摘要

人工智能正在渗透进生产生活的方方面面，未来，人们将离不开大量模型“计算”的结果。脉冲神经网络（Spiking Neural Networks, SNN）以模拟生物神经元的二值脉冲发放特性而节省计算开销闻名，可在神经形态硬件上获得高效部署，有望服务于边缘计算等对高能效比有要求的场景。而网络剪枝能够在维持网络表征能力的情况下移除网络中的冗余连接，从另一个角度提升计算效率。这两者的结合也因此成为研究热点。然而作为神经科学启发的计算模型，SNN剪枝领域的研究却通常忽略了作为高级智能代表的人类大脑，具有着远比当前大多数神经网络更稀疏的权重（突触连接）。事实上，神经系统中广泛存在的**突触修剪（Synaptic Pruning）**现象与深度SNN中的权重剪枝均以去除网络连接中的冗余连接为目的，因此是对应的一体两面。两者的最小修剪单元分别为树突上的微小结构——树突棘，与突触连接权重。因此，树突棘的生长消亡恰好对应了突触连接的产生与消亡现象。基于1）树突棘生长消亡过程中的结构特性；以及2）突触修剪机制的启发，来开发深度脉冲神经网络的剪枝算法，则一起为这一领域提供了全新的研究视角。

本文从生理结构建模、生理机制建模两个层面进行了研究，旨在找到与剪枝最相关的生理结构与机制信息，并最终建模为脉冲神经网络剪枝算法。随着研究的深入，本文渐进地解决了树突棘生长消亡过程中的**树突棘与权重关联建模问题、连接结构重布线机制建模问题、权重符号灵活性欠佳问题、树突棘消亡（权重修剪）速度控制问题**，在深度SNN上验证了解决方案的有效性，并最终在基于Vidar脉冲相机的脉冲目标检测系统中进行了应用。本文的主要创新成果如下：

第一，针对剪枝中连接存在与否的建模问题，提出了一种基于梯度重布线机制的脉冲神经网络剪枝，借鉴树突棘不同形态功能差异的突触权重建模与相应的剪枝算法，增强了剪枝过程中出现的连续的网络连接拓扑变换能力。本方法通过引入一个描述树突棘尺寸的参数，对权重的剪除状态与连接状态进行连续建模，将权重学习与结构学习统一为树突棘参数的学习。使得网络在训练过程中能够自动调节树突棘的生长与消亡。这一方法在不显著降低脉冲神经网络中脉冲传输效率的情况下，极大地降低SNN在极高稀疏度（>99%）下的性能损失到仅3.5%，为发表时深度SNN上剪枝的最佳结果。

第二，针对剪枝中权重符号缺乏灵活性的问题，提出了一种模拟双神经支配棘动态符号特性的脉冲神经网络剪枝算法，通过建模树突棘上受体比例变化的突触权重建模方式，扩大了网络的假设空间，并提升了剪枝后网络的性能。本方法在先前方法的

基础上，增加了树突棘在兴奋-抑制之间转换的建模，使权重的符号可学习，并引入了变化的阈值来控制网络稀疏化的速度。在18层SNN上的结果表明，其效果进一步超越了前述工作，并成为首个在该层数（含以上）深度的SNN有效剪枝的工作，并且也是首个在大型基准数据集上进行的SNN剪枝工作。

第三，针对剪枝中普遍存在的稀疏速度控制问题，提出了一种精细控制树突棘消亡速度的剪枝算法统一框架，避免了剪枝速度异常导致的性能不理想。本方法通过理论分析给出了树突棘状态变化建模剪枝这一背景下，树突棘参数的消亡速率，即突触修剪速度与对应的最优化问题的显式对应关系，成功地将其转换为目标函数的约束超参数选择问题，并基于剪枝速度逐渐降低这一原则提出了适应学习率的阈值调度器。发表时在基准数据集上的SNN中获得了前沿方法中多个稀疏度下的最好剪枝后性能，在深度SNN与高稀疏度（93%）下相比先前的最好剪枝后分类精度提升3%。这一方法能够进一步拓展到一般的ANN剪枝中，并且也在深度ANN与高稀疏度（95%）下相比先前的最好剪枝后分类精度提升2%。该方法为后续可能的剪枝速度优化方面的改进提供了坚实的理论基础。

第四，构建了剪枝压缩后的脉冲目标检测网络，并应用在基于Vidar脉冲相机的脉冲检测系统中，节省了SNN参数量以及计算开销。本方法将所提出的剪枝算法进行修正，适配为硬件友好的通道剪枝算法，并针对检测SNN网络的独特跳连接结构进行了调整。这一算法剪枝后的网络在使用Vidar脉冲相机采集的脉冲数据流上进行了目标检测任务，在检测精度不明显下降的情况下，该方法可以将深度SNN检测网络的参数量节省至稠密网络的约50%，同时不带来明显的检测错误。而对于常见的检测场景，SNN网络的参数量可以进一步压缩至稠密网络的20%。这些结果提升了SNN在现实场景检测任务上的计算效率，为检测网络后续在硬件上的高效部署提供了便利。

综上所述，本文基于神经系统中树突棘生长消亡过程的启发，逐步构建了具有生物合理性、计算有效性的脉冲神经网络剪枝算法。本文从树突棘的生理结构建模、树突棘修剪的生理机制建模两个层面进行了研究，挖掘出对SNN剪枝任务最为关键的生理结构指标和最有效的突触修剪机制，最终在脉冲目标检测系统上验证了其实际应用价值。本文也为关于树突棘与SNN剪枝的进一步类比研究奠定了基础。

关键词：树突棘，模型剪枝，脉冲神经网络，模型压缩，深度学习

Pruning of Deep Spiking Neural Networks Inspired by Formation and Elimination of Dendritic Spines

Yanqi Chen (Computer Applied Technology)
Directed by Prof. Yonghong Tian

ABSTRACT

Artificial intelligence is permeating every aspect of production and life. In the future, people will be inseparable from the results of extensive model calculations. Spiking neural networks (SNNs) are famous for simulating the binary spike firing characteristics of biological neurons to save computing overhead. They can be efficiently deployed on neuromorphic hardware and are expected to serve scenarios such as edge computing that require high energy efficiency. Network pruning can remove redundant connections in the network while maintaining network representation capabilities, improving computing efficiency from another perspective. The combination of the two has therefore become a research hotspot. However, as a computational model inspired by neuroscience, research in the field of SNN pruning usually ignores human brain as a representative of advanced intelligence, which has far sparser weights (synaptic connections) than most current neural networks. In fact, the **synaptic pruning** phenomenon that is widespread in the nervous system and the weight pruning in deep SNN both aim to remove redundant connections in networks, and they are thus two sides of the same coin. The minimum pruning units of aforementioned both are the tiny structures on dendrites - dendritic spines, and synaptic weights, respectively. Therefore, the formation and elimination of dendritic spines exactly corresponds to the emergence and pruning of synaptic connections. The development of pruning algorithms for deep spiking neural networks based on 1) the structural characteristics of dendritic spines during their formation and elimination; and 2) synaptic pruning mechanisms, provides a novel research perspective in this field.

This thesis conducts research from two levels: physiological structure modeling and physiological mechanism modeling, aiming to find the physiological structure and mechanism information most relevant to pruning, and finally model it as a spiking neural network pruning algorithm. With the deepening of the research, this thesis gradually resolves **the problem of modeling the correlation between dendritic spines and weights**, **the modeling problem of**

the connection structure rewiring mechanism, **the concern of poor flexibility of weight symbols**, and **the speed control problem of dendritic spine elimination (weight pruning)**, in the formation and elimination processes of dendritic spine. The effectiveness of the solution is verified on deep SNNs and finally applied in a spiking object detection system based on the Vidar spiking cameras. The main innovations of this thesis are as follows:

Firstly, to address the modeling problem of the presence or absence of connections in pruning, a pruning algorithm for spiking neural networks (SNNs) based on the gradient rewiring mechanisms is proposed. It draws upon differences, both morphological and functional, in synaptic modeling of dendritic spines and corresponding pruning algorithms, and enhances the continuous transformation process of network topology during pruning. This method introduces an extra parameter describing the volume of dendritic spines, modeling both the pruned state and the connected state of weights. This enables the network to automatically regulate the formation and elimination of dendritic spines during the training process. The results on SNNs show that this method greatly reduces the performance loss of SNN under extremely high sparsity (>99%) to only 3.5% without significantly reducing the efficiency of spike transmission throughout the network, which is the best result for pruning on deep SNNs at time of publication.

Secondly, to address the lack of sign flexibility of weight in pruning, a spiking neural network pruning algorithm is proposed to simulate dually innervated dendritic spines. By modeling the changes in proportion of receptors, this method also broadens the hypothetical space of the networks, and thus improves the performance of pruned networks. Based on the previous method, this work takes the modeling of the transition between excitation and inhibition of dendritic spines into consideration, enabling the learning of sign of the weight. The results on an 18-layer SNN show that its effect further surpasses the previous work and becomes the first work to effectively prune deep SNNs. This is also the first SNN pruning work validated on a large scale benchmark dataset.

Thirdly, in view of the common concern of speed control of sparsity in pruning, a unified framework for pruning algorithms that finely controls synaptic pruning speed is proposed to avoid unsatisfactory network performance after pruning caused by abnormal pruning speed. Through theoretical analysis, this method provides an explicit correspondence between the elimination speed of dendritic spines, that is, the synaptic pruning speed and the optimization problem in the context of dendritic spine state change modeling and pruning, and successfully converts it into a hyperparameter selection problem of the constraint in objective function. At

the time of publication, this work achieved the best post-pruning performance under multiple sparsities among cutting-edge methods in SNN on benchmark datasets, and the classification accuracy of deep SNN under high sparsity (93%) was improved by 3% compared to the previous best results. This method can be further extended to general ANN pruning, and also improves the classification accuracy by 2% compared with the previous best post-pruning performances in deep ANN under high sparsity (95%). This method provides a solid theoretical foundation for viable subsequent improvements in pruning speed optimization.

Fourthly, a compressed spiking visual detection network is obtained through pruning, and applied in a spiking detection system based on the Vidar spiking camera, which greatly saves the amount of SNN parameters and computational overhead. This work modifies the proposed pruning algorithm, adapts it to a hardware-friendly channel pruning algorithm, and adjusts it to detect the unique jump connection structure of the SNNs. The network pruned by this algorithm performs a target detection task on the spike streams collected by the Vidar spiking camera. This method can save the parameters of the deep detection SNN to 50% of that of a dense counterpart without significantly reducing the detection accuracy. For common detection scenarios, the amount of the SNN parameters can be further compressed to 20% of that of the dense network. These results have greatly improved the computational efficiency of SNN in real-life detection tasks, and facilitated the subsequent efficient deployment of the detection network on hardware.

In summary, inspired by the formation and elimination process of dendritic spines in the nervous system, this thesis gradually constructs a spiking neural network pruning algorithm with biological plausibility and computational efficiency. This thesis conducted research from two levels: modeling of the physiological structure of dendritic spines and modeling of the physiological mechanism of dendritic spine pruning, and unearthed the most critical physiological structure indicators and the most effective synaptic pruning mechanism for the SNN pruning task. Finally, the value in practical application has been verified on the spiking target detection system. This thesis also lays the foundation for further analogy research on dendritic spines and SNN pruning.

KEY WORDS: Dendritic Spine, Model Pruning, Spiking Neural Networks, Model Compression, Deep Learning