摘要

借鉴和模仿大脑计算原理而提出的脉冲神经网络 (Spiking Neural Network, SNN)，被誉为第三代神经网络模型，拥有生物神经元积分发放、阈值触发、稀疏激活的特性。与传统人工神经网络 (Artificial Neural Network, ANN) 相比，SNN 凭借事件驱动的计算机制，在神经形态计算芯片上运行时具有显著的能耗优势，有潜力实现低功耗类脑人工智能。但 SNN 离散不可导的脉冲发放过程和复杂的时空域前向传播使得设计高性能学习算法充满挑战。传统的学习算法大多只能处理单层或单脉冲网络，至多解决 MNIST 分类这样的简单问题，在大规模数据集上尚未取得令人满意的性能。随着深度学习方法的引入，SNN 的性能大幅度提升，但仍然落后于 ANN，这一性能差距限制了 SNN 的实际应用和发展前景。本文针对深度 SNN 学习算法进行研究，主要贡献如下：

第一，针对大多数学习算法仅训练突触权重而忽略神经动态的学习和神经元的异质性问题，提出了突触权重和神经动态的联合学习算法。首先，该方法分析了 Leaky Integrate-and-Fire (LIF) 神经元中膜时间常数的作用，阐述其对性能的重要影响。其次，对脉冲神经元进行了统一的建模，得到离散时间 LIF 神经元模型。最后，使用限幅函数将膜时间常数参数化，通过训练过程中的自动优化，实现突触权重和神经动态的联合学习。实验结果表明，该方法在常用的静态图像和神经形态数据集分类任务上，正确率较主流方法最高提升 14%，同时仿真步数减少 50% 以上，并具备更好的参数鲁棒性和更快的收敛速度。

第二，针对将残差连接引入 SNN 得到的 Spiking ResNet 性能退化问题，提出 Spike-Element-Wise (SEW) ResNet，实现残差学习。该方法分析指出 Spiking ResNet 难以实现恒等变换、容易引发梯度消失或爆炸，并提出 SEW 残差连接，允许前反向传播完全跳过脉冲神经元层以避免上述问题。实验结果表明，SEW ResNet 网络越深性能越高；首次训练出超过 100 层、最高 152 层的深度 SNN；梯度实验数据与理论分析一致；在 ImageNet 数据集上较国际前沿方法性能提升 4%；在神经形态数据集分类任务上，使用主流模型十分之一的参数量且达到相当性能。

第三，针对现有深度 SNN 逐步串行传播导致计算效率低、训练耗时长的问题，提出无状态突触层和有状态神经元层的并行加速算法。对于无状态突触层，以批量并行代替时间串行，使时域计算完全并行化，且计算结果和精度不受影响。对于有状态神经元层，提出并行脉冲神经元模型 (Parallel Spiking Neuron, PSN)，将传统神经元不发放脉冲时的非迭代求和式推广，以支持并行加速的矩阵乘法生成膜电位，替换只能串行的逐步计算，同时将传统脉冲神经元输入和隐状态的间接关系（马尔可夫链）改为

跨时间步直连，有助于学习长期依赖。实验结果表明，该方法对于无状态层，最高可达3倍加速；对于有状态层，PSN在计算速度上最高提升30倍，记忆能力和分类任务正确率均高于传统串行神经元；对于整个网络，以ResNet-18结构为例，该方法对训练速度最高可提升4倍。

第四，针对脉冲深度学习缺乏软件框架的问题，提出了国际上首批SNN深度学习框架之一的SpikingJelly（惊蜇）框架。新兴的脉冲深度学习领域缺乏软件框架，研究者需要从零开始搭建模块，费时费力，效率低下。为解决这一问题，本文提出了SpikingJelly框架，提供了从神经形态数据集预处理、网络构建、模型训练、数据采集、权重量化到芯片部署的脉冲深度学习全栈式解决方案，并针对SNN的特性利用计算图优化和融合操作进行加速，与其他框架相比最高可达11倍的计算速度优势。目前有大量研究使用SpikingJelly框架进行实验，应用范围既包括脉冲深度学习，也有物理、材料、电化学、生物医学等多个领域的前沿交叉研究。《Nature Computational Science》对SpikingJelly框架进行专文报道，指出其有望成为协调脉冲深度学习发展的生态系统。

综上所述，本文针对深度SNN的部分关键学习算法进行研究，从神经元、网络结构、计算效率和软件系统层次分别入手，将神经动态纳入学习范畴以提升网络性能，改进网络结构实现残差学习以构建大规模深度SNN，对突触和神经元并行化以提升计算效率，开发SpikingJelly框架以提供全栈式脉冲深度学习解决方案。本文的工作进一步完善了神经形态计算的生态系统，为发展高性能类脑人工智能奠定了算法基础。

关键词：脉冲神经网络，梯度替代法，循环神经网络，神经形态计算，深度学习框架

Research on Learning Algorithms in Deep Spiking Neural Networks

Wei Fang (Computer Applied Technology)
Directed by Prof. Yonghong Tian

ABSTRACT

Inspired by the brain, Spiking Neural Networks (SNNs) are proposed and regarded as the third generation of neural network models. SNNs are similar to the biological neurons that use integrate-and-fire, threshold-trigger, and sparse-activate mechanisms. Due to the event-driven computation, SNNs have extremely high energy efficiency than Artificial Neural Networks (ANNs) when deployed in neuromorphic chips and have the potential to implement efficient brain-like artificial intelligence. However, due to the non-differentiable firing mechanism and complex spatial-temporal propagation, it remains challenging to formulate efficient and high-performance learning algorithms for SNNs. Most traditional methods can only be applied to single-layer or single-spike SNNs, and the trained SNNs can not solve tasks that are more difficult than MNIST classification. Recently, deep learning methods have been introduced and improved the performance of SNNs greatly. Nevertheless, the task accuracy of SNNs is still lower than that of ANNs, which constricts the applications of SNNs. This thesis focuses on the learning algorithms of deep SNNs, and the contributions are as follows:

Firstly, this thesis proposes a joint learning method for both synapse weights and neural dynamics to solve the issue of most existing algorithms that only train synapses while ignoring the learning of neuronal dynamics and the heterogeneity of neurons. The proposed method first analyzes the effect of the membrane constant of the Leaky Integrate-and-Fire (LIF) neuron and points out its influence on the SNN. Then a general discrete-time spiking neuron model is established and the LIF neuron model is derived. A clamp function is adopted to parameterize the membrane constant of the LIF neuron and enables it to be learnable during training, which implements the joint learning of synapse weights and neural dynamics. The experimental results show the proposed method improves up to 14% accuracy in both static and neuromorphic datasets than traditional methods, with 50% fewer time-steps, better parameter robustness, and faster convergence speed.

Secondly, this thesis proposes the Spike-Element-Wise (SEW) ResNet and achieves the residual learning in SNNs to solve the degeneration of Spiking ResNet. This method finds that Spiking ResNet is hard to achieve the identify mapping and easy to cause the gradient vanishing/exploding problem. Then the residual structure is improved by the proposed SEW ResNet. This method successfully trains the first SNN with more than 100 layers, up t to 150 layers, and achieves the highest accuracy on the ImageNet dataset among surrogate gradient methods with up to 4% accuracy improvement. The experimental results validate that SEW ResNet solves the degeneration problem. The gradient data recorded in the ResNet-152 structure are also consistent with the analysis. Additionally, SEW ResNet achieves close accuracy of other methods by using 10× less parameters.

Thirdly, this thesis proposes parallel acceleration methods to solve the issue that the SNNs are simulated in serial, which causes low computation efficiency and long training times. The proposed method designs acceleration methods for stateless synapses and stateful spiking neurons in SNNs, respectively. For the stateless synapses, the proposed method merges the time-step and the batch dimension, which fully parallelizes the computation over time-steps. Note that this method only changes the computation order and will not cause any accuracy drop. For the stateful spiking neurons, the Parallel Spiking Neuron (PSN) is proposed, which is generalized from the non-iterative formulation of neural dynamics of the traditional spiking neurons without resetting. The PSN replaces the serial computation of membrane potentials by the parallelizable matrix-matrix multiplication. This replacement also changes the generation of hidden states from by indirect Markov chain to direct connections, which enhances the learning ability of long-term dependency. The experiment results show that the proposed method accelerates the stateless synapses up to 3× and the stateful neurons up to 30×. Meanwhile, the PSN achieves higher accuracy than traditional spiking neurons in both memory ability and task accuracy. For the whole SNN, e.g., the Spiking ResNet-18 structure, the proposed method accelerates the training up to 4×.

Fourth, this thesis proposes one of the first spiking deep learning frameworks, SpikingJelly, to solve the issue that there is no mature framework available for deep SNNs. Researchers who want to combine advanced deep learning methods with SNNs have to build basic spiking neurons and synapses from scratch, resulting in repetitive and uncoordinated efforts. To solve the above issues and promote research on spiking deep learning, we present SpikingJelly, an open-source deep learning framework, to bridge deep learning and SNNs. SpikingJelly provides a full-stack solution for spiking deep learning, including neuromorphic dataset

processing, network building, model training, data recording, weight quantizing, and network deployment. SpikingJelly accelerates the simulation of SNNs by computation graph optimization and operation fusion, achieving up to 11× acceleration than other frameworks. SpikingJelly is employed by many researches, the topics of which involve both spiking deep learning and frontier interdisciplines including physics, materials, electrochemistry, and biomedicine, indicating that SpikingJelly extends the boundary of neuromorphic computing. *Nature Computational Science* reports that "SpikingJelly has the potential to serve as an ecosystem for the coordinated development of spiking deep learning".

In conclusion, this thesis studies some learning algorithms of deep SNNs and proposes the solutions in aspect of neurons, networks, computations, and software systems. More specifically, the proposed methods include the joint learning of synapses and neurons to promote accuracy, the SEW ResNet structure to achieve the residual learning for training large-scale deep SNNs, parallelize the synapses and neurons to accelerate training, and open-sources SpikingJelly to provide full-stack solution for spiking deep learning. This thesis enriches the ecology of neuromorphic computing and lays a foundation for developing high-performance brain-like intelligence.

KEYWORDS: Spiking Neural Networks, Surrogate Gradient, Recurrent Neural Networks, Neuromorphic Computing, Deep Learning Frameworks