Rethinking SNN Online Training and Deployment: Gradient-Coherent Learning via Hybrid-Driven LIF Model

Supplementary Material

1. Experimental configuration

For experimental cases in Tabs.1-3, we choose Stochastic Gradient Descent [1] as our optimizer and Cosine Annealing [6] as our scheduler. The initial learning rate , weight decay and batch-size are respectively set to 0.025, 0 and 256 for ImageNet-1k, 0.01, 5×10^{-4} and 64 for other cases. Each case is executed once with 300 training epochs and a specific random seed. We consider various data augmentation techniques, including Auto-Augment [2], Cutout [3] and Mixup [7].

Specifically, the HD-LIF cases related to ImageNet-1k and Tab.2 utilize 1.5-bit weight compression [5], while the remaining cases use 1-bit scheme [4]. The evaluation of performance metrics in Tab.2 is completed on NVIDIA A100-SXM4-40GB. The calculation of SOPs(M) and Energy(mJ) in Tab.2 refers to [8] and is oriented towards convolutional layers.

2. Proof of Theorem

Theorem 4.2 *For HD-LIF model under the condition of online training, combining with Definition 4.1, we will have:* $\forall t,i \in [1,T], t<i, \boldsymbol{\epsilon}^l[i,t] = \boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]$, *here we define* $\boldsymbol{\chi}^l[i,i] = \frac{\partial \mathbf{s}_i^l}{\partial \mathbf{m}_i^l} \in \{0,1\}, \boldsymbol{\chi}^l[j,j-1] = \frac{\partial \mathbf{m}_j^l}{\partial \mathbf{m}_{j-1}^l} \in \{0, \lambda_j^l\}$.
(i) If $\exists k \in [t,T], \boldsymbol{\chi}^l[k,k] = 1$, *then we can directly derive the gradient relationship between STBP and online training:* $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{Online} = \frac{\boldsymbol{\chi}^l[t,t]}{\left(\boldsymbol{\chi}^l[t,t]+\sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]\right)}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP}$.
(ii) Under more general conditions, $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP} \in \left\{1, \prod_{j=t+1}^{t^*}\lambda_j^l\right\}\cdot\left(\frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\right)$, *here* $\boldsymbol{\chi}^l[t^*,t^*] = 1 \wedge \forall k \in [t,t^*), \boldsymbol{\chi}^l[k,k] = 0$.

Proof. According to Eqs.(2-3), we can separately write $\frac{\partial \mathbf{s}_t^l}{\partial \mathbf{m}_t^l}, \frac{\partial \mathbf{m}_t^l}{\partial \mathbf{m}_{(t-1)}^l}$ of the HD-LIF model at the t-th time step for both firing and silence states:

$$\begin{aligned}\boldsymbol{\chi}^l[t,t] &= \frac{\partial \mathbf{s}_t^l}{\partial \mathbf{m}_t^l} = \begin{cases}1, & \mathbf{m}_t^l \geq \theta_t^l \\ 0, & \text{otherwise}\end{cases}, \\ \boldsymbol{\chi}^l[t,t-1] &= \frac{\partial \mathbf{m}_t^l}{\partial \mathbf{m}_{(t-1)}^l} = \lambda_t^l + \frac{\partial \mathbf{m}_t^l}{\partial \mathbf{s}_{(t-1)}^l}\frac{\partial \mathbf{s}_{(t-1)}^l}{\partial \mathbf{m}_{(t-1)}^l} \\ &= \begin{cases}0, & \mathbf{m}_{(t-1)}^l \geq \theta_{(t-1)}^l \\ \lambda_t^l, & \text{otherwise}\end{cases}.\end{aligned} \tag{S1}$$

For (i), combining Eq.(S1) with Definition 4.1, we can rewrite the relationship between $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP}$ and $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{Online}$ in the form of $\forall i \in [t,T]$, $\boldsymbol{\chi}^l[i,i], \boldsymbol{\chi}^l[i+1,i]$:

$$\begin{aligned}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP} &= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\sum_{i=t}^{T}\boldsymbol{\epsilon}^l[i,t] \\ &= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\left(\boldsymbol{\chi}^l[t,t] + \sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]\right).\end{aligned} \tag{S2}$$

As $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{Online} = \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\boldsymbol{\chi}^l[t,t]$ and $\exists k \in [t,T], \boldsymbol{\chi}^l[k,k] = 1$, if $\boldsymbol{\chi}^l[t,t] = 1$, obviously $\boldsymbol{\chi}^l[t,t] + \sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1] > 0$; otherwise, we choose $k \in [t+1,T]$, s.t. $\forall i \in [t,k), \boldsymbol{\chi}^l[i,i] = 0 \wedge \boldsymbol{\chi}^l[k,k] = 1$, then one can note that $\boldsymbol{\chi}^l[t,t] + \sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1] \geq \boldsymbol{\chi}^l[k,k]\prod_{j=t+1}^{k}\boldsymbol{\chi}^l[j,j-1] = \prod_{j=t+1}^{k}\lambda_j^l > 0$. Therefore, we can finally have $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{Online} = \frac{\boldsymbol{\chi}^l[t,t]}{\left(\boldsymbol{\chi}^l[t,t]+\sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]\right)}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP}$.

For (ii), when $\boldsymbol{\chi}^l[t,t] = 1$, we will have $\boldsymbol{\chi}^l[t+1,t] = 0$ according to Eq.(S1). Based on (i), we can further derive $\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{Online} = \frac{1}{\left(1+\sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+2}^{i}\boldsymbol{\chi}^l[j,j-1]\cdot 0\right)}\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP} = \left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP}$.

When $\boldsymbol{\chi}^l[t,t] = 0$, we have $\forall k \in [t,t^*), \boldsymbol{\chi}^l[k,k] = 0, \boldsymbol{\chi}^l[k+1,k] = \lambda_k^l$ and $\boldsymbol{\chi}^l[t^*,t^*] = 1, \boldsymbol{\chi}^l[t^*+1,t^*] = 0$, Eq.(S2) can then be unfolded according to the following

process:

$$
\begin{aligned}
\left(\frac{\partial \mathcal{L}}{\partial \mathbf{m}_t^l}\right)_{STBP} &= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\left(\boldsymbol{\chi}^l[t,t]+\sum_{i=t+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]\right) \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\left(0+\sum_{i=t+1}^{t^*}0\cdot\prod_{j=t+1}^{i}\boldsymbol{\chi}^l[j,j-1]\right)+ \\
&\frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\boldsymbol{\chi}^l[t^*,t^*]\prod_{j=t+1}^{t^*}\boldsymbol{\chi}^l[j,j-1]+ \\
&\frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\sum_{i=t^*+1}^{T}\boldsymbol{\chi}^l[i,i]\prod_{j=t+1\wedge j\neq t^*+1}^{i}\boldsymbol{\chi}^l[j,j-1]\cdot 0 \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\prod_{j=t+1}^{t^*}\boldsymbol{\chi}^l[j,j-1] \\
&= \frac{\partial \mathcal{L}}{\partial \mathbf{s}_t^l}\prod_{j=t+1}^{t^*}\lambda_j^l
\end{aligned}
\tag{S3}
$$

□

References

[1] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012. 1

[2] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 1

[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1

[4] Nianhui Guo, Joseph Bethge, Christoph Meinel, and Haojin Yang. Join the high accuracy club on imagenet with a binary neural network ticket. *arXiv preprint arXiv:2211.12933*, 2022. 1

[5] Fengfu Li, Bin Liu, Xiaoxing Wang, Bo Zhang, and Junchi Yan. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016. 1

[6] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[7] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

[8] Zhaokun Zhou, Yuesheng Zhu, Chao He, Yaowei Wang, Shuicheng Yan, Yonghong Tian, and Yuan Li. Spikformer: When spiking neural network meets transformer. In *ICLR*, 2023. 1