SMixer: Rethinking Efficient-Training and Event-Driven SNNs

Yijie Lu, Xinhao Luo, Yixing Zhang, Zhiyan Wang, Wentao Li, Yanhan Wang, Zhi Liu, Zhaokun Zhou*, Guoqi Li*

Peking University, Institute of Automation, Chinese Academy of Sciences, Shandong University, Tencent

ICLR

Introduction

Motivation:

To address the computational deviations of spike matrix multiplication in Spiking Self-Attention and the high training costs of SNNs, stemming from non-event-driven GPU training and temporal state overhead. we propose that rational SNN architectures should achieve fully event-driven operations, low training overhead, and competitive performance. The main core is the SMixer.

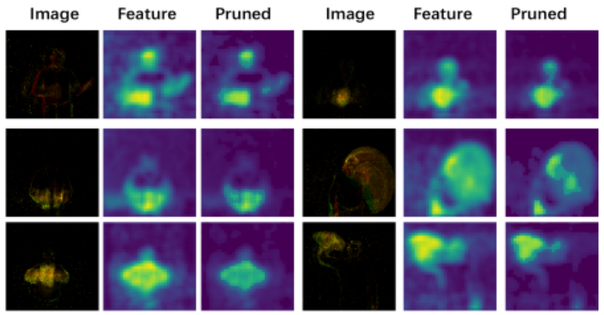
Contribution:

• We analyze the inherent architectural requirements for event-driven SNNs and the need for efficient training methodologies. Based on this, we propose a blueprint for a high-ratio spike feature structured pruning framework built upon the Spiking-token Mixer.

• Based on Spatial-Temporal Redundancy in Spiking-token Mixer, we develop the Dynamic Spatial-Temporal Spike Pruning framework, which integrates dynamic spatial and temporal spike feature pruning methods.

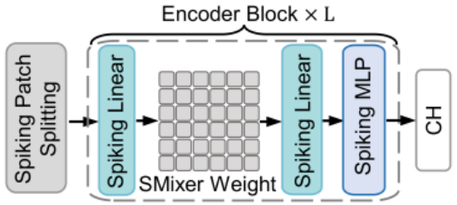
• We demonstrate that the SMixer architecture can achieve performance comparable to that of the Spikformers. Furthermore, we show that our efficient pruning framework built upon SMixer, accelerates training while maintaining performance close to the original model across various neuromorphic and static datasets.

Spatial Temporal Spiking Feature Redundancy:

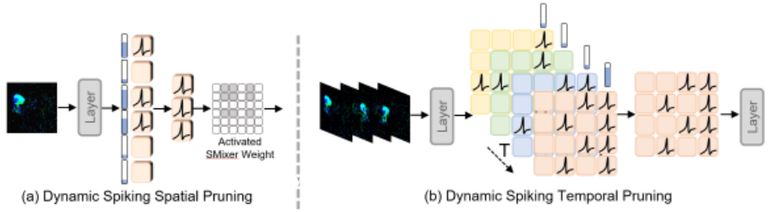


Model Overview

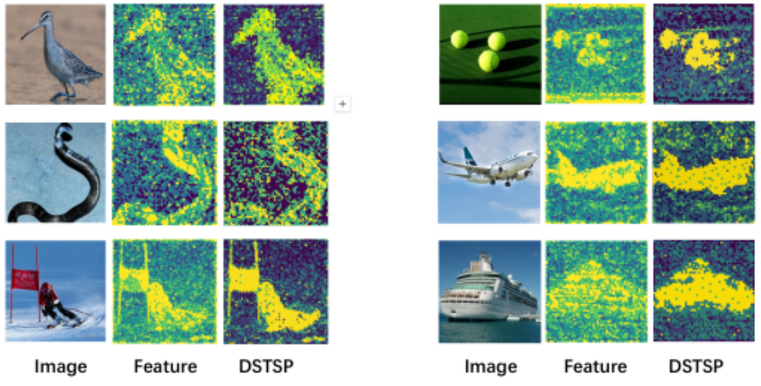
Spiking Token Mixer:



Dynamic Spatial-Temporal Spiking Pruning Method:



Visualization of SMixer:



Experimental Results

ImageNet Classification:

Methods	Architecture	Time Step	Param (M)	TP (im/s)	Memory (MB)	Power (mJ)	Top-1 Acc (%)
STMixer	STMixer-8-512	1	30.12	211	11762	2.20	73.82
	STMixer-8-768	1	61.16	120	17008	4.45	76.68
STMixer + DSTSP	STMixer-8-512	1	27.61	250	9578	1.68	73.56
	STMixer-8-768	1	58.66	162	13002	2.36	76.87
SpikformerV2	Spikformer V2-8-384	1	29.11	130	5260	1.73	75.42
	Spikformer V2-8-512	1	51.55	113	6880	2.84	79.05
	Spikformer V2-8-384	4	29.11	82	12170	4.69	78.80
	Spikformer V2-8-512	4	51.55	67	18786	9.36	80.38
SpikformerV2 → M	Spikformer V2-8-384	1	27.97	156	3490	1.13	76.39
	Spikformer V2-8-512	1	48.56	134	5384	2.12	79.16
	Spikformer V2-8-384	4	27.97	65	9256	3.65	79.12
	Spikformer V2-8-512	4	48.56	56	12982	7.97	80.45
SpikformerV2 → M + DSTSP	Spikformer V2-8-384	1	27.34	245	2584	0.83	76.22
	Spikformer V2-8-512	1	47.93	198	3368	1.98	78.99
	Spikformer V2-8-384	4	27.34	135	7505	1.55	78.03
	Spikformer V2-8-512	4	47.93	98	10387	3.44	79.15
QKFormer	HST-10-384	1	16.47	377	13841	3.38	75.52
	HST-10-512	1	29.08	317	20005	4.79	78.71
QKFormer → M	HST-10-384	1	18.31	365	8385	2.76	76.03
	HST-10-512	1	29.70	325	13051	4.06	78.69
QKFormer → M + DSTSP	HST-10-384	1	14.92	461	6991	2.17	75.13
	HST-10-512	1	25.92	398	9751	3.15	77.39
SDT-V3	Efficient-transformer-S	4	5.1	300	9040	1.70	75.30
	Efficient-transformer-M	4	10.00	267	12388	3.00	78.50
	Efficient-transformer-L	4	19.00	197	16318	5.90	79.80
SDT-V3 → M	Efficient-transformer-S	4	6.37	373	7396	1.42	75.25
	Efficient-transformer-M	4	10.15	280	9988	2.56	78.65
	Efficient-transformer-L	4	19.36	219	13912	4.93	79.25
SDT-V3 → M + DSTSP	Efficient-transformer-S	4	5.05	398	6060	1.01	74.15
	Efficient-transformer-M	4	9.42	338	8210	2.02	76.63
	Efficient-transformer-L	4	18.66	273	10710	3.91	77.75

Coco Detection:

Methods	Architecture	Param (M)	Power (mJ)	Step	mAP@50(%)
ANN2SNN	Spiking-Yolo [95]	10.2	-	3500	25.7
	Spike Calibration [96]	17.1	-	512	45.3
	Fast-SNN [13]	25.1	-	15	46.4
Direct training	Spiking Retina [97]	11.3	-	4	28.5
	EMS-Res-SNN [98]	26.9	-	4	50.1
	Meta-SpikeFormer* [28]	34.9	49.5	1	44.0
		75.0	140.8	1	51.2
Direct training	E-SpikeFormer*	38.7	56.2	2	41.8
		38.7	94.5	4	58.4
		38.7	119.5	8	58.8
Direct training	SMixer	49.7	36.8	4	**58.9**
	SMixer + DSTSP	39.2	21.2	4	57.4