Appendix: On the Role of Temporal Granularity in the Robustness of Spiking Neural Networks

Supplementary Material

1. More Related Work

Currently, mainstream attack methods against deep neural networks (DNNs) can be broadly categorized into two types. The first category comprises gradient-based attack methods. For instance, FGSM [2] perturbs the original data along the sign direction of the gradient of the loss function in a single step, increasing the perturbed linear output to mislead the network. BIM [3] and PGD [4] are iterative attacks. Unlike PGD, BIM updates adversarial samples starting from the original image. The second category includes attack methods based on model loss optimization. Unlike gradient-based attacks, the CW attack [1] relies on model optimization to generate adversarial examples.

In the context of SNNs, researchers typically utilize the temporal *average gradient* of the network to adapt these attack methods effectively. In contrast, this work introduces a temporal granularity attack, which leverages the *gradient at each individual time step*.

2. Detailed Proof of Theorem 1

Proof:

The KL divergence between $f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)$ and $f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon)$ is given by:

$$\begin{aligned}&\mathrm{KL}(f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\|f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon)) = \\ &\mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}[\log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) - \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon)].\end{aligned} \tag{1}$$

Using a second-order Taylor expansion, we approximate:

$$\begin{aligned}\log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon) \approx &\log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) + \epsilon^\top \nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) \\ &+ \frac{1}{2}\epsilon^\top H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\epsilon,\end{aligned} \tag{2}$$

yielding:

$$\begin{aligned}&\mathrm{KL}(f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\|f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon)) \approx \\ &\mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}[-\epsilon^\top \nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) - \frac{1}{2}\epsilon^\top H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\epsilon].\end{aligned} \tag{3}$$

Since the total probability over all possible outcomes must sum to one:

$$\sum_{y_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = 1. \tag{4}$$

Then:

$$\sum_{y_t} \nabla_{s_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = 0. \tag{5}$$

Meanwhile:

$$\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}. \tag{6}$$

Therefore,

$$f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t), \tag{7}$$

and,

$$\sum_{y_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \sum_{y_t} \nabla_{s_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = 0. \tag{8}$$

Finally we have:

$$\begin{aligned}&\mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}\left[\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\right] = \\ &\sum_{y_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = 0.\end{aligned} \tag{9}$$

Then the KL divergence simplifies to:

$$\begin{aligned}&\mathrm{KL}(f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\|f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t+\epsilon)) \approx \\ &\frac{1}{2}\epsilon^\top \mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}\left[-H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\right]\epsilon.\end{aligned} \tag{10}$$

3. Optimization of the Temporal Sensitivity Value

The temporal sensitivity value $TSV(\boldsymbol{s}_t)$ is defined as:

$$TSV(\boldsymbol{s}_t) = \mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}\left[-H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\right]. \tag{11}$$

Expansion of $TSV(\boldsymbol{s}_t)$.

The gradient of the log-likelihood is:

$$\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}. \tag{12}$$

The Hessian of the log-likelihood is:

$$H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \nabla_{\boldsymbol{s}_t}\left(\frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}\right). \tag{13}$$

which simplifies to:

$$\begin{aligned}&H_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) = \\ &\frac{H_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} - \frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^2}.\end{aligned} \tag{14}$$

Thus, the temporal sensitivity value $TSV(\boldsymbol{s}_t)$ becomes:

$$TSV(\boldsymbol{s}_t) = \mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} \left[\frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) \nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^2} - \frac{H_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} \right]. \quad (15)$$

Simplified approximation form.

When the first-order gradient exists and is non-zero (the first term of Equation (15)), it dominates the calculation. Thus, the second-order gradient (Hessian) is ignored to reduce computational cost. In practical applications, we typically approximate by disregarding the second term and retaining only the squared expectation of the gradient term. Meanwhile:

$$\begin{aligned} &\frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) \nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^2} = \\ &\frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} \cdot \frac{\nabla_{\boldsymbol{s}_t} f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top}{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} = \\ &\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) \nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top. \end{aligned} \quad (16)$$

Therefore, a simplified approximation of $TSV(\boldsymbol{s}_t)$ is:

$$S(\boldsymbol{s}_t) \approx \mathbb{E}_{f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)} \left[\nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t) \nabla_{\boldsymbol{s}_t} \log f_{\boldsymbol{w}}(y_t|\boldsymbol{s}_t)^\top \right]. \quad (17)$$

4. Sensitivity to Parameter λ

We set $\lambda = 100$ in our paper, the ablation study of λ is shown in Table 1. It can be seen that $\lambda = 100$ both achieves the best robustness and remarkable original accuracy.

Table 1. The sensitivity to regularization parameter. The dataset is CIFAR10 with VGG11 on $T = 4$. The attack is PGD.

λ	50	80	100	150	200
ACC	92.40	91.90	90.89	88.32	85.23
Robustness	0.31	0.23	4.25	4.04	4.00

5. Practical Motivation and Applicability of TG-Attack

(1) White-box model. In this setting, attacker has access to model architecture, parameters, and input encoding, which is standard in robustness evaluation for ANNs and SNNs. Gradients with respect to the encoded spike sequence are directly available. TG-Attack perturbs each time step using its own gradient rather than averaged one, thus yielding stronger attacks, indicating that temporal gradients reveal vulnerability patterns obscured by temporal aggregation. **(2) Sequence-based and event-based inputs.** For datasets such as DVS-CIFAR10, the temporal sequence is the native input. Time-step-level perturbations are therefore meaningful, while gradient averaging is suboptimal. TG-Attack exploits temporal sensitivities and outperforms standard PGD by up to 5% on DVS-CIFAR10.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[3] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017. 1

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1