

Revisiting Batch Normalization for Training Low-latency Deep Spiking Neural Networks from Scratch

Youngeun Kim
Yale University
youngeun.kim@yale.edu

Priyadarshini Panda
Yale University
priya.panda@yale.edu

Abstract

Spiking Neural Networks (SNNs) have recently emerged as an alternative to deep learning owing to sparse, asynchronous and binary event (or spike) driven processing, that can yield huge energy efficiency benefits on neuromorphic hardware. However, training high-accuracy and low-latency SNNs from scratch suffers from non-differentiable nature of a spiking neuron. *To address this training issue in SNNs, we revisit batch normalization and propose a temporal Batch Normalization Through Time (BNTT) technique.* Most prior SNN works till now have disregarded batch normalization deeming it ineffective for training temporal SNNs. *Different from previous works, our proposed BNTT decouples the parameters in a BNTT layer along the time axis to capture the temporal dynamics of spikes.* The temporally evolving learnable parameters in BNTT allow a neuron to control its spike rate through different time-steps, enabling low-latency and low-energy training from scratch. We conduct experiments on CIFAR-10, CIFAR-100, Tiny-ImageNet and event-driven DVS-CIFAR10 datasets. *BNTT allows us to train deep SNN architectures from scratch, for the first time, on complex datasets with just few 25-30 time-steps.* We also propose an early exit algorithm using the distribution of parameters in BNTT to reduce the latency at inference, that further improves the energy-efficiency. The code has been released at <https://github.com/Intelligent-Computing-Lab-Yale/BNTT-Batch-Normalization-Through-Time>.

1. Introduction

Artificial Neural Networks (ANNs) have shown state-of-the-art performance across various computer vision tasks. Nonetheless, huge energy consumption incurred for implementing ANNs on conventional von-Neumann hardware limits their usage in low-power and resource-constrained Internet of Things (IoT) environment, such as mobile phones, drones among others. In the context of low-power machine intelligence, Spiking Neural Networks (SNNs) have received considerable attention in the recent past [31, 26, 4, 9, 5]. Inspired by biological neuronal mechanisms, SNNs process vi-

sual information with discrete spikes or events over multiple time-steps. Recent works have shown that the event-driven behavior of SNNs can be implemented on emerging neuromorphic hardware to yield 1-2 order of magnitude energy efficiency over ANNs [1, 6]. Despite the energy efficiency benefits, SNNs have still not been widely adopted due to inherent training challenges. The training issue arises from the non-differentiable characteristic of a spiking neuron, generally, Integrate-and-Fire (IF) type [3], that makes SNNs incompatible with gradient descent training.

To address the training issue of SNNs, several methods, such as, *Conversion* and *Surrogate Gradient Descent* have been proposed. In ANN-SNN conversion [34, 13, 10, 32], off-the-shelf trained ANNs are converted to SNNs using normalization methods to transfer ReLU activation to IF spiking activity. The advantage here is that training happens in the ANN domain leveraging widely used machine learning frameworks like, PyTorch, that yield short training time and can be applied to complex datasets. But the ANN-SNN conversion method requires large number of time-steps ($\sim 500 - 1000$) for inference to yield competitive accuracy, which significantly increases the latency and energy consumption of the SNN. On the other hand, directly training SNNs with a surrogate gradient function [24, 19, 39] exploits temporal dynamics of spikes, resulting in lesser number of time-steps ($\sim 100 - 150$). However, the discrepancy between forward spike activation function and backward surrogate gradient function during backpropagation restricts the training capability. Only shallow SNNs (e.g., VGG5) can be trained using surrogate gradient descent and therefore they achieve high performance only for simple datasets (e.g., MNIST and CIFAR-10). Recently, a hybrid method [30] that combines the conversion method and the surrogate gradient-based method shows state-of-the-art performance at reasonable latency (~ 250 time-steps). However, the hybrid method incurs sequential processes, i.e., training ANN from scratch, conversion of ANN to SNN, and training SNNs using surrogate gradient descent, that increases the total computation cost to obtain the final SNN model. Overall, training high-accuracy and low-latency SNNs from scratch

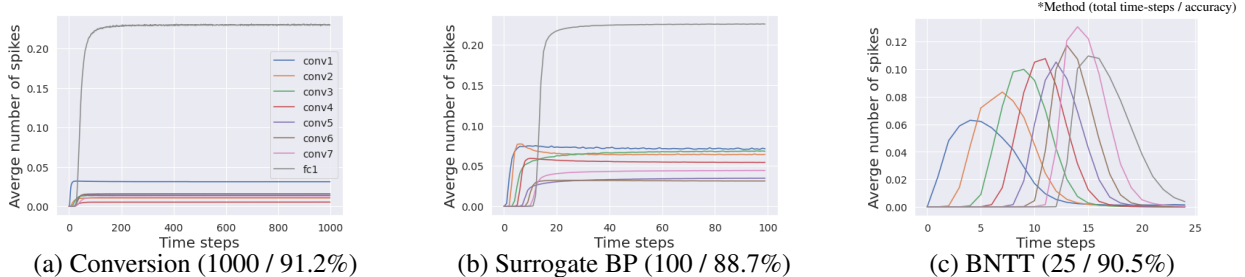


Figure 1. Visualization of the average number of spikes in each layer with respect to time-steps. Compared to (a) ANN-SNN conversion and (b) surrogate gradient-based backpropagation, our (c) BNTT captures the temporal dynamics of spike activation with learnable parameters, enabling low-latency (*i.e.*, small time-steps) and low-energy (*i.e.*, less number of spikes) training. All experiments are conducted on CIFAR-10 with VGG9.

still remains an open problem.

In this paper, we revisit Batch Normalization (BN) for more advanced SNN training. The BN layer [15] has been used extensively in deep learning to accelerate the training process of ANNs. It is well known that BN reduces internal covariate shift (or soothing optimization landscape [33]) mitigating the problem of exploding/vanishing gradients. However, till now, numerous studies on surrogate gradient of SNNs [20] have witnessed that BN does not help with SNN optimization. Moreover, most ANN-SNN conversion methods [34] get rid of BN since time-sequential spikes with BN set the firing threshold of all neurons to non-discriminative/similar values across all inputs, resulting in accuracy decline.

Motivation & Contribution: A natural question then arises: *Can standard BN capture the proper structure of temporal dynamics of spikes in SNNs?* Through this paper, we assert that standard BN hardly captures temporal characteristics as it represents the statistics of total time-steps as one common parameter. Thus, a temporally adaptive BN approach is required. To this end, we propose a new SNN-crafted batch normalization layer called Batch Normalization Through Time (BNTT) that decouples the parameters in the BN layer across different time-steps. BNTT is implemented as an additional layer in SNNs and is trained with surrogate gradient backpropagation. To investigate the effect of our BNTT, we compare the statistics of spike activity of BNTT with previous approaches: Conversion [34] and standard Surrogate Gradient Descent [24], as shown in Fig. 1. Interestingly, different from the conversion method and surrogate gradient method (without BNTT) that maintain reasonable spike activity during the entire time period across different layers, spike activity of layers trained with BNTT follows a gaussian-like trend. BNTT imposes a variation in spiking across different layers, wherein, each layer’s activity peaks in a particular time-step range and then decreases. Moreover, the peaks for early layers occur at initial time-steps and latter layers peak at later time-steps. This phenomenon implies that learnable parameters in BNTT enable the networks

to pass the visual information temporally from shallow to deeper layers in an effective manner.

The newly observed characteristics of BNTT brings several advantages. First, similar to BN, the BNTT layer enables SNNs to be trained stably from scratch even for large-scale datasets. Second, learnable parameters in BNTT enable SNNs to be trained with low latency ($\sim 25 - 50$ time-steps) and impose optimum spike activity across different layers for low-energy inference. Finally, the distribution of the BNTT learnable parameter (*i.e.*, γ) is a good representation of the temporal dynamics of spikes. Hence, relying on the observation that low γ value induces low spike activity and vice-versa, we further propose a temporal early exit algorithm. Here, an SNN can predict at an earlier time-step and does not need to wait till the end of the time period to make a prediction.

In summary, our key contributions are as follows: (i) For the first time, we introduce a batch normalization technique for SNNs, called BNTT. (ii) BNTT allows SNNs to be implemented in a low-latency and low-energy environment. (iii) We further propose a temporal early exit algorithm at inference time by monitoring the learnable parameters in BNTT. (iv) To ascertain that BNTT captures the temporal characteristics of SNNs, we mathematically show that proposed BNTT has similar effect as controlling the firing threshold of the spiking neuron at every time step during inference.

2. Batch Normalization

Batch Normalization (BN) reduces the internal covariate shift (or variation of loss landscape [33]) caused by the distribution change of input signal, which is a known problem of deep neural networks [15]. Instead of calculating the statistics of total dataset, the intermediate representations are standardized with a mini-batch to reduce the computation complexity. Given a mini-batch $\mathcal{B} = \{x_1, \dots, x_m\}$, the BN layer computes the mean and variance of the mini-batch as:

$$\mu_{\mathcal{B}} = \frac{1}{m} \sum_{b=1}^m x_b; \quad \sigma_{\mathcal{B}}^2 = \frac{1}{m} \sum_{b=1}^m (x_b - \mu_{\mathcal{B}})^2. \quad (1)$$

Then, the input features in the mini-batch are normalized with calculated statistics as:

$$\hat{x}_b = \frac{x_b - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}, \quad (2)$$

where, ϵ is a small constant for numerical stability. To further improve the representation capability of the layer, learnable parameters γ and β are used to transform the input features that can be formulated as $BN(x_i) = \gamma\hat{x}_i + \beta$. At inference time, BN uses the running average of mean and variance obtained from training. Previous works show that the BN layer not only improves the performance but also reduces the number of iterations required for training convergence. Therefore, BN is an indispensable training component for all ANN models, such as convolutional neural networks [36] and recurrent neural networks [12]. On the other hand, the effectiveness of BN in bio-plausible SNNs has not been observed yet.

3. Methodology

3.1. Spiking Neural Networks

Different from conventional ANNs, SNNs transmit information using binary spike trains. To leverage the temporal spike information, Leaky-Integrate-and-Fire (LIF) model [7] is widely used to emulate neuronal functionality in SNNs, which can be formulated as a differential equation:

$$\tau_m \frac{dU_m}{dt} = -U_m + RI(t), \quad (3)$$

where, U_m represents the membrane potential of the neuron that characterizes the internal state of the neuron, τ_m is the time constant of membrane potential decay. Also, R and $I(t)$ denote the input resistance and the input current at time t , respectively. Following the previous work [40], we convert this continuous dynamic equation into a discrete equation for digital simulation. For a single post-synaptic neuron i , we can represent the membrane potential u_i^t at time-step t as:

$$u_i^t = \lambda u_i^{t-1} + \sum_j w_{ij} o_j^t. \quad (4)$$

Here, j is the index of a pre-synaptic neuron, λ is a leak factor with value less than 1, o_j is the binary spike activation, and w_{ij} is the weight of the connection between pre- and post-neurons. From Eq. 4, the membrane potential of a neuron decreases due to leak and increases due to the weighted sum of incoming input spikes.

If the membrane potential u exceeds a pre-defined firing threshold θ , the LIF neuron i generates a binary spike output o_i . After that, we perform a soft reset, where the membrane potential u_i is reset by reducing its value by the threshold θ . Compared to a hard reset (resetting the membrane potential u_i to zero after neuron i spikes), the soft reset minimizes

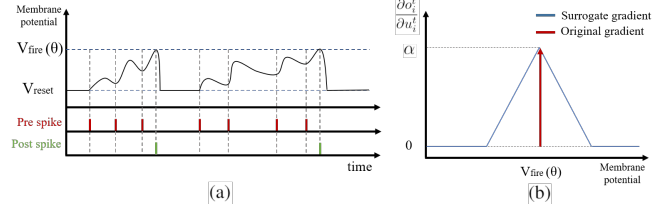


Figure 2. (a) Illustration of spike activities in Leaky-Integrate-and-Fire neurons. (b) The approximated gradient value with respect to the membrane potential.

information loss by maintaining the residual voltage and carrying it forward to the next time step, thereby achieving better performance [13]. Fig. 2(a) illustrates the membrane potential dynamics of a LIF neuron.

For the output layer, we discard the thresholding functionality so that neurons do not generate any spikes. We allow the output neurons to accumulate the spikes over all time-steps by fixing the leak parameter (λ in Eq. 4) as one. This enables the output layer to compute probability distribution after softmax function without information loss. As with ANNs, the number of output neurons in SNNs is identical to the number of classes C in the dataset. From the accumulated membrane potential, we can define the cross-entropy loss for SNNs as:

$$L = - \sum_i y_i \log \left(\frac{e^{u_i^T}}{\sum_{k=1}^C e^{u_k^T}} \right), \quad (5)$$

where, y is the ground-truth label, and T represents the total number of time-steps. Then, the weights of all layers are updated by backpropagating the loss value with gradient descent.

To compute the gradients of each layer l , we use backpropagation through time (BPTT), which accumulates the gradients over all time-steps [24]. These approaches can be implemented with auto-differentiation tools, such as PyTorch [29], that enable backpropagation on the unrolled network. To this end, we compute the loss function at time-step T and use gradient descent optimization. Mathematically, we can define the accumulated gradients at the layer l by chain rule as:

$$\Delta W_l = \sum_t \frac{\partial L}{\partial W_l^t} = \begin{cases} \sum_t \frac{\partial L}{\partial O_l^t} \frac{\partial O_l^t}{\partial U_l^t} \frac{\partial U_l^t}{\partial W_l^t}, & \text{if } l = \text{hidden layer} \\ \sum_t \frac{\partial L}{\partial U_l^T} \frac{\partial U_l^T}{\partial W_l^t}. & \text{if } l = \text{output layer} \end{cases} \quad (6)$$

Here, O_l and U_l are output spikes and membrane potential at layer l , respectively. For the output layer, we get the derivative of the loss L with respect to the membrane potential u_i^T at final time-step T :

$$\frac{\partial L}{\partial u_i^T} = \frac{e^{u_i^T}}{\sum_{k=1}^C e^{u_k^T}} - y_i. \quad (7)$$

This derivative function is continuous and differentiable for all possible membrane potential values. On the other hand, LIF neurons in hidden layers generate spike output only if the membrane potential u_i^t exceeds the firing threshold, leading to non-differentiability. To deal with this problem, we introduce an approximate gradient:

$$\frac{\partial o_i^t}{\partial u_i^t} = \alpha \max\{0, 1 - \frac{u_i^t - \theta}{\theta}\}, \quad (8)$$

where, α is a damping factor for back-propagated gradients. Note, a large α value causes unstable training as gradients are summed over all time-steps. Hence, we set α to 0.3. Overall, we update the network parameters at the layer l based on the gradient value (Eq. 6) as $W_l = W_l - \eta \Delta W_l$.

3.2. Batch Normalization Through Time (BNTT)

The main contribution of this paper is a new SNN-crafted Batch Normalization (BN) technique. **Naively applying BN does not have any effect on training SNNs.** This is because using the same BN parameters (e.g., global mean μ , global variation σ , and learnable parameter γ) for the statistics of all time-steps do not capture the temporal dynamics of input spike trains. For example, an LIF neuron requires at least one time-step to propagate spikes to the next layer. Therefore, input signals for the third layer of an SNN will have a zero value till $t = 2$. Following the initial spike activity in the layer at $t = 2$, the spike signals vary depending upon the weight connections and the membrane potentials of previous layers. Therefore, a fixed global mean from a standard BN layer may not store any time-specific information, resulting in performance degradation at inference.

To resolve this issue, we vary the internal parameters in a BN layer through time, that we define as, BNTT. Similar to the digital simulation of LIF neuron across different time-steps, one BNTT layer is expanded temporally with a local learning parameter associated with each time-step. This allows the BNTT layer to capture temporal statistics (see Section 3.3 for mathematical analysis). The proposed BNTT layer is easily applied to SNNs by inserting the layer after convolutional/linear operations as:

$$\begin{aligned} u_i^t &= \lambda u_i^{t-1} + BNTT_{\gamma^t}(\sum_j w_{ij} o_j^t) \\ &= \lambda u_i^{t-1} + \gamma_i^t \left(\frac{\sum_j w_{ij} o_j^t - \mu_i^t}{\sqrt{(\sigma_i^t)^2 + \epsilon}} \right). \end{aligned} \quad (9)$$

During the training process, we compute the mean μ_i^t and variance σ_i^t from the samples in a mini-batch \mathcal{B} for each time step t , as shown in Algorithm 1. Note, for each time-step t , we apply an exponential moving average to approximate global mean $\bar{\mu}_i^t$ and variance $\bar{\sigma}_i^t$ over training iterations. These global statistics are used to normalize the

test data at inference. Also, we do not utilize β as in conventional BN, since it adds redundant voltage to the membrane potential of SNNs.

Adding the BNTT layer to LIF neurons changes the gradient calculation for backpropagation. Given that $x_i^t = \sum_j w_{ij} o_j^t$ is an input signal to the BNTT layer, we can calculate the gradient value passed through lower layers by the BNTT layer as:

$$\frac{\partial L}{\partial x_b^t} = \frac{1}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left(m \frac{\partial L}{\partial \hat{x}_b^t} - \sum_{k=1}^m \frac{\partial L}{\partial \hat{x}_k^t} - \hat{x}_b^t \sum_{k=1}^m \frac{\partial L}{\partial \hat{x}_k^t} \hat{x}_k^t \right). \quad (10)$$

Here, we omit a neuron index i for simplicity. Also, m and b denote the batch size and batch index (see Appendix A for more detail). Thus, for every time-step t , gradients are calculated based on the time-specific statistics of input signals. This allows the networks to take into account temporal dynamics for training weight connections. Moreover, a learnable parameter γ is updated to restore the representation power of the batch normalized signal. Since we use different γ^t values across all time-steps, γ^t finds an optimum over each time-step for efficient inference. We update gamma $\gamma^t = \gamma^t - \eta \Delta \gamma^t$ where:

$$\Delta \gamma^t = \frac{\partial L}{\partial \gamma^t} = \frac{\partial L}{\partial u^t} \frac{\partial u^t}{\partial \gamma^t} = \sum_{k=1}^m \frac{\partial L}{\partial u_k^t} \hat{x}_k^t. \quad (11)$$

3.3. Mathematical Analysis

In this section, we discuss the connections between BNTT and the firing threshold of a LIF neuron. Specifically, we formally prove that using BNTT has a similar effect as varying the firing threshold over different time-steps, thereby ascertaining that BNTT captures temporal characteristics in SNNs. Recall that BNTT normalizes the input signal using stored approximated global average $\bar{\mu}_i^t$ and standard deviation $(\bar{\sigma}_i^t)^2$ at inference. From Eq. 9, we can calculate a membrane potential at time-step $t = 1$, given that initial membrane potential u_i^0 has a zero value:

$$\begin{aligned} u_i^1 &= \gamma_i^1 \left(\frac{\sum_j w_{ij} o_j^1 - \bar{\mu}_i^1}{\sqrt{(\bar{\sigma}_i^1)^2 + \epsilon}} \right) \\ &\approx \frac{\gamma_i^1}{\sqrt{(\bar{\sigma}_i^1)^2 + \epsilon}} \sum_j w_{ij} o_j^1 = \frac{\gamma_i^1}{\sqrt{(\bar{\sigma}_i^1)^2 + \epsilon}} \tilde{u}_i^1. \end{aligned} \quad (12)$$

Here, we assume $\bar{\mu}_i^1$ can be neglected with small signal approximation due to the spike sparsity in SNNs, and $\tilde{u}_i^1 = \sum_j w_{ij} o_j^1$ is membrane potential at time-step $t = 1$ without BNTT (obtained from Eq. 4). We can observe that the membrane potential with BNTT is proportional to the membrane potential without BNTT at $t = 1$. For time-step $t > 1$, we should take into account the membrane potential from the previous time-step, which is multiplied by leak λ .

Algorithm 1 BNTT layer

Input: mini-batch \mathcal{B} at time step t ($x_{\{1 \dots m\}}^t$), learnable parameter (γ^t), update factor (α)

Output: $\{y^t = \text{BNTT}_{\gamma^t}(x^t)\}$

- 1: $\mu^t \leftarrow \frac{1}{m} \sum_{b=1}^m x_b^t$
 - 2: $(\sigma^t)^2 \leftarrow \frac{1}{m} \sum_{b=1}^m (x_b^t - \mu^t)^2$
 - 3: $\hat{x}^t = \frac{x^t - \mu^t}{\sqrt{(\sigma^t)^2 + \epsilon}}$
 - 4: $y^t \leftarrow \gamma^t \hat{x}^t \equiv \text{BNTT}_{\gamma^t}(x^t)$
 - 5: % Exponential moving average
 - 6: $\bar{\mu}^t \leftarrow (1 - \alpha)\bar{\mu}^t + \alpha\mu^t$
 - 7: $\bar{\sigma}^t \leftarrow (1 - \alpha)\bar{\sigma}^t + \alpha\sigma^t$
-

To this end, by substituting Eq. 12 in the BNTT equation (Eq. 9), we can formulate the membrane potential at $t = 2$ as:

$$\begin{aligned}
 u_i^2 &\approx \lambda u_i^1 + \frac{\gamma_i^2}{\sqrt{(\sigma_i^2)^2 + \epsilon}} \sum_j w_{ij} o_j^2 \\
 &= \left(\frac{\lambda \gamma_i^1}{\sqrt{(\sigma_i^1)^2 + \epsilon}} \right) \tilde{u}_i^1 + \frac{\gamma_i^2}{\sqrt{(\sigma_i^2)^2 + \epsilon}} \sum_j w_{ij} o_j^2 \\
 &\approx \frac{\gamma_i^2}{\sqrt{(\sigma_i^2)^2 + \epsilon}} \{ \lambda \tilde{u}_i^1 + \sum_j w_{ij} o_j^2 \} = \frac{\gamma_i^2}{\sqrt{(\sigma_i^2)^2 + \epsilon}} \tilde{u}_i^2.
 \end{aligned} \tag{13}$$

In the third line, the learnable parameter γ_i^t and σ_i^t have similar values in adjacent time intervals ($t = 1, 2$) because of continuous time property. Hence, we can approximate γ_i^1 and σ_i^1 as γ_i^2 and σ_i^2 , respectively. Finally, we can extend the equation of BNTT to the time-step t :

$$u_i^t \approx \frac{\gamma_i^t}{\sqrt{(\sigma_i^t)^2 + \epsilon}} \tilde{u}_i^t. \tag{14}$$

Considering that a neuron produces an output spike activation whenever the membrane potential \tilde{u}_i^t exceeds the pre-defined firing threshold θ , the spike firing condition with BNTT can be represented $u_i^t \geq \theta$. Comparing with the threshold of a neuron without BNTT, we can reformulate the firing condition as:

$$\tilde{u}_i^t \geq \frac{\sqrt{(\sigma_i^t)^2 + \epsilon}}{\gamma_i^t} \theta. \tag{15}$$

Thus, we can infer that using a BNTT layer changes the firing threshold value by $\sqrt{(\sigma_i^t)^2 + \epsilon}/\gamma_i^t$ at every time-step. In practice, BNTT results in an optimum γ during training that improves the representation power, producing better performance and low-latency SNNs. This observation allows us to consider the advantages of time-varying learnable parameters in SNNs. This implication is in line with previous

Algorithm 2 Training process with BNTT

Input: mini-batch (X); label set (Y); max_timestep (T)

Output: updated network weights

- 1: **for** $i \leftarrow 1$ to max_iter **do**
 - 2: fetch a mini batch X
 - 3: **for** $t \leftarrow 1$ to T **do**
 - 4: $O \leftarrow \text{PoissonGenerator}(X)$
 - 5: **for** $l \leftarrow 1$ to $L - 1$ **do**
 - 6: $(O_l^t, U_l^t) \leftarrow (\lambda, U_l^{t-1}, \text{BNTT}_{\gamma^t}(W_l, O_{l-1}^{t-1}))$
 - 7: **end for**
 - 8: % For the final layer L , stack the voltage
 - 9: $U_L^t \leftarrow (U_L^{t-1}, \text{BNTT}_{\gamma^t}(W_L, O_{L-1}^{t-1}))$
 - 10: **end for**
 - 11: % Calculate the loss and back-propagation
 - 12: $L \leftarrow (U_L^T, Y)$
 - 13: **end for**
-

work [13], which insists that manipulating the firing threshold improves the performance and latency of the ANN-SNN conversion method. However, Han et al. change the threshold value in a heuristic way without any optimization process and fix the threshold value across all time-steps. On the other hand, our BNTT yields time-specific γ^t which can be optimized via back-propagation.

3.4. Early Exit Algorithm

The main objective of early exit is to reduce the latency during inference [38, 27]. Most previous methods [39, 19, 34, 30, 13] accumulate output spikes till the end of the time-sequence, at inference, since all layers generate spikes across all time-steps as shown in Fig. 1(a) and Fig. 1(b). On the other hand, learnable parameters in BNTT manipulate the spike activity of each layer to produce a peak value, which falls again (a gaussian-like trend), as shown in Fig. 1(c). This phenomenon shows that SNNs using BNTT convey little information at the end of spike trains.

Inspired by this observation, we propose a temporal early exit algorithm based on the value of γ^t . From Eq. 15, we know that a low γ^t value increases the firing threshold, resulting in low spike activity. A high γ^t value, in contrast, induces more spike activity. It is worth mentioning that $(\sigma_i^t)^2$ shows similar values across all time-steps and therefore we only focus on γ^t . Given that the intensity of spike activity is proportional to γ^t , we can infer that spikes will hardly contribute to the classification result once γ^t values across every layer drop to a minimum value. Therefore, we measure the average of γ^t values in each layer l at every time-step, and terminate the inference when γ^t value in every layer is below a pre-determined threshold. For example, as shown in Fig. 3, we observe that all averaged γ^t values are lower than threshold 0.1 after $t > 20$. Therefore, we define the early exit time at $t = 20$. Note that we can determine the optimum time-step for early exit before forward propagation without

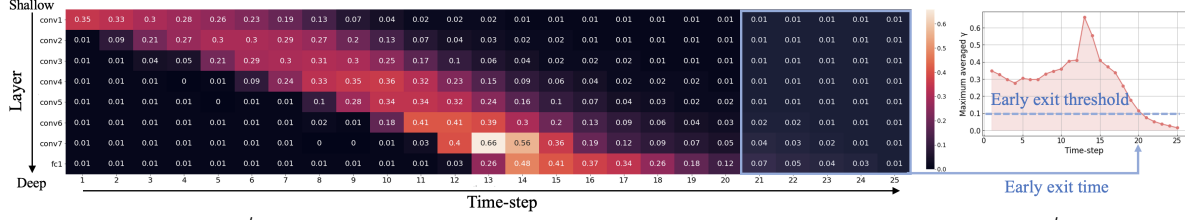


Figure 3. The average value of γ^t at each layer over all time-steps. Early exit time can be calculated as $t = 20$ since γ^t values at every layer have lower value than threshold 0.1 after time-step 20 (blue shaded area). Here, we use a VGG9 architecture on CIFAR-10.

any additional computation. In summary, the temporal early exit method enables us to find the earliest time-step during inference that ensures integration of crucial information, in turn reducing the inference latency without significant loss of accuracy.

3.5. Overall Optimization

Algorithm 2 summarizes the whole training process of SNNs with BNTT. Our proposed BNTT acts as a regularizer, unlike previous methods [19, 34, 20, 30] that use dropout to perform regularization. Our training scheme is based on widely used rate coding where the spike generator produces a Poisson spike train (see Appendix B) for each pixel in the image with frequency proportional to the pixel intensity [31]. For all layers, the weighted sum of the input signal is passed through a BNTT layer and then is accumulated in the membrane potential. If the membrane potential exceeds the firing threshold, the neuron generates an output spike. For last layer, we accumulate the input voltage over all time-steps without leak, that we feed to a softmax layer to output a probability distribution. Then, we calculate a cross-entropy loss function and gradients for weight of each layer with the approximate gradient function. During the training phase, a BNTT layer computes the time-dependent statistics (*i.e.*, μ^t and σ^t) and stores the moving-average global mean and variance. At inference, we first define the early exit time-step based on the value of γ in BNTT. Then, the networks classify the test input (note, test data normalized with pre-computed global $\bar{\mu}^t, \bar{\sigma}^t$ BNTT statistics) based on the accumulated output voltage at the pre-computed early exit time-step.

4. Experiments

In this section, we carry out comprehensive experiments on public classification datasets. Till now, training SNNs from scratch with surrogate gradient has been limited to simple datasets, *e.g.*, CIFAR-10, due to the difficulty of direct optimization. In this paper, for the first time, we train SNNs with surrogate gradients from scratch and report the performance on large-scale datasets including CIFAR-100 and Tiny-ImageNet with multi-layered network architectures. We first compare our BNTT with previous SNNs training methods. Then, we quantitatively and qualitatively demonstrate the effectiveness of our proposed BNTT.

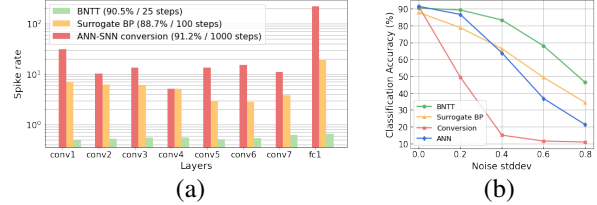


Figure 4. (a) Visualization layer-wise spike activity (log scale) in VGG9 on CIFAR-10 dataset. (b) Performance change with respect to the standard deviation of the Gaussian noise.

4.1. Experimental Setup

We evaluate our method on three static datasets (*i.e.*, CIFAR-10, CIFAR-100, Tiny-ImageNet) and one neuromorphic dataset (*i.e.*, DVS-CIFAR10). **CIFAR-10** [17] consists of 60,000 images (50,000 for training / 10,000 for testing) with 10 categories. All images are RGB color images whose size are 32×32 . **CIFAR-100** has the same configuration as CIFAR-10, except it contains images from 100 categories. **Tiny-ImageNet** is the modified subset of the original ImageNet dataset [8], with 100,000 training and 10,000 validation images. The resolution of the images is 64×64 pixels. **DVS-CIFAR10** [21] has the same configuration as CIFAR-10. This discrete event-stream dataset is collected by moving the event-driven camera. We follow the similar data pre-processing protocol and a network architecture used in previous work [40] (details in Appendix C). Our implementation is based on Pytorch [29]. We train the networks with standard SGD with momentum 0.9, weight decay 0.0005 and also apply random crop and horizontal flip to input images. The base learning rate is set to 0.3 and we use step-wise learning rate scheduling with a decay factor 10 at 50%, 70%, and 90% of the total number of epochs. Here, we set the total number of epochs to 120, 240, 90, and 60 for CIFAR-10, CIFAR-100, Tiny-ImageNet, and DVS-CIFAR10, respectively.

4.2. Comparison with Previous Methods

On public datasets, we compare our proposed BNTT method with previous rate-coding based SNN training methods, including ANN-SNN conversion [13, 34, 4], surrogate gradient back-propagation [19], and hybrid [30] methods. From Table 1, we can observe some advantages and disadvantages of each training method. The ANN-SNN con-

Table 1. Classification Accuracy (%) on CIFAR-10, CIFAR-100, and Tiny-ImageNet.

	Dataset	Training Method	Architecture	Time-steps	Accuracy(%)
Cao <i>et al.</i> [4]	CIFAR-10	ANN-SNN Conversion	3Conv, 2Linear	400	77.4
Sengupta <i>et al.</i> [34]	CIFAR-10	ANN-SNN Conversion	VGG16	2500	91.5
Lee <i>et al.</i> [19]	CIFAR-10	Surrogate Gradient	VGG9	100	90.4
Rathi <i>et al.</i> [30]	CIFAR-10	Hybrid	VGG16	200	92.0
Han <i>et al.</i> [13]	CIFAR-10	ANN-SNN Conversion	VGG16	2048	93.6
w.o. BNTT	CIFAR-10	Surrogate Gradient	VGG9	100	88.7
BNTT (ours)	CIFAR-10	Surrogate Gradient	VGG9	25	90.5
BNTT + Early Exit (ours)	CIFAR-10	Surrogate Gradient	VGG9	20	90.3
Sengupta <i>et al.</i> [34]	CIFAR-100	ANN-SNN Conversion	VGG16	2500	70.9
Rathi <i>et al.</i> [30]	CIFAR-100	Hybrid	VGG16	125	67.8
Han <i>et al.</i> [13]	CIFAR-100	ANN-SNN Conversion	VGG16	2048	70.9
w.o. BNTT	CIFAR-100	Surrogate Gradient	VGG11	n/a	n/a
BNTT (ours)	CIFAR-100	Surrogate Gradient	VGG11	50	66.6
BNTT + Early Exit (ours)	CIFAR-100	Surrogate Gradient	VGG11	30	65.8
Sengupta <i>et al.</i> [34]	Tiny-ImageNet	ANN-SNN Conversion	VGG11	2500	54.2
w.o. BNTT	Tiny-ImageNet	Surrogate Gradient	VGG11	n/a	n/a
BNTT (ours)	Tiny-ImageNet	Surrogate Gradient	VGG11	30	57.8
BNTT + Early Exit (ours)	Tiny-ImageNet	Surrogate Gradient	VGG11	25	56.8

Table 2. Classification Accuracy (%) on DVS-CIFAR10.

Method	Type	Accuracy (%)
Orchard <i>et al.</i> [25]	Random Forest	31.0
Lagorce <i>et al.</i> [18]	HOTS	27.1
Sironi <i>et al.</i> [37]	HAT	52.4
Sironi <i>et al.</i> [37]	Gabor-SNN	24.5
Wu <i>et al.</i> [40]	Surrogate Gradient	60.5
w.o. BNTT	Surrogate Gradient	n/a
BNTT (ours)	Surrogate Gradient	63.2

version method performs better than the surrogate gradient method across all datasets. However, they require large number of time-steps for training and testing, which is energy-inefficient and impractical in a real-time application. The hybrid method aims to resolve this high-latency problem, but it still requires over hundreds of time-steps. The surrogate gradient method suffers from poor optimization and hence cannot be scaled to larger datasets such as CIFAR-100 and Tiny-ImageNet. Our BNTT is based on the surrogate gradient method, however, it enables SNNs to achieve high performance even for more complicated datasets. At the same time, we dramatically reduce the latency due to the inclusion of learnable parameters and temporal statistics in the BNTT layer. As a result, BNTT can be trained with 25 time-steps on a simple CIFAR-10 dataset, while preserving state-of-the-art accuracy. For CIFAR-100, we achieve about $40\times$ and $2\times$ faster inference speed compared to the conversion methods and the hybrid method, respectively. Interestingly, for Tiny-ImageNet, BNTT achieves better performance and shorter latency compared to previous conversion method. Note that ANN with VGG11 architecture used for ANN-SNN conversion achieves 56.3% accuracy. Moreover, using an early exit algorithm further reduces the latency by $\sim 20\%$, which enables the networks to be implemented with lower-latency and energy-efficiency. It is worth mentioning

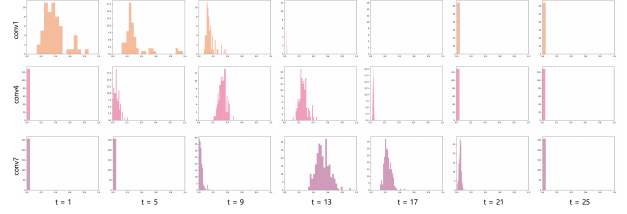


Figure 5. Histogram visualization (x axis: γ value, y axis: frequency) at conv1 (row1), conv4 (row2), and conv7 (row3) layers in VGG9 across all time-steps. The experiments are conducted on CIFAR-10 with 25 time-steps.

Table 3. Energy efficiency comparison.

Method	Latency	Accuracy (%)	E_{ANN}/E_{method}
VGG9 (ANN)	1	91.5	$1\times$
Conversion	1000	91.2	$0.32\times$
Conversion	500	90.9	$0.55\times$
Conversion	100	89.3	$2.71\times$
Surrogate Gradient	100	88.7	$1.05\times$
BNTT	25	90.5	$9.14\times$

that surrogate gradient method without BNTT (w.o. BNTT in Table 1) only converges on CIFAR-10. For neuromorphic DVS-CIFAR10 dataset (Table 2), ANN-SNN Conversion methods are not applicable since ANNs hardly capture the temporal dynamic of a spike train. Using BNTT improves the stability of training compared to a surrogate gradient baseline (*i.e.*, w.o. BNTT), and achieves state-of-the-art performance. These results show that our BNTT technique is very effective on event-driven data and hence well-suited for neuromorphic applications.

4.3. Energy Comparison

We compare the layer-wise spiking activities of our BNTT with two widely-used methods, *i.e.*, ANN-SNN conversion method [34] and surrogate gradient method (w.o. BNTT) [24]. Note, we refer to our approach as *BNTT* and standard

surrogate approach w.o. BNTT as *surrogate gradient* in the remainder of the text. Specifically, we calculate the spike rate of each layer l , which can be defined as the total number of spikes at layer l over total time-steps T divided by the number of neurons in layer l (see Appendix D for the equation of spike rate). In Fig. 4(a), converted SNNs show a high spike rate for every layer as they forward spike trains through a larger number of time-steps compared to other methods. Even though the surrogate gradient method uses less number of time-steps, it still requires nearly hundreds of spikes for each layer. Compared to these methods, we can observe that BNTT significantly improves the spike sparsity across all layers.

More precisely, as done in previous works [28, 20], we compute the energy consumption for SNNs in standard CMOS technology [14] as shown in Appendix D by calculating the net multiplication-and-accumulate (MAC) operations. As the computation of SNNs are event-driven with binary $\{1, 0\}$ spike processing, the MAC operation reduces to just a floating point (FP) addition. On the other hand, conventional ANNs still require one FP addition and one FP multiplication to conduct the same MAC operation (see Appendix D for more detail). Table 3 shows the energy efficiency of ANNs and SNNs with a VGG9 architecture [36] on CIFAR-10. As expected, ANN-SNN conversion yields a trade-off between accuracy and energy efficiency. For the same latency, the surrogate gradient method expends higher energy compared to the conversion method. It is interesting to note that even though our BNTT is trained based on the surrogate gradient method, we get $\sim 9\times$ improvement in energy efficiency compared to ANNs. In addition, we conduct further energy comparison on Neuromorphic architecture in Appendix E.

4.4. Analysis on Learnable Parameters in BNTT

The key observation of our work is the change of γ across time-steps. To analyze the distribution of the learnable parameters in our BNTT, we visualize the histogram of γ in conv1, conv4, and conv7 layers in VGG9 as shown in Fig. 5. Interestingly, all layers show different temporal evolution of gamma distributions. For example, conv1 has high γ values at the initial time-steps which decrease as time goes on. On the other hand, starting from small values, the γ values in conv4 and conv7 layers peak at $t = 9$ and $t = 13$, respectively, and then shrink to zero at later time-steps. Notably, the peak time is delayed as the layer goes deeper, implying that the visual information is passed through the network sequentially over a period of time similar to Fig. 1(c). This gaussian-like trend with rise and fall of γ across different time-steps can support the explanation of overall low spike activity compared to other methods (Fig. 4(a)).

4.5. Analysis on Early Exit

Recall that we measure the average of γ values in each layer at every time-step, and stop the inference when all γ

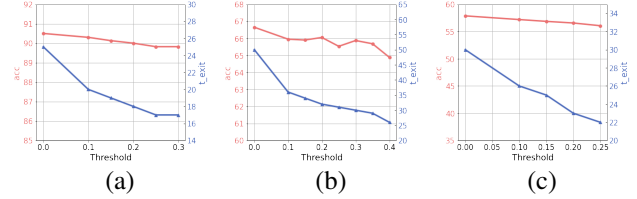


Figure 6. Visualization of accuracy and early exit time with respect to the threshold value for γ . (a) CIFAR-10. (b) CIFAR-100. (c) Tiny-ImageNet.

values in every layer is lower than a predetermined threshold. To further investigate this, we vary the predetermined threshold and show the accuracy and exit time T_{exit} trend. As shown in Fig. 6, we observe that high threshold enables the networks to infer at earlier time-steps. Although we use less number time-steps during inference, the accuracy drops marginally. This implies that BNTT rarely sends crucial information at the end of spike train (see Fig. 1(c)). Note that the temporal evolution of learnable parameter γ with our BNTT allows us to exploit the early exit algorithm that yields a huge advantage in terms of reduced latency at inference. Such strategy has not been proposed or explored in any prior works that have mainly focused on reducing the number of time-steps during training without effectively using temporal statistics.

4.6. Analysis on Robustness

Finally, we highlight the advantage of BNTT in terms of the robustness to noisy input. To investigate the effect of our BNTT for robustness, we evaluate the performance change in the SNNs as we feed in inputs with varying levels of noise. We generate the noisy input by adding Gaussian noise ($0, \sigma^2$) to the clean input image. From Fig. 4 (b), we observe the following: i) The accuracy of conversion method degrades considerably for $\sigma > 0.4$. ii) Compared to ANNs, SNNs trained with surrogate gradient back-propagation shows better performance at higher noise intensity. Still, they suffer from large accuracy drops in presence of noisy inputs. iii) BNTT achieves significantly higher performance than the other methods across all noise intensities. This is because using BNTT decreases the overall number of time-steps which is a crucial contributing factor towards robustness [35]. These results imply that, in addition to low-latency and energy-efficiency, our BNTT method also offers improved robustness for suitably implementing SNNs in a real-world scenario. We further analyze the robustness regarding adversarial attack [11] in Appendix F.

5. Conclusion

In this paper, we revisit the batch normalization technique and propose a novel mechanism for training low-latency, energy-efficient, robust, and accurate SNNs from scratch. Our key idea is to extend the effect of batch normalization to the temporal dimension with time-specific learnable parameters and statistics. We discover that optimizing learnable

parameters γ during the training phase enables visual information to be passed through the layers sequentially. For the first time, we directly train SNNs on large datasets such as Tiny-ImageNet, which opens up the potential advantage of surrogate gradient-based backpropagation for future practical research in SNNs.

A. Appendix: Backward Gradient of BNTT

Here, we calculate the backward gradient of a BNTT layer. Note that we omit the neuron index i for simplicity. For one sample x_b in a mini-batch, we compute the backward gradient of BNTT at time-step t :

$$\frac{\partial L}{\partial x_b^t} = \frac{\partial L}{\partial \hat{x}_b^t} \frac{\partial \hat{x}_b^t}{\partial x_b^t} + \frac{\partial L}{\partial \mu^t} \frac{\partial \mu^t}{\partial x_b^t} + \frac{\partial L}{\partial (\sigma^t)^2} \frac{\partial (\sigma^t)^2}{\partial x_b^t}. \quad (16)$$

where,

$$\hat{x}_b^t = \frac{x_b^t - \mu^t}{\sqrt{(\sigma^t)^2 + \epsilon}}. \quad (17)$$

It is worth mentioning that we accumulate input signals at the last layer in order to remove information loss. Then we convert the accumulated voltage into probabilities by using a softmax function. Therefore, we calculate backward gradient with respect to the loss L , following the previous work [24]. The first term of R.H.S in Eq. (16) can be calculated as:

$$\frac{\partial L}{\partial \hat{x}_b^t} \frac{\partial \hat{x}_b^t}{\partial x_b^t} = \frac{1}{\sqrt{(\sigma^t)^2 + \epsilon}} \frac{\partial L}{\partial \hat{x}_b^t}. \quad (18)$$

For the second term of R.H.S in Eq. (16),

$$\begin{aligned} \frac{\partial L}{\partial \mu^t} \frac{\partial \mu^t}{\partial x_b^t} &= \left\{ \frac{\partial L}{\partial \hat{x}_b^t} \frac{\partial \hat{x}_b^t}{\partial \mu^t} + \frac{\partial L}{\partial (\sigma^t)^2} \frac{\partial (\sigma^t)^2}{\partial \mu^t} \right\} \frac{\partial \mu^t}{\partial x_b^t} \\ &= \left\{ \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \frac{-1}{\sqrt{(\sigma^t)^2 + \epsilon}} + \frac{\partial L}{\partial (\sigma^t)^2} \frac{1}{m} \sum_{j=1}^m -2(x_j^t - \mu^t) \right\} \frac{\partial \mu^t}{\partial x_b^t} \\ &= \left\{ \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \frac{-1}{\sqrt{(\sigma^t)^2 + \epsilon}} - 2 \frac{\partial L}{\partial (\sigma^t)^2} (\mu^t - \frac{\mu^t m}{m}) \right\} \frac{\partial \mu^t}{\partial x_b^t} \\ &= \left\{ \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \frac{-1}{\sqrt{(\sigma^t)^2 + \epsilon}} \right\} \frac{\partial \mu^t}{\partial x_b^t} \\ &= \frac{1}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \right\}. \end{aligned} \quad (19)$$

For the third term of R.H.S in Eq. (16),

$$\begin{aligned} \frac{\partial L}{\partial (\sigma^t)^2} \frac{\partial (\sigma^t)^2}{\partial x_b^t} &= \left\{ \frac{\partial L}{\partial \hat{x}_b^t} \frac{\partial \hat{x}_b^t}{\partial (\sigma^t)^2} \right\} \frac{\partial (\sigma^t)^2}{\partial x_b^t} \\ &= \left\{ -\frac{1}{2} \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} (x_j^t - \mu^t) ((\sigma^t)^2 + \epsilon)^{-1.5} \right\} \frac{\partial (\sigma^t)^2}{\partial x_b^t} \\ &= \left\{ -\frac{1}{2} \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} (x_j^t - \mu^t) ((\sigma^t)^2 + \epsilon)^{-1.5} \right\} \frac{2(x_b - \mu^t)}{m} \\ &= \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \frac{(x_j^t - \mu^t)}{\sqrt{(\sigma^t)^2 + \epsilon}} ((\sigma^t)^2 + \epsilon)^{-0.5} \right\} \frac{(x_b - \mu^t)}{m \sqrt{(\sigma^t)^2 + \epsilon}} \\ &= \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \hat{x}_j^t ((\sigma^t)^2 + \epsilon)^{-0.5} \right\} \frac{\hat{x}_b}{m} \\ &= \frac{\hat{x}_b}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \hat{x}_j^t \right\}. \end{aligned} \quad (20)$$

Based on Eq (18), Eq (19), and Eq (20), we can reformulate Eq. (16) as:

$$\begin{aligned} \frac{\partial L}{\partial x_b^t} &= \frac{\partial L}{\partial \hat{x}_b^t} \frac{\partial \hat{x}_b^t}{\partial x_b^t} + \frac{\partial L}{\partial \mu^t} \frac{\partial \mu^t}{\partial x_b^t} + \frac{\partial L}{\partial (\sigma^t)^2} \frac{\partial (\sigma^t)^2}{\partial x_b^t} \\ &= \frac{1}{\sqrt{(\sigma^t)^2 + \epsilon}} \frac{\partial L}{\partial \hat{x}_b^t} + \frac{1}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \right\} \\ &\quad + \frac{\hat{x}_b}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left\{ - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \hat{x}_j^t \right\} \\ &= \frac{1}{m \sqrt{(\sigma^t)^2 + \epsilon}} \left\{ m \frac{\partial L}{\partial \hat{x}_b^t} - \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} - \hat{x}_b \sum_{j=1}^m \frac{\partial L}{\partial \hat{x}_j^t} \hat{x}_j^t \right\}. \end{aligned} \quad (21)$$

To summarize, for every time-step t , gradients are calculated based on the time-specific statistics of input signals. This allows the networks to take into account temporal dynamics for training weight connections.

B. Appendix: Rate Coding

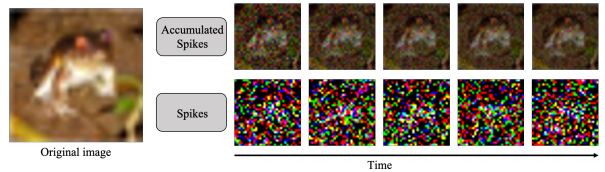


Figure 7. Example of rate coding. As time goes on, the accumulated spikes represent similar image to original image. We use image in the CIFAR-10 dataset.

Spiking neural networks process multiple binary spikes. Therefore, for training and inference, a static image needs to be converted. There are various spike coding schemes such as rate, temporal, and phase [23, 16]. Among them, we use rate coding due to its reliable performance across various tasks. Rate coding provides spikes proportional to the pixel intensity of the given image. In order to implement this, following previous work [31], we compare each pixel value with a random number ranging between $[I_{min}, I_{max}]$ at every time-step. Here, I_{min}, I_{max} correspond to the minimum and maximum possible pixel intensity. If the random number is greater than the pixel intensity, the Poisson spike generator outputs a spike with amplitude 1. Otherwise, the Poisson spike generator does not yield any spikes. We visualize rate coding in Fig. 7. We see that the spikes generated at a given time-step is random. However, as time goes on, the accumulated spikes represent a similar result to the original image.

C. Appendix: DVS-CIFAR10 dataset

On DVS-CIFAR10, following [40], we downsample the size of the 128×128 images to 42×42 . Also, we divide the total number of time-steps available from the original time-frame data into 20 intervals and accumulate the spikes within each interval. We use a similar architecture as previous work [40], which consists of a 5-layered feature extractor and a classifier. The detailed architecture is shown in Fig. 8 in this appendix.

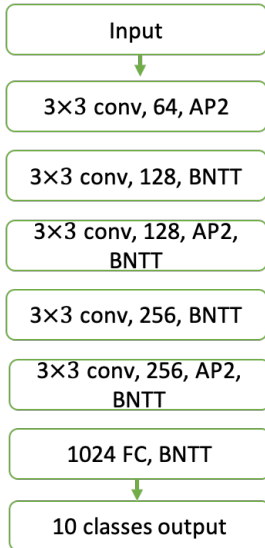


Figure 8. Illustration of network structures for DVS dataset. Here, AP denotes average pooling, FC denotes fully connected configuration.

D. Appendix: Energy Calculation

In this appendix section, we provide the details of energy calculation discussed in Section 4.3 in the main paper. The total computational cost is proportional to the total number of floating point operations (FLOPS). This is approximately the same as the number of Matrix-Vector Multiplication (MVM) operations. For layer l in ANNs, we can calculate FLOPS as:

$$FLOPS_{ANN}(l) = \begin{cases} k^2 \times O^2 \times C_{in} \times C_{out}, & \text{if } l: \text{Conv,} \\ C_{in} \times C_{out}, & \text{if } l: \text{Linear.} \end{cases} \quad (22)$$

Here, k is kernel size. O is output feature map size. C_{in} and C_{out} are input and output channels, respectively. For SNNs, we first define spiking rate $R_s(l)$ at layer l which is the average firing rate per neuron.

$$R_s(l) = \frac{\text{\#spikes of layer } l \text{ over all timesteps}}{\text{\#neurons of layer } l}. \quad (23)$$

Since neurons in SNNs only consume energy whenever the neurons spike, we multiply the spiking rate $R_s(l)$ with FLOPS to obtain the SNN FLOP count.

$$FLOPS_{SNN}(l) = FLOPS_{ANN}(l) \times R_s(l). \quad (24)$$

Finally, total inference energy of ANNs (E_{ANN}) and SNNs (E_{SNN}) across all layer can be obtained.

$$E_{ANN} = \sum_l FLOPS_{ANN}(l) \times E_{MAC}. \quad (25)$$

$$E_{SNN} = \sum_l FLOPS_{SNN}(l) \times E_{AC}. \quad (26)$$

The E_{AC} , E_{MAC} values are calculated using a standard 45 nm CMOS process [14] as shown in Table 1.

Table 4. Energy table for 45nm CMOS process.

Operation	Energy(pJ)
32bit FP MULT (E_{MULT})	3.7
32bit FP ADD (E_{ADD})	0.9
32bit FP MAC (E_{MAC})	4.6 ($= E_{MULT} + E_{ADD}$)
32bit FP AC (E_{AC})	0.9

E. Appendix: Energy Comparison in Neuromorphic Architecture

We further show the energy-efficiency of BNTT in a neuromorphic architecture, TrueNorth [1]. Following the previous work [28, 22], we compute the normalized energy, which can be classified into dynamic energy (E_{dyn}) and static energy (E_{sta}). The E_{dyn} value corresponds to the computing cores and routers, and E_{sta} is for maintaining the state of the CMOS circuit. The total energy consumption

Table 5. Normalized energy comparison on neuromorphic architecture: TrueNorth[1]. We set conversion as a reference for normalized energy comparison. We conduct experiments on CIFAR-10 with a VGG9 architecture.

Method	Time-steps	#Spikes (10^4)	Energy [1]
Conversion	1000	419.30	1
Surrogate	100	141.96	0.3384
BNTT	25	13.106	0.0312

can be calculated as $\#Spikes \times E_{dyn} + \#Time-step \times E_{sta}$, where (E_{dyn}, E_{sta}) are (0.4, 0.6). In Table 5, we show that our BNTT has a huge advantage in terms of energy efficiency in neuromorphic hardware.

F. Appendix: Adversarial Robustness

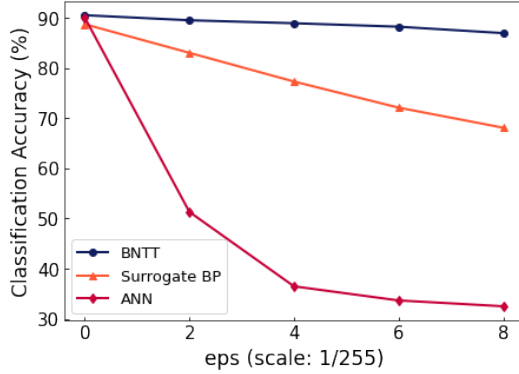


Figure 9. Classification accuracy with respect to the intensity of FGSM attack (eps).

In order to further validate the robustness of BNTT, we conduct experiments on adversarial inputs. We use FGSM [11] to generate adversarial samples for ANN. For a given image x , we compute the loss function $\mathcal{L}(x, y)$ with the ground truth label y . The objective of FGSM attack is to change the pixel intensity of the input image that maximizes the cost function:

$$x_{adv} = x + \epsilon \times \text{sign}(\nabla_x \mathcal{L}(x, y)). \quad (27)$$

We call x_{adv} as “adversarial sample”. Here, ϵ denotes the strength of the attack. To conduct the FGSM attack for SNN, we use the SNN-crafted FGSM method proposed in [35]. In Fig. 9, we show the classification performance for varying intensities of FGSM attack. The SNN approaches (e.g., BNTT and Surrogate BP) show more robustness than ANN due to the temporal dynamics and stochastic neuronal functionality. We highlight that our proposed BNTT shows much higher robustness compared to others. Thus, we assert that BNTT improves robustness of SNNs in addition to energy efficiency and latency.

Table 6. Comparison with Layer Normalization on CIFAR-10 dataset.

Method	Acc (%)
Layer Normalization [2]	75.4
BNTT	90.5

G. Appendix: Comparison with Layer Norm

Layer Normalization (LN) [2] proposed the optimization method for recurrent neural networks (RNNs). They asserted that directly applying BN layers is hardly applicable since RNNs vary with the length of the input sequence. To this end, a LN layer calculates the mean and variance for each single layer. As SNNs also take the time-sequence data as an input, we compare our BNTT with Layer Normalization in Table 6. For all experiments, we use a VGG9 architecture. Also, we set a base learning rate to 0.3 and we use step-wise learning rate scheduling as described in Section 4.1 of our main manuscript. The results show that BNTT is more suitable structure to capture the temporal dynamics of Poisson encoding spikes.

References

- [1] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE transactions on computer-aided design of integrated circuits and systems*, 34(10):1537–1557, 2015. 1, 10, 11
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 11
- [3] Anthony N Burkitt. A review of the integrate-and-fire neuron model: I. homogeneous synaptic input. *Biological cybernetics*, 95(1):1–19, 2006. 1
- [4] Yongqiang Cao, Yang Chen, and Deepak Khosla. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113(1):54–66, 2015. 1, 6, 7
- [5] Iulia M Comsa, Thomas Fischbacher, Krzysztof Potempa, Andrea Gesmundo, Luca Versari, and Jyrki Alakuijala. Temporal coding in spiking neural networks with alpha synaptic function. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8529–8533. IEEE, 2020. 1
- [6] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 38(1):82–99, 2018. 1
- [7] Peter Dayan, Laurence F Abbott, et al. Theoretical neuroscience, vol. 806, 2001. 3

- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [9] Peter U Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, 9:99, 2015. 1
- [10] Peter U Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ieee, 2015. 1
- [11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 8, 11
- [12] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 3
- [13] Bing Han, Gopalakrishnan Srinivasan, and Kaushik Roy. Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13558–13567, 2020. 1, 3, 5, 6, 7
- [14] Mark Horowitz. 1.1 computing’s energy problem (and what we can do about it). In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 10–14. IEEE, 2014. 8, 10
- [15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 2
- [16] Jaehyun Kim, Heesu Kim, Subin Huh, Jinho Lee, and Kiy-oung Choi. Deep neural networks with weighted spikes. *Neurocomputing*, 311:373–386, 2018. 10
- [17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [18] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(7):1346–1359, 2016. 7
- [19] Chankyu Lee, Syed Shakib Sarwar, Priyadarshini Panda, Gopalakrishnan Srinivasan, and Kaushik Roy. Enabling spike-based backpropagation for training deep neural network architectures. *Frontiers in Neuroscience*, 14, 2020. 1, 5, 6, 7
- [20] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016. 2, 6, 8
- [21] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017. 6
- [22] Saber Moradi and Rajit Manohar. The impact of on-chip communication on memory technologies for neuromorphic systems. *Journal of Physics D: Applied Physics*, 52(1):014003, 2018. 10
- [23] Hesham Mostafa. Supervised learning based on temporal coding in spiking neural networks. *IEEE transactions on neural networks and learning systems*, 29(7):3227–3235, 2017. 10
- [24] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36:61–63, 2019. 1, 2, 3, 7, 9
- [25] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. Hfirst: a temporal approach to object recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2028–2040, 2015. 7
- [26] Priyadarshini Panda, Sai Aparna Aketi, and Kaushik Roy. Toward scalable, efficient, and accurate deep spiking neural networks with backward residual connections, stochastic softmax, and hybridization. *Frontiers in Neuroscience*, 14, 2020. 1
- [27] Priyadarshini Panda, Abhronil Sengupta, and Kaushik Roy. Conditional deep learning for energy-efficient and enhanced pattern recognition. In *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 475–480. IEEE, 2016. 5
- [28] Seongsik Park, Seijoon Kim, Byunggook Na, and Sungroh Yoon. T2fsnn: Deep spiking neural networks with time-to-first-spike coding. *arXiv preprint arXiv:2003.11741*, 2020. 8, 10
- [29] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 3, 6
- [30] Nitin Rath, Gopalakrishnan Srinivasan, Priyadarshini Panda, and Kaushik Roy. Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation. *arXiv preprint arXiv:2005.01807*, 2020. 1, 5, 6, 7
- [31] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019. 1, 6, 10
- [32] Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, Michael Pfeiffer, and Shih-Chii Liu. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017. 1
- [33] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Advances in Neural Information Processing Systems*, pages 2483–2493, 2018. 2
- [34] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in neuroscience*, 13:95, 2019. 1, 2, 5, 6, 7
- [35] Saima Sharmin, Nitin Rath, Priyadarshini Panda, and Kaushik Roy. Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations. *arXiv preprint arXiv:2003.10399*, 2020. 8, 11
- [36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8

- [37] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. Hats: Histograms of averaged time surfaces for robust event-based object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1731–1740, 2018. 7
- [38] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469. IEEE, 2016. 5
- [39] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018. 1, 5
- [40] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, Yuan Xie, and Luping Shi. Direct training for spiking neural networks: Faster, larger, better. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1311–1318, 2019. 3, 6, 7, 10