NEURAL INFORMATION PROCESSING SYSTEMS

Enhanced Self-Distillation Framework for Efficient Spiking Neural Network Training

Xiaochen Zhao†, Chengting Yu†, Kairong Yu, Lei Liu, Aili Wang*

Zhejiang University

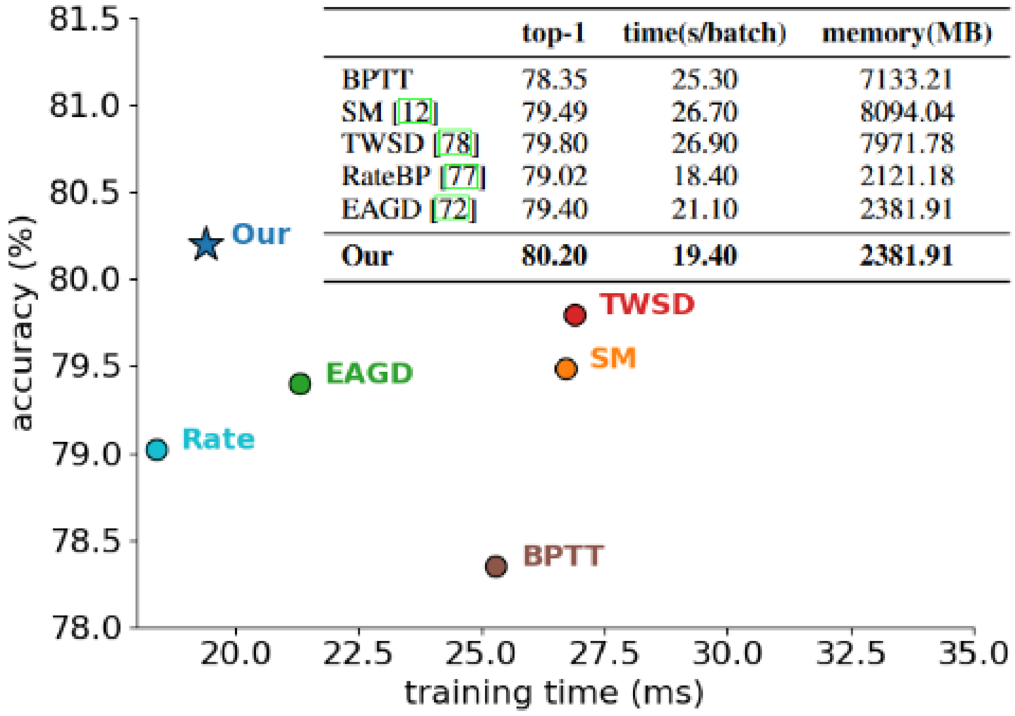
Haining, Zhejiang, China

Code link: https://github.com/Intelli-Chip-Lab/enhanced-self-distillation-framework-for-snn

Motivation

- SNNs are brain-inspired models
  - Offer a potential **energy efficiency** advantage on neuromorphic hardware
  - An alternative to traditional ANNs

- Major limitations of SNNs:
  - **Lower accuracy** compared to ANNs
  - **Gradient approximation error exists**
  - **Training cost is linear with the time dimension**

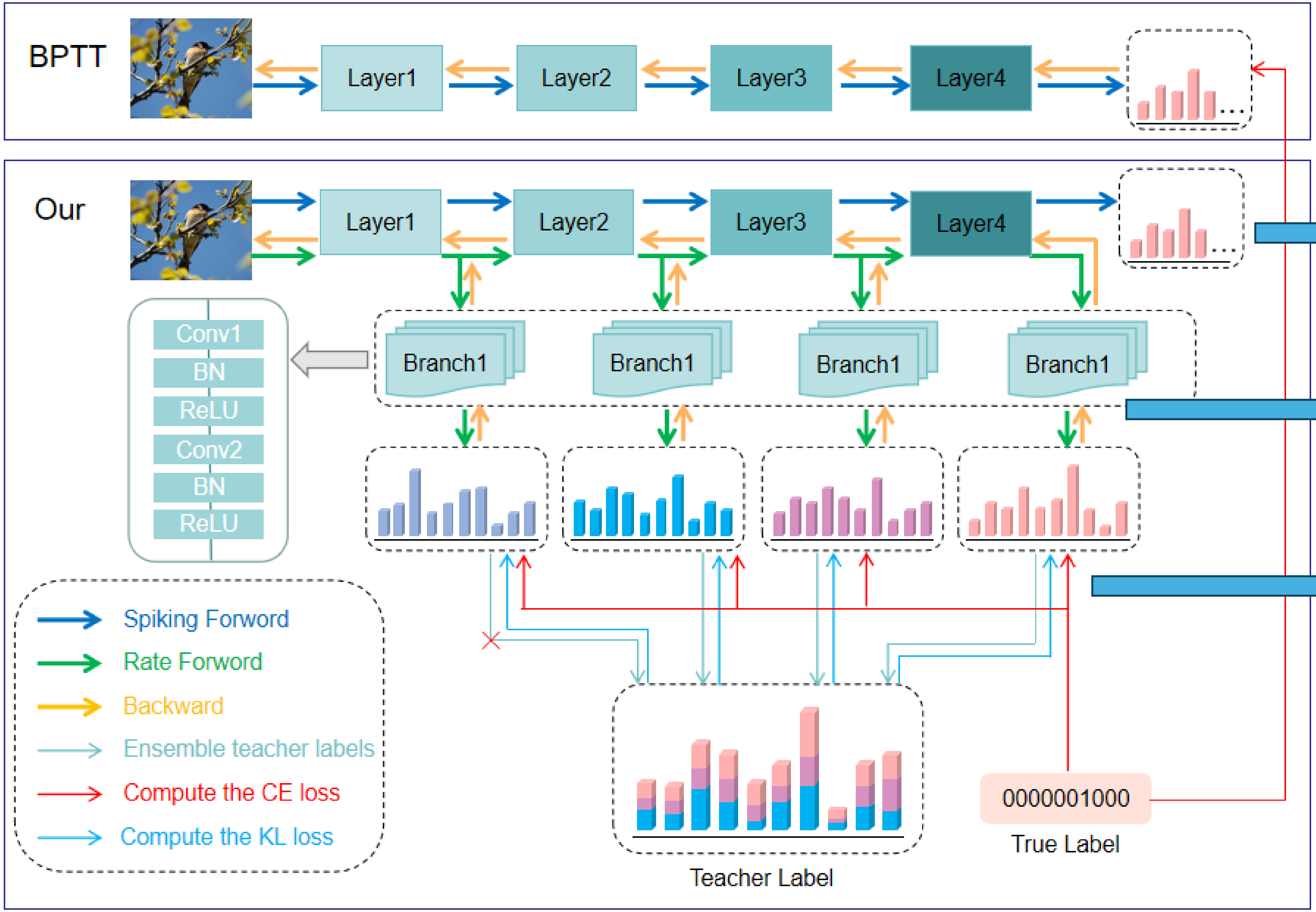


	top-1	time(s/batch)	memory(MB)
BPTT	78.35	25.30	7133.21
SM [12]	79.49	26.70	8094.04
TWSD [78]	79.80	26.90	7971.78
RateBP [77]	79.02	18.40	2121.18
EAGD [72]	79.40	21.10	2381.91
Our	**80.20**	**19.40**	**2381.91**

Key Innovation

- **Approach**:
  - Auxiliary module mitigates gradient error
  - Propose the concept of label reliability
  - Decouple label reliability to boost distillation efficiency
- **Empirical Results**:
  - Show outstanding performance on standard datasets.
  - The self-distillation framework boasts superior scalability.

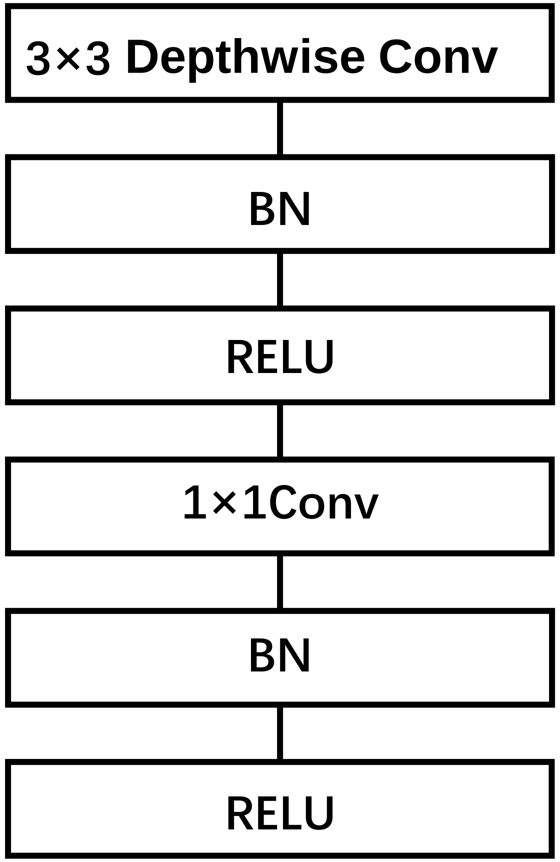
Framework Overview



- Rate-based backpropagation
- Auxiliary module corrects accuracy
- Teacher labels Decoupling and integration

Design of the auxiliary module

- **Depthwise separable convolution:**
  - Depthwise Convolution
    - Extract spatial local features per channel.
    - Reduce parameters/computational cost, focus on spatial dimension.
  - Pointwise Convolution
    - Fuse cross-channel features, adjust output channels.
    - Keep feature map size, optimize channel dimension only.



Alleviate the Gradient Error Problem

- **Baseline:**
  - Correct gradient： $\frac{\partial y_c}{\partial \theta} = G$,
  - True gradient from RateBP approximation： $R\left(\frac{\partial y_c}{\partial \theta}\right) = G + \delta$,
  - Error ratio： $K_{rate} = \frac{\delta}{G}$.
- **Our Work:**
  - We define the total loss as： $L_{total} = \alpha \cdot L_1 + L_c$,
  - Our gradient： $\frac{\partial L_{total}}{\partial \theta} = \frac{\partial L_c}{\partial y_c} \cdot (G + \alpha \cdot S)$,
  - Auxiliary gradient cuts error ratio： $K_{esd} = \frac{\delta}{G+\alpha \cdot S} \leq \frac{\delta}{G} = K_{rate}$.

Design of Self-Distillation Loss

- Filter incorrectly predicted labels

$$y_{\text{teacher}} = \frac{\sum_{l=1}^{L} p_l \cdot \mathbb{I}(\text{argmax} p_l = \text{argmax} y)}{\sum_{l=1}^{L} \mathbb{I}\,(\text{argmax} p_l = \text{argmax} y) + \epsilon}$$

- Hard loss against the ground truth

$$L_{ce} = \sum_{l=1}^{L} \left[-\sum_{c=1}^{C} y^{(c)} \log\left(p_l^{(c)}\right) \right]$$

- Total loss function

$$L_{esd} = \sum_{l=1}^{L} \left\{ \left[\sum_{c=1}^{C} p_{\text{teacher}}^{(c)} \log\left(\frac{p_{\text{teacher}}^{(c)}}{p_i^{(c)}} \right) \right] \cdot \mathbb{I}\left(\sum_{c=1}^{C} \left| p_{\text{teacher}}^{(c)} \right| \neq 0 \right) + \eta \cdot \mathcal{R}_l \cdot \mathbb{I}\left(\sum_{c=1}^{C} \left| p_{\text{teacher}}^{(c)} \right| = 0 \right) \right\}$$

Results--Performance Comp. on Benchmarks

Results on CIFAR-10, CIFAR-100, and ImageNet Datasets

Table 1: Comparison of top-1 accuracy (%) averaged over three runs on CIFAR-10, CIFAR-100, and ImageNet datasets. *indicates the use of an additional pre-trained ANN model for distillation. For all experiments on ImageNet, the ResNet-34 model is consistently used for training.

Datasets	Training	Method	Architecture	Timestep	CIFAR10 Top-1 Acc (%)	CIFAR100 Top-1 Acc (%)	ImageNet Top-1 Acc (%)
Direct-training	OTTT [68]	online	VGG-11	6	93.52	71.05	65.15
	OS [85]	online	ResNet-19	4	95.20	77.86	67.54
	Dspike [41]	BPTT	ResNet-19	6	94.25	74.24	68.19
				4	93.66	73.35	
				2	93.13	71.68	
	TET [11]	BPTT	ResNet-19	6	94.50	74.72	64.79
				4	94.44	74.47	
				2	94.16	72.87	
	SEW-ResNet [17]	BPTT	ResNet-34	4	-	-	67.04
	DSR [47]	one-step	PreAct-ResNet-18	20	95.10	78.50	67.74
	RateBP [76]	one-step	ResNet-18	6	95.9	79.02	70.01
				4	95.61	78.26	
				2	94.75	75.97	
			ResNet-19	6	96.38	80.83	
				4	96.26	80.71	
				2	96.23	79.87	
w/ distillation	BKDSNN* [70]	BPTT	ResNet-19	4	94.64	74.95	67.21
	TKS [15]	BPTT	ResNet-19	4	96.35	79.89	69.60
	SM [12]	BPTT	ResNet-18	4	96.04	79.49	68.25
			ResNet-19	4	96.82	81.70	
	TWKD* [77]	BPTT	ResNet-18	6	95.96	79.80	71.04
				4	95.57	79.10	
	EAGD* [71]	one-step	ResNet-18	6	96.14	79.40	70.64
				4	95.92	78.85	
				2	95.19	77.06	
			ResNet-19	2	96.56	81.44	
	ours	one-step	ResNet-18	6	96.19 ± 0.12	80.20 ± 0.17	70.72
				4	95.92 ± 0.03	79.30 ± 0.21	
				2	95.29 ± 0.10	77.46 ± 0.17	
			ResNet-19	6	96.46 ± 0.11	82.14 ± 0.07	
				4	96.39 ± 0.01	81.90 ± 0.20	
				2	96.31 ± 0.07	80.97 ± 0.05	

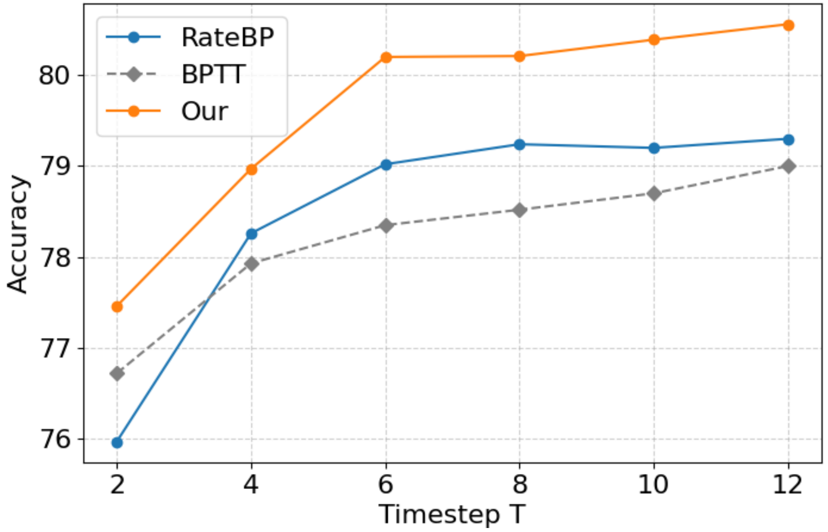
Results on DVS-CIFAR-10

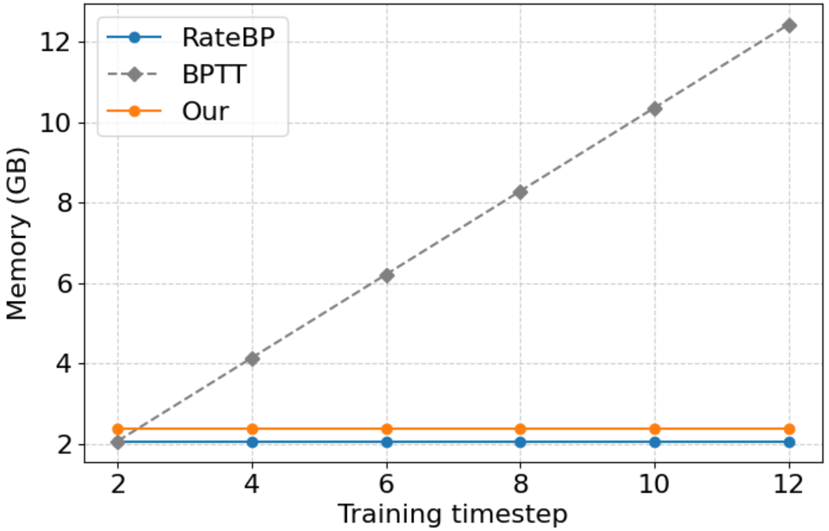
Training	Method	Architecture	Timestep	Top-1 ACC(%)
OTTT [69]	online	VGG-11	10	76.63
RateBP [77]	one-step	ResNet-18	10	80.40
EAGD [72]	one-step	ResNet-19	4	80.54
TET [11]	BPTT	VGGSNN	10	83.17
SM [12]	BPTT	ResNet-18	10	83.19
Enof [22]	BPTT	ResNet-19	10	80.10
TWKD [78]	BPTT	ResNet-19	10	83.80
ours	one-step	ResNet-18	10	81.40
		ResNet-19	10	81.90
	BPTT	ResNet-18	10	85.70
		ResNet-19	10	85.90

- **Performance**:
  - Excellent performance on standard datasets
  - Our self-distillation framework is also applicable to BPTT

Temporal Step Scalability and Firing Rate

- Achieves constant training overhead.
- T increases, the accuracy continues to improve.





- Lower spike firing rate.

	Method	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$	$T=6$	avg
Trained for 4 time steps	BPTT	0.1799	0.2137	0.2045	0.2091	-	-	0.2018
	ours	0.1591	0.1709	0.1715	0.1706	-	-	0.1680
Trained for 6 time steps	BPTT	0.1761	0.2034	0.2023	0.1966	0.2060	0.1941	0.1964
	ours	0.1548	0.1560	0.1550	0.1516	0.1550	0.1532	0.1543

Conclusion

- **Problem Addressed:** Training Costs and Performance Issues in SNN
- **Proposed Method**: Enhanced Self-Distillation Framework for Efficient SNNs Training
  - An auxiliary module is introduced to mitigate gradient errors.
  - A reliability-decoupled self-distillation strategy is proposed.
  - Provides both theoretical analysis and empirical experiments
- **Experimental Results**:
  - A constant training cost is achieved.
  - Achieves a breakthrough in accuracy compare with standard self-distillation.
  - Could extend to ANNs and BPTT.

Acknowledgement & Resources

- **Funding Grants**
  - NSFC with Grant No. 62304203
  - International campus of ZJU international research collaboration seed project.
  - The ZJU-YST joint research center for fundamental science.

- **Resources**

arXiv

Thank you!

GitHub

Contact us: {xiaochen.24, chengting.21, ailiwang}@intl.zju.edu.cn