量	低	白盒
[122]	图像	—	—	l_∞	离散张量	低	灰盒
[124]	图像	BPTT	FGSM/PGD/MIM/BIM	l_∞	连续张量	低	白盒
[22]	图像	BPTT	FGSM/PGD	l_∞	连续张量	低	黑盒/白盒
[24]	图像	BPTT	FGSM/PGD	l_∞	连续张量	低	黑盒/白盒
[131]	图像	BPTT	FGSM/PGD	l_∞	连续张量	低	黑盒/白盒
[25]	图像	CBA/BPTT/BPTR	FGSM/PGD/RFGSM/BIM	l_∞	连续张量	低/高/低	黑盒/白盒
[26]	图像	CBA/BPTT	FGSM/RFGSM	l_∞	连续张量	高/低	黑盒/白盒
[120]	图像	BPTT	FGSM/PGD	l_∞/l_1	连续张量	低	白盒
[42]	图像	BPTT	EOT+FGSM/PGD	l_∞/l_2	连续张量	低	黑盒/白盒
[130]	图像	—	—	改进 l_1	连续张量	低	黑盒
[37]	事件	BPTT	Deepfool	l_0	离散张量	低	白盒
[123]	事件	BPTT	—	l_1	离散张量	低	白盒
[126]	事件	BPTT/—	—	l_1	离散张量	低	黑盒/白盒
[39]	事件	BPTT	PGD	l_∞	连续张量	低	白盒
[132]	图像/事件	—	—	l_0/l_2	离散张量	低	黑盒/白盒
[29]	图像/事件	BPTT	—	l_2	离散张量	低	白盒
[133]	图像/事件	—	PGD	l_2	连续张量	低	黑盒
[23]	图像/事件	BPTT	FGSM	l_∞	连续张量	低	白盒
[30]	图像/事件	BPTR	FGSM/PGD	l_∞	连续张量	高	黑盒/白盒
[40]	图像/事件	HART	FGSM/PGD	l_∞	连续张量	高	黑盒/白盒

(1)目前大部分攻击方法仍以图像数据为主要攻击对象[22,24-26,42]。跨模态攻击方法数量排在第二[29-30,40,129-130]。而针对事件数据的攻击研究相对较少[37,39,126,128]。因此，未来可以围绕事件数据设计更多攻击方法。

(2)BPTT 的可微近似是当前使用最广泛的可微近似方法[22,24-25,39,120]，原因是其可精准构建时间维度上的梯度信息，但计算开销大。未来可积极研究计算量较小的方法，如脉冲发放率的攻击方法。

(3)受到深度学习影响，FGSM 与 PGD 是最常见的基础攻击方法[22-24,39,42]，可结合多种范数生成对抗样本。目前有少量研究使用多步攻击、基于优化的攻击等方法实施进阶攻击[37,42,122]，提高了攻击方法的多样性。梯度无关攻击[41,123,124,129,130]适用于黑盒条件，但通常攻击效率较低。未来可在进阶攻击和梯度无关攻击上提高攻击性能以及改善计算复杂度。

(4)图像攻击大多直接使用连续张量扰动[22-25]，便于直接优化像素值，范数约束通常为 l_∞。事件数据和概率 SNN 攻击多为离散张量扰动[37,126,128,129]，需考虑脉冲数量、位置等整数约束，范数约束通常为 l_1，l_2。

(5)计算效率大多较低，主要原因是采用 BPTT 可微近似或梯度无关攻击方法[37,124]；脉冲发放率类方法在计算效率上[30]表现更好。目标模型的透明度上，攻击以白盒攻击为主[24-25,122]，但部分工作支持黑盒或灰盒设定[42,124,129,130]。

当前 SNN 攻击和防御工作目前主要围绕识别任务展开。常用数据集包括静态图像数据如 MNIST、FMNIST、CIFAR-10/100 和神经形态数据如 DVS-Gesture、N-MNIST、DVS-CIFAR10、N-Caltech101。目前常用的性能量化指标主要包括分类准确率与攻击成功率(Attack Success Rate，ASR)。分类准确率表示模型在对抗样本上的正确分类比例。攻击成功率表示在所有用于测试的对抗样本中，能够成功欺骗模型使其输出错误预测的比例，如式(8)：

$$ASR=\frac{N_{\text{success}}}{N_{\text{total}}} \tag{8}$$

其中，N_{success} 表示攻击成功的样本数，N_{total} 表示攻击样本总数。

攻击方法的能力越强，分类准确率越低，ASR 越高。表 3 显示，在强度范围为 0.03 至 0.1 的攻击下，多种方法在上述数据集上取得了超过 90%的攻击成功率，部分方法如 Büchel、Liang 和 Hao 等甚至

表 3　脉冲神经网络攻击性能

攻击方法	攻击强度	网络结构	数据集	攻击成功率
Liang 等人[29]	—	8 层 SNN	MNIST	91.31%
Bu 等人[30]	8/255	VGG-11	CIFAR10	93.74%
Liang 等人[29]	—	8 层 SNN	CIFAR10	98.68%
Ding 等人[25]	8/255	VGG-11	CIFAR10	99.63%
Hao 等人[40]	8/255	VGG-11	CIFAR10	100.00%
Bu 等人[30]	8/255	ResNet-17	CIFAR100	92.06%
Bu 等人[30]	8/255	VGG-11	CIFAR100	94.72%
Ding 等人[25]	8/255	VGG-11	CIFAR100	99.59%
Hao 等人[40]	8/255	VGG-11	CIFAR100	99.96%
Marchisio 等人[126]	—	—	N-MNIST	74.41%
Liang 等人[29]	—	7 层 SNN	N-MNIST	97.38%
Büchel 等人[37]	0.5	LeNet-5	N-MNIST	99.88%
Bu 等人[30]	8/255	VGG-11	DVS-CIFAR10	59.56%
Liang 等人[29]	—	5 层 SNN	CIFAR10-DVS	100%
Ding 等人[25]	8/255	VGG-DVS	DVS-CIFAR10	87.11%
Hao 等人[40]	8/255	VGG-DVS	DVS-CIFAR10	93.03%
Marchisio 等人[126]	—	—	DVS-Gesture	92.44%
Büchel 等人[37]	0.1	LeNet-5	DVS-Gesture	99.87%
Lee 等人[39]	0.1	ResNet	N-Caltech101	95.50%

接近 100%,显示出端到端可微近似策略在攻击图像与事件数据方面的强大效果,进一步凸显了 SNN 在安全性方面所面临的严峻挑战。

4　脉冲神经网络防御方法

第 3 节说明 SNN 在面对对抗攻击时表现出脆弱性。这些攻击方法大部分借鉴了 ANN 中的梯度攻击技术,实施有效的攻击。目前已有研究工作指出 SNN 在某些方面具有比传统 ANN 更强的鲁棒性。因此,如何发挥 SNN 在鲁棒性方面的优势,以及如何进一步增强这种优势,成为了研究者们关注的问题。本节首先将分析 SNN 鲁棒性原因,并且从输入层面、网络层面以及输出层面分别介绍 SNN 防御方法。

4.1　鲁棒性分析

SNN 与传统 ANN 使用的数据表征和信息处理方式有区别,因此在面对噪声时表现出不同的计算性质。本节将介绍 SNN 鲁棒性分析的理论框架,为后文 SNN 具体防御方法作铺垫。

(1)脉冲神经元

脉冲神经元具有滤波功能。其原因有二。其一,从神经元动力学角度看,深度 SNN 最常用的 LIF 神经元因衰减漏电等行为可以被视为滤波器,能对输入信号进行滤波。例如脉冲响应模型(Spike Response Model,SRM)就存在两个具有滤波功能的滤波器。响应核描述神经元如何对来自其他神经元的输入脉冲反应。不应期核描述神经元在发放脉冲后对后续脉冲活动的抑制作用[9]。其二,脉冲神经元在达到阈值后发放脉冲并进行重置操作。重置操作在清除残余膜电位的同时也清除了历史噪声,从而保持对后续输入的准确响应。

(2)网络编解码

SNN 输出离散的脉冲序列。假设推理时间步为 T,使用频率编码实际能表征的数据范围应在集合 $\left\{\frac{1}{T},\frac{2}{T},\cdots,1\right\}$ 中取到。即使扰动影响输入落在离散表征范围中间,编码结果不受影响。

目前,频率编码中的泊松编码被认为具有天然鲁棒性。研究人员已经提出了用于解释这种鲁棒性的理论。泊松分布具有一定的随机性,能够帮助神经网络适应了带有噪声的输入。Sharmin 和 Marchisio 等人进行了 ANN 与 SNN 的对比研究,发现 SNN 相较于 ANN 在对抗性攻击方面具有一定程度的固有鲁棒性。论文论证了泊松编码的能力[22]。延迟编码以及首次脉冲时间编码等方法也在其他论文中被证实具有一定鲁棒性[28,134]。因此,选择合适的输入编码方式可以显著提高 SNN 对随机噪声和对抗噪声的鲁棒性。Bhaskar 等人使用随机平滑分析讨论泊松编码下 SNN 的鲁棒性[120]。随机平滑可以确定输入带噪声情况下分类器对于输入变化的容忍度。Bhaskar 等人提出针对 SNN 的随机平滑定理。该定理将图像像素的 0－1 张量视为即将输入到一个平滑分类器的伯努利随机变量的采样。该分类器能够给出每个类别的概率分布。框架认为对于任意输入扰动 δ,扰动 δ 的 l_1 范数要小于带有扰动输出中最大概率和次大概率之间差的一半。该理论表明 SNN 视为的平滑分类器在一定程度上对扰动具有鲁棒性,能够保持分类结果的稳定性。

(3)网络信息处理机制

SNN 的脉冲编码过程导致其与 ANN 信息处理机制存在差异。目前有两套理论框架可以帮助理解 SNN 相比于 ANN 在鲁棒性上的优势。

①误差放大理论

Ding 等人提出 SNN 误差放大理论[25]。ANN 中的扰动放大主要考虑扰动前后激活值的距离 $\|a^l-\widetilde{a}^l\|$ 和神经网络输出 $\|f(a^l)-f(\widetilde{a}^l)\|$ 的关系。SNN 误差放大理论借鉴 Lipschitz 分析将

整个推理过程的脉冲表征视为激活，并分析其前后层脉冲距离的放大情况。记 SNN 第 l 层的输出脉冲序列为 $S^l=\{s^l(t)\mid t=1,2,\cdots,T\}\in\chi^{T\times N_l}$（$\chi\in\{0,1\}$），其中 T 是总时间步，N_l 是第 l 层中神经元的数量。扰动距离可以定义为式(9)：

$$d(S^l,\widetilde{S}^l)=\|S^l-\widetilde{S}^l\|_2 \tag{9}$$

理论得到脉冲序列距离 Lipschitz 分析结果为式(10)：

$$d(S^l,\widetilde{S}^l)^2\leqslant\frac{1}{\theta^2}\Lambda^{l2}d(S^{l-1},\widetilde{S}^{l-1})^2+\Gamma^l \tag{10}$$

其中，Λ^l 是脉冲表征的 Lipschitz 系数，Γ^l 是关于第 l 层参数的常数。Ding 等人在论文中论证 SNN 的 Lipschitz 常数要小于 ANN 的 Lipschitz 常数。因此认为 SNN 的鲁棒性理论上会优于 ANN 的鲁棒性。

②输入边界传播分析

Liang 等人从可证鲁棒性的角度剖析了 SNN 鲁棒性[23]。该理论方法在确定输入上下限后确定膜电位和脉冲函数的上下限。假定 u 为膜电位，s 为脉冲，H 为 Heaviside 阶跃函数，脉冲激活函数生成的上下限 u^{lb} 和 u^{ub} 可以由式(11)给出：

$$H(u^{lb}[t]-\theta)\leqslant s[t]\ H(u^{ub}[t]-\theta) \tag{11}$$

通过更加精细的线性上下限控制，作者将脉冲产生函数约束在一个由膜电位上下限划分的三角形区域中(如式(12)～(13))：

$$0\leqslant s[t]\leqslant\frac{u[t]-u^{lb}[t]}{\theta-u^{lb}[t]} \tag{12}$$

$$0\leqslant u^{ub}[t]-\theta<\theta-u^{lb}[t] \tag{13}$$

以此类推，可以将所有层的膜电位上下界都确定好并得到输出的上下界。SNN 输出不会超过上下限的约束范围。

Calaim 等人[135]进一步确定，SNN 的输出边界由多种因素决定，例如解码器矩阵的属性、神经元调整、阈值的异质性以及网络内的连接性。网络参数的变化。作者发现负反馈、组件异质性和模块化等都能够明显改变网络鲁棒性。

4.2 输入防御方法

既然 SNN 可以用于静态图像任务以及神经形态数据任务。相应地，防御方法也会涉及两种数据模态，静态图像输入到 SNN 需要脉冲编码，因此可以设计脉冲编码从源头改善鲁棒性。同时，神经形态事件数据也可通过滤波方法进行防御。下面介绍基于脉冲编码的方法和基于滤波的方法。图 9 中给出了这两类方法的示意图。

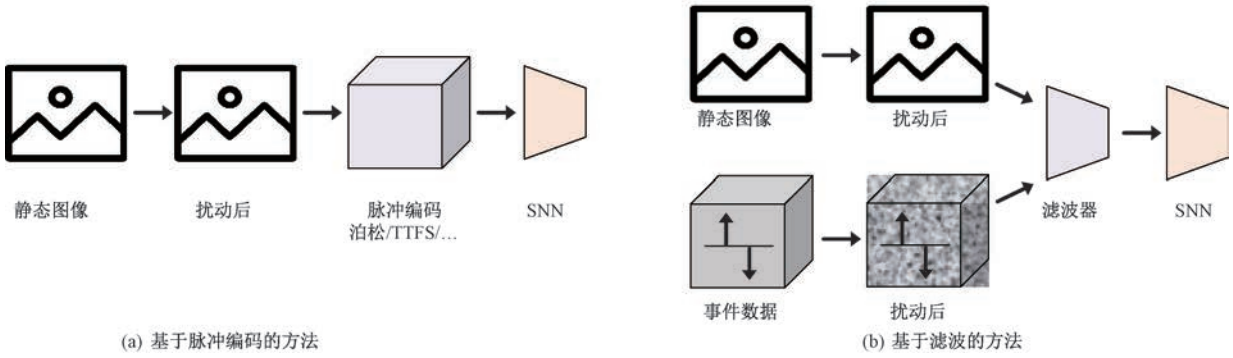

图 9 输入防御方法

(1)基于脉冲编码的方法

Sharmin 和 Marchisio 等人研究表明，特定的输入编码方法可以提升 SNN 的鲁棒性。目前，泊松编码、延迟编码以及首次脉冲时间编码被实验证明在小数据集上展示出了一定程度的鲁棒性[28,43,135]。同时新兴的一些编码方法也被研究人员提出，它们有的借鉴类脑机制，有的利用了 SNN 时序特性。

Sharmin 等人通过实验发现，鲁棒性提升与编码和训练的结合有关。泊松编码引入的随机性[136]可提升鲁棒性，尤其是黑盒攻击[26]。作者发现由于 SNN 直接训练相比转换保留了时序上的内在随机性与动力学特性，因此在多数攻击场景下表现出更强鲁棒性。该方法的优点在于泊松编码实现简单，却能为输入带来扰动提升了网络对扰动的抵抗力。然而，编码的随机性会给推理准确率带来负面影响。作者在实验中还结合泊松编码验证了 SNN 中神经元的动态特性，如漏电率和时间步长，对鲁棒性的影响。Sharmin 等人绘制了量化后的激活函数在使用泊松编码时带来的误差示意图，并指出若考虑泊松编码和神经元结合可以减弱扰动强度。

Mukhoty 等人系统论证了基于泊松编码的 SNN 具备可证对抗鲁棒性[120]。作者将泊松编码视为对输入 Bernoulli 采样的随机平滑。在此基础上，

作者构造平滑分类器,证明在 l_1 范数下一定强度的输入扰动会导致输出不变,从而得到理论上的扰动可证半径。该方法为泊松编码的首个 l_1 范数下的理论可证鲁棒性框架。作者的分析兼容转换后的SNN模型,只要平滑后的分类器满足鲁棒性判据即可。然而,作者提出的方法也有一定缺点。方法给出的可证半径受时间步数制约,随着时间增加,模型在长时间推理下的鲁棒性无法保证。

Ding 等人设计了一种基于编码机制的鲁棒性增强策略[42],通过引入随机门控对脉冲信号进行编码调控。该方法模拟了生物神经系统中突触传输和离子通道的随机性,使得每一层的脉冲传输在时间上具有不确定性,从而增强模型对输入扰动的鲁棒性。该方法指出泊松编码实质上是随机门控在输入层的特例,并从理论上为泊松编码提升 SNN 鲁棒性的现象给出了数学解释。该方法机制简单,具有合理理论支持,并可在不改变网络结构的情况下与其他防御策略联合使用,然而该方法会导致网络分类准确率略有下降,且训练复杂度提高。

除了泊松编码,研究人员还探索了很多脉冲编码方法的鲁棒性。首次脉冲时间编码(Time-to-First-Spike,TTFS)不依赖于脉冲频率而是通过神经元第一次脉冲的时间表示信息。Nomura 等人的论文通过引入时间惩罚项并发现,降低参考时刻有助于提高 SNN 的鲁棒性[28]。Park 等人分析了SNN在噪声环境下使用频率编码、相位编码等,并提出时间平均脉冲编码用于改善 SNN 的鲁棒性[134]。Leontev 等人发现使用延迟编码的 SNN 在对抗攻击下表现出较强的防御力[43]。延迟编码引入时间维度增强信息的表征。Ding 等人提出基于同步的编码方法,通过控制脉冲发放的起始和终止时刻,改善了低时间步下的鲁棒性[137]。为了支持编码策略的效果,作者结合多种反向传播训练算法与攻击场景对鲁棒性进行了验证,证明了编码中时间结构信息是 SNN 鲁棒性的关键。上述这些方法说明了编码对于鲁棒性的重要性,然而,这些方法在对比时通常还是与频率编码进行比较,防御较为被动。

Zhang 等人提出类视网膜编码[133]。该方法模拟了人眼在视觉感知中的动态扫描特性,通过模拟眼动生成一系列非重复的时间步图像输入,代替传统直接编码方式,充分展示了 SNN 的时序处理能力。然而,相比于泊松编码和直接编码,该方法引入了额外图像预处理开销。

在此基础上,Wu 等人提出了一种基于随机平滑编码的防御方法,以提升 SNN 在对抗攻击下的鲁棒性[138]。该方法将高斯随机扰动引入输入编码过程,从而实现与传统随机平滑方法等价的对抗防御机制。与泊松编码相比,该方法具有协方差稳定的优势。同时,通过控制噪声方差与训练目标权重,可灵活调节鲁棒性与性能。然而,该方法在未受扰动数据上性能会有所下降,需引入额外教师网络与损失,训练流程较为复杂。

综上,引入具有时间结构和随机性的脉冲编码方法可以提升 SNN 在对抗攻击下的鲁棒性,但同时也存在准确率下降和训练复杂度增加等权衡问题。

(2)基于滤波的方法

滤波方法在深度学习中可以过滤掉任务无关信息提高网络对于重要信息的特征提取,以此改善鲁棒性。在事件数据集上,滤波方法主要过滤事件。而在图像数据集上,滤波方法处理图像中的特征。

Marchisio 等人提出了一种基于滤波的方法 R-SNN 以增强 DVS 输入下 SNN 的鲁棒性[127]。R-SNN 通过分析 DVS 信号时空相关性,发现噪声事件更稀疏且缺乏时空聚集性。若事件与邻域内其他事件的时间差超过一定阈值,则该事件被视为噪声并被删除。该方法计算开销小,滤波仅在局部域上进行,便于部署在资源受限的神经形态硬件上。然而,该方法防御能力受限于参数选择,缺乏通用设计。对于模拟真实事件动态扰动的攻击,该方法可能失效。

除此之外,Marchisio 等人还针对 DVS 背景噪声提出了背景活动滤波器(Background Activity Filter, BAF)和掩码滤波器(Mask Filter, MF)[126]。BAF 基于事件的时空相关性,设定一定的邻域窗口和时间阈值,仅保留相关性较强的事件;而 MF 则根据每个像素在指定时间窗口内的脉冲频率设定掩码,过滤掉脉冲频率过高的像素点所产生的事件。上述方法都属于数据预处理手段,具有良好的计算效率,适用于边缘设备。然而,当噪声不存在或参数设置不当时误滤掉原始有用事件而导致性能降低。

上述方法主要适用于 DVS 数据,对于图像数据,研究人员也进行了大量研究。Li 等人提出的混合注意力机制本质上是一种滤波方法[139]。该方法通过注意力图对编码层输入进行动态调控,从而实现对背景干扰的抑制与有用信息的保留。该方法通过弱监督目标鼓励注意力图具有稀疏性与平滑性,

使得 SNN 在时空上集中注意力上保持连续性。该方法的优点在于其是数据驱动的动态滤波过程，无需手动设计滤波方法。另外，该方法还能注意力机制节省能耗与计算开销。然而，该方法训练时引入的 ANN 模块可能带来额外计算成本。

Xu 等人提出的 FEEL-SNN 方法涉及一种基于滤波思想的鲁棒性增强机制，将输入图像转换至频域，通过逐时间步使用不同频率掩码抑制高频至低频信息，从而模拟生物神经系统的选择性视觉注意力并去除干扰[140]。该机制相比时空滤波能更准确保留图像有效低频结构，同时滤除攻击引入的中高频成分。然而，该机制对 DVS 数据的时空处理能力有待拓展。

Chen 等人提出了一种图像滤波的防御方法[141]。该方法通过图像去噪实现对对抗攻击的防御，是一种显式的图像空间滤波策略。方法滤波主要由图像净化模块完成。该净化模块由噪声估计网络 NeSNN 和图像重建网络 RecSNN 组成。两个网络都是 SNN。作者设计了滤波前后图片检测方法，比较图像净化前后的变化程度判断图像是否为对抗样本。该方法可有效去除对抗扰动，且对原始图像内容影响较小。图像重建网络 RecSNN 可作为预处理步骤集成到任意 SNN 架构，而无需修改主干网络。然而，该方法会存在图像信息丢失风险，一些细节信息可能被误删，进而影响分类准确性。

Cheng 等人提出了基于侧向抑制的 SNN 以增强 SNN 的抗干扰能力[44]。LISNN 通过引入生物学的侧向抑制连接，使其在面对干扰时能够有效地过滤无关信息，从而更接近大脑的处理方式。LISNN 网络中有卷积层、池化层和全连接层，其中卷积层负责基础特征提取，池化采用平均池化替代最大池化以减少信息损失。网络通过动态神经场模型增强显著特征区域并抑制噪声区域，模拟生物学中的侧向交互机制的功能，进一步提高了对噪声干扰的抵御能力。该方法生物启发性强，提升了 SNN 的生物合理性。且方法仅对局部神经元进行修正，容易集成到现有 SNN 架构中。然而，局部性机制受感受野限制，可能对跨大尺度扰动的处理能力有限。

Zhang 等人在神经元层面抑制由对抗扰动引起的异常激活，从而实现对输入扰动的动态过滤[142]。方法建模同一通道内神经元间的侧抑制关系，等同于高频抑制滤波器防御对抗扰动的高频分量，并保持其他激活结构。该方法可无缝集成到多种 SNN 主干网络中，能够有效抑制高频分量扰动。然而，由于使用固定侧抑制结构，该方法对一些复杂或非均匀扰动可能适应性不足。

Dapello 等人提出了 VOneNet 网络，旨在通过模拟灵长类动物初级视觉皮层的处理过程来提高图像识别任务中的鲁棒性[143]。该网络关键设计体现了 SNN 的特点：VOneNet 的核心 VOneBlock 结合了 Gabor 滤波器组、简单和复杂细胞的非线性特性以及 V1 神经元的随机编码。该方法能够有效抵御白盒对抗攻击和常见图像缺损。随机编码的引入不仅提高了模型对噪声的鲁棒性，还减少了对特定噪声模式的过度适应。

Perez-Nieves 提出了一种利用神经异质性来提升 SNN 鲁棒性的方法[45]。通过神经元在不同时间尺度上的动态响应，起到对输入扰动的时域过滤作用。该方法通过引入个体神经元独立的膜电位衰减常数，使得不同神经元对输入脉冲的响应在时间上具备不同的保留能力。因此，在处理具有复杂时序结构的任务时，SNN 能更稳定地从扰动中提取关键信息。该方法具备生物合理性，无需增加神经元或连接，仅通过增加少量神经元参数，即可获得鲁棒性提升。然而，该方法主要受制于神经动态的训练调整，没有显式滤波模型解析或量化滤波频段或响应特性。

综上，滤波方法可通过时空特征筛选、频率抑制或神经动态调控等机制，有效去除任务无关干扰，从而提升 SNN 在事件数据和图像数据上的对抗鲁棒性，但其效果往往依赖参数设计、训练策略及模型结构的适配性。

4.3 网络防御方法

网络层面提高鲁棒性的方法更为丰富，可以借鉴深度学习防御方法或者设计 SNN 特定方法。借鉴深度学习方法有对抗训练、正则训练、网络轻量化等方法。而 SNN 特定方法有区间边界传播方法、基于神经元设计的方法等。图 10 中给出了这些类型的示意图。下面将介绍这些方法。

(1)基于神经元设计的方法

神经元设计可以通过调整神经元膜时间常数、阈值等参数改变神经元动力学性质，进而改善模型性能。

El-Allami 等人认为，SNN 的鲁棒性受到一些关键神经元结构参数的显著影响。因此作者提出了一种系统化的方法，通过调整膜电位阈值和推理时

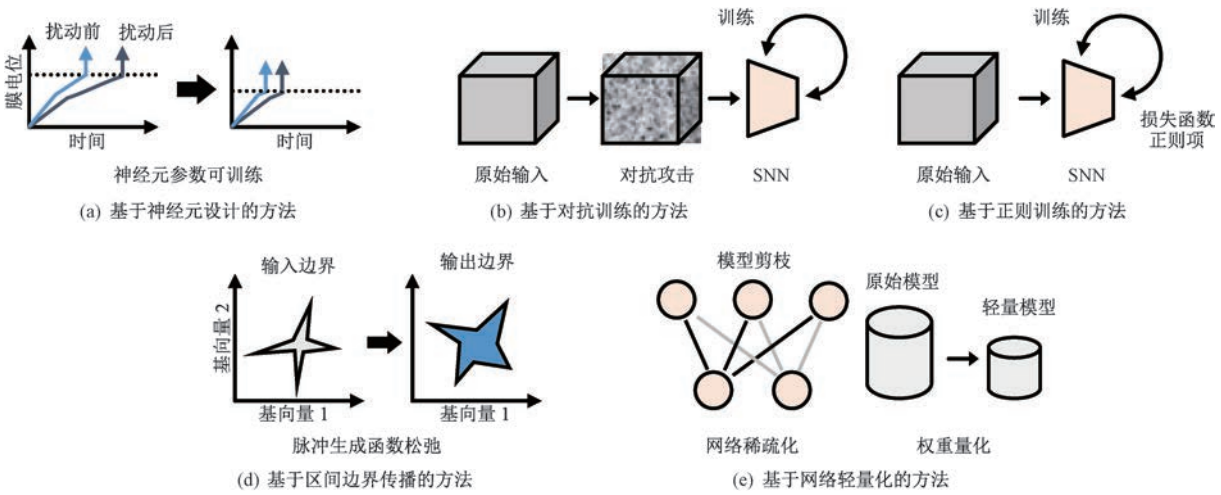

图 10　网络防御方法

间长度改善 SNN 鲁棒性[46]。作者提出搜索算法，通过在不同的膜电位阈值和推理时间长度组合下训练 SNN，评估它们在对抗攻击中的表现。作者还指出，某些初始表现良好的参数组合并不一定在对抗攻击中表现鲁棒。该方法首次揭示结构参数与鲁棒性的关系，为 SNN 领域提供了结构性设计的调优方向。另外，调节神经元阈值与时间窗口无需大量额外计算。然而，该方法需人工调参，缺乏自动化鲁棒性最优化流程，因此 SNN 设计中需要更加全面的参数调优方法。

Xu 等人提出的 FEEL-SNN 方法给 SNN 神经元加入进化泄漏因子。每个神经元在每个时间步上都能自主学习最优泄漏强度，既保留有效信号，又过滤扰动传播[140]。进化泄漏因子通过动态学习保持信息表达能力，在提升鲁棒性的同时保证未受扰动样本的高准确率。然而，训练时进化泄漏因子对每个神经元和时间步引入可训练参数，计算代价略增。

Geng 等人提出了一种基于神经元自适应阈值的鲁棒性增强方法[144]。作者受到生物神经系统中稳态调节机制的启发，提出了一个改进的 LIF 神经元模型。在该模型中，神经元的阈值不再是固定的，而是按照其膜电位的扰动误差与其历史期望值动态调整。作者从理论上证明改进神经元满足有界输入—有界输出稳定性，避免扰动效应逐层累积。该方法有较好理论支持，模拟大脑中调控神经活动强度的稳态调节机制。然而，目前实验使用静态图像数据集，对事件数据任务的适配用有待探索。

除此之外，研究人员还提出很多利用神经元动力学上的调整影响和改善鲁棒性的方法。Chen 等人提出的噪声估计网络 NeSNN[141]中，使用的是具有多阈值脉冲神经元增强对噪声的响应灵敏度。Ding 等人从非线性动力系统的稳定性出发分析了 LIF 神经元在输入扰动下的膜电位演化过程，并提出动态 LIF 神经元以实现对神经元时间演化过程的主动调控[145]。动态神经元中逐步可训练参数通过最小化膜电位扰动的均方差调整，以增强 SNN 的鲁棒性。这些方法与标准 SNN 训练相比，需记录神经元浮点变量，增加存储和训练复杂度。

综上，通过调整神经元的膜电位阈值、泄漏因子等参数影响神经元结构与动力学设计可以提升 SNN 对抗鲁棒性，但也面临参数调优复杂、计算成本上升等挑战。

(2)基于对抗训练的方法

对抗训练能够有效提高深度网络的鲁棒性，对 SNN 也不例外。Bagheri 等人提出将对抗训练用于训练 SNN 的方法[41]。然而该方法仅在浅层 SNN 结构上进行验证，深层能力为止。因此，基于上述方法，Kundu 等人将对抗训练应用到深层 SNN 上[24]。作者提出了 HIRE-SNN 训练方法，旨在不显著增加额外训练时间的前提下通过引入构造扰动来增强 SNN 的鲁棒性。时间步被划分为相等长度的周期，在每个周期结束时为输入图像添加噪声。然而，该方法仅针对图像数据集验证，尚未在事件数据或时序任务中拓展。该方法涉及的对抗训练思想还在后续其他方法体现[42,121,140,144]。

Ding 等人拓展了对抗训练的扰动样本使用[25]。在 Kundu 等人工作中，只有 FGSM 被使用。在 Ding 等人工作中，作者构建了多种梯度近似攻击的扰动样

本,在训练过程中随机采样混合扰动样本进行对抗训练,使模型能够泛化到不同攻击方法上。然而,采用多种梯度近似攻击也会导致训练时间大大增加。

Özdenizci 等人提出了一种基于 ANN 转换 SNN 的对抗训练防御方法,通过整合预训练 ANN 中的鲁棒性,提升 SNN 的鲁棒性[132]。作者引入深度学习中增强的对抗训练方法 TRADES 和 MART 优化 ANN,并将网络转换为 SNN。随后,该方法对 SNN 进行对抗微调,联合优化突触连接权重和各层发放阈值。该方法的不足在于需要训练 ANN,不适用于纯事件数据的任务,且训练过程复杂,整体训练开销较大。

Mukhoty 等人使用对抗训练框架时对网络中的随机性采用直通估计器进行梯度估计,从而生成有效扰动样本[120]。

Ding 等人提出新的对抗训练目标以及辅助损失函数,以提高 SNN 鲁棒性[145]。作者从 LIF 神经元的动力学方程出发,推导出扰动前后膜电位的演化差值公式,并构建可微的扰动度量指标即膜电位扰动的均方差。训练时,作者采用未受扰动样本和扰动后样本的平均任务损失的基础上,引入膜电位扰动的均方差作为对抗训练辅助损失函数,鼓励网络对输入扰动在特征层面的响应更加一致。该方法使用简单,不会增加太多计算量。然而,膜电位扰动的均方差损失依赖动力学假设,实际中可能受到数据分布等其他因素的干扰。

综上,对抗训练已被广泛用于提升 SNN 的鲁棒性,引入扰动样本、辅助损失函数或结合 ANN 训练策略均可有效增强 SNN 抗干扰能力,但是会带来训练开销增加等问题。

(3)基于正则训练的方法

正则训练方法通过向模型加入正则化改善模型的能力。正则化的对象可以为权重、脉冲激活以及梯度等。

Ding 等人通过理论分析认为 SNN 具备一定相比于 ANN 的鲁棒性优势[25]。作者提出了正则化对抗训练 RAT 方法,通过 SNN 的理论分析结果增强 SNN 的鲁棒性。作者推导出脉冲 Lipschitz 常数刻画扰动在脉冲传播中的放大效应,并引入谱范数约束,减小扰动距离的扩张。该方法采用谱范数正则化从结构约束角度防止扰动放大,且可以与对抗训练互补增强鲁棒性。然而,该方法对不同架构需调参评估鲁棒性性能,且当前调控脉冲 Lipschitz 常数的方式是间接方式。

Liu 等人发现 SNN 对于随机扰动的鲁棒性远高于对抗扰动下的鲁棒性,并证明了两种鲁棒性之间的差异可以被网络梯度的稀疏性所约束[121]。作者提出基于梯度稀疏性正则化的 SR 训练方法,旨在通过减小对抗鲁棒性与随机鲁棒性之间的差异,来提升 SNN 的对抗鲁棒性。为了避免将梯度的 l_0 范数直接作为正则化项导致的二次反传问题,作者使用了有限差分来估计梯度的范数。该方法主要在梯度上进行操作,不依赖于损失函数的具体形式,也不依赖于攻击方式,且理论基础较好。然而,该方法在面对非梯度攻击或梯度遮蔽等场景下可能存在问题,且引入 SR 后训练时间增加幅度较大。

综上,正则化方法通过引入结构性约束抑制脉冲放大和梯度异常,可有效提升 SNN 的对抗鲁棒性,但仍面临调参复杂、训练耗时等挑战。

(4)基于区间边界传播的方法

区间边界传播方法可以确定网络扰动后可证明的输出边界范围。然而 ANN 中方法无法直接使用在 SNN 中,研究人员为 SNN 设计并修改了区间边界传播方法以支持 SNN 的脉冲激活传播。

Liang 等人基于区间边界传播的框架探讨了 SNN 的鲁棒性问题[23]。作者提出了两种方法:S-IBP(Spiking Interval Bound Propagation)和 S-CROWN(Spiking CROWN),通过为 SNN 的神经元建模和输入形式设定边界,提升其对对抗攻击的抵抗力。S-IBP 方法主要针对 SNN 中的脉冲产生函数,通过线性松弛计算神经元发放脉冲时的上下界,确保在扰动较小时模型的输出依然稳定。该方法将 SNN 的复杂动态行为转化为可计算的上下界,确保模型能够应对各种扰动。S-CROWN 方法基于 CROWN,最初是为权重噪声鲁棒性设计的验证方法。S-CROWN 利用线性方程来表述输入对输出的影响,从而处理 SNN 中的复杂神经元动态行为。作者以线性形式描述 SNN 的输出边界,进而为不同输入类型设定边界条件。训练时作者使用下界作为损失函数,逐步增加训练中的边界半径提高模型的鲁棒性。该方法首次实现 SNN 的可认证鲁棒训练,并支持脉冲输入、图像输入等格式。然而,该方法计算成本较高,需分别进行前向和反向边界传播。脉冲产生函数附近的松弛带来较大误差,使得可证半径不准确也会影响模型性能。

基于上述工作,Mukhoty 等人将发放率编码过程视为一种对输入加入伯努利噪声的随机平滑,建立了 SNN 可证防御模型的框架[120]。为了构建可证边界,作者首先建立了单伯努利噪声扰动下的理

论鲁棒性界限,然后推广到多变量伯努利噪声编码的情况。作者还结合对抗训练策略,在伯努利平滑分类器的基础上进一步增强鲁棒性。通过使用直通估计器处理非可导的随机采样节点。该方法有可证半径保证,但是需要大量蒙特卡洛采样得出可证边界,计算开销较大。

综上,区间边界传播方法通过为 SNN 建模上下输出边界,实现对扰动下模型行为的可证保障,以此增强鲁棒性表现,但仍受限于边界松弛误差、脉冲函数不可导及较高计算成本等实际问题。

(5)基于网络轻量化的方法

稀疏性和量化可以通过移除冗余连接或正则化改善过拟合,以此增强神经网络泛化能力和鲁棒性。在 SNN 中也有相关工作。

Li 等人提出将感知量化与对抗训练及推理相融合,以缓解量化引起的性能下降并提升对抗鲁棒性[47]。训练过程中,作者在每轮训练迭代中引入对抗攻击以增强模型鲁棒性,提出将对抗训练与感知量化训练相结合的综合框架,使得模型能够在感知量化过程中进行对抗训练,减少攻击影响。作者通过设计融合感知量化的推理过程,在推理阶段采用感知量化后的模型,提高对抗攻击下的推理精度。该方法的优势是作者在 PKU-NC64C 类脑芯片上进行硬件验证,证明了量化方法在实际硬件部署中的有效性。然而,方法存在以下不足。方法实验主要集中在较小规模模型和数据集,缺乏大规模、复杂任务的系统验证。另外,模型需要在训练时精细调整比特宽度和量化比例,缺乏自动化调整手段。

改善稀疏性有助于降低 SNN 计算能耗。Schmolli 等人提出了一种基于稀疏结构的对抗鲁棒 SNN 转换方法,旨在提高鲁棒性的同时兼顾稀疏性与能效[48]。该方法首先在 ANN 训练阶段采用 TRADES 鲁棒训练框架进行对抗训练,并通过稀疏剪枝方法生成稀疏但鲁棒的 ANN 权重和连接结构。作者在转换后的稀疏 SNN 上进行对抗微调,微调阶段使用基于代替梯度的反向传播方法,并结合正则化的对抗训练目标,稳定神经元的输出响应。该方法得到的 SNN 不仅比传统端到端稀疏 SNN 训练方法具有更高的对抗鲁棒性,还显著降低了计算成本。然而该方法也存在不足,方法需要多阶段训练流程,训练复杂度较高:整体训练流程复杂、周期长,对于资源受限的应用仍然有较大负担。

综上,网络轻量化方法通过稀疏性剪枝与量化等手段在改善 SNN 能效的同时增强其鲁棒性,但仍面临训练流程复杂等问题。

4.4　输出防御方法

Ding 等人探讨了通过调整 SNN 的输出解码方法来增强鲁棒性的方法[137]。当前常用的解码策略是通过平均脉冲发放率来确定模型的输出类别。作者引入了首次脉冲时间解码(Time-to-first-spike,TTFS)这种时间依赖的解码方式。该方法不依赖输出脉冲的数量,更加关注输出神经元首次发放脉冲的时间点,作为模型决策和误差信号的依据。作者发现采用 TTFS 解码的模型整体鲁棒性更优,可以在面对黑盒攻击时表现出较好的鲁棒性。作者也指出 TTFS 输出防御方法的局限性。为了保证 TTFS 解码的可微分性,训练过程中需要对首次脉冲时间进行可微分处理以配合交叉熵损失函数。这带来了额外的计算开销和复杂度。

输出防御方法利用时间信息提升 SNN 对抗鲁棒性,但需对解码进行可微处理以适配训练,增加了计算复杂度。

4.5　防御方法对比与总结

本文将防御方法分为输入防御方法、网络防御方法以及输出防御方法三个大类。

输入防御方法主要针对扰动样本在进入网络前的处理,可分为基于脉冲编码与基于滤波两类,前者从编码源头提升鲁棒性,后者从数据预处理角度过滤扰动。这类方法的优势在于实现灵活且易于部署,能够直接作用于输入数据而不改变网络结构。然而,其鲁棒性提升常与参数选择或随机性有关,存在降低准确率或增加训练复杂度的情况。

网络防御方法更注重从模型本体出发提升鲁棒性,可分为五类:神经元设计、对抗训练、正则训练、区间边界传播与网络轻量化策略。在神经元设计方面,研究者通过调整神经元膜电位阈值、泄漏因子等参数,改善 SNN 的鲁棒性。然而在计算成本和调参复杂性上存在挑战。对抗训练方法通过引入扰动数据和辅助损失函数提升鲁棒性。这些方法改善了泛化性但训练耗时增加。正则训练方法利用结构或梯度正则约束防止扰动放大,可不依赖攻击方法但存在额外调参与训练开销。区间边界传播方法建模脉冲激活的上下界实现可证鲁棒性,但受限于边界松弛误差与高计算代价。网络轻量化如稀疏剪枝与感知量化,则兼顾鲁棒性与能效,适合边缘部署,但训练流程复杂。整体而言,这些方法通过模型设计与训练策略的多层优化通常理论基础扎实,适

合在训练阶段提升模型鲁棒性，但也带来了效率与性能的权衡问题。

输出防御方法通过改变解码方法提升 SNN 的对抗鲁棒性。当前常用的输出方式多基于平均脉冲发放率，研究者提出使用首次脉冲时间解码增强鲁棒性，强化了 SNN 对时间信息的利用。输出防御方法无需修改模型结构，适合作为后处理模块集成进模型；但实现时需权衡优化时的计算代价。

表 4 总结了文中提到的防御方法所涉及的防御方案、防御类型、进行了验证的攻击类型以及测试用的数据集。可以观察到以下趋势：

表 4 脉冲神经网络防御方法总结

参考文献	输入	网络						输出	防御类型	验证攻击	数据集
	①	②	③	④	⑤	⑥	⑦	⑧	主动/被动	黑盒/白盒	
[144]		√							被动	白盒	MNIST/CIFAR10/CIFAR100
[127]			√						被动	白盒	MNIST
[28]	√								被动	白盒	MNIST/FashionMNIST
[31]	√								被动	白盒	MNIST/CIFAR10/CIFAR100
[46]	√	√							被动	白盒	ImageNetC
[134]	√								被动	白盒	MNIST/CIFAR10/CIFAR100
[146]		√							被动	黑盒	MNIST/FashionMNIST/NMNIST/DVSGesture
[140]	√								被动	黑盒	CIFAR10/ImageNetC
[137]		√							被动	黑盒/白盒	MNIST/CIFAR10
[26]	√								被动	黑盒/白盒	CIFAR10
[22]	√								被动	黑盒/白盒	CIFAR10/CIFAR100
[147]	√							√	被动	黑盒/白盒	MNIST/FashionMNIST/CIFAR10
[41]	√				√				被动	黑盒/白盒	CIFAR10
[143]	√								被动	黑盒/白盒	MNIST
[23]						√			主动	白盒	MNIST/FashionMNIST/NMNIST
[120]	√					√	√		主动	白盒	CIFAR10/CIFAR100/SVHN/ImageNet100
[48]							√		主动	白盒	USPS
[45]		√							主动	黑盒	MNIST/FashionMNIST
[25]				√			√		主动	黑盒/白盒	CIFAR10/CIFAR100
[24]							√		主动	黑盒/白盒	CIFAR10/CIFAR100
[125]					√		√		主动	黑盒/白盒	CIFAR10/CIFAR100/TinyImageNet
[142]			√	√			√		主动	黑盒/白盒	CIFAR10/CIFAR100
[138]		√	√						主动	黑盒/白盒	SVHN/CIFAR10/CIFAR100/TinyImageNet
[145]			√				√		主动	黑盒/白盒	SVHN/FMNIST/CIFAR10/CIFAR100
[141]		√	√				√		主动	黑盒/白盒	CIFAR10/CIFAR100/TinyImageNet
[139]							√		主动	黑盒/白盒	SVHN/CIFAR10/CIFAR100
[131]				√			√		主动	黑盒/白盒	CIFAR10/CIFAR100
[47]	√								主动	黑盒/白盒	CIFAR10/CIFAR100/TinyImageNet/ImageNet
[42]	√		√				√		主动	黑盒/白盒	CIFAR10/CIFAR100
[44]		√							主动/被动	黑盒	NMNIST/DVSGesture
[126]		√							主动/被动	黑盒	NMNIST/DVSGesture

注：表中序号对应防御技术：①脉冲编码；②滤波；③神经元设计；④正则训练；⑤网络轻量化；⑥区间边界传播；⑦对抗训练；⑧脉冲解码。

(1)脉冲编码是目前研究最广泛的输入防御技术,说明从输入提升鲁棒性是主流研究方向。其次是滤波方法,尤其在神经形态数据中应用广泛,说明这类方法在处理事件噪声上具有实际价值。在网络层面,神经元设计、对抗训练和正则训练被多项研究高频采用,表明 SNN 内部结构与训练策略是网络层面鲁棒性增强的核心。区间边界传播与网络轻量化现频次较低,未来研究可在此方向继续发展。输出防御方法目前较少,只有脉冲解码唯一子类。鉴于其防御效果,未来研究可在此方向进一步研究。

(2)防御类型上,表中被动防御主动防御几乎参半。被动防御指防御方法不了解攻击方法的细节,也无法利用攻击后的数据提高鲁棒性,例如编码、滤波方法;而主动防御指防御方法可利用攻击后的数据提高鲁棒性,例如对抗训练。主动防御方法思想上与深度学习对抗鲁棒性增强方法一致性较高,能够利用的是 SNN 网络层面的独特性设计方法。而被动防御能够主要体现 SNN 本征鲁棒性并进行发扬。因此,两个防御类型都是未来重要研究方向。

(3)黑盒攻击是目前评估 SNN 防御方法的主要方法。多数编码与滤波方法适用于黑盒攻击场景,说明输入上的操作具有良好的通用性。区间边界传播、对抗训练与正则训练方法则主要在白盒攻击下验证鲁棒性。未来研究需要设计能同时抵御黑盒与白盒攻击的方法。

(4)CIFAR10 是当前 SNN 鲁棒性研究的主要评估数据集,几乎所有防御方法都基于该数据集进行了实验验证。其次是 MNIST、CIFAR100 与 FashionMNIST 等中小规模数据集。用于事件数据验证的方法相对较少,事件模态下的防御策略仍处于发展初期,未来需要更多面向 DVS 等任务的防御方法,以及面向大规模图像数据集的防御方法。

防御方法评估时使用的性能指标与攻击方法相同。模型的防御能力或鲁棒性越好,攻击后分类准确率越高,攻击成功率越低。表 5 通过 CIFAR10 数据集下的白盒 FGSM 攻击实验,展示了近年来部分 SNN 防御方法在攻击前后准确率的变化情况。整体趋势表明,SNN 防御方法的鲁棒性自 2020 年起逐年有所提升,反映出该领域的持续进展。然而,近年来提升趋于缓慢。部分研究指出,尽管某些防御方法能够显著提升 SNN 的鲁棒性,但在原始数据上的性能依然较低。这一现象体现了深度学习领域中普遍存在的鲁棒性与准确率之间的性能权衡问题,提示未来防御设计需兼顾扰动前后的整体表现。通过进一步对比表 3 与表 5 可以发现,当前 SNN 模型在整体上仍面临攻击强度高于其防御能力的严峻挑战。

表 5　脉冲神经网络防御性能

防御方法	网络结构	年份	攻击前准确率	攻击后准确率
Sharmin 等的方法[26]	VGG5	2020	89.3%	15%
Sharmin 等的方法[26]	ResNet20	2020	86.1%	31.3%
Kundu 等的方法[24]	ResNet12	2021	91.9%	21.1%
Kundu 等的方法[24]	VGG5	2021	87.9%	35.5%
Ding 等的方法[25]	VGG-11	2022	90.74%	45.23%
Bhaskar 等的方法[120]	VGGSNN	2023	79.55%	75.57%
Ding 等的方法[25]	VGG-11	2024	90.13%	45.75%

5　未来研究方向

在未来的研究中,SNN 的攻击与防御方法的探索应当从多层次展开,以应对其安全挑战。同时需要针对 SNN 特点及特定应用设计相应防御方法。下面列举了一些未来攻击方法和防御方法可以涉及的研究方向:

对于攻击方法:

(1)针对防御方法漏洞的攻击方法:未来的攻击方法需要分析现有 SNN 防御方法中的弱点。通过对这些防御方法的脆弱性分析,开发针对性的精细攻击方法,提升攻击成功率。这一研究方向需要动态且持续地设计新型攻击手段,以应对未来防御方法的发展。

(2)针对数据特点的攻击方法:目前 SNN 兼容处理静态数据、动态事件等。事件数据由于其隐私性好、动态范围广等已成为研究热点。未来可围绕事件数据集设计攻击方法。同时,针对运动物体的物理世界攻击[146]也将成为关键研究方向,这涉及如何对数据施加合理的范数约束,使网络做出误判断。

(3)针对脉冲编码特点的攻击方法:未来攻击方法可以针对 SNN 将浮点输入转换为脉冲的编码机制展开攻击。脉冲编码既涉及神经元层面的编码方式,也体现在整个网络结构的脉冲信息传递过程中。未来的攻击方法应针对不同层面的编码方式展开,探索如何通过干扰神经元的发放模式或操纵网络内的脉冲传递路径,破坏 SNN 输出信号。这类攻击不仅可以在单一神经元层面进行微观操作,还可以在网络层面进行操控,从而影响整个脉冲信息处理过程,提升攻击的有效性。

(4)结合现有深度学习技术的攻击方法:未来攻击方法可以继续借鉴深度学习中已经成熟的攻击方法,尤其是基于梯度的攻击方法,并结合 SNN 的特性进行脉冲版本的变体改进。未来还可以进一步引入深度学习中的前沿技术,如自监督学习、度量学习和生成式学习等,以增强对 SNN 的攻击能力。通过这些技术,构建更加适应性强的攻击方法,能够在更广泛的场景和模型上保持高效,提升攻击的可迁移性和泛化能力,从而使攻击方法在不同的 SNN 架构和任务中均可表现良好。

(5)探索其他梯度或梯度无关的有效攻击方法:在没有明确梯度信息的情况下,传统基于梯度的攻击方法可能失效,因此需要探索新的攻击路径。未来的研究应开发无梯度或代替梯度的攻击方法,如基于黑盒优化、进化方法、随机搜索等方法,这些技术可以通过操控输入数据或网络结构,破坏 SNN 的学习和决策过程。

(6)探索计算量小的有效攻击方法:目前有效的 SNN 攻击计算量均较大。SNN 梯度攻击需要借助可微近似施加扰动。而梯度无关的算法需要大量迭代和推理次数。目前有趋势可以借助计算量较小的脉冲发放率可微近似方法实现有效攻击方法。未来应设计计算量较小且有效的攻击方法。

对于防御方法:

(1)结合脉冲编码特点的防御方法:未来防御方法需要基于对 SNN 脉冲编码机制的深刻理解设计防御方法。通过优化脉冲生成和传输的过程,增强网络的鲁棒性。可以通过随机化脉冲的发放间隔或动态调整脉冲频率,增加系统对攻击的抗扰动能力。此类防御方法应在防御攻击上表现出通用性,能有效应对基于时序和频率特征的高级攻击。

(2)针对 SNN 时序数据特点的防御方法:未来防御方法将提升 SNN 在处理时序数据时的鲁棒性。由于 SNN 在时序数据处理上具有显著优势,未来的防御方法应设计出能够适应和抵御时序扰动的机制。方法应具备对输入数据的时间偏移的鲁棒性。针对噪声干扰或脉冲丢失问题,防御方法应包含去噪机制或丢失信息的重建能力。

(3)结合现有深度学习技术的防御方法:未来防御方法将借鉴深度学习中成熟的防御技术,增强 SNN 的鲁棒性。可发展对抗样本检测技术识别威胁,并改进对抗训练方法。未来可引入基于范数约束的正则化手段提高网络的防御能力。通过将这些深度学习中的防御技术与 SNN 的特点相结合,有望显著提升 SNN 在复杂攻击环境下的鲁棒性。

(4)新型脉冲架构的防御方法:未来防御方向需要围绕近年来的创新 SNN 架构进行深入探索。例如,基于 Transformer 架构的 SNN 和 attention 机制的 SNN 在处理复杂时序数据和长范围依赖关系方面表现出色[148-151],因此需要研究如何为这些新型架构设计有效的防御方法。针对这些新架构的防御方法应考虑其独特的网络结构和信息处理方式,开发针对防御方法。

(5)生物启发的防御方法:未来防御方法将从生物神经系统机制中汲取灵感,以开发 SNN 的防御方法。SNN 的设计初衷便是模仿生物神经系统的工作方式,因此,深入探索生物系统中的信息处理和调控机制可以为 SNN 提供宝贵的防御思路。生物神经系统中的自适应学习、冗余编码等特性,可以启发开发出新型的防御方法,以提高 SNN 的鲁棒性和安全性。这类防御方法能推动生物学和人工智能领域的交叉研究,拓展防御技术的应用边界。

(6)神经形态范式下的防御方法:未来防御方向需要在硬件资源限制的条件下提升 SNN 的防御能力。SNN 应用于神经形态计算时,依赖专门的硬件平台。这些平台通常具有特定的资源约束,如计算能力、存储空间和能耗限制。因此,未来的防御研究需要在这些硬件限制下开发高效的防御方法。研究应着重于设计能够在有限硬件资源下实现的防御机制。

(7)SNN 典型应用场景测试和任务验证:未来防御方法需要在 SNN 的实际应用环境中评估其防御能力。SNN 在事件驱动的视觉计算、脑机接口等典型应用场景中展现出独特优势。因此,未来的研究需要在这些真实场景中进行深入测试。此外,还需要保证 SNN 在典型场景中的实际效果,例如在物体分类、目标检测和时序预测等任务中验证 SNN 的性能,评估其在处理不同类型输入数据时的鲁棒性和稳定性。

6 总 结

近年国内外在 SNN 对抗扰动和防御研究领域取得了显著进展。本文总结了当前在攻击与防御方法领域的研究成果,并探讨了未来可能的研究方向。SNN 在一定程度上由于其离散的内部信息表示和时序滤波的信息处理方式展现出一定的鲁棒性。然而,多模态数据输入和端到端梯度训练的普及也带

来了负面影响,使得 SNN 同样容易受到攻击。SNN 在实际应用中实现更安全的部署需要神经形态计算技术的不断进步以及对抗方法和防御方法的持续升级。推动 SNN 在实际应用中的安全部署不仅对提升其自身的安全性至关重要,也为整个深度学习领域的安全研究提供了新的视角和启示。

致　谢　衷心感谢国家自然科学基金(优秀青年基金,项目批准号 62422601;联合重点项目,项目批准号 U24B20140;面上项目,项目批准号 62176003)以及北京市科技新星(项目批准号 20230484362,20240484703)对本研究的资助,使得本研究得以顺利进行。

参　考　文　献

[1] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[2] Fei Nanyi, Lu Zhiwu, Gao Yizhao, Yang Guoxing, Huo Yuqi, Wen Jingyuan, Lu Haoyu, Song Ruihua, Gao Xin, Xiang Tao, et al. Towards artificial general intelligence via a multimodal foundation model. Nature Communications, 2022, 13(1): 3094

[3] Jumper John, Evans Richard, Pritzel Alexander, Green Tim, Figurnov Michael, Ronneberger Olaf, Tunyasuvunakool Kathryn, Bates Russ, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583-589

[4] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and harnessing adversarial examples//International Conference on Learning Representations. San Diego, USA, 2015

[5] Kurakin Alexey, Goodfellow Ian J, Bengio Samy. Adversarial examples in the physical world//International Conference on Learning Representations Workshop. Toulon, France, 2017: 1-14

[6] Kaviani Sara, Han Ki Jin, Sohn Insoo. Adversarial attacks and defenses on AI in medical imaging informatics: A survey. Expert Systems with Applications, 2022, 198: 116815

[7] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Virtual, 2020: 119-126

[8] Maass Wolfgang. Networks of spiking neurons: the third generation of neural network models. Neural networks, 1997, 10(9): 1659-1671

[9] Gerstner Wulfram, Kistler Werner M, Naud Richard, Paninski Liam. Neuronal dynamics: From single neurons to networks and models of cognition. Cambridge University Press, 2014

[10] Zenke Friedemann, Neftci Emre O. Brain-inspired learning on neuromorphic substrates. Proceedings of the IEEE, 2021, 109(5): 935-950

[11] Pei Jing, Deng Lei, Song Sen, Zhao Mingguo, Zhang Youhui, Wu Shuang, Wang Guanrui, Zou Zhe, Wu Zhenzhi, He Wei, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature, 2019, 572(7767): 106-111

[12] Roy Kaushik, Jaiswal Akhilesh, Panda Priyadarshini. Towards spike-based machine intelligence with neuromorphic computing. Nature, 2019, 575(7784): 607-617

[13] Zenke Friedemann, Bohté Sander M, Clopath Claudia, et al. Visualizing a joint future of neuroscience and neuromorphic engineering. Neuron, 2021, 109(4): 571-575

[14] Zheng Hanle, Wu Yujie, Deng Lei, Hu Yifan, Li Guoqi. Going deeper with directly-trained larger spiking neural networks//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2021: 11062-11070

[15] Fang Wei, Yu Zhaofei, Chen Yanqi, Huang Tiejun, Masquelier Timothée, Tian Yonghong. Deep residual learning in spiking neural networks//Advances in Neural Information Processing Systems, Virtual, 2021, 34: 21056-21069

[16] Fang Wei, Yu Zhaofei, Chen Yanqi, Masquelier Timothée, Huang Tiejun, Tian Yonghong. Incorporating learnable membrane time constant to enhance learning of spiking neural networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 2661-2671

[17] Wu Yujie, Deng Lei, Li Guoqi, Zhu Jun, Shi Luping. Spatio-temporal backpropagation for training high-performance spiking neural networks. Frontiers in Neuroscience, 2018, 12: 331

[18] Su Jiawei, Vargas Danilo Vasconcellos, Sakurai Kouichi. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841

[19] Dong Yinpeng, Liao Fangzhou, Pang Tianyu, Su Hang, Zhu Jun, Hu Xiaolin, Li Jianguo. Boosting adversarial attacks with momentum//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[20] Li Yujie, Xu Xing, Xiao Jinhui, Li Siyuan, Shen Heng Tao. Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. IEEE Internet of Things Journal, 2020, 8(8): 6337-6347

[21] Xiao Zihao, Gao Xianfeng, Fu Chilin, Dong Yinpeng, Gao Wei, Zhang Xiaolu, Zhou Jun, Zhu Jun. Improving transferability of adversarial patches on face recognition with generative models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 11845-11854

[22] Sharmin Saima, Rathi Nitin, Panda Priyadarshini, Roy Kaushik. Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 399-414

[23] Liang Ling, Xu Kaidi, Hu Xing, Deng Lei, Xie Yuan. Toward robust spiking neural network against adversarial perturbation//Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 10244-10256

[24] Kundu Souvik, Pedram Massoud, Beerel Peter A. HIRE-SNN: Harnessing the inherent robustness of energy-efficient deep spiking neural networks by training with crafted input noise//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 5209-5218

[25] Ding Jianhao, Bu Tong, Yu Zhaofei, Huang Tiejun, Liu Jian. SNN-RAT: Robustness-enhanced spiking neural network through regularized adversarial training//Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 24780-24793

[26] Sharmin Saima, Panda Priyadarshini, Sarwar Syed Shakib, Lee Chankyu, Ponghiran Wachirawit, Roy Kaushik. A comprehensive analysis on adversarial robustness of spiking neural networks//Proceedings of the International Joint Conference on Neural Networks. Budapest, Hungary, 2019: 1-8

[27] Massa Riccardo, Marchisio Alberto, Martina Maurizio, Shafique Muhammad. An efficient spiking neural network for recognizing gestures with a DVS camera on the Loihi neuromorphic processor//Proceedings of the International Joint Conference on Neural Networks. Virtual, 2020: 1-9

[28] Nomura Osamu, Sakemi Yusuke, Hosomi Takeo, Morie Takashi. Robustness of spiking neural networks based on time-to-first-spike encoding against adversarial attacks. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(9): 3640-3644

[29] Liang Ling, Hu Xing, Deng Lei, Wu Yujie, Li Guoqi, Ding Yufei, Li Peng, Xie Yuan. Exploring adversarial attack in spiking neural networks with spike-compatible gradient. IEEE Transactions on Neural Networks and Learning Systems, 2021, 34(5): 2569-2583

[30] Bu Tong, Ding Jianhao, Hao Zecheng, Yu Zhaofei. Rate gradient approximation attack threats deep spiking neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7896-7906

[31] Li Yaxin, Shen Jiangrong, Xu Qi. A summary of image recognition-relevant multi-layer spiking neural networks learning algorithms. Journal of Image and Graphics, 2023, 28(2): 385-400 (in Chinese)
李雅馨, 申江荣, 徐齐. 面向图像识别的多层脉冲神经网络学习算法综述. 中国图象图形学报, 2023, 28(2): 385-400

[32] Zhang Huigang, Xu Guizhi, Guo Jiarong, Guo Lei. A review of brain-like spiking neural network and its neuromorphic chip research. Journal of Biomedical Engineering, 2021, 38(5): 986-994(in Chinese)
张慧港, 徐桂芝, 郭嘉荣, 郭磊. 类脑脉冲神经网络及其神经形态芯片研究综述. 生物医学工程学报, 2021, 38(5): 986-994

[33] Liu Xiaode, Guo Yufei, Huang Xuhui, Ma Zhe. Research advance in intelligent control based on spiking neural networks. Control Theory & Applications, 2024, 41(12): 2189-2206 (in Chinese)
刘晓德, 郭宇飞, 黄旭辉, 马喆. 基于脉冲神经网络的智能控制研究进展. 控制理论与应用, 2024, 41(12): 2189-2206

[34] Lin Xianghong, Wang Xiangwen, Zhang Ning, Ma Huifang. Supervised learning algorithms for spiking neural networks: A review. Acta Electronica Sinica, 2015, 43(3): 577-586 (in Chinese)
蔺想红, 王向文, 张宁, 马慧芳. 脉冲神经网络的监督学习算法研究综述. 电子学报, 2015, 43(3): 577-586

[35] Sun Hao, Chen Jin, Lei Lin, Ji Kefeng, Kuang Gangyao. Adversarial robustness of deep convolutional neural network-based image recognition models: A review. Journal of Radars, 2021, 10(4): 571-594 (in Chinese)
孙浩, 陈进, 雷琳, 计科峰, 匡纲要. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述. 雷达学报, 2021, 10(4): 571-594

[36] Wang Wenxun, Wang Chenglei, Qi Huihui, Ye Menghao, Zhang Yanning. Survey on adversarial attack and adversarial defense technologies for deep learning models. Journal of Signal Processing, 2025, 41(2): 198-223 (in Chinese)
王文萱, 汪成磊, 齐慧慧, 叶梦昊, 张艳宁. 面向深度模型的对抗攻击与对抗防御技术综述. 信号处理, 2025, 41(2): 198-223

[37] Büchel Julian, Lenz Gregor, Hu Yalun, Sheik Sadique, Sorbaro Martino. Adversarial attacks on spiking convolutional neural networks for event-based vision. Frontiers in Neuroscience, 2022, 16: 1068193

[38] Gehrig Daniel, Loquercio Antonio, Derpanis Konstantinos G, Scaramuzza Davide. End-to-end learning of representations for asynchronous event-based data//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5633-5643

[39] Lee Wooju, Myung Hyun. Adversarial attack for asynchronous event-based data//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual, 2022: 1237-1244

[40] Hao Zecheng, Bu Tong, Shi Xinyu, Huang Zihan, Yu Zhaofei, Huang Tiejun. Threaten spiking neural networks through combining rate and temporal information//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-17

[41] Bagheri Alireza, Simeone Osvaldo, Rajendran Bipin. Adversarial training for probabilistic spiking neural networks//Proceedings of the IEEE International Workshop on Signal Processing Advances in Wireless Communications. Kalamata, Greece, 2018: 1-5

[42] Ding Jianhao, Yu Zhaofei, Huang Tiejun, Liu Jian K. Enhancing the robustness of spiking neural networks with stochastic gating mechanisms//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 492-502

[43] Leontev Mikhail, Antonov Dmitry, Sukhov Sergey. Robustness of spiking neural networks against adversarial attacks//Proceedings of the International Conference on Information Technology and Nanotechnology. Samara, Russia, 2021: 1-6

[44] Cheng Xiang, Hao Yunzhe, Xu Jiaming, Xu Bo. LISNN: Improving spiking neural networks with lateral interactions for robust object recognition.//Proceedings of the International Joint Conference on Artificial Intelligence. Yokohama, Japan, 2020: 1519-1525

[45] Perez-Nieves Nicolas, Leung Vincent CH, Dragotti Pier Luigi, Goodman Dan FM. Neural heterogeneity promotes robust learning. Nature Communications, 2021, 12(1): 5791

[46] El-Allami Rida, Marchisio Alberto, Shafique Muhammad, Alouani Ihsen. Securing deep spiking neural networks against adversarial attacks through inherent structural parameters//Design, Automation & Test in Europe Conference & Exhibition. Virtual, 2021: 774-779

[47] Li Ying, Li Yanjie, Cui Xiaoxin, Ni Qinglong, Zhou Yinhao. Weight quantization method for spiking neural networks and analysis of adversarial robustness. Journal of Electronics & Information Technology, 2023, 45(9): 3218-3227 (in Chinese)
(李莹, 李艳杰, 崔小欣, 倪庆龙, 周鉴灏. 脉冲神经网络权重量化方法与对抗鲁棒性分析. 电子与信息学报, 2023, 45(9): 3218-3227)

[48] Schmolli Mathias, Baronig Maximilian, Legenstein Robert, Ozdenizci Ozan. Adversarially robust spiking neural networks with sparse connectivity//Proceedings of the Conference on Parsimony and Learning. California, USA, 2025: 865-883

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus. Intriguing properties of neural networks//Proceedings of the International Conference on Learning Representations. Banff, Canada, 2014

[50] Moosavi-Dezfooli Seyed-Mohsen, Fawzi Alhussein, Fawzi Omar, Frossard Pascal. Universal adversarial perturbations//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1765-1773

[51] Schott Lukas, Rauber Jonas, Bethge Matthias, Brendel Wieland. Towards the first adversarially robust neural network model on MNIST//International Conference on Learning Representations. Vancouver, Canada, 2018: 1-17

[52] Yang Yuzhe, Zhang Guo, Katabi Dina, Xu Zhi. ME-Net: Towards effective adversarial robustness with matrix estimation//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019: 7025-7034

[53] Guo Chuan, Rana Mayank, Cisse Moustapha, van der Maaten Laurens. Countering adversarial images using input transformations//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[54] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards deep learning models resistant to adversarial attacks//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[55] Tramér Florian, Kurakin Alexey, Papernot Nicolas, Goodfellow Ian, Boneh Dan, McDaniel Patrick. Ensemble adversarial training: Attacks and defenses//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-22

[56] Kannan Harini, Kurakin Alexey, Goodfellow Ian. Adversarial logit pairing. arXiv preprint arXiv:1803.06373, 2018

[57] Cisse Moustapha, Bojanowski Piotr, Grave Edouard, Dauphin Yann, Usunier Nicolas. Parseval networks: Improving robustness to adversarial examples//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 854-863

[58] Shixiang Gu, Luca Rigazio. Towards deep neural network architectures robust to adversarial examples//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-9

[59] Papernot Nicolas, McDaniel Patrick, Wu Xi, Jha Somesh, Swami Ananthram. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2016: 582-597

[60] Xie Cihang, Wu Yuxin, Maaten Laurens van der, Yuille Alan L, He Kaiming. Feature denoising for improving adversarial robustness//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 501-509

[61] Dhillon Guneet S, Azizzadenesheli Kamyar, Lipton Zachary C, Bernstein Jeremy D, Kossaifi Jean, Khanna Aran, Anandkumar Animashree. Stochastic activation pruning for robust adversarial defense//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018

[62] Bu Tong, Fang Wei, Ding Jianhao, DAI PENGLIN, Yu Zhaofei, Huang Tiejun. Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks//Proceedings of the International Conference on Learning Representations. Virtual, 2022: 1-19

[63] Rueckauer Bodo, Lungu Iulia-Alexandra, Hu Yuhuang, Pfeiffer Michael, Liu Shih-Chii. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Frontiers in Neuroscience, 2017, 11: 682

[64] Rathi Nitin, Srinivasan Gopalakrishnan, Panda Priyadarshini, Roy Kaushik. Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation//Proceed-

ings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-14

[65] Lee Chankyu, Panda Priyadarshini, Srinivasan Gopalakrishnan, Roy Kaushik. Training deep spiking convolutional neural networks with STDP-based unsupervised pre-training followed by supervised fine-tuning. Frontiers in Neuroscience, 2018, 12: 435

[66] Wu Jibin, Chua Yansong, Zhang Malu, Yang Qu, Li Guoqi, Li Haizhou. Deep spiking neural network with spike count based learning rule//Proceedings of the International Joint Conference on Neural Networks. Budapest, Hungary, 2019: 1-6

[67] Grün Sonja, Rotter Stefan. Analysis of Parallel Spike Trains. New York, USA: Springer Science + Business Media, LLC. 2010

[68] Sengupta Abhronil, Ye Yuting, Wang Robert, Liu Chiao, Roy Kaushik. Going deeper in spiking neural networks: VGG and residual architectures. Frontiers in Neuroscience, 2019, 13: 95

[69] Zhang Shao-Qun, Zhou Zhi-Hua. Theoretically provable spiking neural networks//Advances in Neural Information Processing Systems. New Orleans, USA, 2022, 35: 19345-19356

[70] Knight Bruce W., Omurtag Ahmet, Sirovich Lawrence. The approach of a neuron population firing rate to a new equilibrium: an exact theoretical result. Neural Computation, 2000, 12(5): 1045-1055

[71] Hu Yifan, Li Guoqi, Wu Yujie, Deng Lei. Spiking neural networks: A survey on recent advances and new directions, 2021, 36(1):1-26 (in Chinese)
(胡一凡, 李国齐, 吴郁杰, 邓磊. 脉冲神经网络研究进展综述. 控制与决策, 2021, 36(1):1-26)

[72] Zhang Tielin, Xu Bo. Research advances and perspectives on spiking neural network. Chinese Journal of Computers, 2021, 44(9): 1767-1785 (in Chinese)
(张铁林, 徐波. 脉冲神经网络研究现状及展望. 计算机学报, 2021, 44(9): 1767-1785)

[73] Yan Xu, Xiaoqin Zeng, Lixin Han, Jing Y. A supervised multi-spike learning algorithm based on gradient descent for spiking neural networks. Neural Network, 2013, 43: 99-113

[74] Sander M. Bohté, Joost N. Kok, Johannes A. La Poutré. SpikeProp: backpropagation for networks of spiking neurons//European Symposium on Artificial Neural Networks. Bruges, Belgium, 2000: 419-424

[75] Ponulak Filip. Supervised learning in spiking neural networks with ReSuMe method[Ph. D. Thesis]. Poznan University of Technology, 2006, 46: 47

[76] Ghosh-Dastidar Samanwoy, Adeli Hojjat. A new supervised learning algorithm for multiple spiking neural networks with application in epilepsy and seizure detection. Neural networks, 2009, 22(10): 1419-1431

[77] Deng Shikuang, Gu Shi. Optimal conversion of conventional artificial neural networks to spiking neural networks//Proceedings of the International Conference on Learning Representations. Virtual, 2021: 1-14

[78] Xu Qi, Li Yaxin, Shen Jiangrong, Liu Jian K, Tang Huajin, Pan Gang. Constructing deep spiking neural networks from artificial neural networks with knowledge distillation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 7886-7895

[79] Pérez-Carrasco José Antonio, Zhao Bo, Serrano Carmen, Acha Begona, Serrano-Gotarredona Teresa, Chen Shouchun, Linares-Barranco Bernabé. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing-application to feedforward ConvNets. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2706-2719

[80] Cao Yongqiang, Chen Yang, Khosla Deepak. Spiking deep convolutional neural networks for energy-efficient object recognition. International Journal of Computer Vision, 2015, 113: 54-66

[81] Diehl Peter U, Neil Daniel, Binas Jonathan, Cook Matthew, Liu Shih-Chii, Pfeiffer Michael. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing//Proceedings of the International Joint Conference on Neural Networks. Killarney, Ireland, 2015: 1-8

[82] Rueckauer Bodo, Lungu Iulia-Alexandra, Hu Yuhuang, Pfeiffer Michael. Theory and tools for the conversion of analog to spiking convolutional neural networks. arXiv preprint arXiv:1612.04052, 2016

[83] Kim Seijoon, Park Seongsik, Na Byunggook, Yoon Sungroh. Spiking-YOLO: spiking neural network for energy-efficient object detection//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020: 11270-11277

[84] Han Bing, Roy Kaushik. Deep spiking neural network: Energy efficiency through time based coding//Proceedings of the European Conference on Computer Vision. Virtual, 2020: 388-404

[85] Kim Jaehyun, Kim Heesu, Huh Subin, Lee Jinho, Choi Kiyoung. Deep neural networks with weighted spikes. Neurocomputing, 2018, 311: 373-386

[86] Comsa Iulia M, Potempa Krzysztof, Versari Luca, Fischbacher Thomas, Gesmundo Andrea, Alakuijala Jyrki. Temporal coding in spiking neural networks with alpha synaptic function//IEEE International Conference on Acoustics, Speech and Signal Processing. Barcelona, Spain, 2020: 8529-8533

[87] Deng Shikuang, Li Yuhang, Zhang Shanghang, Gu Shi. Temporal efficient training of spiking neural network via gradient re-weighting//International Conference on Learning Representations. Virtual, 2022: 1-15

[88] Guo Yufei, Tong Xinyi, Chen Yuanpei, Zhang Liwen, Liu Xiaode, Ma Zhe, Huang Xuhui. Recdis-snn: Rectifying membrane potential distribution for directly training spiking

neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 326-335

[89] Huh Dongsung, Sejnowski Terrence J. Gradient descent for spiking neural networks//Advances in Neural Information Processing Systems. Montréal, Canada, 2018: 1440-1450

[90] Duan Chaoteng, Ding Jianhao, Chen Shiyan, Yu Zhaofei, Huang Tiejun. Temporal effective batch normalization in spiking neural networks//Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 34377-34390

[91] Jiang Haiyan, Zoonekynd Vincent, De Masi Giulia, Gu Bin, Xiong Huan. TAB: Temporal accumulated batch normalization in spiking neural networks//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 1-27

[92] Guo Yufei, Zhang Yuhan, Chen Yuanpei, Peng Weihang, Liu Xiaode, Zhang Liwen, Huang Xuhui, Ma Zhe. Membrane potential batch normalization for spiking neural networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 19420-19430

[93] Li Yuhang, Guo Yufei, Zhang Shanghang, Deng Shikuang, Hai Yongqing, Gu Shi. Differentiable spike: Rethinking gradient-descent for training spiking neural networks//Advances in Neural Information Processing Systems. Virtual, 2021: 23426-23439

[94] Xiao Mingqing, Meng Qingyan, Zhang Zongpeng, He Di, Lin Zhouchen. Online training through time for spiking neural networks//Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 20717-20730

[95] Zhang Wenrui, Li Peng. Temporal spike sequence learning via backpropagation for deep spiking neural networks//Advances in Neural Information Processing Systems. Virtual, 2020: 12022-12033

[96] Zhu Yaoyu, Yu Zhaofei, Fang Wei, Xie Xiaodong, Huang Tiejun, Masquelier Timothée. Training spiking neural networks with event-driven backpropagation//Advances in Neural Information Processing Systems. New Orleans, USA, 2022: 30528-30541

[97] Zhu Yaoyu, Ding Jianhao, Huang Tiejun, Xie Xiaodong, Yu Zhaofei. Online stabilization of spiking neural networks//International Conference on Learning Representations. Vienna, Austria, 2024: 1-20

[98] Batllori Robert, Laramee Craig B, Land W, Schaffer J David. Evolving spiking neural networks for robot control. Procedia Computer Science, 2011, 6: 329-334

[99] Hagras Hani, Pounds-Cornish Anthony, Colley Martin, Callaghan Victor, Clarke Graham. Evolving spiking neural network controllers for autonomous robots//IEEE International Conference on Robotics and Automation. New Orleans, USA, 2004: 4620-4626

[100] Foderaro Greg, Henriquez Craig, Ferrari Silvia. Indirect training of a spiking neural network for flight control via spike-timing-dependent synaptic plasticity//Proceedings of the IEEE Conference on Decision and Control. Atlanta, USA, 2010: 911-917

[101] Gu Weibin, Valavanis Kimon P, Rutherford Matthew J, Rizzo Alessandro. UAV model-based flight control with artificial neural networks: A survey. Journal of Intelligent & Robotic Systems, 2020, 100(3): 1469-1491

[102] Wang Wei, Zhou Shibo, Li Jingxi, Li Xiaohua, Yuan Junsong, Jin Zhanpeng. Temporal pulses driven spiking neural network for time and power efficient object recognition in autonomous driving//Proceedings of the International Conference on Pattern Recognition. Milan, Italy, 2021: 6359-6366

[103] Viale Alberto, Marchisio Alberto, Martina Maurizio, Masera Guido, Shafique Muhammad. Carsnn: An efficient spiking neural network for event-based autonomous cars on the Loihi neuromorphic research processor//Proceedings of the International Joint Conference on Neural Networks. Shenzhen, China, 2021: 1-10

[104] Dethier Julie, Gilja Vikash, Nuyujukian Paul, Elassaad Shauki A, Shenoy Krishna V, Boahen Kwabena. Spiking neural network decoder for brain-machine interfaces//Proceedings of the International IEEE/EMBS Conference on Neural Engineering. Cancún, Mexico, 2011: 396-399

[105] Gong Peiliang, Wang Pengpai, Zhou Yueying, Zhang Daoqiang. A spiking neural network with adaptive graph convolution and LSTM for EEG-based brain-computer interfaces. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 1440-1450

[106] Hsieh Yung-Ting, Li Zhile, Pompili Dario. A lightweight hybrid analog-digital spiking neural network for IoT//Proceedings of the International Conference on Distributed Computing in Smart Systems and the Internet of Things. Abu Dhabi, UAE, 2024: 249-253

[107] Mead Carver. Neuromorphic electronic systems. Proceedings of the IEEE, 1990, 78(10): 1629-1636

[108] Rathi Nitin, Roy Kaushik. STDP based unsupervised multimodal learning with cross-modal processing in spiking neural networks. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 5(1): 143-153

[109] Rebecq Henri, Ranftl René, Koltun Vladlen, Scaramuzza Davide. High speed and high dynamic range video with an event camera. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(6): 1964-1980

[110] Huang Tiejun, Zheng Yajing, Yu Zhaofei, Chen Rui, Li Yuan, Xiong Ruiqin, Ma Lei, Zhao Junwei, Dong Siwei, Zhu Lin, et al. 1000x faster camera and machine vision with ordinary devices. Engineering, 2023, 25: 110-119

[111] Chakravarthi Bharatesh, Verma Aayush Atul, Daniilidis Kostas, Fermuller Cornelia, Yang Yezhou. Recent event camera innovations: A survey. arXiv preprint arXiv: 2408.13627, 2024

[112] Zheng Yajing, Yu Zhaofei, Wang Song, Huang Tiejun. Spike-based motion estimation for object tracking through bio-inspired unsupervised learning. IEEE Transactions on Image Processing, 2022, 32: 335-349

[113] Khodagholy Dion, Gelinas Jennifer N, Thesen Thomas, Doyle Werner, Devinsky Orrin, Malliaras George G, et al. NeuroGrid: recording action potentials from the surface of the brain. Nature Neuroscience, 2015, 18(2): 310-315

[114] Merolla Paul A, Arthur John V, Alvarez-Icaza Rodrigo, Cassidy Andrew S, Sawada Jun, Akopyan Filipp, Jackson Bryan L, Imam Nabil, Guo Chen, Nakamura Yutaka, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. Science, 2014, 345(6197): 668-673

[115] Furber Steve B, Galluppi Francesco, Temple Steve, Plana Luis A. The spinnaker project. Proceedings of the IEEE, 2014, 102(5): 652-665

[116] Ma De, Shen Juncheng, Gu Zonghua, Zhang Ming, Zhu Xiaolei, Xu Xiaoqiang, Xu Qi, Shen Yangjing, Pan Gang. Darwin: A neuromorphic hardware co-processor based on spiking neural networks. Journal of systems architecture, 2017, 77: 43-51

[117] Brandli Christian, Berner Raphael, Yang Minhao, Liu Shih-Chii, Delbruck Tobi. A 240×180 130db 3μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits, 2014, 49(10): 2333-2341

[118] Dong Siwei, Zhu Lin, Xu Daoyuan, Tian Yonghong, Huang Tiejun. An efficient coding method for spike camera using inter-spike intervals//Proceedings of the Data Compression Conference. Snowbird, USA, 2019: 568-568

[119] Li Jianing, Tianyong Hong. Recent advances in neuromorphic vision sensors: A survey. Chinese Journal of Computers, 2021, 44(6): 1258-1286 (in Chinese)
(李家宁, 田永鸿. 神经形态视觉传感器的研究进展及应用综述. 计算机学报, 2021, 44(6): 1258-1286)

[120] Mukhoty Bhaskar, AlQuabeh Hilal, De Masi Giulia, Xiong Huan, Gu Bin. Certified adversarial robustness for rate encoded spiking neural networks//Proceedings of the International Conference on Learning Representations. Vienna, Austria, 2024: 1-17

[121] Liu Yujia, Bu Tong, Ding Jianhao, Hao Zecheng, Huang Tiejun, Yu Zhaofei. Enhancing adversarial robustness in SNNs with sparse gradients//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 30738-30754

[122] Jiang Chunming, Zhang Yilei. Adversarial defense in spiking neural networks via neural oscillation inspired gradient masking. SSRN, 2024

[123] Venceslai Valerio, Marchisio Alberto, Alouani Ihsen, Martina Maurizio, Shafique Muhammad. Neuroattack: Undermining spiking neural networks security through externally triggered bit-flips//Proceedings of the International Joint Conference on Neural Networks. Virtual, 2020: 1-8

[124] Marchisio Alberto, Nanfa Giorgio, Khalid Faiq, Hanif Muhammad Abdullah, Martina Maurizio, Shafique Muhammad. Is spiking secure? A comparative study on the security vulnerabilities of spiking and deep neural networks//Proceedings of the International Joint Conference on Neural Networks. Virtual, 2020: 1-8

[125] Xu Qi, Deng Jie, Shen Jiangrong, Tang Huajin, Pan Gang. A review of image reconstruction based on event cameras. Journal of Electronics & Information Technology, 2023, 45(8): 2699-2709 (in Chinese)
(徐齐, 邓洁, 申江荣, 唐华锦, 潘纲. 基于事件相机的图像重构综述. 电子与信息学报, 2023, 45(8): 2699-2709)

[126] Marchisio Alberto, Pira Giacomo, Martina Maurizio, Masera Guido, Shafique Muhammad. DVS-Attacks: Adversarial attacks on dynamic vision sensors for spiking neural networks//Proceedings of the International Joint Conference on Neural Networks. Shenzhen, China, 2021: 1-9

[127] Marchisio Alberto, Pira Giacomo, Martina Maurizio, Masera Guido, Shafique Muhammad. R-SNN: An analysis and design methodology for robustifying spiking neural networks against adversarial attacks through noise filters for dynamic vision sensors//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, Czech, 2021: 6315-6321

[128] Yao Yanmeng, Zhao Xiaohan, Gu Bin. Exploring vulnerabilities in spiking neural networks: Direct adversarial attacks on raw event data//European Conference on Computer Vision. Milan, Italy, 2024: 412-428

[129] Lin Xuanwei, Dong Chen, Liu Ximeng, Zhang Yuanyuan. SPA: An efficient adversarial attack on spiking neural networks using spike probabilistic//Proceedings of the 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing. Taormina, Italy, 2022: 366-375

[130] Lin Xuanwei, Dong Chen, Liu Ximeng. SFTA: Spiking neural networks vulnerable to spiking feature transferable attack//Proceedings of the IEEE 21st International Conference on Ubiquitous Computing and Communications. Chengdu, China, 2022: 140-149

[131] Krithivasan Sarada, Sen Sanchari, Rathi Nitin, Roy Kaushik, Raghunathan Anand. Efficiency attacks on spiking neural networks//Proceedings of the 59th ACM/IEEE Design Automation Conference. San Francisco, USA, 2022: 373-378

[132] Ozdenizci Ozan, Legenstein Robert. Adversarially robust spiking neural networks through conversion. Transactions on Machine Learning Research, 2024

[133] Zhang Jiahong, Wang Kexin, Yao Man, Xu Han, Zhou Peng, Xu Bo, Li Guoqi. Enhancing robustness of spiking neural networks through retina-like coding and memory-based neurons. Authorea Preprints. 2024: 1-15

[134] Park Seongsik, Lee Dongjin, Yoon Sungroh. Noise-robust deep spiking neural networks with temporal information//

Proceedings of the ACM/IEEE Design Automation Conference. San Francisco, USA, 2021: 373-378

[135] Calaim Nuno, Dehmelt Florian A, et al. The geometry of robustness in spiking neural networks. Elife, 2022, 11: e73276

[136] Van Rullen Rufin, Thorpe Simon J. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. Neural computation, 2001, 13(6): 1255-1283

[137] Ding Jianhao, Yu Zhaofei, Huang Tiejun, Liu Jian K. Spike timing reshapes robustness against attacks in spiking neural networks. arXiv preprint arXiv:2306.05654, 2023

[138] Wu Keming, Yao Man, Chou Yuhong, Qiu Xuerui, Yang Rui, Xu Bo, Li Guoqi. RSC-SNN: Exploring the trade-off between adversarial robustness and accuracy in spiking neural networks via randomized smoothing coding//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia, 2024: 2748-2756

[139] Liu Faqiang, Zhao Rong. Enhancing spiking neural networks with hybrid top-down attention. Frontiers in Neuroscience, 2022, 16: 949142

[140] Xu Mengting, Ma De, Tang Huajin, Zheng Qian, Pan Gang. FEEL-SNN: Robust spiking neural networks with frequency encoding and evolutionary leak factor. Advances in Neural Information Processing Systems, Vancouver, Canada, 2024, 37: 91930-91950

[141] Chen Weiran, Xu Qi. Robust and efficient adversarial defense in SNNs via image purification and joint detection//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Hyderabad, India, 2025: 1-5

[142] Zhang Yan, Chen Cheng, Shen Dian, Wang Meng, Wang Beilun. Take CARE: Improving inherent robustness of spiking neural networks with channel-wise activation recalibration module//Proceedings of the IEEE International Conference on Data Mining. Shanghai, China, 2023: 828-837

[143] Dapello Joel, Marques Tiago, Schrimpf Martin, Geiger Franziska, Cox David, DiCarlo James J. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations//Advances in Neural Information Processing Systems. Virtual, 2020: 13073-13087

[144] Geng Hejia, Li Peng. HoSNN: Adversarially-robust homeostatic spiking neural networks with adaptive firing Thresholds. Transactions on Machine Learning Research, 2025

[145] Ding Jianhao, Pan Zhiyu, Liu Yujia, Yu Zhaofei, Huang Tiejun. Robust stable spiking neural networks//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2024: 11016-11029

[146] Eykholt Kevin, Evtimov Ivan, Fernandes Earlence, Li Bo, Rahmati Amir, Xiao Chaowei, Prakash Atul, Kohno Tadayoshi, Song Dawn. Robust physical-world attacks on deep learning visual classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1625-1634

[147] Kim Youngeun, Park Hyoungseob, Moitra Abhishek, Bhattacharjee Abhiroop, Venkatesha Yeshwanth, Panda Priyadarshini. Rate coding or direct coding: Which one is better for accurate, robust, and energy-efficient spiking neural networks? //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore, 2022: 71-75

[148] Yao Man, Hu Jiakui, Zhou Zhaokun, Yuan Li, Tian Yonghong, Xu Bo, Li Guoqi. Spike-driven transformer//Advances in Neural Information Processing Systems. Vancouver, Canada, 2023: 64043-64058

[149] Zhou Zhaokun, Zhu Yuesheng, He Chao, Wang Yaowei, Shuicheng Yan, Tian Yonghong, Yuan Li. Spikformer: When Spiking Neural Network Meets Transformer//Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 64043-64058

[150] Yao Man, Zhao Guangshe, Zhang Hengyu, Hu Yifan, Deng Lei, Tian Yonghong, Xu Bo, Li Guoqi. Attention spiking neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 9393-9410

[151] Xu Qi, Gao Yuyuan, Shen Jiangrong, Li Yaxin, Ran Xuming, Tang Huajin, Pan Gang. Enhancing adaptive history reserving by spiking convolutional block attention module in recurrent neural networks//Advances in Neural Information Processing Systems, New Orleans, USA, 2023, 36: 58890-58901

DING Jian-Hao, Ph. D., assistant research fellow. His research interests include neuromorphic computing, spiking neural networks, neuromorphic sensors and neural coding.

LIU Yu-Jia, Ph. D., assistant research fellow. Her main research interests include spiking neural networks, adversarial attacks, and robustness learning.

BU Tong, Ph. D. candidate. His current research interests include machine learning, brain-inspired AI and neuromorphic computing.

HAO Ze-Cheng, Ph. D. candidate. His current research interests include include visual information processing and neuromorphic computing.

YU Zhao-Fei, Ph. D., assistant professor. His research interests include brain-like computing and neural networks.

HUANG Tie-Jun, Ph. D., professor. His research interests include visual information processing and neuromorphic computing.

Background

Spiking neural networks (SNNs) are expected to be widely used in real-time scenarios and safety-critical applications due to their less sensitivity to small random perturbations compared to artificial neural networks (ANNs). However, SNNs face security and privacy challenges due to their training using differential approximation techniques. Current attack methods include gradient-based and gradient-independent approaches. Researchers have also explored SNNs' performance on event streams and images, as they are naturally suited for processing event streams generated by event cameras. SNNs have distinct encoding and learning mechanisms, offering significant potential in defending against adversarial perturbations. To boost the application of SNNs, lots of researchers work on attacking and defending SNNs, including input defense methods, network defense methods, output defense methods, etc.

In this paper, we review the existing attack and defense methods for SNNs. First, we briefly overview the basics of SNNs, hardware, and their safety-critical applications. Second, we review the adversarial attack methods for SNNs. Third, we review the existing defense methods for SNNs. Finally, we give the prospect of future work on adversarial attack and defense for SNNs.

This work was supported by NSFC (Grant Nos. 62422601, U24B20140, 62176003) and Beijing Nova program (Grant Nos. 20230484362, 20240484703).