Rate Gradient Approximation Attack Threats Deep Spiking Neural Networks

Tong Bu, Jianhao Ding, Zecheng Hao, Zhaofei Yu*

Peking University

Code link

Paper link

Quick Summary

1. Converted SNN and surrogate-trained SNN are rate encoded.

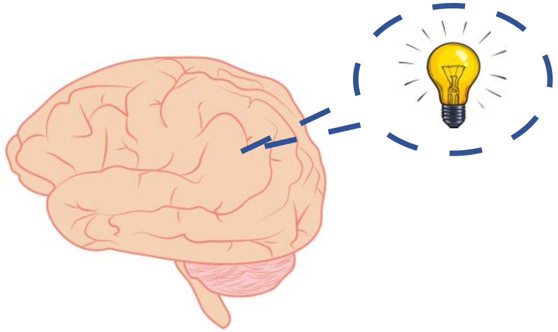
2. Rate Gradient Approximation Attack is a strong and robust attack.

3. SNNs composed of LIF cannot provide strong enough security.

Overview

What is Spiking Neural Network?

- Bio-inspired neural networks
- Convey Information through discrete spike train
- Discrete representation and event-driven



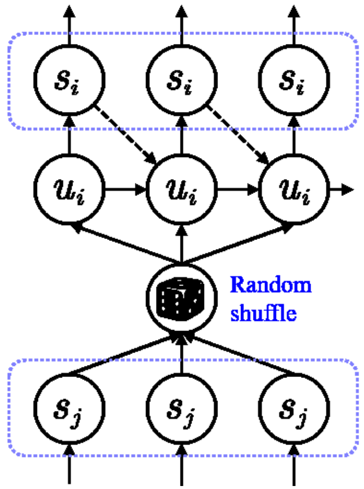
What is the motivation?

- When SNNs are applied to safety-critical systems, the reliability of SNNs should be a major concern. The adversarial attack is one of the most significant categories that threatens model security.
- Previous researches believed that SNNs have the natural ability to defense adversarial attacks (Nonlinearity of LIF neuron and sparsity of Poisson coding [2]).
- Do such spiking neural networks contain temporal information? Attacks may be performed over firing rates.
- For SNNs, the BPTT based attack through a surrogate function may give a false sense of security for SNNs.

Brainstorm

Does well-trained SNNs contain timing information?

- We random shuffled each neuron’s output spike firing order so that the spike trains contain no temporal information.
- We then compare whether the performance is influenced after spike shuffle.



Results

- On both CIFAR and DVS-CIFAR dataset, the performance will not degrade after spike shuffle
- We found that both converted SNNs and surrogate trained SNNs are rate encoded.

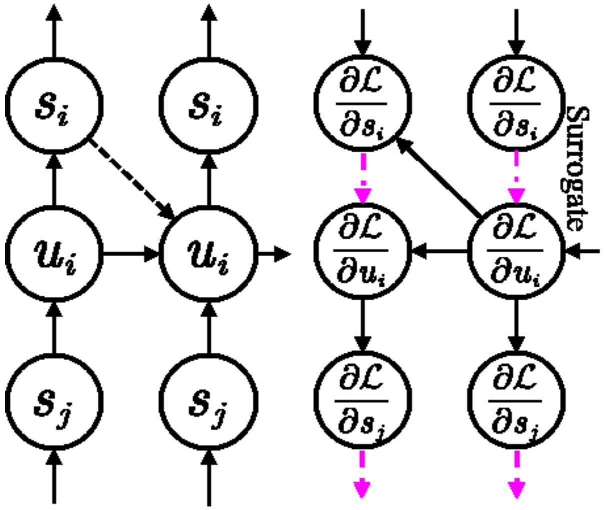
Table 1. Performance before and after the spike shuffle

Dataset	Training Method	T	λ	Reset	Clean Acc.	Shuffled Acc.	Rate
CIFAR-10	ANNSNN	16	1.0	soft	93.25	93.358	✓
CIFAR-10	STBP	8	1.0	soft	92.75	92.086	✓
CIFAR-10	STBP	8	1.0	hard	93.06	92.214	✓
CIFAR-10	STBP	8	0.9	hard	93.03	92.545	✓
CIFAR-10	STBP	8	0.5	hard	91.48	91.225	✓
CIFAR10-DVS	STBP	10	0.9	hard	77.00	75.400	✓

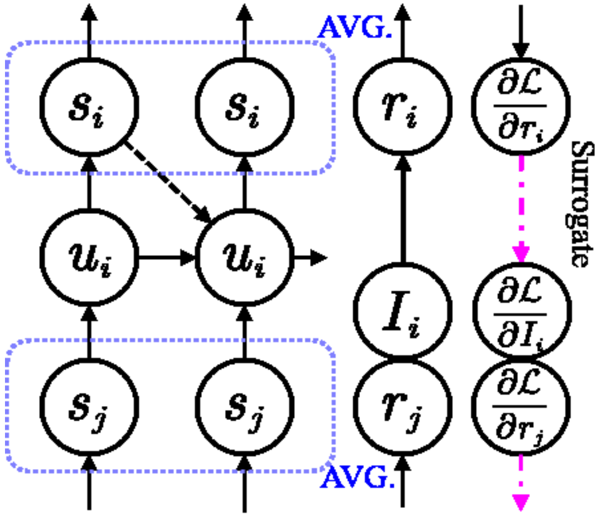
Method

Rate Gradient Approximation Attack

- Since well-trained SNNs are all rate-encoded at each layer, we can approximate the backward pass of SNNs using only the average firing rate over time-steps to generate effective gradients.



BPTT-based Gradient Backward [1]



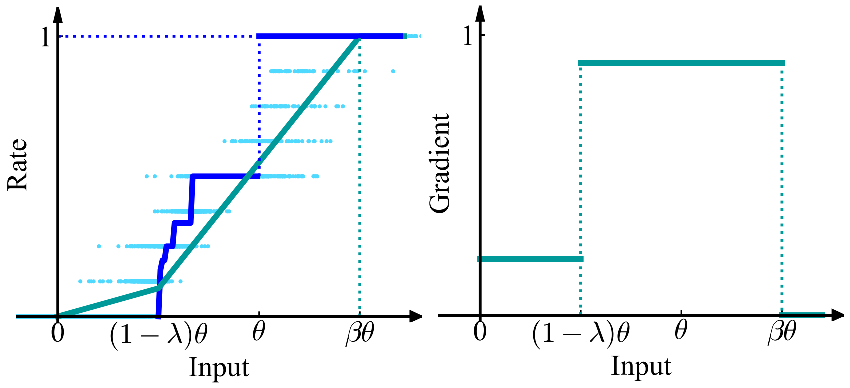
RGA-based Gradient Backward

Method

Rate Gradient Surrogate function

We use the static R-I curve, which refers to the relationship between the input current and output firing rate when the input is constant, as the approximation function.

$$\frac{\partial r_i}{\partial I_i} = \begin{cases} \gamma, & 0 \leqslant I_i \leqslant (1-\lambda)\theta \\ \frac{1-\gamma\theta+\gamma\theta\lambda}{(\beta+\lambda-1)\theta}, & (1-\lambda)\theta < I_i \leqslant \beta\theta \\ 0, & I_i > \beta\theta \text{ or } I_i < 0 \end{cases}$$



Surrogate function and deviation for LIF neuron

λ is the leaky parameter, β and γ are the smooth parameter which prevent the gradient to be zero or infinity. When λ is set to 1, this function will degenerate into the surrogate function for IF neurons.

Method

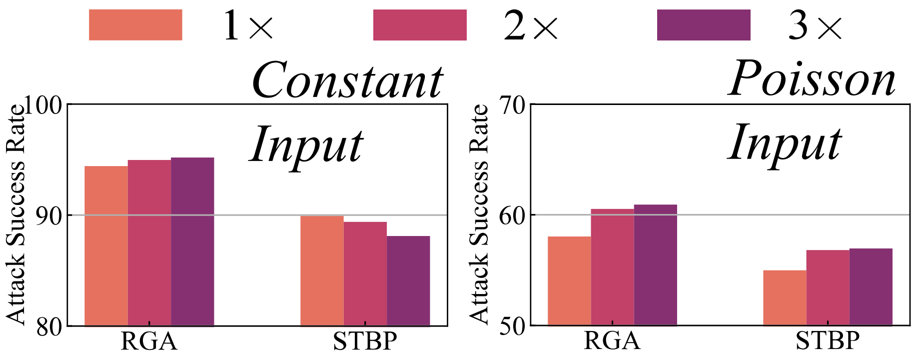
Time extended Attack

- Time Extended attack is to generate more effective adversarial samples by increasing the inference time of SNNs. Time Extended Attack can generate stronger adversarial examples.

Possion Attack

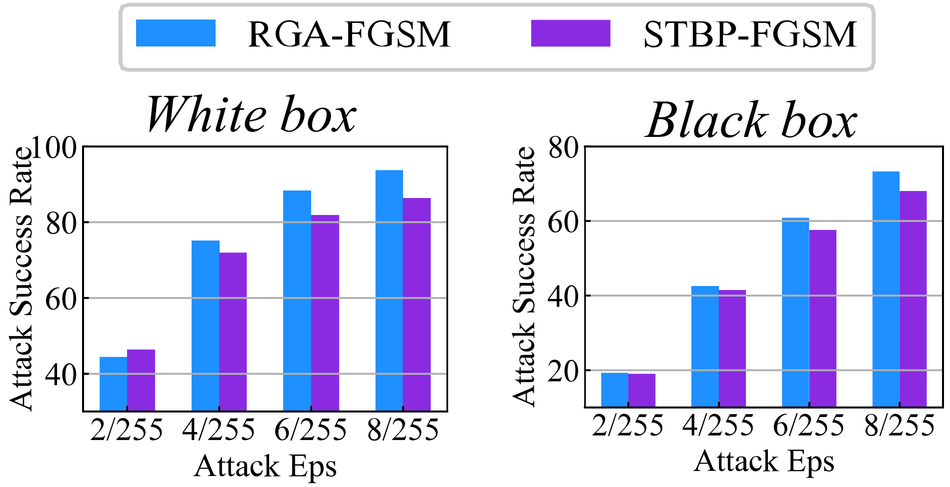
- For Poisson input SNN, we can regard it as a combined structure of a Poisson encoder and an end-to-end SNN receives spike input. We can consider the Poisson encoder as a random transformation and use a straight through estimator to attack this random transformation.

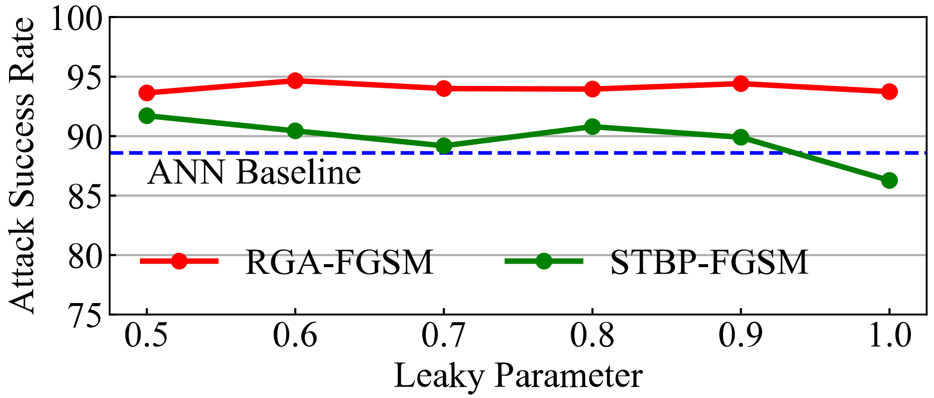
$$\frac{\partial \mathrm{Poisson}(x)}{\partial x} \approx \frac{\partial \mathbb{E}_x\left(\mathrm{Poisson}(x)\right)}{\partial x} = 1$$



Time extended attack increase the attack success rate

Experiments





The attack success rate change with respect to the attack strength for VGG-11 model on the CIFAR-10 dataset. The RGA attack is more effective than STBP attack.

The white box attack success rate changes with respect to the leaky parameter of the spiking neuron. This experiment is conducted on the CIFAR-10 dataset with VGG-11.

Experiments

Architecture	Dataset	Input	T	λ	TE	Attack	Clean Acc.	**White Box Attack**		**Black Box Attack**	
								ASR. (STBP)	ASR. (RGA)	ASR. (STBP)	ASR. (RGA)
VGG-11	CIFAR-10	Direct	8	1.0	-	FGSM	93.06	86.2777	**93.7352**	68.0314	**73.2646**
VGG-11	CIFAR-10	Direct	8	1.0	2×	FGSM	93.06	86.0735	**94.7346**	64.8399	**73.6192**
VGG-11	CIFAR-10	Direct	8	1.0	-	PGD	93.06	99.4949	**99.8281**	86.4604	**87.1266**
VGG-11	CIFAR-10	Direct	16	1.0	-	FGSM	93.03	85.3273	**92.4218**	65.7960	**73.1269**
VGG-11	CIFAR-10	Direct	16	1.0	-	PGD	93.03	99.3658	**99.8388**	85.5853	**87.4234**
VGG-11	CIFAR-10	Poisson	16	1.0	-	FGSM	86.72	54.9798	**58.0328**	40.8673	**44.2259**
VGG-11	CIFAR-10	Poisson	16	1.0	2×	FGSM	86.72	56.8106	**60.5296**	42.8440	**46.9085**
VGG-11	CIFAR-10	Poisson	16	1.0	-	PGD	86.72	51.9022	**57.1412**	37.0917	**41.1887**
VGG-11	CIFAR-10	Direct	8	0.5	-	FGSM	91.48	91.7140	**93.6270**	77.7656	**79.6458**
VGG-11	CIFAR-10	Direct	8	0.5	-	PGD	91.48	**99.8251**	99.7704	**93.6817**	93.0367
VGG-11	CIFAR-10	Direct	8	0.9	-	FGSM	93.03	89.9065	**94.4104**	73.4494	**77.2761**
VGG-11	CIFAR-10	Direct	8	0.9	-	PGD	93.03	99.7313	**99.8280**	**91.7661**	91.3899
ResNet-17	CIFAR-10	Direct	8	0.9	-	FGSM	93.04	84.2433	**92.9278**	67.1109	**80.1053**
ResNet-17	CIFAR-10	Direct	8	0.9	-	PGD	93.04	99.9248	**100.000**	92.0034	**97.5172**
VGG-11	CIFAR-100	Direct	8	0.9	-	FGSM	73.28	92.8766	**94.7189**	80.8952	**84.2658**
VGG-11	CIFAR-100	Direct	8	0.9	-	PGD	73.28	99.7544	**99.8499**	**92.2353**	92.0579
ResNet-17	CIFAR-100	Direct	8	0.9	-	FGSM	72.05	85.6627	**92.0611**	74.2956	**81.1936**
ResNet-17	CIFAR-100	Direct	8	0.9	-	PGD	72.05	99.5836	**99.8890**	87.6336	**95.2949**
VGG-11	CIFAR10-DVS	Frame	10	0.9	-	FGSM	77.00	59.5084	**59.5607**	**48.4967**	47.9275

Results of RGA based attack and STBP based attack on different type of SNNs. The better of the two is bolded.

Conclusion

- Benchmark for future research on SNN adversarial robustness.
- Lower time cost property showing potential on adversarial training.
- The current rate-coded SNN is not secure, highlighting the need for exploring SNNs utilizing complex neurons and other neuronal codings.

References

[1] HIRE-SNN: Harnessing the Adversarial Robustness of Energy-Efficient Deep Spiking Neural Networks via Training with Crafted Input Noise. *ICCV*. 2021.
[2] Inherent Adversarial Robustness of Deep Spiking Neural Networks: Effects of Discrete Input Encoding and Non-linear Activations. *ECCV*. 2020.

Thanks for your attention

Peking University

Code link

Paper link

Email: putong30@pku.edu.cn
yuzf12@pku.edu.cn