摘要

脉冲神经网络作为第三代神经网络，因其独特的脉冲表示、低能耗优势以及时序特性，受到了广泛研究关注。与传统的深度学习方法相比，脉冲神经网络提供了一种更加接近生物信息处理机制的计算方式，为人工智能开辟了新的道路。目前，以深度学习为代表的人工智能已在计算机视觉等领域取得了较大进展，但仍面临鲁棒性挑战，在输入受到扰动时会导致错误输出。本文旨在研究脉冲神经网络如何提高鲁棒性，以应对深度学习的鲁棒性不足的问题。

脉冲神经网络鲁棒性研究主要有两方面动机。首先，深度神经网络易受对抗攻击，威胁无人驾驶、工业制造等安全关键型应用。脉冲神经网络依靠异步脉冲传递信息，已有研究表明其对小范围随机扰动不敏感，但主流端到端近似梯度训练方法仍可能使其面临攻击风险。因此，提高脉冲神经网络的鲁棒性对构建可靠神经形态系统至关重要。其次，从应用角度看，神经形态相机在视觉任务中易受噪声影响，会导致进行后端处理的脉冲神经网络在目标识别等任务中的性能降低。因此，提升脉冲神经网络在复杂噪声环境下的鲁棒性，有助于提高神经形态计算系统的实用性。

脉冲神经网络鲁棒性增强研究存在一些亟待解决的关键问题。首先，深度脉冲神经网络中的神经元面临脉冲扰动逐层放大的问题，无法直接使用连续激活的传统人工神经网络的扰动分析方法，这增大了脉冲扰动调控的难度。其次，脉冲神经网络内部膜电位和脉冲状态在扰动下会偏离未受扰动时的正常计算轨迹，导致输出脉冲模式异常，影响网络的整体决策能力。最后，脉冲神经网络主要依赖频率编码，其性能趋同于传统人工神经网络，而扰动的并发性会影响网络的泛化性能。目前，尚无系统性研究分析时间编码对脉冲神经网络鲁棒性的影响。

为解决上述关键问题，本文将从脉冲神经网络的特点出发，分别从神经元、网络及其学习方法逐步探究脉冲神经网络鲁棒性的机理，并设计有效的模型或方法来增强其鲁棒性。本文的主要创新点包括以下几个方面：

第一，针对脉冲神经元扰动逐层放大问题，提出了脉冲扰动抑制的神经元模型，解决了脉冲神经元扰动放大调控的难题。本文结合理论分析，探究了脉冲神经网络在突触连接层面的扰动容忍能力，并导出了针对离散脉冲表示的 Lipschitz 常数。在此基础上，本文提出了一种权重的正则化方法。此外，受到精细生物神经元的启发，本文还提出了带有随机门控的脉冲神经元模型，有效限制了逐层传播中的脉冲扰动放大。所提出脉冲扰动抑制神经元构成的网络在 CIFAR10 和 CIFAR100 数据集上实现了 70.63% 和 55.95% 的扰动后分类准确率，甚至能使较强的投影梯度下降的黑盒攻击几乎失效。

第二，针对受扰网络动力学输出异变问题，提出了膜电位扰动自适应网络，有效克

服了脉冲神经网络内部状态非稳态带来的随时间输出偏移的挑战。本文从非线性系统稳定性的角度出发，揭示了网络鲁棒性与膜电位随时间变化之间的关系。基于膜电位时间扰动统计量能够可靠描述扰动强度的观察，设计了辅助损失函数约束方法，以确保扰动后动力学满足网络输入-输出稳定性的要求。在网络训练过程中，可以自适应地调节逐步膜时间常数，以实现更强的输入-输出稳定能力。该网络在 CIFAR10 和 CIFAR100 数据集上进行高斯噪声训练和对抗训练后，能在现有脉冲网络鲁棒性增强算法中取得最佳抗扰动性能。

第三，针对扰动并发时脉冲网络泛化性能降低的问题，提出了融合脉冲时间编码的鲁棒学习方法，并论证了脉冲时间可显著提高脉冲神经网络的抗扰动能力，解决了异步脉冲网络与传统网络在扰动性能上趋同的矛盾。本文利用脉冲神经网络的时空异步特性，设计了基于脉冲同步的时间编码方法，确保任务关键信息在时间编码中优先表示，并采用首次脉冲时间解码方法，以减少时间上后续扰动对网络任务性能的影响。借助脉冲网络时序训练算法的泛化能力，进一步提升了网络的抗干扰能力。为实现网络在自然数据泛化能力与扰动输入鲁棒性之间的平衡，本文还提出了多路融合编码方法。在 CIFAR10 数据集上的实验验证了该鲁棒学习方法的有效性，采用上述方法训练的网络在处理扰动数据时的分类准确率相比同结构传统人工神经网络提高了约一倍。

第四，本文整合了所述鲁棒性增强技术的核心，设计并开发了基于脉冲相机的鲁棒性增强系统。该系统在 SpiReco 数据集上进行训练，并在高斯噪声和对抗攻击下测试其识别性能。实验结果表明，设计的正则自适应训练方法能够有效提高脉冲神经网络在数据集上的鲁棒性。最后，开发了识别结果展示界面，有效展示了脉冲神经网络在脉冲流识别任务中的表现。

综上所述，针对脉冲神经网络鲁棒性挑战，本文设计了多层次的模型与方法以增强其抗扰动能力，并通过神经形态数据集进行了系统验证。本文研究为未来在真实场景中构建脉冲神经网络系统奠定了基础。此外，本文研究也将促进脉冲神经网络在安全关键型应用中的推广，并为理解大脑如何在噪声条件下实现高鲁棒性提供了新的视角。

关键词：脉冲神经网络，神经编码，鲁棒性，神经形态计算，对抗攻击

Research on Robustness Enhancement for Spiking Neural Networks

Jianhao Ding (Computer Application Technology)
Supervised by Prof. Tiejun Huang

ABSTRACT

As the third generation of neural networks, spiking neural networks (SNNs) have attracted extensive research attention due to their unique spike representation, low energy consumption advantages and timing characteristics. Compared with traditional deep learning methods, SNNs provide a computing method that is closer to the biological information processing mechanism, and lead to a new path for the development of artificial intelligence (AI). At present, AI represented by deep learning has made great progress in fields such as computer vision, but it still faces robustness challenges, which will lead to erroneous output when the input is disturbed. This thesis aims to study how to improve the robustness of SNNs address the problem of insufficient robustness of deep learning.

There are two main motivations for studying the robustness of SNNs. First, deep traditional artificial neural networks (ANNs) are vulnerable to adversarial attacks, threatening safety-critical applications such as autonomous driving and industrial manufacturing. SNNs rely on asynchronous spikes to transmit information. Studies have shown that they are insensitive to small crafted perturbations, but mainstream end-to-end approximate gradient training strategies may still expose them to attack risks. Therefore, improving the robustness of SNNs is crucial to building reliable neuromorphic systems. Secondly, from an application perspective, neuromorphic cameras are susceptible to noise in visual tasks, which will reduce the performance of SNNs in tasks such as object recognition. Improving the robustness of SNNs in complex noise environments will help improve the practicality of neuromorphic computing systems.

There are some key issues that need to be addressed in the study of robustness enhancement of SNNs. First, neurons in deep SNNs face the problem of layer-by-layer amplification of spike perturbations, and cannot directly use the perturbation analysis method of traditional ANNs with continuous activation, which increases the difficulty of regulating spike perturbations. Second, the internal membrane potential and spike state of SNNs will deviate from the

normal trajectory when perturbed under perturbations, resulting in abnormal output spike patterns and affecting the overall decision-making ability of the network. Finally, SNNs mainly rely on rate coding, and their performance reduces to traditional ANNs. The concurrency of perturbations will affect the generalization performance of the network. At present, there is no systematic study analyzing the impact of time coding on the robustness of SNNs.

To solve the above key problems, this thesis starts from analyzing the characteristics of SNNs and then explores the robustness mechanism of SNNs from neurons and networks to their learning methods and designs effective models or methods to enhance their robustness. The main innovations of this thesis include the following aspects:

First, in view of the problem of layer-by-layer amplification of spike perturbations, a spike perturbation suppression neuron model is proposed, which solves the problem of spike neuron perturbation amplification regulation. Combined with theoretical analysis, this thesis explores the perturbation tolerance of SNNs at the synaptic connection level and derives the Lipschitz constant for spike representation. Based on this theoretical result, this thesis proposes a regularized training scheme for SNN. In addition, inspired by fine biological neurons, this thesis proposes a spike neuron model with stochastic gating, which effectively limits the amplification of spike perturbations in layer-by-layer propagation. This method achieves 70.63% and 55.95% post-perturbation classification accuracy on CIFAR10 and CIFAR100 datasets and can even make the black-box attack of projected gradient descent attacks almost ineffective.

Secondly, in view of the problem of abnormal output of perturbed network dynamics, a membrane potential perturbation adaptive network is proposed, which effectively overcomes the challenge of output deviation over time caused by the non-steady internal state of SNNs. From the perspective of nonlinear system stability, this thesis reveals the relationship between network robustness and membrane potential change over time. Based on the observation that the statistics of the membrane potential perturbation dynamics can reliably describe the perturbation intensity, an auxiliary loss function is designed to ensure that the post-perturbation dynamics meet the requirements of network input-output stability. During network training, the membrane time constant can be adaptively adjusted to achieve stronger input-output stability. The network structure of this method performs well in image classification tasks after Gaussian noise training and adversarial training on CIFAR10 and CIFAR100 datasets, and can achieve the best post-perturbation performance among the existing SNN robustness improvement algorithms.

Third, in view of the problem of reduced generalization performance of SNNs when

perturbations are concurrent, a robust learning method integrating spike time coding is proposed, which demonstrates that spike time has the ability to significantly improve the invulnerable ability of SNNs, solving the contradiction between the reduction of asynchronous SNNs and traditional ANNs in perturbation performance. This thesis uses the spatiotemporal asynchronous characteristics of SNNs to design a time coding scheme based on spike synchronization, ensuring that task-critical information is represented first in the time coding, and adopts the first spike time decoding strategy to reduce the impact of subsequent perturbations in time on the network task performance. With the generalization ability of the SNN timing-based training algorithm, the invulnerable ability of the system is further improved. In order to achieve a balance between the network's generalization ability in natural data and the robustness of perturbed input, this thesis also proposes a multi-channel fusion coding strategy. The robust learning method verifies its effectiveness on the CIFAR10 dataset. The classification accuracy of the network trained using the above strategy is increased by about two times when processing perturbed data, which lays a good foundation for the development of robust intelligent systems based on spike neural networks.

Fourth, this thesis integrates the core of the network robustness enhancement technology and designs a robustness enhancement system based on spike cameras. The system is trained on the SpiReco dataset, and its recognition performance is tested under Gaussian noise and adversarial attacks. Experimental results show that the regularized adaptive training method can effectively improve the robustness of the spiking neural network on the dataset. Finally, a recognition result display system is constructed to effectively demonstrate the performance of the spiking neural network in the spike stream recognition task.

In summary, in response to the robustness challenge of spiking neural networks, this thesis designed a multi-level solution to enhance the robustness of SNN and systematically verified it through a neuromorphic dataset. This thesis lays the foundation for building SNN systems in real scenarios in the future. In addition, this thesis will also promote the promotion of SNNs in safety-critical applications and provide a new perspective for understanding how the brain achieves high robustness under noisy conditions.

KEY WORDS: Spiking neural network, Neural coding, Robustness, Neuromorphic computing, Adversarial attack