# Improving Generalization and Robustness in SNNs Through Signed Rate Encoding and Sparse Encoding Attacks

Bhaskar Mukhoty, Hilal AlQuabeh, Bin Gu

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

## A Spiking Neuron

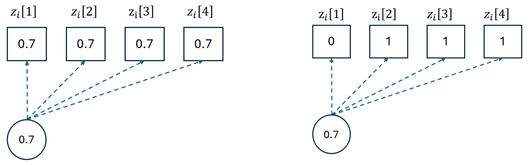
- The neuron cell membrane accumulates the input spikes over time and sends the output spike whenever the potential exceeds a predetermined threshold.
- A reset follows the spiking activity in the membrane potential. The Heaviside function abstracts a spike $s_i^{(l)}[t]$, in the $i$-th neuron at $l$-th layer at time-step $t$.

$$u_i^{(l)}[t] = \beta(u_i^{(l)}[t-1] - s_i^{(l)}[t-1]u_{th}) + \sum_j w_{ij} s_j^{(l-1)}[t],$$

$$s_i^{(l)}[t] = H(u_i^{(l)}[t] - u_{th}) = \begin{cases} 1 & \text{if } u_i^{(l)}[t] > u_{th} \\ 0 & \text{otherwise} \end{cases}$$

## Input Encoding: Constant (a.k.a. direct) vs. Rate

- An SNN takes input over T time steps. The constant encoding replicates the input for T times.
- In rate, input $x \in [0,1]$ is encoded by T independent Bernoulli samples.




- Rate-encoded SNNs are more robust to adversarial perturbation than constant-encoded SNNs, while the latter offers higher clean accuracy.
- Since randomness in rate encoding introduces the accuracy-robustness tradeoff, we intend to reduce the randomness in rate encoding.

## Adversarial Attack

- Given an input $\mathbf{x}$, an untargeted adversarial attack finds a perturbation $\boldsymbol{\delta}$, such that the classifier output is altered, i.e., $h(\mathbf{x}) \neq h(\mathbf{x}+\boldsymbol{\delta})$. Maximizing the loss over the input perturbations: $\max_{\|\boldsymbol{\delta}\| \leq \epsilon} \mathcal{L}(h_\theta(\mathbf{x}+\boldsymbol{\delta}), y)$ can give us such perturbations.
- As perturbation in the input must pass through the encoding space, attack in the encoding spaces is supposed to be stronger. We explore sparse adversarial attack in the encoding space $\mathbf{z} \in \{0,1\}^{d \times T}$, instead of the input space $\mathbf{x} \in \{0,1\}^d$.

## Contributions

- We proposed Signed Rate Encoding, which effectively reduces randomness introduced by Rate Encoding and helps improve clean accuracy.
- We introduce Sparse Encoding Attack (SEA) that performs a sparse adversarial attack on encoding space that jointly optimizes domain and sparsity constraints. We establish theoretical connections with input space attacks to compare the budgets.
- Adversarial training with SEA improves the robustness of the rate-encoded SNNs. The empirical results reveal that SEA under signed rate encoding offers significantly higher robust accuracy than the existing methods.

## Signed Rate Encoding

Allows input $x_i \in [-1,1]$ and generates Signed Bernoulli spike according to the sign of the input.

$$z_i = \text{sBer}(x_i) := \begin{cases} \text{Ber}(x_i) & \text{if } x_i \geq 0 \\ -\text{Ber}(-x_i) & \text{if } x_i < 0 \end{cases} \tag{1}$$

so that, $z_i \in \{-1, 0, 1\}$, with,

$$\mathbb{P}(z_i = 1) = x_i \mathbb{I}[x_i \geq 0] = x_i^+, \qquad \mathbb{P}(z_i = -1) = -x_i \mathbb{I}[x_i < 0] = x_i^-$$
$$\mathbb{P}(z_i = 0) = 1 - x_i^+ - x_i^- = 1 - |x_i| \tag{2}$$

## Reduction in Randomness

An input image $\mathbf{x} \in [0,1]^d$, **Bernoulli** encoded twice independently at a particular time step, i.e., $\mathbf{z}, \hat{\mathbf{z}} \sim Ber(\mathbf{x})$, where, $\mathbf{z}, \hat{\mathbf{z}} \in \{0,1\}^d$, we have,

$$\mathbb{E}_{\mathbf{z},\hat{\mathbf{z}}}[\|\mathbf{z} - \hat{\mathbf{z}}\|_0] = \sum_{i=1}^d \mathbb{E}_{z_i,\hat{z}_i \sim Ber(x_i)}[|z_i - \hat{z}_i|] = 2\sum_{i=1}^d x_i(1-x_i) = 2\langle \mathbf{x}, 1-\mathbf{x} \rangle =: k_1(\mathbf{x}) \tag{3}$$

In contrast, input $\mathbf{x} \in [-1,1]^d$, **Signed Bernoulli** encoded twice independently $\mathbf{z}, \hat{\mathbf{z}} \sim \text{sBer}(\mathbf{x})$, with, $\mathbf{z}, \hat{\mathbf{z}} \in \{-1,0,1\}^d$, we have,

$$\begin{aligned} \mathbb{E}_{\mathbf{z},\hat{\mathbf{z}} \sim \text{sBer}(\mathbf{x})}[\|\mathbf{z} - \hat{\mathbf{z}}\|_0] &= \sum_{i=1}^d \mathbb{E}_{z_i,\hat{z}_i \sim \text{sBer}(x_i)}[\|z_i - \hat{z}_i\|_0] = \sum_{i=1}^d \mathbb{P}(z_i \neq \hat{z}_i) = \sum_{i=1}^d (1 - \mathbb{P}(z_i = \hat{z}_i)) \\ &= \sum_{i=1}^d (1 - \mathbb{P}(z_i = 1, \hat{z}_i = 1) - \mathbb{P}(z_i = -1, \hat{z}_i = -1) - \mathbb{P}(z_i = 0, \hat{z}_i = 0)) \\ &= \sum_{i=1}^d \left(1 - (x_i^+)^2 - (x_i^-)^2 - (1-|x_i|)^2\right) = 2\langle |\mathbf{x}|, \mathbf{1} - |\mathbf{x}| \rangle =: k_2(\mathbf{x}) \end{aligned} \tag{4}$$

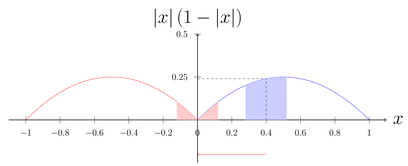



Figure 1. Signed Bernoulli reducing randomness after mean-centering of CIFAR-10 dataset.

Let $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i \in [0,1]^d$, we compute the factor $k_3(\mathbf{x}) = \frac{k_1(\mathbf{x})}{k_2(\mathbf{x}-\boldsymbol{\mu})}$, representing the factor by which the randomness reduces by mean centering of a particular input $\mathbf{x}$.

| | CIFAR-10 | CIFAR-100 | SVHN | ImageNet-100 |
|---|---|---|---|---|
| $d$ | $32 \times 32 \times 3$ | $32 \times 32 \times 3$ | $32 \times 32 \times 3$ | $224 \times 224 \times 3$ |
| Avg. $k_1(\mathbf{x})$ | 994.18 | 953.09 | 1106.59 | 54701.22 |
| Avg. $k_2(\mathbf{x}-\mu)$ | 942.39 | 964.14 | 830.33 | 43998.53 |
| Avg. $k_3(\mathbf{x})$ | 1.11 | 1.05 | 1.44 | 1.366 |

Table 1. Randomness reduction factor ($k_3$) due to signed rate encoding averaged over training data points.

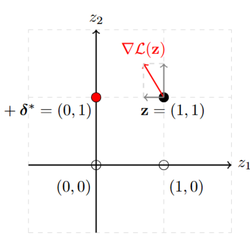
## Sparse Encoding Attacks

With gradient $\mathbf{g} := \nabla_{\mathbf{z}} \mathcal{L}(h_\theta(\mathbf{z}+\boldsymbol{\delta}), y) \in \mathbb{R}^{d \times T}$, and $\boldsymbol{\delta} = [\boldsymbol{\delta}^{(1)}, \boldsymbol{\delta}^{(2)}, \ldots, \boldsymbol{\delta}^{(T)}]$, we perform untargeted $l_0$ adversarial attack on the rate encoding $\mathbf{z} \in \{0,1\}^{d \times T}$, where at most $k$ coordinates can change in each frame:

$$\boldsymbol{\delta}^* := \underset{\boldsymbol{\delta}: \forall t \, \|\boldsymbol{\delta}^{(t)}\|_0 \leq k, \mathbf{z}^{(t)} + \boldsymbol{\delta}^{(t)} \in \{0,1\}^d}{\arg\max} \langle \boldsymbol{\delta}, \mathbf{g} \rangle \tag{5}$$

The restriction $\mathbf{z}^{(t)} + \boldsymbol{\delta}^{(t)} \in \{0,1\}^d$ implies $\boldsymbol{\delta}^{(t)} \in \{-1,0,1\}^d$. We first solve the problem without the sparsity constraint:

$$\hat{\boldsymbol{\delta}} := \underset{\boldsymbol{\delta}: \forall t \, \mathbf{z}^{(t)} + \boldsymbol{\delta}^{(t)} \in \{0,1\}^d}{\arg\max} \langle \boldsymbol{\delta}, \mathbf{g} \rangle \tag{6}$$



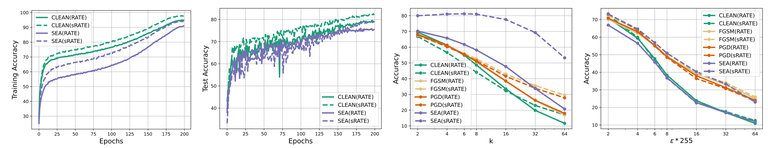

where, the solution can be given as:

$$\hat{\delta}_i^{(t)} = \begin{cases} 1 & g_i^{(t)} > 0, z_i^{(t)} = 0 \\ -1 & g_i^{(t)} < 0, z_i^{(t)} = 1 \\ 0 & \text{otherwise} \end{cases}$$

| $\text{sign}(g_i)$ | -1 | -1 | 1 | 1 | 0 | 0 |
|---|---|---|---|---|---|---|
| $z_i$ | 0 | 1 | 0 | 1 | 0 | 1 |
| $\hat{\delta}_i$ | 0 | -1 | 1 | 0 | 0 | 0 |

Solution for Eqn. (5) can be found by choosing the top k no-zero coordinates of $\hat{\boldsymbol{\delta}}^{(t)}$, according to the magnitude $\hat{\delta}_i^{(t)} g_i^{(t)}$. Similar results can be obtained for Signed Rate Encoding.

## Empirical Evaluation




(a) Train Acc. vs. Epoch (b) Test Acc. vs. Epoch (c) SEA attack vs. $k$ (d) PGD attack vs. $\epsilon$

Figure 2. Comparison of training and test accuracies and sensitivity of attack strength for SEA and PGD.

| T=4 | CIFAR-10, Rate Encoding | | | | | CIFAR-10, Signed Rate Encoding | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Attack | Clean | GN | FGSM | PGD | SEA | Clean | GN | FGSM | PGD | SEA |
| clean | 78.79 | 78.87 | 75.54 | 77.07 | 75.3 | **82.05** | 82.01 | 78.51 | 79.78 | 79.06 |
| gn | 79.23 | 78.58 | 75.92 | 77.51 | 75.12 | **82.13** | 81.83 | 78.26 | 79.7 | 78.01 |
| fgsm | 44.83 | 44.35 | 53.93 | 53.16 | 42.31 | 44.3 | 43.59 | 55.1 | 53.81 | **56.58** |
| pgd | 38.87 | 38.41 | 49.68 | 48.83 | 37.19 | 36.88 | 37.03 | 50.21 | 48.47 | **51.87** |
| sea, k=10 | 44.31 | 43.89 | 47.37 | 46.94 | 55.33 | 39.61 | 39.94 | 49.05 | 48.07 | **80.18** |
| sea, k=20 | 28.74 | 28.15 | 34.67 | 34.52 | 43.16 | 29.06 | 29.23 | 40.36 | 38.75 | **75.77** |
| Avg | 52.46 | 52.04 | 56.19 | 56.34 | 54.74 | 52.34 | 52.27 | 58.58 | 58.10 | **70.25** |
| T=4 | SVHN, Rate Encoding | | | | | SVHN, Signed Rate Encoding | | | | |
| clean | 85.66 | 85.84 | 85.87 | 86.04 | 85.01 | 89.44 | 89.59 | 89.64 | **89.87** | 83.98 |
| gn | 85.44 | 85.73 | 85.97 | 85.65 | 85.01 | 89.14 | 89.38 | 89.44 | **89.53** | 81.93 |
| fgsm | 44.09 | 44.15 | 50.26 | 48.33 | 41.94 | 43.37 | 44.06 | 49.94 | 48.23 | **67.57** |
| pgd | 38.82 | 38.77 | 45.03 | 43.31 | 35.82 | 35.46 | 35.43 | 42.45 | 40.53 | **65.76** |
| sea, k=10 | 37.81 | 37.61 | 38.13 | 37.89 | 54.06 | 34.62 | 35.10 | 35.26 | 35.13 | **92.27** |
| sea, k=20 | 21.46 | 21.75 | 22.36 | 22.63 | 32.66 | 23.55 | 23.72 | 24.93 | 24.97 | **87.21** |
| Avg | 52.21 | 52.31 | 54.60 | 53.98 | 55.75 | 52.60 | 52.88 | 55.28 | 54.71 | **79.79** |

Table 2. A comparison of adversarially trained models (in columns) and attacks (in rows). Adv. training with SEA attack under sRate, offers superior robust accuracy across all attacks.