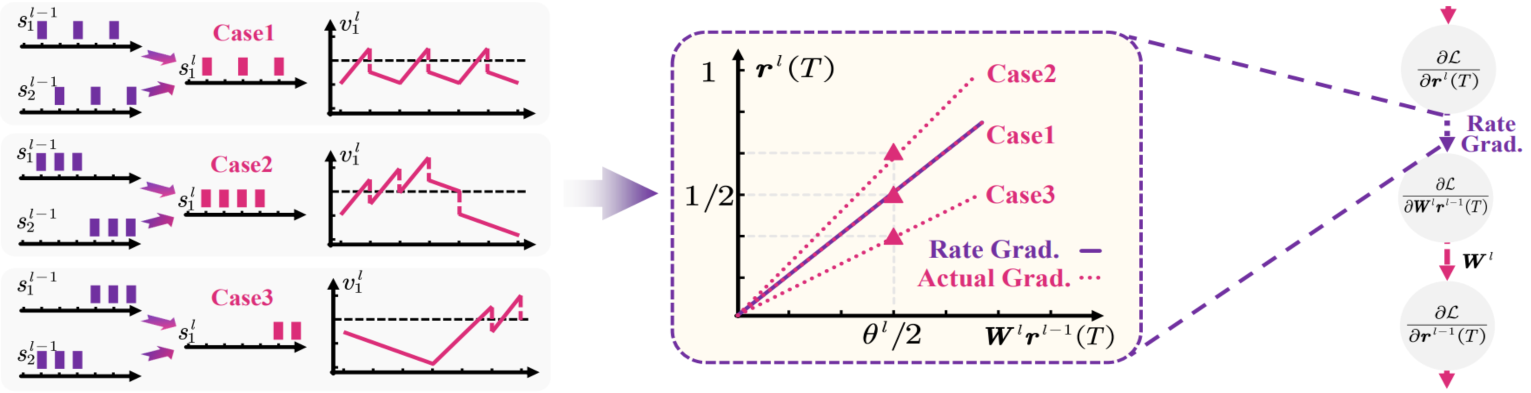
Threaten Spiking Neural Networks through Combining Rate and Temporal Information

Zecheng Hao, Tong Bu, Xinyu Shi, Zihan Huang, Zhaofei Yu, Tiejun Huang
Peking University

Rethinking Rate Information Gradient in SNNs

- **Rate information** in SNNs mainly refers to an approximate linear transformation relationship, similar to ANNs, between the average firing rate of adjacent layers.
- The same average input current corresponds to multiple different average firing rates
- Consider adversarial attacks similar to ANN gradient calculation mode



$$\boldsymbol{r}^l(T) = \boldsymbol{W}^l \boldsymbol{r}^{l-1}(T) - \frac{\boldsymbol{v}^l(T) + \sum_{t=1}^{T-1}(1-\lambda^l)\boldsymbol{v}^l(t)}{T}.$$

$$\boldsymbol{g}^l_{\text{rate}} = \left(\frac{\partial \boldsymbol{r}^l(T)}{\partial \boldsymbol{W}^l \boldsymbol{r}^{l-1}(T)}\right)_{\text{rate}} = \begin{cases} \mathbb{E}\left(\frac{\boldsymbol{r}^l(T)}{\boldsymbol{W}^l \boldsymbol{r}^{l-1}(T)}\right), & \boldsymbol{W}^l \sum_{t=1}^{T} \boldsymbol{s}^{l-1}(t) > 0 \\ 0, & \text{otherwise} \end{cases}.$$

Impact factors about the retention degree of temporal information in SNNs

$$\chi^l = \int_{-\infty}^{+\infty} \mathbf{Var}\left(\frac{\boldsymbol{r}^l(T)}{\boldsymbol{W}^l\boldsymbol{r}^{l-1}(T)} \middle| \boldsymbol{W}^l\boldsymbol{r}^{l-1}(T) = x \right) \mathbf{P}\left(\boldsymbol{W}^l\boldsymbol{r}^{l-1}(T) = x\right) \mathrm{d}x.$$

Measure the difference degree in spike firing sequences under the same average input current

Theorem 1. *If* $\boldsymbol{W}^l\boldsymbol{r}^{l-1}(T) \sim \mathbf{U}(-c,c)$*, for the soft-reset mechanism, we have* $\chi^l = \int_{-c}^{c} \frac{[(T-1)(1-\lambda^l)^2+1]h^2(x,\lambda^l)}{6cT^2x^2}\mathrm{d}x$*. Moreover, assuming* $h(x,\lambda^l) = ax+b$*, we will further have* $\chi^l = \frac{a^2c^2-b^2}{3c^2}\frac{(T-1)(1-\lambda^l)^2+1}{T^2}$.

1. Leakage degree of membrane potential
2. Time-steps
3. Input data types: static, neuromorphic

Table 1: Attack success rate of CBA and Ours under white-box attack.

Datasets	Time-steps	FGSM, λ=0.5	FGSM, λ=1.0	PGD, λ=0.5	PGD, λ=1.0
CIFAR-10	4	59.95/**86.42**	64.95/**90.28**	41.51/**99.08**	52.65/**98.89**
	8	60.40/**88.34**	71.76/**92.56**	42.13/**99.47**	67.94/**99.90**
CIFAR10-DVS	5	42.44/**49.92**	37.39/**55.80**	44.58/**55.57**	42.46/**62.90**
	10	36.05/**51.18**	45.39/**74.47**	38.95/**58.03**	54.74/**89.61**

Hybrid adversarial attack based on both rate and temporal information

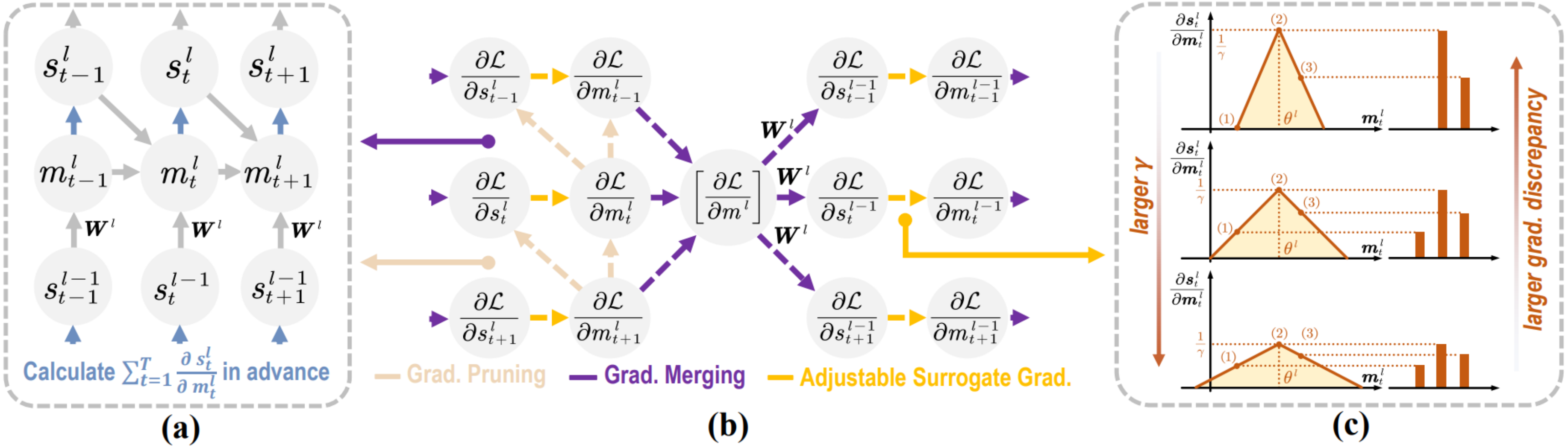


Figure 2: Overall algorithm framework for HART. (a): the property of pre-calculation, (b): back-propagation design, (c): adjustable temporal attribute.

$$\nabla_{\boldsymbol{W}^l}\mathcal{L} = \sum_{t=1}^{T}\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{m}^l}\right]\frac{\partial\boldsymbol{m}^l(t)}{\partial\boldsymbol{W}^l}, \quad \frac{\partial\mathcal{L}}{\partial\boldsymbol{s}^{l-1}(t)} = \left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{m}^l}\right]\frac{\partial\boldsymbol{m}^l(t)}{\partial\boldsymbol{s}^{l-1}(t)}.$$

$$\left[\frac{\partial\mathcal{L}}{\partial\boldsymbol{m}^l}\right] = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial\mathcal{L}}{\partial\boldsymbol{s}^l(t)}\frac{\partial\boldsymbol{s}^l(t)}{\partial\boldsymbol{m}^l(t)}.$$

Rate attribute: By pruning and merging gradients, we have:

$\mathbb{E}\left(\nabla_{\boldsymbol{W}^l}\mathcal{L}\right) = \left(\nabla_{\boldsymbol{W}^l}\mathcal{L}\right)_{rate}$ *and* $\mathbb{E}\left(\sum_{t=1}^{T}\frac{\partial\mathcal{L}}{\partial\boldsymbol{s}^{l-1}(t)}\right) = \left(\frac{\partial\mathcal{L}}{\partial\boldsymbol{r}^{l-1}(T)}\right)_{rate}$

Pre-calculation property: calculate $\sum_{t=1}^{T}\frac{\partial s_t^l}{\partial m_t^l}$ in advance to reduce the overhead of back-propagation from O(T) to O(1)

Hybrid adversarial attack based on both rate and temporal information

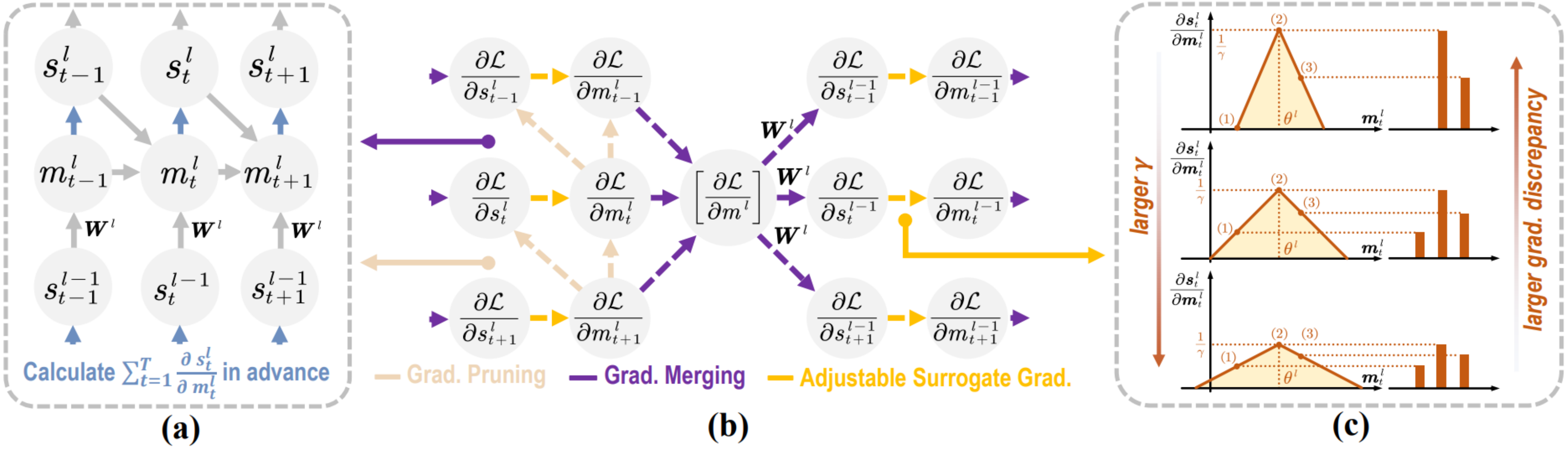
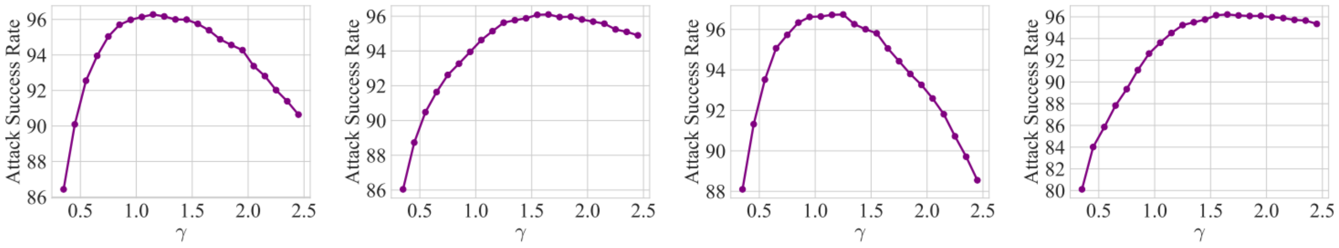


Figure 2: Overall algorithm framework for HART. (a): the property of pre-calculation, (b): back-propagation design, (c): adjustable temporal attribute.



(a) λ=0.5, FGSM (b) λ=1.0, FGSM (c) λ=0.5, PGD (d) λ=1.0, PGD

Figure 3: The performance of HART under different γ on CIFAR-10.

Temporal attribute: dynamically regulate the surrogate gradient curve through γ

Empirical principles for selecting γ:

1. a smaller γ corresponds to a gradient with more temporal attributes
2. ASR-γ curve approximately follows an unimodal distribution

Experiments: white-box attack

Table 2: Comparison between HART and previous works under white-box attack (WBA). * denotes robust target models.

Dataset	Architecture	λ	Clean Acc.	Attack	**CBA**	**BPTR**	**STBP**	**RGA**	**Ours**
CIFAR-10	VGG-11	0.5	91.48	FGSM	60.40	82.67	91.71	93.63	**96.28**
				PGD	42.13	99.21	99.95	99.92	**100.00**
		0.9	93.03	FGSM	70.58	88.36	89.91	94.41	**97.24**
				PGD	55.29	99.45	99.94	99.97	**99.98**
		0.9*	89.99	FGSM	25.49	41.77	55.41	56.76	**58.70**
				PGD	20.77	61.45	78.55	74.42	**83.54**
		1.0	93.06	FGSM	71.76	88.76	86.28	93.74	**96.22**
				PGD	67.94	99.63	99.70	99.94	**99.97**
	ResNet-17	0.9	93.04	FGSM	44.29	85.06	84.24	92.93	**94.80**
				PGD	29.76	99.86	99.91	100.00	**100.00**
CIFAR-100	VGG-11	0.9	73.28	FGSM	83.73	92.47	92.88	94.72	**96.06**
				PGD	82.91	99.59	99.86	99.92	**99.96**
		0.9*	67.21	FGSM	32.69	57.19	70.42	70.24	**72.41**
				PGD	27.57	71.98	86.56	83.35	**87.68**
	ResNet-17	0.9	72.05	FGSM	65.34	86.94	85.66	92.06	**94.54**
				PGD	45.17	99.65	99.69	99.90	**99.96**
CIFAR10-DVS	VGG-DVS	0.5	76.0	FGSM	36.05	50.39	59.08	53.95	**61.05**
				PGD	38.95	60.00	71.05	62.11	**74.08**
		1.0	76.0	FGSM	45.39	69.74	76.97	76.05	**78.42**
				PGD	54.74	87.11	92.63	89.08	**93.03**

Experiments: black-box attack

Table 3: Comparison between HART and previous works under black-box attack (BBA). * denotes robust target models.

Dataset	Architecture	λ	Clean Acc.	Attack	**CBA**	**BPTR**	**STBP**	**RGA**	**Ours**
CIFAR-10	VGG-11	0.5	91.48	FGSM	43.04	63.44	77.77	79.65	**82.68**
				PGD	23.50	84.21	95.99	95.36	**96.74**
		0.9	93.03	FGSM	43.45	66.72	73.45	77.28	**85.82**
				PGD	23.98	84.72	95.04	94.69	**97.62**
		0.9*	89.99	FGSM	14.08	25.26	35.83	35.44	**38.26**
				PGD	10.63	31.10	46.06	44.42	**47.83**
		1.0	93.06	FGSM	43.28	64.25	68.03	73.26	**80.34**
				PGD	24.75	80.55	90.91	91.36	**96.22**
	ResNet-17	0.9	93.04	FGSM	36.07	69.53	67.11	80.11	**84.95**
				PGD	15.57	93.72	94.30	98.36	**99.28**
CIFAR-100	VGG-11	0.9	73.28	FGSM	68.33	80.10	80.90	84.27	**88.51**
				PGD	42.45	88.91	93.65	93.91	**97.32**
		0.9*	67.21	FGSM	22.59	37.58	47.20	47.94	**50.78**
				PGD	18.24	41.73	54.40	54.78	**57.66**
	ResNet-17	0.9	72.05	FGSM	61.22	75.65	74.30	81.19	**85.31**
				PGD	32.59	91.07	89.13	95.66	**98.06**
CIFAR10-DVS	VGG-DVS	0.5	76.0	FGSM	34.87	44.08	47.89	48.55	**49.74**
				PGD	35.13	47.63	50.53	50.92	**53.16**
		1.0	76.0	FGSM	43.03	62.50	66.32	65.79	**69.74**
				PGD	52.11	70.92	76.45	75.66	**78.03**

Experiments: time-steps & perturbation degrees

Table 4: Attack success rate for STBP/RGA/HART with different time-steps on CIFAR-10/VGG-11.

λ	Time-steps	FGSM, WBA	FGSM, BBA	PGD, WBA	PGD, BBA
0.5	4	90.07/93.24/**95.68**	76.22/78.52/**80.10**	99.88/99.85/**99.98**	94.59/94.21/**94.96**
	8	91.71/93.63/**96.28**	77.77/79.65/**82.68**	99.92/99.92/**100.00**	95.99/95.36/**96.74**
	16	91.86/93.48/**95.82**	77.49/79.66/**83.49**	99.95/99.91/**99.99**	96.12/95.98/**97.29**
1.0	4	81.89/91.03/**92.67**	65.52/71.24/**76.43**	99.17/99.23/**99.40**	87.48/89.37/**92.71**
	8	86.28/93.74/**96.22**	68.03/73.26/**80.34**	99.70/99.94/**99.97**	90.91/91.36/**96.22**
	16	87.49/95.24/**96.65**	66.89/75.07/**81.41**	99.88/99.97/**99.99**	90.86/92.67/**97.14**

Table 5: Attack success rate for STBP/RGA/HART with different perturbation degrees on CIFAR-10/VGG-11.

λ	ϵ	FGSM, WBA	FGSM, BBA	PGD, WBA	PGD, BBA
0.5	2/255	49.15/45.76/**55.91**	24.67/22.87/**26.41**	66.32/62.08/**78.33**	29.30/28.42/**30.50**
	4/255	76.30/76.86/**83.06**	51.28/50.05/**54.31**	96.99/95.14/**98.95**	69.43/68.12/**71.54**
	8/255	91.71/93.63/**96.28**	77.77/79.65/**82.68**	99.92/99.92/**100.00**	95.99/95.36/**96.74**
1.0	2/255	46.41/44.46/**46.76**	19.19/19.62/**21.89**	**65.58**/61.44/65.26	21.89/21.96/**24.75**
	4/255	71.82/75.17/**78.56**	41.48/42.76/**47.80**	95.28/95.27/**96.39**	57.29/56.78/**64.08**
	8/255	86.28/93.74/**96.22**	68.03/73.26/**80.34**	99.70/99.94/**99.97**	90.91/91.36/**96.22**

Discussion & Conclusion

- We revisit the gradient calculation mode based on average spike firing rate, and quantitatively analyzed the retention degree of temporal information in SNNs.
- We propose a hybrid attack framework based on two types of information and analyze its **rate and temporal attributes**. We point out that the **pre-calculation property** of this framework and **empirical rules for determining gamma** can further reduce the computational overhead.
- Our method achieves state-of-the-art attack success rate (ASR) across various hyper-parameter settings for both static and neuromorphic datasets.

Thanks for Listening!