DOI:10.20176/j.cnki.nxdz.000096

一种生物视觉启发的高鲁棒性脉冲循环神经网络模型

陈林果[1], 黄 荣[1,2], 韩 芳[1,3*]

(1.东华大学 信息科学与技术学院,上海 201620; 2.东华大学 数字化纺织服装技术教育部工程研究中心,上海 201620;
3.宁夏大学 数学统计学院,宁夏 银川 750021)

摘 要:针对脉冲神经网络(spike neural networks,SNN)在对抗攻击下表现出的低鲁棒性问题,提出一种受生物视觉启发的高鲁棒性脉冲循环神经网络模型。该模型引入了初级视觉皮层V1区域的生物机制,设计了一个受到生物约束的卷积SNN前端。此外,通过结合视觉信息在皮层中的反馈连接,构建了具有内部循环机制的SNN后端。在无对抗训练的情况下,该模型在基于脉冲频率的快速梯度符号法(fast gradient sign method,FGSM)攻击下,分别在SVHN,CIFAR10,CIFAR100数据集上实现了显著提升的对抗准确率,分别提升了31.6%,22.11%和20.99%。在对抗训练的情况下,其对抗准确率分别提升了20.64%,8.79%和6.89%。随着扰动因子ε和时间窗口T的增加,该模型的准确率始终优于基准模型。实验结果表明,在面对对抗攻击时,融入生物视觉机制的脉冲循环神经网络模型的准确率显著提升,展现出更强的对抗鲁棒性。

关键词:脉冲神经网络;鲁棒性;对抗攻击;生物视觉

分类号:(中图)TP183;TP391 **文献标志码:**A

近年来,脉冲神经网络(spike neural networks,SNN)因其神经动力学特性和事件驱动的二进制脉冲传输等优势,引起了学者们的广泛关注[1-2]。研究者们不断探索将人工神经网络(artificial neural networks,ANN)的成功经验应用于SNN,以提升其性能和实用性[3-5]。然而,随着SNN在图像识别、语音识别、自然语音处理以及工业故障检测等领域的深入应用,其安全性问题日益凸显,尤其是对抗攻击下的稳定性问题。例如,在工业电机故障检测中,对抗攻击可能导致SNN将正常运行状态误判为故障状态,从而引发不必要的设备停机;或者将故障状态误判为正常状态,产生潜在的安全隐患。因此,研究SNN的对抗鲁棒性不仅具有重要的理论意义,也对其实际应用和部署至关重要。

早期关于SNN鲁棒性的研究主要聚焦于脉冲神经网络本身的特性。例如,Sharmin等[6]指出,输入的泊松编码离散化和泄漏-积累-发放(leaky integrate-and-fire, LIF)神经元的非线性激活是SNN鲁棒性的来源。他们还验证了基于梯度反向传播的模型比基于ANN2SNN方法的模型更具鲁棒性。El-Allami等[7]研究了不同放电电压阈值和时间窗口对SNN对抗鲁棒性的影响,指出脉冲结构赋予层间数据天然的稀疏性。同时,他们还强调高基线准确率并不等同于高鲁棒性,SNN的鲁棒性是相对的,取决于结构参数的选择。Kundu等[8]对视觉几何组(visual geometry group,VGG)网络结构的鲁棒性进行了多维度评估。此外,Kim等[9]在低延迟训练中比较了频率编码和直接编码的特性及优缺点,指出直接编码的准确率更高,而频率编码在功耗和鲁棒性方面更具优势。这些研究尽管揭示了SNN的天然优势,但主要集中在SNN模型本身的算法分析层面,缺乏对如何进一步提升SNN鲁棒

收稿日期:2025-03-16
基金项目:国家自然科学基金资助项目(12272092)
作者简介:陈林果(1999—),男,硕士研究生,主要从事脉冲神经网络研究,(电子信箱)chenlg2363@163.com。
*通信联系人:韩芳(1981—),女,教授,博士研究生导师,主要从事神经动力学、类脑智能及智能控制研究,(电子信箱)yadiahan@dhu.edu.cn。
引用格式:陈林果,黄荣,韩芳.一种生物视觉启发的高鲁棒性脉冲循环神经网络模型[J].宁夏大学学报(自然科学版中英文),2026,47(1):24-32.

性的深入探讨。

同时，在SNN的研究中，对抗训练虽处于起步阶段，但其潜力已开始显现。例如，Kundu等[8]在训练中加入噪声干扰，从而提高了模型在攻击下的分类精度。Chen等[10]基于过滤理论，提出了一种基于图像净化的防御方法，用于保护脉冲神经网络免受对抗攻击。Ding等[11]首次将Lipschitz分析理论扩展到SNN，设计了RAT正则化对抗训练方法，显著提升了SNN的鲁棒性。然而，对抗样本训练虽然提升了SNN的鲁棒性，但可能导致训练过程不稳定，且通常仅针对特定攻击进行优化，从而限制了模型的泛化能力[12]。

生物视觉系统在漫长的进化过程中，面对噪声干扰形成了一套独特的鲁棒性机制[13-14]。Dapello等[15]的研究表明，模型对初级视觉皮层（V1）神经反应模式的解释能力与对抗鲁棒性正相关。然而，目前关于生物视觉对模型鲁棒性影响的研究大多局限于ANN领域[16-18]。相比之下，SNN的研究主要集中在训练算法和误差分析方面，尚未充分探索初级视觉皮层对模型鲁棒性的潜在影响。此外，Kar等[19]的研究揭示了循环回路在灵长类腹侧流视觉通路中的重要性，但现有的SNN大多采用前馈架构[20-21]，未考虑生物循环回路对鲁棒性的潜在贡献。

因此，基于前人研究，针对脉冲神经网络（SNN）在对抗攻击下鲁棒性不足的问题，本文创新性地将初级视觉皮层（V1）区域的生物学特性和生物循环机制融入SNN的架构中，提出了一种生物视觉启发的高鲁棒性脉冲循环神经网络模型（V1-spike recurrent neural networks，V1-SRNN）。该模型在CIFAR-10[22]，CIFAR-100[22]和SVHN[23]（street view house numbers）数据集上进行了训练和测试。结果显示，与传统SNN相比，本文提出的V1-SRNN模型在对抗攻击下展现出更强的对抗鲁棒性和抗干扰能力。

1　构建V1-SRNN网络模型

1.1　LIF（leaky integrate-and-fire）神经元

SNN主要通过脉冲进行信息传输。多年来，研究者们已经构造出多种脉冲神经元模型[24-26]。LIF神经元在保持模型复杂性与计算需求平衡的同时，成功保留了足够的生物神经元特性，因此，本文选择LIF神经元模型作为后续脉冲神经网络研究的基础。其动力学方程表述为

$$v^l(t^-)=v^l(t-1)+W^l s^{l-1}(t), \tag{1}$$

$$s^l(t)=H(v^l(t^-)-\theta), \tag{2}$$

$$v^l(t)=\lambda v^l(t^-)(1-s^l(t))。 \tag{3}$$

其中：$v^l(t)$和$s^l(t)$分别为第l层神经元在t时刻的膜电位及其发放的二进制脉冲；$v^l(t^-)$表示触发脉冲前膜电位的瞬时状态；W^l表示第$l-1$层与第l层之间的突触权重；H是单位阶跃函数；θ为神经元设定的阈值电压；$\lambda(0<\lambda<1)$为衰减因子。对于一个脉冲神经元，若膜电位$v^l(t)$超过激活阈值，则该神经元发放脉冲后，膜电位复位为0；否则，膜电位将以λ的速率衰减。

1.2　基于初级视觉皮层V1区域的脉冲神经网络前端

构建生物视觉启发的脉冲循环神经网络模型的第一步是模拟初级视觉皮层V1区域的方位选择性。为此，本文设计了一个受生物约束的卷积脉冲特征提取模块——SpikeV1，用于提取图像输入特征，其架构见图1。

计算机视觉领域的研究显示，椭圆形各向异性高斯核更贴近部分V1区细胞方位选择感受野的特性。同时，Gabor滤波器的数学形式与神经元感受野的同心圆结构相似，能有效捕捉图像中的空间频率信息和特定相位信息，这与V1区神经元对特定空间频率及相位位置变化的敏感性高度吻合。因此，SpikeV1的第一层采用参数化的Gabor滤波器（GFB），其参数根据灵长类动物V1神经元的反应特性进行优化调整。该层通过多方向、多尺度和多空间频率的Gabor滤波器处理RGB输入图像，以提取关键特征。具体来说，该层由二维Gabor函数构成，其数学方程为

$$\boldsymbol{G}_{\theta,f,\phi,n_x,n_y}(x,y)=\frac{\cos(2\pi f+\phi)}{2\pi\sigma_x\sigma_y}\exp\left[-0.5\left(\frac{x_{\text{rot}}^2}{\sigma_x^2}+\frac{y_{\text{rot}}^2}{\sigma_y^2}\right)\right], \tag{4}$$

$$x_{\text{rot}}=x\cos\theta+y\sin\theta,\ \sigma_x=\frac{n_x}{f}, \tag{5}$$

$$y_{\text{rot}}=-x\sin\theta+y\cos\theta,\ \sigma_y=\frac{n_y}{f}。 \tag{6}$$

其中：x和y是像素在图像中的位置坐标；x_{rot}和y_{rot}分别表示相对于光栅的正交和平行方向；θ是角度参数，表示光栅的方向；f是光栅的空间频率，表示滤波器对图像中不同频率成分的敏感性；ϕ是相位参数，表示光栅相对于高斯包络的位移；σ_x和σ_y是正交和平行于光栅的高斯包络的标准偏差，可以定义为光栅周期（n_x和n_y）的倍数。

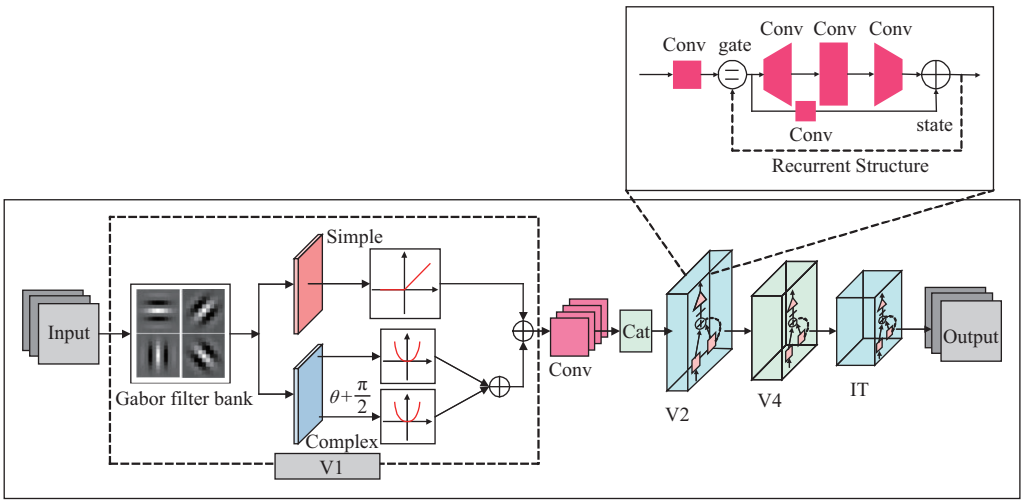

图 1 V1-SRNN 模型架构

Fig. 1 Architecture of the V1-SRNN model

Gabor 滤波器作为一种二维的带通滤波器,其参数设计基于灵长类动物 V1 神经元的生物学特性。光栅方向 θ 是通过均匀分布采样得到的,采样范围基于灵长类动物 V1 神经元的方向选择性实验结果[27],从而能够捕捉图像中的多方向特征。空间频率参数 f 被限制为 $f<5.6$ cpd(每度的周期数),这是基于灵长类动物 V1 神经元的空间频率选择性实验结果[28],确保滤波器能够捕捉到与灵长类动物视觉系统相似的频率范围。相位参数 ϕ 是通过在 [0,360] 度范围内均匀采样确定的,以确保滤波器对不同相位的刺激都有响应。高斯包络的标准差 σ_x 和 σ_y 是基于灵长类动物 V1 神经元的实验结果[29],这种设计确保滤波器的空间分辨率与频率相关,从而更好地模拟灵长类动物 V1 神经元的特性。同时,将 Gabor 滤波器的步长设置为 1,以生成一个 28×28 的激活空间,并将通道数设置为 256,平均分配给后续的简单细胞和复杂细胞。每个 Gabor 滤波器通道对输入图像的一个颜色通道进行卷积操作。

在 SpikeV1 模块的第二层,采用了两种不同的非线性层,它们根据单元类型分别作用于不同的通道。第一种非线性层用于简单细胞,以校正线性变化;第二种非线性层则用于处理正交相位的复杂细胞。假定模型的输入信息为 $\boldsymbol{I}_{\theta,f,\phi,n_x,n_y}$,经过 SpikeV1 模块第一层 GFB 滤波器处理后的线性响应可由下式给出。即

$$\boldsymbol{E}_{\theta,f,\phi,n_x,n_y}=\boldsymbol{I}_{\theta,f,\phi,n_x,n_y}*\boldsymbol{G}_{\theta,f,\phi,n_x,n_y}(x,y)。\tag{7}$$

因此,简单细胞的整流矫正的非线性响应为

$$\boldsymbol{S}_{\theta,f,\phi,n_x,n_y}=\begin{cases}\boldsymbol{E}_{\theta,f,\phi,n_x,n_y}, & \boldsymbol{E}_{\theta,f,\phi,n_x,n_y}>0,\\ 0, & \text{otherwise}。\end{cases}\tag{8}$$

而对于复杂细胞的处理,可采用能量模型进行建模,由两个相位相互正交的平方和表示。即

$$\boldsymbol{C}_{\theta,f,\phi,n_x,n_y}=\frac{1}{\sqrt{2}}\sqrt{\left(\boldsymbol{E}_{\theta,f,\phi,n_x,n_y}\right)^2+\left(\boldsymbol{E}_{\theta,f,\phi+\frac{\pi}{2},n_x,n_y}\right)^2}。\tag{9}$$

在真实的视觉处理过程中,初级视觉皮层(V1)的神经元并非仅传递输入信息,而是展现出一定程度的随机活动。这种特性被称作“随机性”或“噪声”。这种随机活动有助于扩展神经元对刺激的反应范围,提升系统对微弱信号的敏感度。然而,脉冲神经网络(SNN)的非线性动力学特性可能在长时间运行过程中引发震荡、随机噪声或信号饱和等问题。尽管生物神经网络具有随机性,但通常希望探究完全确定性网络是否能实现精确处理。因此,为了构建高精度、低错误率的模型,本文在 SpikeV1 模块中暂未对“随机发生器”进行建模。

为了确保 SpikeV1 模块提取的特征能够有效传递至后端脉冲神经网络,并进行进一步的特征分析,模块后需设置一个特定的过渡层以降低通道数量。在本文的研究中,通道数量通常设置为 64。

1.3 基于内部循环机制的脉冲神经网络后端

生物视觉循环机制的生理学基础是一个复杂而精妙的过程,涉及多个脑区的协同工作。视觉信息首先在视网膜中被光感受器转换为电信号,经过视网膜内的初步处理后,通过视神经传递到外侧膝状体(lateral geniculate nucleus,LGN)。LGN 作为中继站,将信号传递到初级视觉皮层(V1)。在初级视觉皮层,简单细胞和复杂细胞分别对特定方向和位置的边缘以及边缘的位置变化进行检测,然后将信号传递到高级视觉皮层(如 V2、V4)。高级视觉皮层区域负责整合局部特征,形成对物体形状和颜

色的初步理解。最终，信号到达下颞叶皮层（inferior temporal cortex，IT），该区域的神经元对特定的视觉对象具有高度选择性，能够准确预测核心视觉识别任务中的表现。Kar等[19]的研究表明，循环连接在这一过程中扮演重要角色。例如，视觉信息从初级视觉皮层（V1）逐步传递到更高级的视觉皮层（如V2、V4和IT），并在每个阶段通过反馈连接进行动态调整。这种反馈连接允许系统在时间维度上整合信息，从而提高处理的精确性和鲁棒性。

然而，现有的脉冲神经网络大多采用有监督学习的深层前馈架构。这种过多的层级使其难以与大脑视觉系统的腹侧通路相对应，从而降低了模型的生物可解释性；相反，若模型层数过少，则缺乏具有生物意义的大脑结构，难以模拟大脑复杂的神经动态。因此，本文的第二项工作是基于生物视觉循环机制，构建具有内循环机制的脉冲神经网络后端，并手动定性地将模型的重要区域映射到特定的大脑区域。

如图1所示，本文的模型被划分为V1、V2、V4和IT 4个区域。其中，V1区域由SpikeV1模块实现，而V2、V4和IT区域则采用了类似ResNet[30]的内循环门控连接。这种连接方式使得这些区域的输出能够反馈至自身，形成动态响应轨迹。这种机制与大脑皮层中的反馈连接相吻合，对视觉信息的处理和整合至关重要。V2、V4区域的循环连接方式与特征处理流程相似。具体来说，每个区域均包含1×1、3×3和1×1的卷积层，每个卷积层后均进行响应归一化和非线性激活。网络内部循环次数的问题将在2.2节中详细讨论。SpikeV1前端模块仅包含单向的前馈连接，负责输入的预处理，而V2、V4和IT区域的循环连接则负责更高层次的特征提取和物体识别。

2　实验结果与分析

2.1　实验平台和数据

本文实验基于配备24 GB显存的NVIDIA 3090 GPU和Intel i9-10920x处理器的平台。鉴于所提模型在信息流传递过程中的最长路径包含16个卷积层和全连接层，本文选取VGG16作为基准模型进行比较。在训练过程中，设定时间窗口为$T=6$，采用随机梯度下降（stochastic gradient descent，SGD）作为优化器，并通过L2正则化防止过拟合。模型的初始学习率为0.1，动量为0.9，并采用余弦退火学习率调度器，其周期等于训练总轮数（epochs）。

为了验证基于生物视觉通路特性的脉冲神经网络的对抗鲁棒性，选择在CIFAR-10，CIFAR-100和Street View House Numbers（SVHN）3个数据集上进行测试。其中，CIFAR-10包含10个图像类别，CIFAR-100包含100个类别。这两个数据集均包含6万张32 px×32 px像素的彩色图像，划分为5万张训练图像和1万张测试图像。SVHN是一个基于真实世界场景的10分类数字图像数据集，包含训练集、测试集和额外集。本文仅使用训练集中的73 257张和测试集中的26 032张带标签的数字图像进行实验。

2.2　神经网络后端循环次数的确定

循环次数的选取依据是基于模型动态响应轨迹的优化需求。循环次数过少可能导致特征提取不足，而过多则可能引起过拟合或信息冗余。为了确定最优循环次数，本文在2.1节所述实验设置的基础上，选择CIFAR-10数据集，分别在干净样本和对抗样本上训练了模型，并记录了不同循环次数下的准确率。结果见表1，其中“1-1-1”表示V2、V4和IT区域的内部结构各进行1次，其余循环次数表示的含义类似。结果显示，随着循环次数的增加，模型性能呈现出先升后降的趋势。当V2、V4和IT区域均循环3次时，模型性能达到最优。因此，后续实验均采用这一循环次数。

表1　V1-SRNN模型循环次数对样本准确率的影响分析

Table 1　Analysis of the effect of V1-SRNN model iteration number on sample accuracy

循环次数	样本准确率/%	
	干净样本	对抗样本
1-1-1	88.89	60.49
2-2-2	87.26	69.53
2-3-2	88.22	71.82
2-4-2	88.71	81.57
3-3-3	88.90	81.69
4-4-4	85.73	79.52

2.3　对抗攻击能力测试

（ⅰ）无对抗训练下的鲁棒性。在未进行对抗训练的情况下，针对各个数据集采用快速梯度符号法（fast gradient sign method，FGSM）[31]和投影梯度下降法（projected gradient descent，PGD）[32]两种典型对抗攻击方法，分别基于时域反向传播（backward pass through time，BPTT）和频域反向传播（backward pass through rate，BPTR）机制，对VGG16模型及本文提出的V1-SRNN模型进行对抗鲁棒性评估。其中，攻击参数设置如下：扰动因子ε为8/255，步长α为2/255，迭代次数steps为4。

无对抗训练下模型的准确率见表2。在3种数据集和4种攻击情况下,尽管本文模型在干净数据集上的准确率(ACC)始终不及原始VGG16模型,但在面对未知攻击时,本文模型的准确率普遍高于VGG16,显示出更好的稳定性。例如,在SVHN数据集上,针对基于脉冲频率的FGSM攻击,本文模型的准确率比VGG16提高了31.6%;在PGD攻击下,文本模型的准确率比VGG16提升了43.15%。

表2 无对抗训练条件下模型的准确率比较

Table 2 Accuracy comparison of models under non-adversarial training conditions

%

数据	模型	ACC	FGSM		PGD	
			BPTT	BPTR	BPTT	BPTR
SVHN	VGG16	95.95	37.68	35.02	12.43	20.14
	V1-SRNN	95.06	53.99	66.62	41.32	63.29
CIFAR10	VGG16	93.32	9.14	18.12	0.16	3.08
	V1-SRNN	88.90	34.24	40.23	16.89	30.84
CIFAR100	VGG16	72.22	8.76	6.57	0.62	0.87
	V1-SRNN	56.40	22.13	27.56	18.42	27.73

(ⅱ)对抗训练下的鲁棒性。评估了对抗训练下V1-SRNN模型的鲁棒性,结果见表3。在每次前向传播过程中,训练过程全部使用基于时序的FGSM方法生成对抗样本,扰动因子ε设置为2/255,测试条件与之前无对抗训练的测试条件相同。可以看出,对抗训练方法在各种攻击情况下均能提升模型的鲁棒性。例如,在CIFAR-10数据集上,在基于脉冲频率的FGSM攻击下,VGG16的准确率从18.12%提升至46.82%,而本文模型的准确率从40.23%提升至55.61%。但实验结果总体表明,本文模型在各种攻击条件下的表现始终优于原始VGG16模型。

值得注意的是,根据表2、表3中的数据分析可知,无论是采用无对抗训练还是对抗训练,V1-SRNN模型的基本准确率始终低于VGG16基准模型。例如,在SVHN,CIFAR10,CIFAR-100三种数据集上,无对抗训练下,V1-SRNN模型的基本准确率分别比基准模型的准确率低1.89%、4.42%和15.82%;而在对抗训练下,V1-SRNN模型的基本准确率分别比基准模型的准确率低2.3%、8.19%和8.18%。主要原因在于循环神经网络后端结构的设计以及脉冲神经网络时间窗口的重复,这增加了时间成本并变相加大了脉冲神经网络的延迟。由于长时间的延迟以及脉冲神经元的非线性特性,随着脉冲神经网络深度的增加,后续网络模块获得的重要输入特征相对减少。虽然这种特性能够对对抗样本的噪声起过滤作用,使得基于循环神经网络后端的V1-SRNN模型在各种对抗攻击下表现得比VGG16网络更好,但在干净样本上,重要输入特征的减少导致V1-SRNN模型的基本准确率始终低于基准模型。这表明,模型获得高鲁棒性的代价往往是模型在正常样本上的性能退化,这也是目前整个深度学习领域普遍存在的问题。因此,需要选择合适的方法或创新算法,以在对抗鲁棒性与基本准确率之间实现自主平衡,从而使模型在这两方面均能达到高性能表现。

表3 对抗训练条件下模型的准确率比较

Table 3 Accuracy comparison of models under adversarial training conditions

%

数据	模型	ACC	FGSM		PGD	
			BPTT	BPTR	BPTT	BPTR
SVHN	VGG16	95.15	48.13	53.66	48.65	55.11
	V1-SRNN	92.85	63.45	74.30	64.74	76.03
CIFAR10	VGG16	89.88	40.53	46.82	42.15	50.48
	V1-SRNN	81.69	48.10	55.61	49.50	58.93
CIFAR100	VGG16	61.41	22.06	24.27	23.28	27.21
	V1-SRNN	53.23	26.74	31.16	28.99	33.22

2.4 其他参数影响

在脉冲神经网络中,时间窗口T决定了网络收集和传递脉冲信息的时间长度。通过调整时间窗口的大小,可以检验模型对时间变化的鲁棒性。在SVHN和CIFAR-10数据集上对VGG16和本文V1-SRNN模型进行了无对抗攻击训练,并在FGSM(ε=8/255)攻击下测试了模型的对抗鲁棒性,结果见图2。

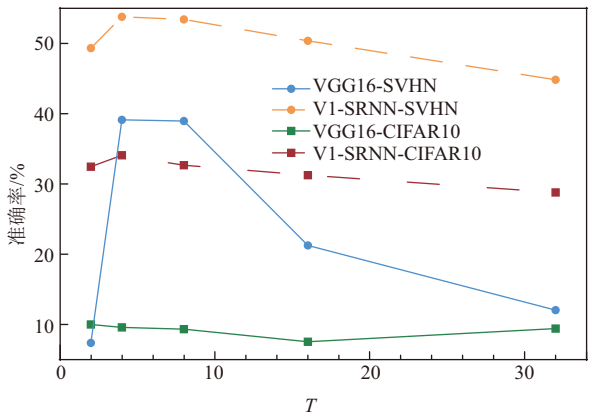

图2 时间窗口T对模型的影响

Fig. 2 Effect of the time window T on the model performance

如图 2 所示，在时间窗口较小时，VGG16 模型在两个数据集上的准确率均低于 10%，无法有效学习图像特征，而本文提出的 V1-SRNN 模型表现出较好的稳定性。随着时间窗口接近训练模型的设定值，两个模型的性能均有所提升，但 V1-SRNN 模型的表现始终优于 VGG16。然而，当时间窗口继续增大时，由于脉冲神经元的非线性特性，两个模型的稳定性均呈下降趋势，但对 VGG16 网络的影响更大，而 V1-SRNN 模型受影响较小。例如，在 SVHN 数据集上，当时间窗口 T 从 8 增大到 32 时，VGG16 模型的准确率下降了 26.92%，而 V1-SRNN 模型的准确率仅下降了 8.58%。

在对抗性攻击中，ε 值的变化反映了攻击的强度或扰动的大小。通过观察 ε 值的变化，可以了解模型对不同攻击强度的抵抗能力。为了验证模型在 ε 攻击增强时的性能变化，在 SVHN 和 CIFAR-10 数据集上对 VGG16 和 V1-SRNN 模型进行了无对抗攻击训练，并在 FGSM（ε=8/255）攻击下测试了它们在不同攻击强度下的对抗鲁棒性，结果见图 3。

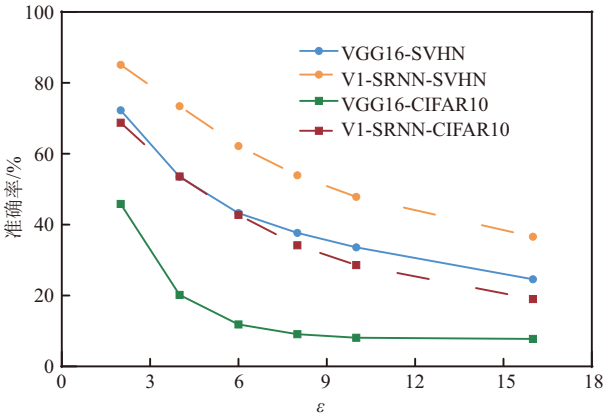

图 3　攻击参数 ε 对模型的影响

Fig. 3 Effect of the attack parameter ε on the model performance

如图 3 所示，随着 ε 值的增加，所有模型的准确率均呈下降趋势。尽管在 CIFAR-10 数据集上，V1-SRNN 模型与 VGG16 模型的准确率差距不如在 SVHN 数据集上显著，但 V1-SRNN 模型仍保持了较高的准确率。在相同的 ε 值下，V1-SRNN 模型的准确率普遍高于 VGG16 模型。

2.5　消融实验

为验证本文提出的两部分模块对原始脉冲神经网络鲁棒性的提升效果，在 CIFAR-10 数据集上设计了以下消融实验。表 4 展现了模型中某部分组件对模型整体鲁棒性的影响，其中“×”表示未包含该模块，“√”表示包含该模块。实验中，修改后的模型分别在干净数据集和对抗样本上进行训练，并采用基于时序的 FGSM 攻击（ε= 8/255）进行测试，记录模型准确率。实验结果显示，只有当两部分模块共同作用时，模型的准确率才会显著提高，展现出最佳的对抗鲁棒性。

表 4　不同模块对模型的影响

Table 4 Effect of different modules on the model performance

SpikeV1	循环后端	CLEAR	ATTACK
×	×	9.14	40.53
√	×	12.98	37.91
×	√	16.31	39.58
√	√	34.24	48.10

2.6　与现有工作的对比

为深入探究文中所提出的网络模型相较于现有先进模型的优越性，在 CIFAR-10 数据集上进行了无对抗训练和对抗训练。随后，使用基于 BPTT 的 FGSM 和 PGD 攻击，评估了各模型在不同攻击强度下的对抗鲁棒性，实验结果见表 5。

在不同攻击强度下，本文提出的模型性能均优于其他现有模型。例如，在面对 FGSM 攻击时，无对抗训练情况下，本文模型的准确率比 ResNet-SNN 模型高 8.92%；在对抗训练条件下，本文模型的准确率比 SNN-RAT 模型提升了 5.86%。这表明，模型对初级视觉皮层神经反应的拟合度越高，其在对抗性攻击中的鲁棒性越强。

表 5　V1-SRNN 模型与最新 SNN 模型的对抗鲁棒性比较

Table 5 Comparison of adversarial robustness between V1-SRNN model and the state-of-the-art SNN models

Model	FGSM		PGD	
	CLEAR	ATTACK	CLEAR	ATTACK
VGG[33]	6.49	29.86	1.32	35.46
ResNet[34]	25.32	42.09	5.80	46.05
Wide ResNet[35]	10.50	39.38	0.10	42.80
SNN-RAT[11]	15.52	42.24	6.08	42.86
V1-SRNN	34.24	48.10	16.89	49.50

2.7　模型可视化

为观察模型在决策过程中所关注的区域，使用 Grad-CAM[36] 技术对模型决策过程进行可视化解

释。随机选择2张CIFAR-10数据集的图片和1张CIFAR-100数据集的图片,使用无对抗训练得到的模型在基于BPTR的FGSM和PGD攻击下进行可视化展示。实验结果见图4,其中"VGG16/FGSM"表示该图是基于VGG16模型在FGSM对抗攻击下的可视化效果展示。

由图4可以看出,在相同的对抗攻击条件下,V1-SRNN模型相较于VGG16模型能够更稳定地关注核心目标的核心区域。然而,随着对抗扰动强度的增加,两种模型的热力图显著性区域均呈减弱趋势,表明模型的对抗鲁棒性随着攻击强度的增大而逐渐降低。这一现象与2.4节中关于ε参数的分析结论一致。尽管如此,V1-SRNN模型在对抗鲁棒性方面仍表现出优于VGG16模型的特性。

进一步分析热力图的显著性分布变化,从CIFAR-10数据集到CIFAR-100数据集的实验结果表明,随着数据集复杂度的增加,两种模型的对抗鲁棒性均有所下降。这一趋势与表2中展示的分类准确率下降趋势一致,表明模型在面对更复杂的任务时,其对抗鲁棒性会受到显著影响。

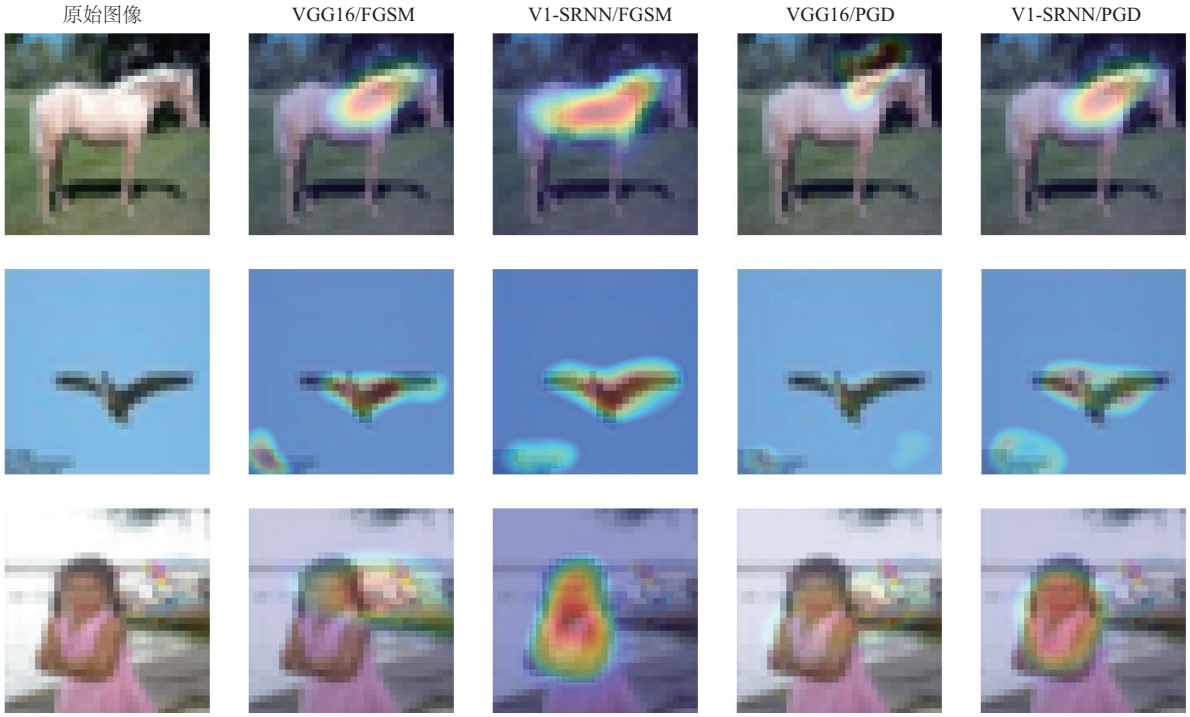

图4 基于无对抗训练模型决策的可视化

Fig. 4 Visualization of decision-making based on the non-adversarially trained model

3 结论与展望

本文针对脉冲神经网络(SNN)在对抗攻击下鲁棒性不足的问题,通过融合生物视觉通路特性和内部循环机制,提出了一种新型的生物视觉启发的高鲁棒性脉冲循环神经网络模型(V1-SRNN)。该模型在CIFAR-10,CIFAR-100和SVHN数据集上进行了训练和测试。实验结果表明,相较于传统SNN模型,V1-SRNN模型在面对对抗攻击时,准确率显著提升,展现出更强的对抗鲁棒性和抗干扰能力。本文为SNN在图像识别领域的应用提供了新思路,并为提升其鲁棒性研究提供了参考。

未来的研究可进一步探讨复杂的生物视觉通路,特别是视网膜和外侧膝状体(LGN)结构的特性,并将其融入SNN模型构建中,以研究其对脉冲神经网络性能的影响。同时,值得注意的是,尽管V1-SRNN在对抗样本中表现出较强的鲁棒性,但其本身准确率始终低于基线模型。因此,未来工作也可将受生物约束的卷积脉冲特征提取模块有效融合到现有SNN的先进结构中,以解决模型本身准确率不足的问题。

参考文献:

[1] ROY K, JAISWAL A, PANDA P. Towards spike-based machine intelligence with neuromorphic computing[J]. Nature, 2019, 575(7784): 607-617.

[2] CHEN Guangyao, PENG Peixi, LI Guoqi, et al. Training full spike neural networks via auxiliary accumulation pathway[EB/OL]. (2023-01-26)[2025-03-16]. https://doi.org/10.48550/arXiv.2301.11929.

[3] ZHU Zulun, PENG Jiaying, LI Jintang, et al. Spiking

graph convolutional networks[EB/OL]. (2022-05-05)[2025-03-16]. https://doi.org/10.48550/arXiv.2205.02767.

[4] NA B, MOK J, PARK S, et al. AutoSNN: Towards energy-efficient spiking neural networks [EB/OL]. (2022-01-30)[2025-03-16]. https://doi.org/10.48550/arXiv.2201.12738.

[5] CHENG Xiang, HAO Yunze, XU Jiaming, et al. LISNN: Improving spiking neural networks with lateral interactions for robust object recognition[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama, Japan. International Joint Conferences on Artificial Intelligence Organization, 2020: 1519-1525.

[6] SHARMIN S, RATHI N, PANDA P, et al. Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations[C]//Computer Vision-ECCV 2020, 16th European Conference. Glasgow, UK:[s.n.],2020:399-414.

[7] EL-ALLAMI R, MARCHISIO A, SHAFIQUE M, et al. Securing deep spiking neural networks against adversarial attacks through inherent structural parameters[C]//Design, Automation & Test in Europe Conference & Exhibition (DATE). Grenoble, France: IEEE, 2021: 774-779.

[8] KUNDU S, PEDRAM M, BEEREL P A. HIRE-SNN: Harnessing the inherent robustness of energy-efficient deep spiking neural networks by training with crafted input noise[C]//IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada:IEEE, 2021: 5189-5198.

[9] KIM Y, PARK H, MOITRA A, et al. Rate coding or direct coding: Which one is better for accurate, robust, and energy-efficient spiking neural networks? [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 71-75.

[10] CHEN Weiran, SUN Qi, XU Qi. Defending spiking neural networks against adversarial attacks through image purification[EB/OL]. (2024-04-26)[2025-03-16]. https://doi.org/10.1109/ICASSP49660.2025.10888581.

[11] DING Jianhao, BU Tong, YU Zhaofei, et al. SNNRAT: Robustness-enhanced spiking neural network through regularized adversarial training[EB/OL]. (2023-06-19)[2025-03-16]. https://www.cnblogs.com/lucifer1997/p/17491592.html.

[12] BAI Tao, LUO Jinqi, ZHAO Jun, et al. Recent advances in adversarial training for adversarial robustnes[EB/OL]. (2021-02-02)[2025-03-16]. https://arxiv.org/pdf/2102.01356.pdf.

[13] TIAN Yang, SUN Pei. Percolation may explain efficiency, robustness, and economy of the brain[J]. Network Neuroscience, 2022, 6(3): 765-790.

[14] CHEN Guozhang, SCHERR F, MAASS W. A data-based large-scale model for primary visual cortex enables brain-like robust and versatile visual processing[J]. Science Advances, 2022, 8(44): eabq7592.

[15] DAPELLO J, MARQUES T, SCHRIMPF M, et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations[J]. Advances in Neural Information Processing Systems, 2020, 33: 13073-13087.

[16] REDDY M V, BANBURSKI A, PANT N, et al. Biologically inspired mechanisms for adversarial robustnes[J]. Advances in neural information processing systems, 2020, 33: 2135-2146.

[17] 孟媛,汪西原.一种特征融合的双流深度检测伪造人脸方法[J].宁夏大学学报(自然科学版),2024,45(3):99-306.

[18] 姚宗亮,黄荣,董爱华,等.基于多模态融合和自适应剪枝Transformer的脑肿瘤图像分割算法[J].宁夏大学学报(自然科学版),2024,45(1):16-24.

[19] KAR K, KUBILIUS J, SCHMIDT K, et al. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior[J]. Nature Neuroscience, 2019, 22(6): 974-983.

[20] 肖云发,韩芳,王青云.基于脉冲强化学习和CPG的四足机器人分层运动控制[J/OL].控制与决策,2025,40(7):2070-2078. DOI:10.13195/j.kzyjc.2024.1029.

[21] KONAR D, DAS SARMA A, BHANDARY S, et al. A shallow hybrid classical - quantum spiking feedforward neural network for noise-robust image classification[J]. Applied Soft Computing, 2023, 136: 110099. DOI: 10.1016/j.asoc.2023.110099.

[22] KRIZHEVSKY A. Learning multiple layers of features from tiny images[EB/OL]. (2009-04-08)[2025-03-16]. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[23] NETZER Y, WANG Tao, COATES A, et al. Reading digits in natural images with unsupervised feature learning[C]//Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning. Granada, Spain: Neural Information Processing Systems Foundation, 2011: 5-16.

[24] HODGKIN A L, HUXLEY A F. Propagation of electrical signals along giant nerve fibers[J]. Proceedings of the Royal Society of London Series B:Biological Sciences, 1952, 140(899): 177-183.

[25] KABILAN R, MUTHUKUMARAN N. A neuromorphic model for image recognition using SNN[C]// 6th International Conference on Inventive Computation Technologies (ICICT). Coimbatore, India: IEEE, 2021: 720-725.

[26] IZHIKEVICH E M. Simple model of spiking neurons [J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1569-1572.
[27] DE VALOIS R L, WILLIAM YUND E, HEPLER N. The orientation and direction selectivity of cells in macaque visual cortex[J]. Vision Research, 1982, 22(5): 531-544.
[28] DE VALOIS R L, ALBRECHT D G, THORELL L G. Spatial frequency selectivity of cells in macaque visual cortex [J]. Vision Research, 1982, 22(5): 545-559.
[29] RINGACH D L. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex[J]. Journal of Neurophysiology, 2002, 88(1): 455-463.
[30] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.
[31] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and harnessing adversarial examples [EB/OL]. (2014-12-20)[2025-03-16]. https://doi.org/10.48550/arXiv.1412.6572.
[32] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. (2019-09-04) [2025-03-16]. https://arxiv.org/pdf/1706.06083.pdf.
[33] SENGUPTA A, YE Yuting, WANG R, et al. Going deeper in spiking neural networks: VGG and residual architectures[J]. Frontiers in Neuroscience, 2019,13:95. DOI: 10.3389/fnins.2019.00095.
[34] HU Yangfang, TANG Huajin, PAN Gang. Spiking deep residual networks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 5200-5205.
[35] ZAGORUYKO S, KOMODAKIS N. Wide residual networks [EB/OL]. (2017-06-14) [2025-03-16]. https://arxiv.org/pdf/1605.07146.
[36] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization [J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

A Highly Robust Spiking Recurrent Neural Network Model Inspired by Biological Vision

CHEN Linguo[1], *HUANG Rong*[1,2], *HAN Fang*[1,3*]
(1. College of Information Science and Technology, Donghua University, Shanghai 201620, China;
2. Engineering Research Center of Digitized Textile and Apparel Technology, Ministry of Education, Donghua University, Shanghai 201620, China;
3. School of Mathematics and Statistics, Ningxia University, Yinchuan 750021, China)

Abstract: To address the low robustness of Spiking Neural Networks (SNN) against adversarial attacks, a highly robust Spiking Recurrent Neural Network model inspired by biological vision was proposed. This model incorporates the biological mechanisms of the primary visual cortex (V1) and features a convolutional SNN front end designed with biological constraints. Additionally, by integrating feedback connections from the cortical visual information, an SNN back end with an internal recurrent mechanism was constructed. In the absence of adversarial training, this model demonstrates significant improvements in adversarial accuracy of 31.6%, 22.11%, and 20.99% on the SVHN, CIFAR10, and CIFAR100 datasets, respectively. With adversarial training, the adversarial accuracy improves by 20.64%, 8.79%, and 6.89%, respectively. Furthermore, as the perturbation factor (ε) and the time window (T) increase, the accuracy of this model consistently surpasses that of the baseline model. Experimental results show that the Spiking Recurrent Neural Network model, which incorporates biological vision mechanisms, shows significantly improved accuracy when faced with adversarial attacks, demonstrating enhanced adversarial robustness.

Key words: spiking neural networks; robustness; adversarial attacks; biological vision

(责任编辑 张 娣)