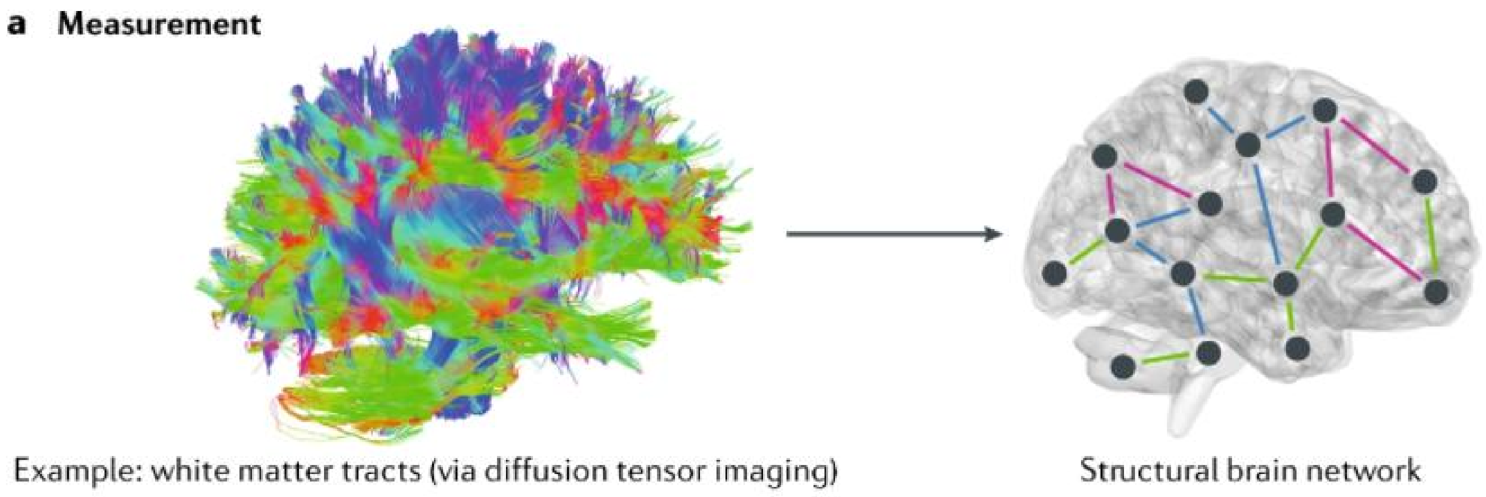
Towards Efficient Deep Spiking Neural Networks Construction with Spiking Activity based Pruning

Yaxin Li [1 2], Qi Xu [✉][1], Jiangrong Shen [3], Hongming Xu [1], LongChen [4], GangPan [3]

[1]School of Computer Science and Technology, Dalian University of Technology, [2]Guangdong Laboratory of Artificial Intelligence and Digital Economy, [3]College of Computer Science and Technology, Zhejiang University, [4]Centre for Systems Engineering and Innovation, Imperial College London

Introduction

Motivation: joint learning the structure and parameters of SNN models



Neural **plasticity** of the brain:
- The brain can autonomously learn the strength and structure of synaptic connections.
- Brain networks can adaptively generate various complex neural network structures.

Neuromorphic computing is a novel computing technology inspired by the way the human brain stores and processes information.

Introduction

Related Work: unstructured pruning is not hardware friendly

Unstructured Pruning Methods

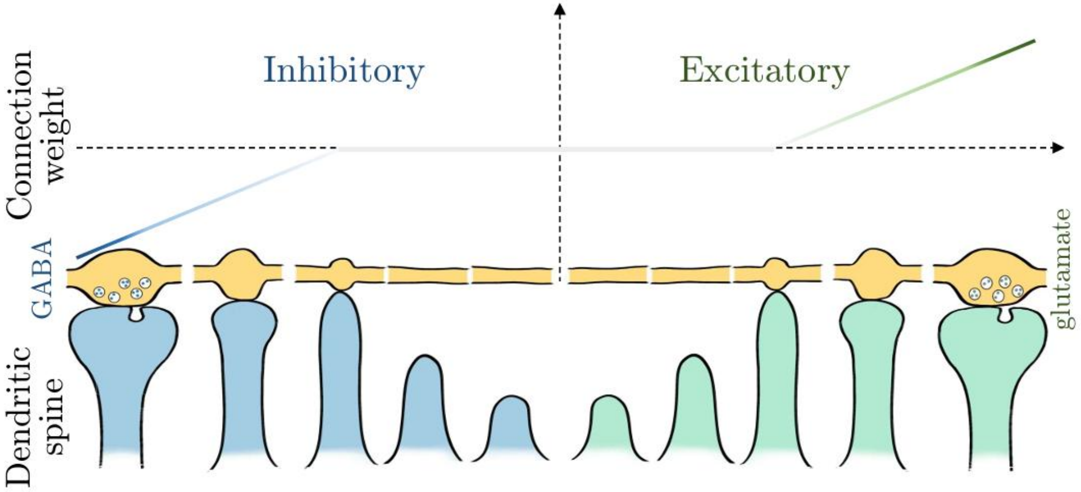


Figure 1. This method model different states of SNN weights, facilitating weight optimization for pruning.[1]

Unstructured Pruning Methods

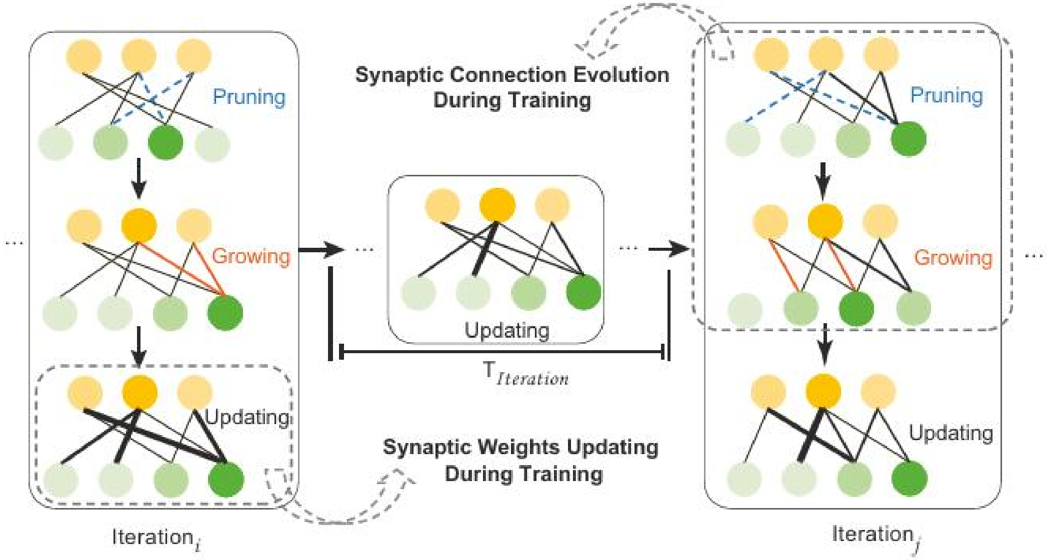
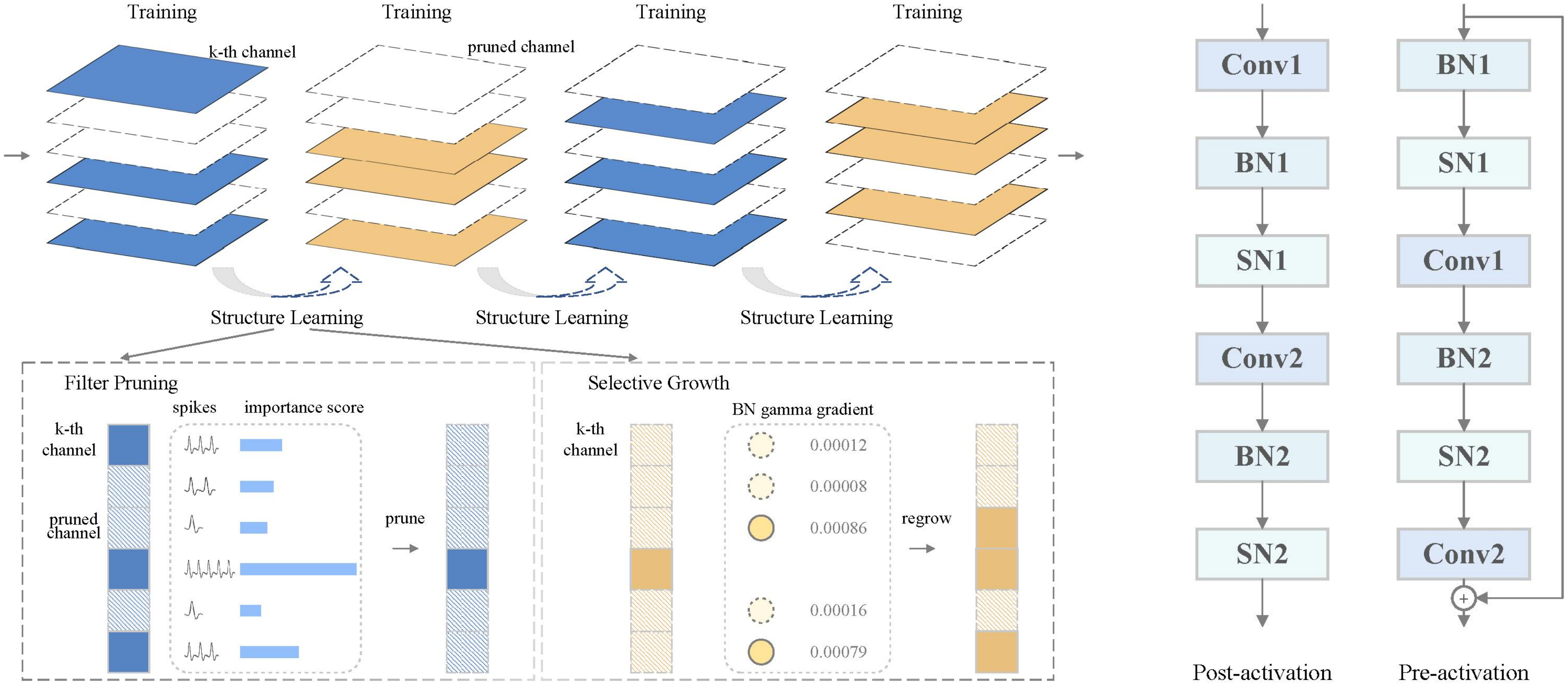


Figure 2. The synaptic connection pruning and growth procedures proceed every iteration.[2]

[1] Chen, Y., Yu, Z., Fang, W., Ma, Z., Huang, T., and Tian, Y. State transition of dendritic spines improves learning of sparse spiking neural networks. In International Conference on Machine Learning, pp. 3701–3715. PMLR, 2022.
[2] Shen, J., Xu, Q., Liu, J. K., Wang, Y., Pan, G., and Tang, H. Esl-snns: An evolutionary structure learning strategy for spiking neural networks. arXiv preprint arXiv:2306.03693, 2023.

Methodology--overall training algorithm

Method: The **Spiking Channel Activity-based (SCA)** Structure Learning Framework



(a) The SCA structure learning framework.

(b) Network structures.

Figure 3. The schematic illustration of the SCA structure learning framework.

Methodology

Channel Importance Score:

$$r_k^l = \frac{1}{N}\frac{1}{T}\left(\sum_{n=1}^{N}\sum_{t=1}^{T}\|H_k^l(t)\|\right) \quad (\|H_k^l(t)\| = \left[H_t^{ij}\right]_{h\times w})$$

Excitation: Membrane potential increases, termed depolarization.

Inhibition: Membrane potential decreases, termed hyperpolarization.

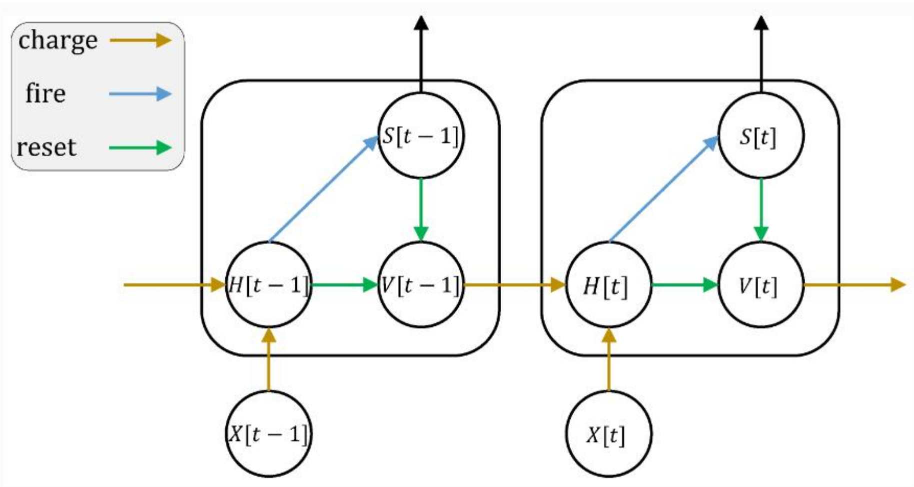


Figure 4. Spiking Neuron Model.

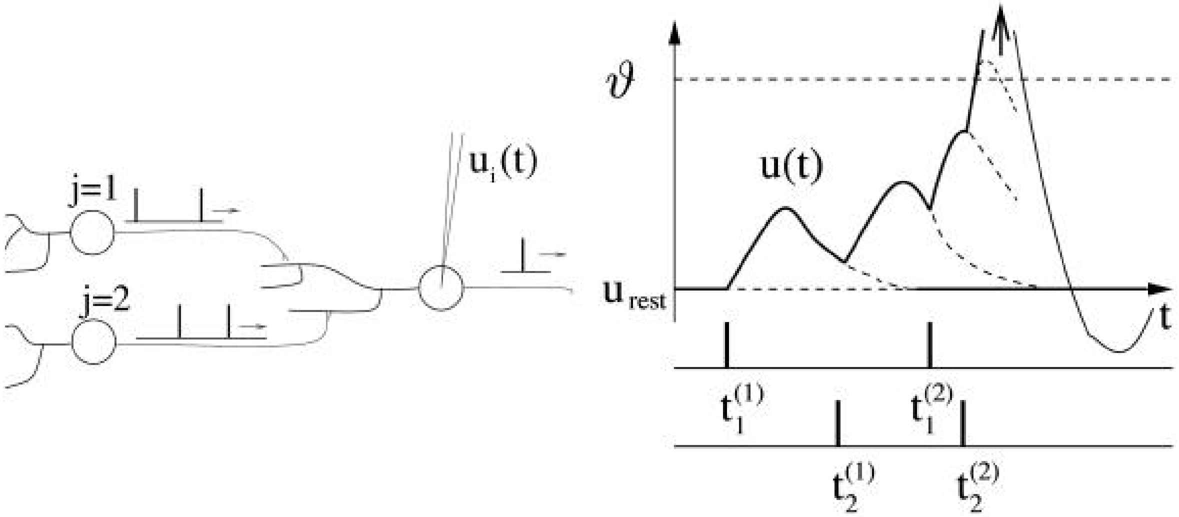


Figure 5. Excitatory postsynaptic potential (EPSP).

Methodology

Filter Pruning: a certain proportion of convolutional kernels are pruned based on channel importance scores.

Selective Growth: some channels are reactivated based on the gradient magnitudes of the Gamma parameters in BN layers.

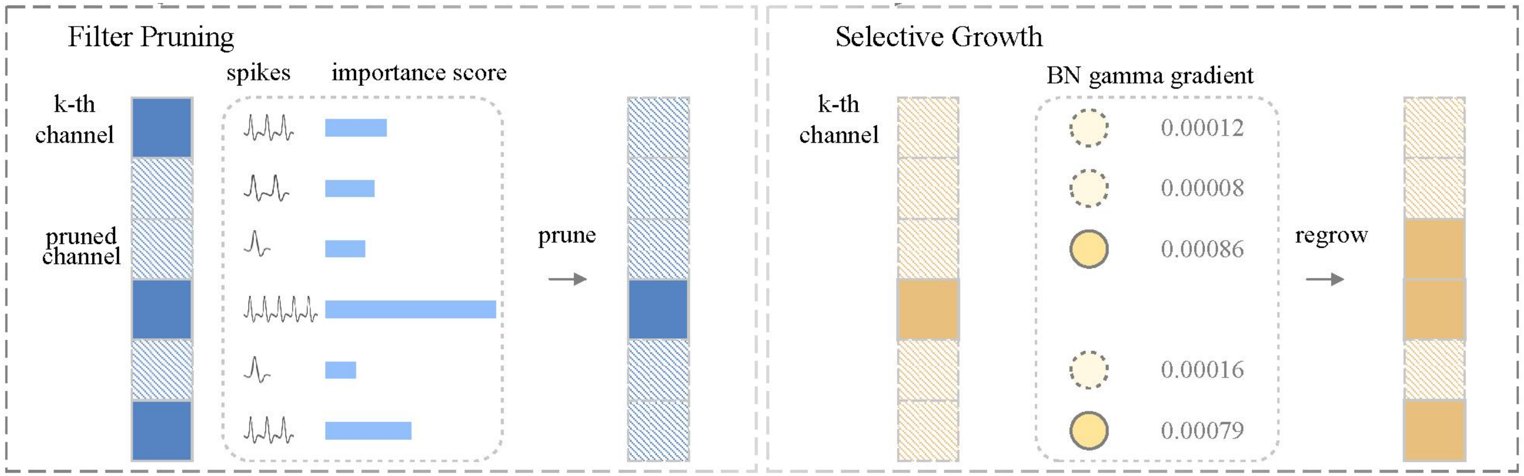


Figure 6. Filter pruning and selective growth.

Algorithm 1 The Overall Training Framework.

Input: Input data X. Labels Y.
Output: Pruning channels ratio $p\%$. Mask update ratio $q\%$. The weight W. The weight mask M.
1: Initialize the weight W;
2: **for** i in $[1, epoch]$ **do**
3: Learn weight parameters based on surrogate gradient under L1 regularization;
4: Learn weight connection based on structure learning rule:
5: (1) Prune $q\%$ channels, resulting in the removal of $(p+q)\%$ channels;
6: (2) Regrow $q\%$ channels, maintaining a pruning ratio of $p\%$;
7: **for** each layer of the model **do**
8: $W = M \odot W$;
9: **end for**
10: **end for**
11: Completely remove the channels corresponding to zero positions in the mask to obtain the compressed network model.
12: **return:** The lightweight SNN.

Experiments--Performance

Results: Evaluation under **different pruning ratios**

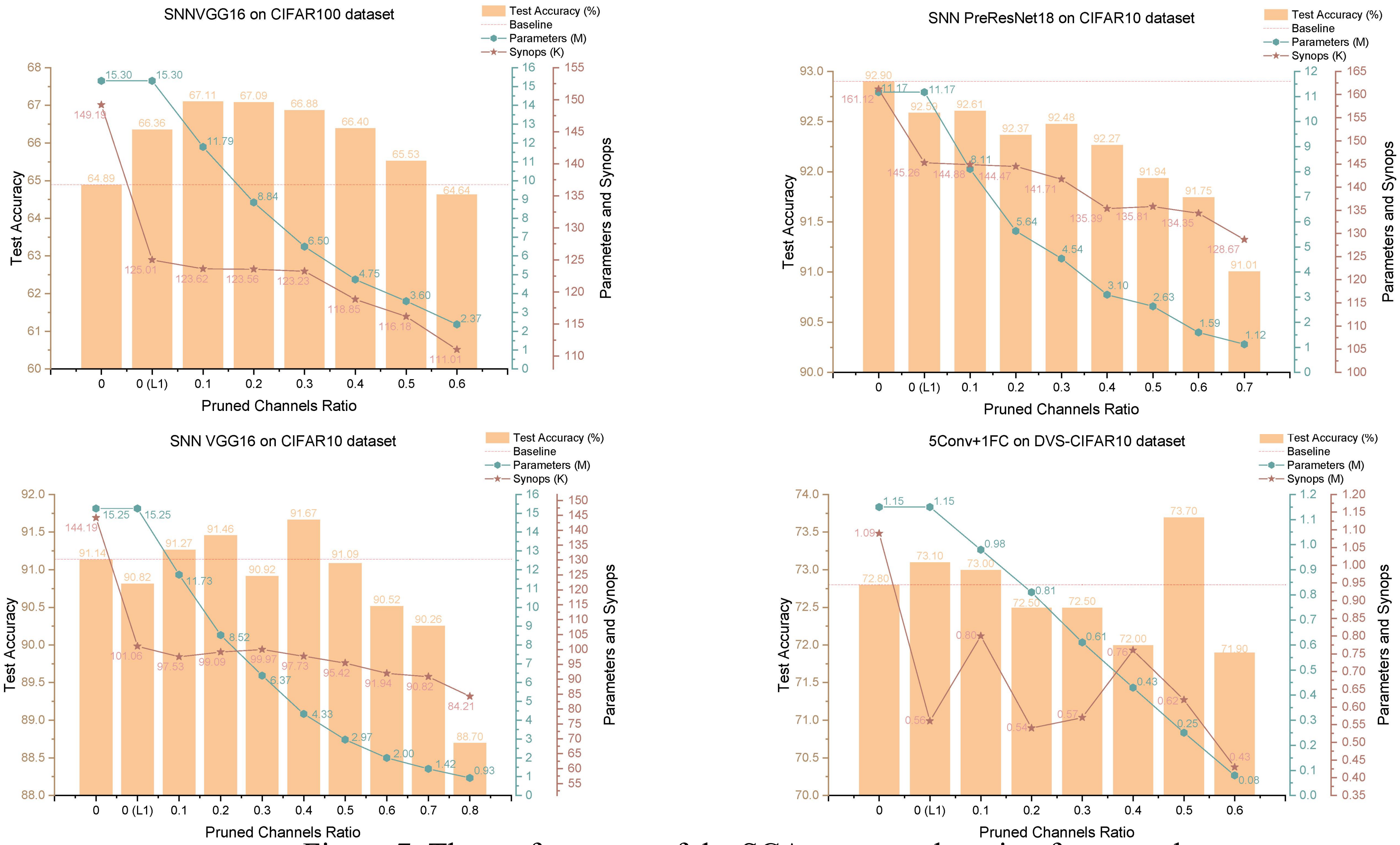
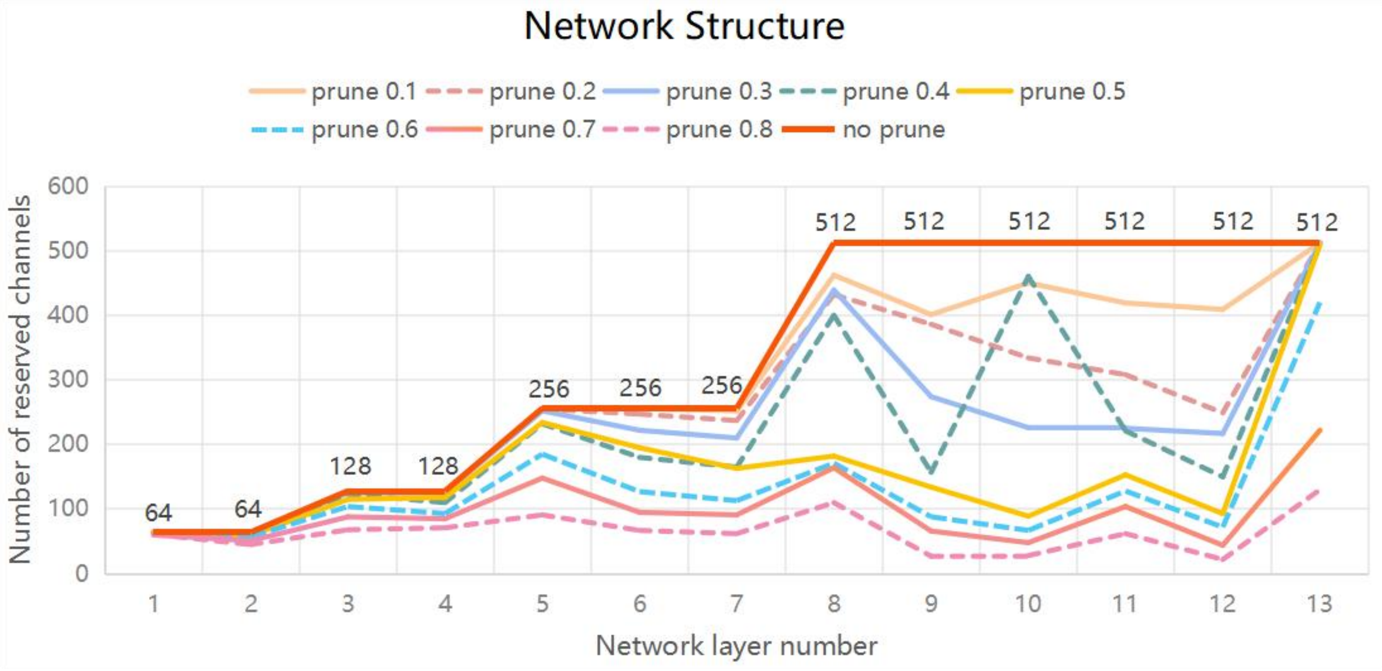


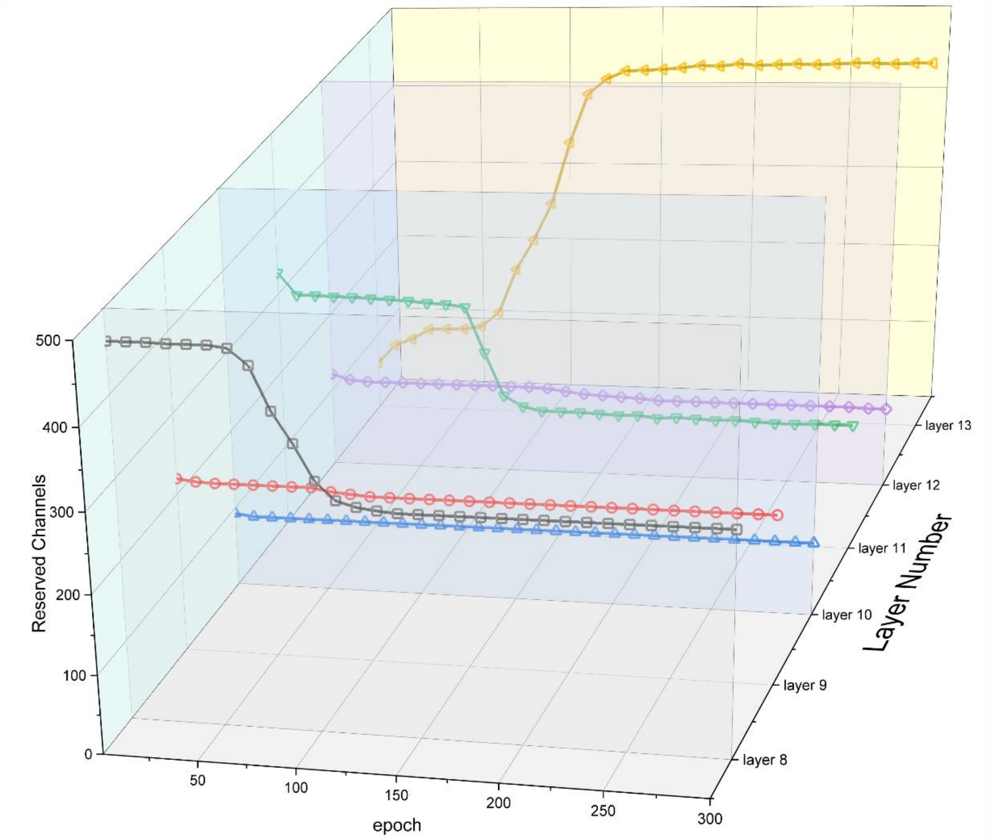
Figure 7. The performance of the SCA structure learning framework.

Experiments--Structure Learning

Results: This indicates that the structural learning framework, as the network structure evolves, autonomously **adapts to an appropriate structure**.



(a) Network structures under various pruning ratios.

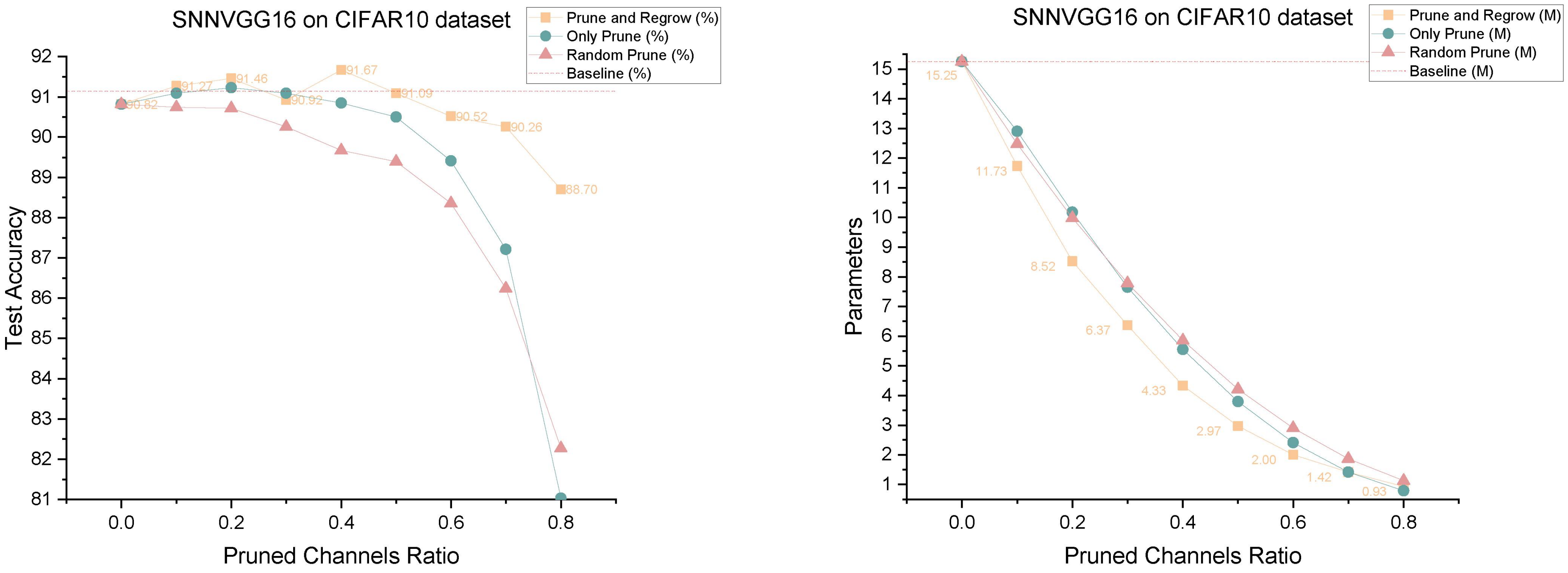


(b) The evolution of network structure during the training process.

Figure 8. Analysis of structural changes in the SCA framework.

Experiments--Ablation

Results: The SCA framework's mechanism for identifying redundant channels and reactivating pruned channels **autonomously searches** for an appropriate network structure.



(a) The accuracy of ablation experiments.

(b) The parameters of ablation experiments.

Figure 9. Analysis of the effectiveness of the SCA structure learning framework.

Conclusion

Summary and Future Outlook:

- Lightweight and high-performance SNNs can better leverage their advantages of low power consumption. The structured pruning methods can result in regular, sparse SNN models, making them more hardware-friendly.
- The approach proposed in this paper starts from the perspective of biological plasticity, combining pruning and regrowth in an adaptive manner during training to explore suitable lightweight network structures.
- This is of significant value for deploying high-performance, low-memory SNNs on neuromorphic chips.

Thank you!