Masked Spiking Transformer

Ziqing Wang[1, 2*] Yuetong Fang[1*] Jiahang Cao[1] Qiang Zhang[1]
Zhongrui Wang[3, 4†] Renjing Xu[1†] Correspondence: zrwang@eee.hku.hk, renjingxu@ust.hk

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]North Carolina State University [3]The University of Hong Kong [4]ACCESS - AI Chip Center for Emerging Smart System;

ICCV23 PARIS

Motivation

- **Performance Gap between ANNs and SNNs:** SNNs are efficient at processing sparse data on neuromorphic hardware but currently lag behind ANNs in performance on complex tasks.
- **High delay and energy consumption of ANN-to-SNN conversion method:** ANN-to-SNN conversion methods convert pre-trained ANNs into SNNs for better performance while requiring more simulation time steps, with increased power consumption to reduce conversion error.

Contributions

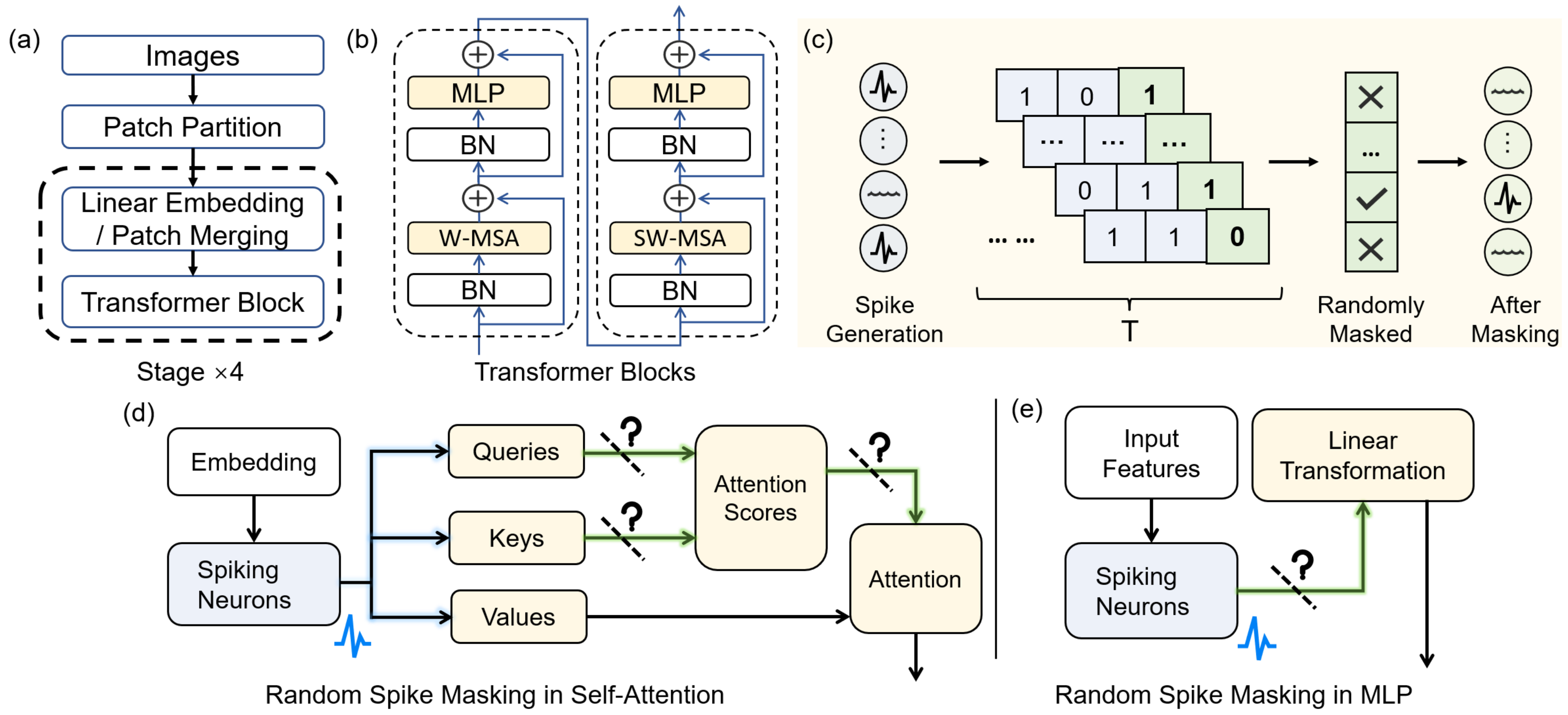
- **Masked Spiking Transformer based on ANN-to-SNN conversion methods:** the first exploration of fully implementing the self-attention mechanism in SNNs using ANN-to-SNN conversion methods
- **Random Masking Method for improving energy efficiency:** bio-inspired spike pruning to reduce redundant synaptic operations

Evaluation

- The performance of **the Masked Spiking Transformer model** was evaluated on both static and neuromorphic datasets.
- The effectiveness of **the Random Spike Masking method** was evaluated across various masking configurations and model architectures, using spike count as an indicator of energy efficiency.

Methods

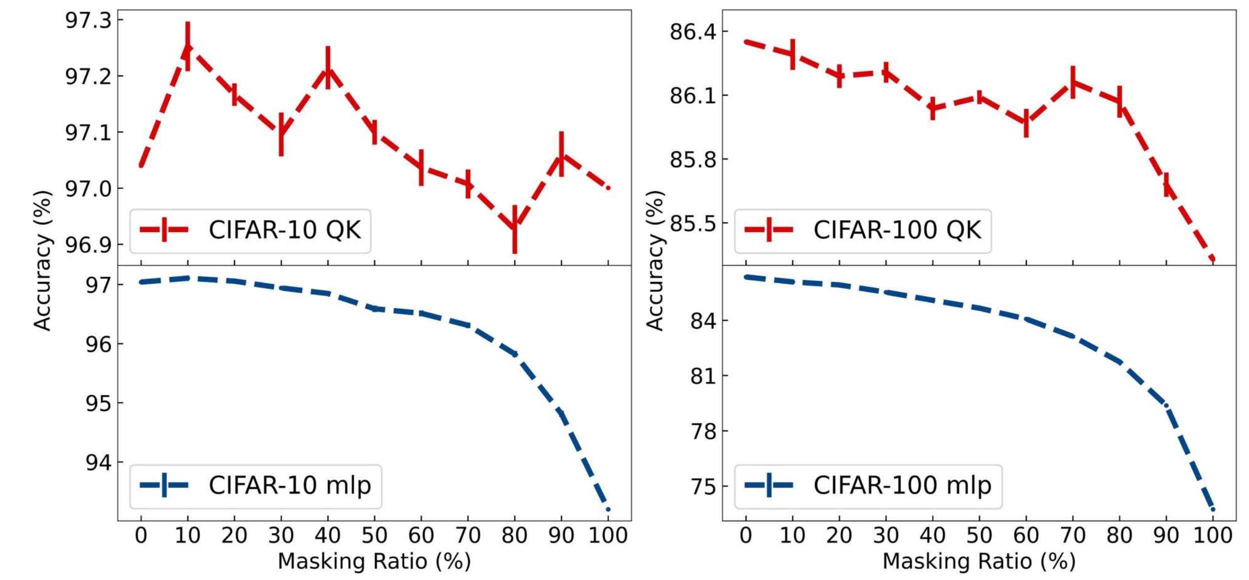
Masked Spiking Transformer Architecture with Random Spike Masking Method



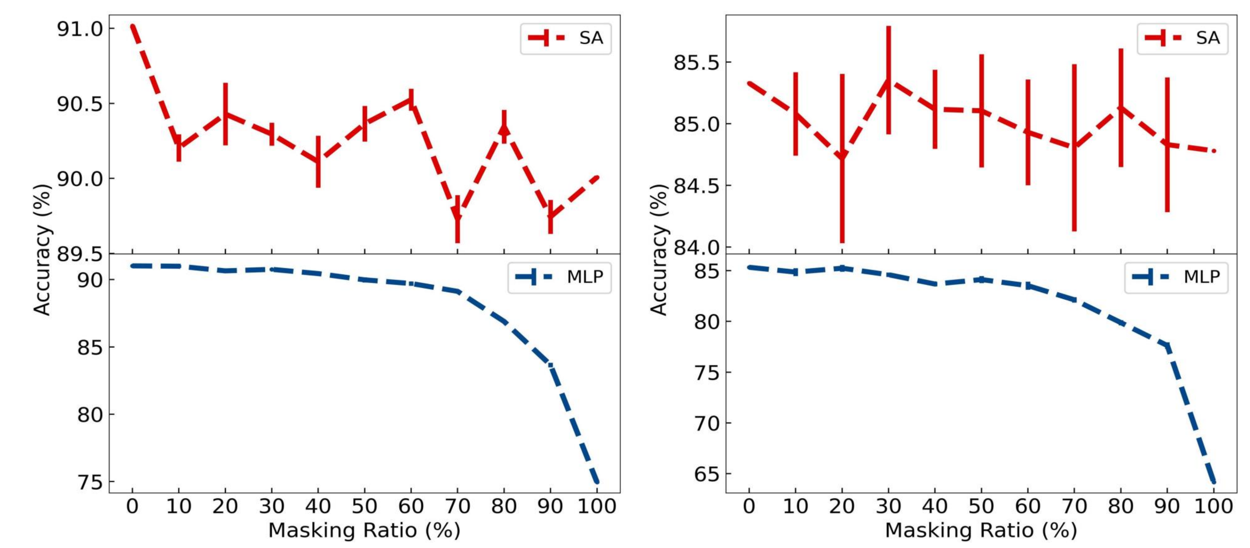
Results

1) **The MST model that combines self-attention mechanism and ANN-to-SNN conversion methods achieves SOTA top-1 accuracy on both static and neuromorphic dataset.**

- Static Dataset



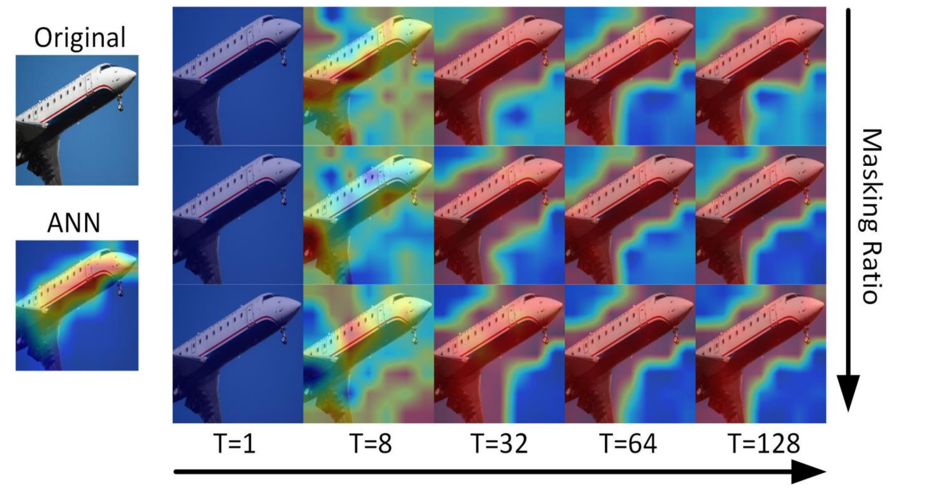
- Neuromorphic Dataset



2) **The RSM method reduces redundant spike operations while keeping model performance over a certain range of mask rates.** For instance, the RSM method reduces MST model power by 26.8% at a 75% mask rate with no performance drop.

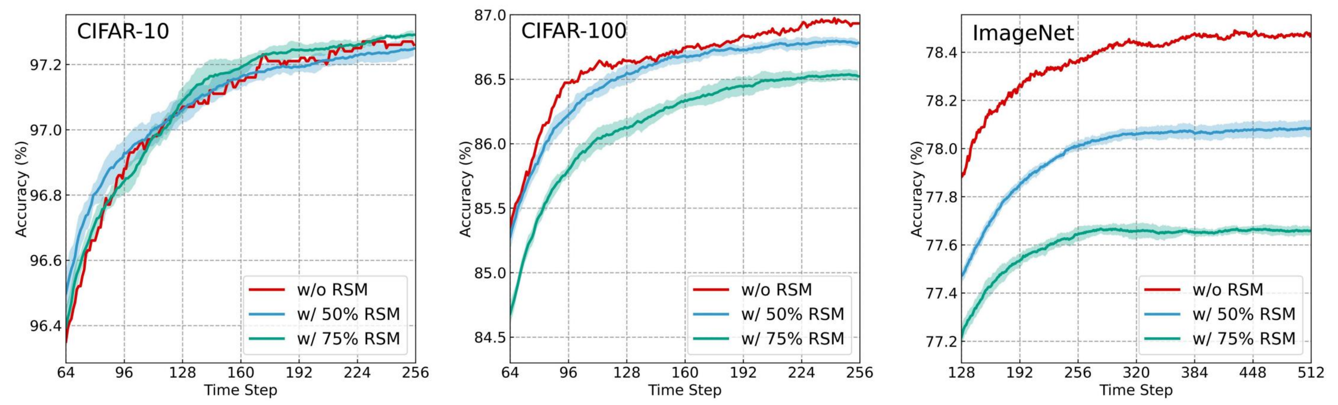
Model	Random Ratio	P (α Watts)	Accuracy (%)
MST	0%	3.9G (×1)	97.27 (+0)
	50%	3.2G (×0.82)	97.25 (-0.02)
	75%	2.9G (×0.74)	97.29 (+0.02)
ResNet-18	0%	58.2M (×1)	96.48 (+0)
	50%	40.7M (×0.70)	92.88 (-3.60)
	75%	34.1M (×0.58)	82.68 (-13.80)
VGG-16	0%	24.4M (×1)	95.46 (+0)
	50%	18.9M (×0.77)	89.56 (-5.90)
	75%	16.7M (×0.68)	79.09 (-16.37)

Models with varying mask rates focus on similar object regions at the same time step, as shown by the red outlines.

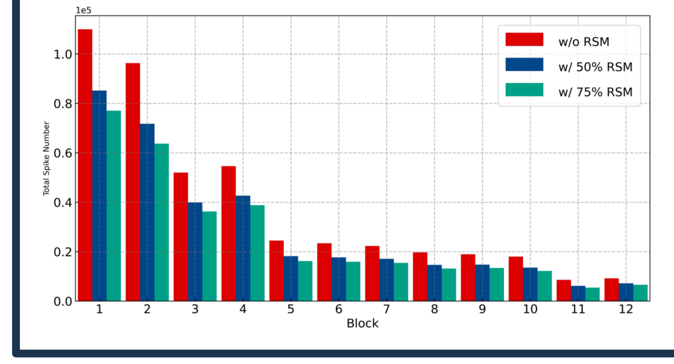


Conclusion

1) Implementing self-attention via ANN-to-SNN conversion achieves SOTA accuracy on CIFAR-10, CIFAR-100 and ImageNet, surpassing existing methods by 1.21%, 7.3% and 3.7%.



2) The MST model with RSM for randomly prunes input spike reduces power, while differentiating the layer-wise masking ratios can more effectively remove redundancy, as blocks have varying impact to overall performance.



Module	SC1	Q	K	A	V	soft(A)*V	Proj	SC2	MLP1	MLP2	Accuracy (%)
Masking Ratio (%)	0	0.5	0.5	0.5	0	0	0	0	0	0	86.7
	0	0.5	0.5	0.5	0.5	0	0	0	0	0	83.1
	0	0.5	0.5	0.5	0.5	0.5	0	0	0	0	80.5
	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	76.7
	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0.1	75.2
	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0.1	0	75.7
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0	74.46
	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0	0	62.36
	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0.5	0	68.67
	0	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0.5	67.91

Future work

1) **Investigating the methodology for reducing conversion errors at extremely low time steps could help improve the performance of converted SNNs.**
2) **Optimizing the masking policies could expand the acceptable masking range without compromising performance.**
3) **Applying the RSM method to directly train SNNs could further optimize energy use, making SNNs more practical for real-world deployment.**

Acknowledgements

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0682), HK RGC (Grant Nos. 27206321, 17205922, 17212923), NSFC (Grant No. 62122004), National Key R&D Program of China (Grant No. SQ2022YFB3600159). This research is also partially supported by ACCESS - AI Chip Center for Emerging Smart Systems, sponsored by Innovation and Technology Fund (ITF), Hong Kong SAR.