Supplementary Materials: The Name of the Title is Hope

Anonymous Authors

1 DETAILS OF GROUP-LIF

Fig.1 shows the group G-HLIF mechanism proposed in this section. First, the features are grouped vertically, taking the G-HLIF neuron mechanism with the total number of feature groups as $groups = 4$ (i.e., the time step is 4) as an example. The feature information of the patches in the 4n-th column index is first sent to the G-HLIF neurons. The membrane potential values above the threshold V_{th} are retained, while the rest are set to V_{reset}, and at the same time, the preserved membrane potential values are accumulated to the corresponding elements in the next column, thus facilitating group interaction; the interaction of the patches with index number $4n + i$ follows the same principle. In summary, after determining the direction of information transmission, the features are grouped according to the direction of the patches. The indices of the grouped patches serve as the time steps, and then the VLIF or HLIF neuron mechanisms are employed. This feature grouping mechanism further optimizes VLIF and HLIF neurons to obtain G-VLIF and G-HLIF, which helps to enhance the transfer and extraction capability of short-distance spatial information.

The combination of the group-based G-VLIF and G-HLIF has implemented a multi-directional combination of LIF neuron mechanisms in the spatial domain, G-MLIF. It can preserve and sum the contextual information in both the horizontal and vertical directions occurring at a feature location at the same time. Through grouping, the correlation of distant information is increased, and the local spatial relationship in different directions at the same location is obtained, enhancing spatial consistency, as shown in Fig.2.

Furthermore, this chapter also attempts to use the Threshold-Dependent Batch Normalization (TDBN) method as a batch normalization operation, which considers both spatial and temporal domains dependencies, replacing equation (3.13) with equation (3.15) as shown during the training process. Here, μ_{ci} and σ_{ci}^2 are the mean and variance of each channel of the mini-batch input sequence I_i^{t+1}, ϵ is a small constant to prevent division by zero, λ_i and β_i are two trainable parameters, and α is a hyperparameter that depends on the threshold to control the pulse firing rate.

$$V_i^{t+1,n} = -\frac{1}{\tau} V_{i,t,n}(1 - X_{i,t,n}) + TDBN(I_i^{t+1,n-1}) \tag{1}$$

$$TDBN(I_i^{t+1}) = \lambda_i \alpha V_{th} \frac{(I_i^{t+1} - \mu_{ci})}{\sqrt{\sigma_{ci}^2 + \epsilon}} + \beta_i \tag{2}$$

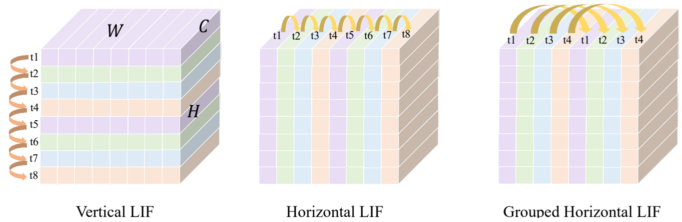

Figure 1: Visualization of three improved forms of spatial LIF neurons.

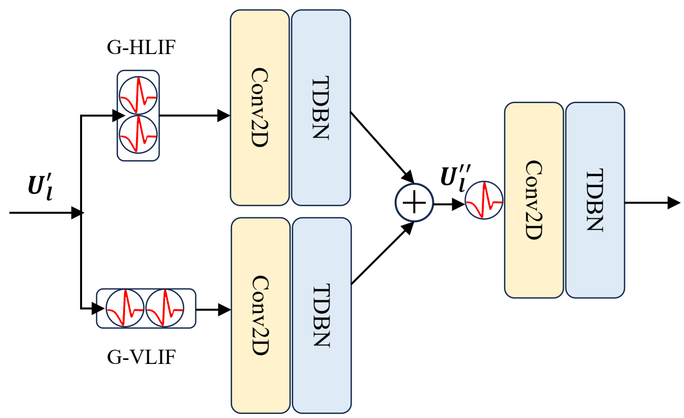

Figure 2: GLIF schematic diagram, where vertically combined LIF neurons are denoted as G-HLIF and horizontally combined LIF neurons are denoted as G-VLIF.

2 EXPERIMENTS

2.1 Dataset Description

We use the widely adopted datasets in this study for training and validation, specifically ImageNet1K, for preliminary validation of the proposed spiking neural network architecture, as well as CIFAR10 for auxiliary validation. ImageNet1K dataset consists of 1,000 categories, providing an official training set and validation set, including a training set with 1,281,167 images and a validation set with 50,000 images, validating the performance of the proposed spiking neural network architecture. The CIFAR10 dataset contains 10 categories, with a total of 60,000 32×32 RGB color images, including 50,000 training images and 10,000 validation images, which are used to further validate the effectiveness of the proposed spiking neural network in processing high-frequency input signals and its ability to handle large-scale data. In the preliminary validation on ImageNet1K, we train the spiking neural network to compare its performance with that of CIFAR10, focusing on its performance in short-term spike signal processing.

ADE20K is a dataset provided by MIT, including various scenes, objects, and categories. It consists of 25,000 images for training and validation, with an average image size of 19.5 × 10.5 megapixels. In the experiments with the visual attention mechanism ADE20K, the visual attention mechanism significantly improved performance, especially in tasks involving high-frequency signal processing.

2.2 Experiment Setup

The experiments were conducted using PyTorch on NVIDIA TITAN X (Pascal) GPUs. For large-scale data processing, the PSSD-Transformer was used to pre-process the data before feeding it into the ImageNet1K to train the spiking neural network. The PSSD-TSeg model was then used to process ADE20K to validate the effectiveness of spatial segmentation.

During the training of the spiking neural network, the initial learning rate was set to $1e^{-4}$, and a warm-up linear learning rate decay strategy was employed. In ImageNet1K, images of size 224×224 were used, and the batch size was set to 128 for 400 epochs, with the ILamb optimizer being employed. The network was trained to detect and classify images, effectively capturing high-frequency spike signals from the visual system and processing them through a 384x384 spiking neural network architecture, running batches of 80 and 20, respectively. In addition, AdamW was used as the optimizer for the spiking neural network.

The visual attention mechanism experiments in ADE20K used a multi-directional LIF neuron mechanism, saving and summing contextual information in both horizontal and vertical directions of feature locations. Through grouping, the association of distant information was increased, and the local spatial relationship of different directions at the same location was captured, enhancing spatial consistency.