Energy-Efficient Models for High-Dimensional Spike Train Classification using Sparse Spiking Neural Networks

Hang Yin
Worcester Polytechnic Institute
Worcester, USA
hyin@wpi.edu

John Boaz Lee
Worcester Polytechnic Institute
Worcester, USA
jtlee@wpi.edu

Xiangnan Kong
Worcester Polytechnic Institute
Worcester, USA
xkong@wpi.edu

Thomas Hartvigsen
Worcester Polytechnic Institute
Worcester, USA
twhartvigsen@wpi.edu

Sihong Xie
Lehigh University
Bethlehem, USA
six316@lehigh.edu

ABSTRACT

Spike train classification is an important problem in many areas such as healthcare and mobile sensing, where each spike train is a high-dimensional time series of binary values. Conventional research on spike train classification mainly focus on developing Spiking Neural Networks (SNNs) under resource-sufficient settings (e.g., on GPU servers). The neurons of the SNNs are usually densely connected in each layer. However, in many real-world applications, we often need to deploy the SNN models on resource-constrained platforms (e.g., mobile devices) to analyze high-dimensional spike train data. The high resource requirement of the densely-connected SNNs can make them hard to deploy on mobile devices. In this paper, we study the problem of energy-efficient SNNs with sparsely-connected neurons. We propose an SNN model with sparse spatio-temporal coding. Our solution is based on the re-parameterization of weights in an SNN and the application of sparsity regularization during optimization. We compare our work with the state-of-the-art SNNs and demonstrate that our sparse SNNs achieve significantly better computational efficiency on both neuromorphic and standard datasets with comparable classification accuracy. Furthermore, compared with densely-connected SNNs, we show that our method has a better capability of generalization on small-size datasets through extensive experiments.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computer systems organization** → *Neural networks*; • **Computing methodologies** → *Supervised learning*.

KEYWORDS

spiking neural networks, supervised learning, spatio-temporal coding, sparsity, hard-concrete distribution

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467252

ACM Reference Format:
Hang Yin, John Boaz Lee, Xiangnan Kong, Thomas Hartvigsen, and Sihong Xie. 2021. Energy-Efficient Models for High-Dimensional Spike Train Classification using Sparse Spiking Neural Networks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467252

1 INTRODUCTION

Motivation. Our brains have about a hundred billion neurons that fire signals to communicate with each other all the time. Each signal is electrochemical in nature and is referred to as a spike, or an action potential. The most popular way to think of spike *trains* is as a digital sequence of events: 1 for a spike, and 0 for no spike. Such spike trains arise during physical sensory stimuli such as vision and motion, or abstract stimuli such as memory. Recently, spike train classification has attracted much attention in the field of data mining [17, 20, 25, 26]. Unlike tradition classification, classifying spike trains is a task with sequences of spikes as both inputs and outputs. By assuming that all spikes are discrete characteristic events, the processing of information is reduced to the timing and counting of said spikes. Designing machine learning algorithms for spike train classification is very important in many high-impact fields such as sensor systems for disease diagnosis and human activity monitoring.

Knowledge Gap. Spiking Neural Network (SNN) show great potential for dealing with spike train classification [21–23, 25, 26]. Originally proposed to imitate biological information processing [9], the neurons transfer information between one another via spike trains. Unlike Recurrent Neural Networks (RNN), which use continuous value as inputs and outputs, SNNs take sparse spike trains as inputs and outputs, building large-scale neural networks with far less energy and memory on neuromorphic hardware systems, which operate on principles that are fundamentally different from standard digital computers. Thus, SNNs are clear candidates for spike train classification.

However, opportunities are always accompanied by challenges. Due to significant advances in miniaturization of sensor systems, more and more smart devices such as wearable sensors and smart phones for elderly care and aerial robots appear around us, which can produce high-dimensional data in the form of spike trains. These devices require high quality pattern recognition to meet

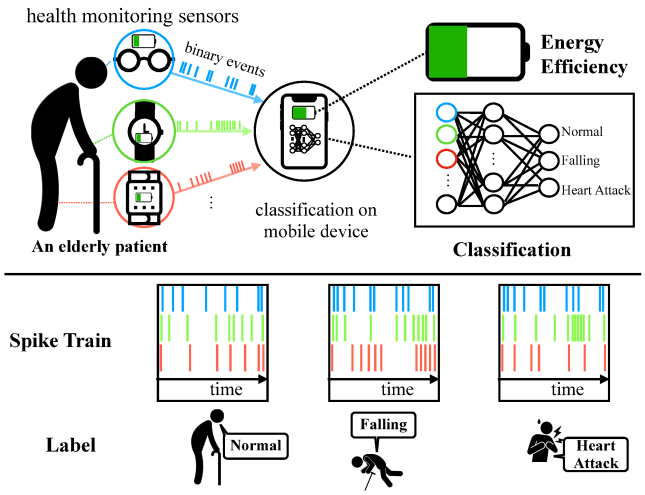

Figure 1: An example of energy-efficient spike train classification problem.

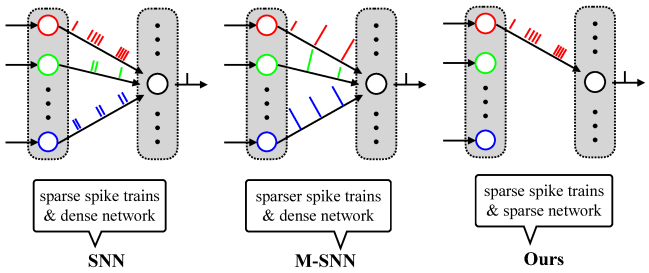

Figure 2: Comparison of the key differences between SNNs [21, 25], M-SNNs [5], and proposed sparse SNN.

their design requirements. At the same time, they are often limited by available energy and thus low computational cost is required during inference. As illustrated in Figure 1, wearable devices incorporated with varieties of motion sensors are used to monitor different physical conditions of seniors for identifying their body conditions. They generate data that cover massive measurements, including heart rate (HR), blood pressure (BP), and oxygen saturation (SpO2), among others, and are expected to be collected by smart devices subject to limited power.To run SNNs efficiently on such high-dimensional data, we need to ensure both high classification accuracy *and* low computational cost during inference.

Regarding computational cost, however, modern SNNs [25, 26] perform many unnecessary computations due to their dense network architectures. These unnecessary computations are caused by *weak* connections between neurons. Weak connections play a limited role in model performance during inference, as shown in the example in Figure 1. On one hand, inferring *Falling* needs activity-related signals given by a person's intertial measurement units (IMU), HR, and BP. On the other hand, inferring *Heart Disease* doesn't consider signals from IMUs, but needs BP, SpO2 and could use more measurements from their photoplethysmograms. As a result, current SNNs are still not suitable for spike train classification, especially when the data is high dimensional and comes from devices with limited power.

Challenges. In this paper, we propose an energy-efficient method for high-dimensional spike train classification. To solve this problem, there are two main challenges:

- *Sparse SNN vs Sparse RNN*: Sparsification techniques have been employed in RNNs [18] to reduce computational costs. However, RNNs model sequences via continuous values, and are outmatched by SNNs for spike train classification. By sparsifying SNNs, we can avoid unnecessary computations caused by weak connections between neurons. A sparsified model has far fewer non-zero parameters and so performs fewer computations during inference, making it more power efficient. Thus, instead of using sparse RNNs, we must find a way to sparsify the network structure of SNNs.
- *Sparsifying Inputs vs Networks*: SNNs have been accelerated by either sparsifying the *inputs* [5] or using stochastic computing [1]. However, by only focusing on spike *rate*, as opposed to spike *timing*, they disregard a major component of the problem. A successful method must consider both rate and timing together to successfully perform high-dimensional spike train classification.

Proposed Method. Inspired by the success of Artificial Neural Networks (ANN) with a sparse structure [7, 13], we propose an SNN model with sparse spatio-temporal coding. We reparameterize the connection between each neuron in an SNN by multiplying each original weight by a binary "gate". Each gate is considered to be a Bernoulli random variable. As a result, our proposed approach allows each neuron in the SNN to consider the necessity of coming into contact with each neuron in the next layer. Therefore, it allows us to penalize the possibility of each gate for being different nonzero with no further restrictions, thereby pruning weak links. This reduces the overall computational cost and adds the benefit of regularization, reducing overfitting. We show through empirical evaluation on multiple real-world datasets that, compared to baselines, the sparse SNN we propose greatly speeds up computation while incurring only a negligible deterioration in classification performance. Meanwhile, we also show improved generalizability by varying the size of the training set.

Contributions. Our contributions in this paper can be summarized as follows:

- We define the problem of spike train classification, which is very important for smart devices with limited energy.
- We propose a sparse SNN for high-dimensional spike train classification to be performed on energy-limited smart devices.
- We demonstrate that our model outperforms recent state-of-the-art alternatives by achieving lower computational cost when tested on both neuromorphic as well as standard datasets with very negligible degradation in classification performance.
- We also demonstrate that our method has excellent generalization capability on small datasets.

2 RELATED WORK

In spike train classification, "indirect" learning methods, such as ANN-to-SNN conversion [4, 6, 8, 15], have been proposed to high

dimensional inputs. These are indirect learning methods because a regular non-spiking ANN (e.g., a multi-layer perceptron) is initially used during the training phase. At inference-time, the trained model is then converted to an SNN. However, there are several disadvantages associated with such indirect training. First, it doesn't align well with how an SNN operates. In ANNs, it does not matter if activations are negative, but firing rates in SNNs are always positive. Furthermore, many limiting constraints are typically added while training the ANN models. These include not using bias terms, only supporting average pooling, and only using ReLU activation functions.

In response, methods for directly training an SNN have recently been proposed [12, 21, 26]. These approaches are mainly based on conventional gradient descent. Most notably, different from previous techniques based only on spatial back-propagation [10, 12], SNNs trained directly using back-propagation in both the spatial as well as the temporal domains [21, 26] have achieved state-of-the-art accuracy on the MNIST and N-MNIST datasets. However, although these methods perform better than the others described above on many real-world datasets, from the perspective of computational efficiency, they are still far from power-efficient in solving high-dimensional spike trains classification. Therefore there have been some recently-proposed power-efficient SNNs [1, 5]. [5] aims to enforce more neurons silence by making input spikes of each neuron sparser. [1] introduces a stochastic SNN by exploiting the benefits of stochastic computing to generate input spike trains and reduce the connection complexity. However, both of them are only applicable to standard datasets, but not to neuromorphic datasets

3 PRELIMINARY

To describe the SNN models with a sparse structure, we first introduce the baseline framework for SNNs, as proposed by [26]. We begin by describing the simplest possible SNN, one which comprises a single neuron with one input entry. This neuron is a recurrent unit that is affected by the current input, and the previous input and output. For each timestep t, it combines the current input with the previous input and output to compute a new value. This value can be referred to as the *membrane* potential in biological neural network. If the membrane is greater than a threshold, the neuron fires and outputs 1 to indicate a spike, otherwise, it outputs 0 to indicate silence. Therefore, for each timestep t, the membrane and output are expressed as follows:

$$u_t = \tau u_{t-1}(1 - z_{t-1}) + w x_t + b, \tag{1}$$

$$z_t = \Theta(u_t - \vartheta), \tag{2}$$

where we write u_t, x_t, and z_t to denote the membrane potential, input, and output of the neuron on timestep t, respectively. $\tau \in [0, 1]$ is the time decay constant hyperparameter, and w and b are the connection and bias between input and this neuron, respectively. $\Theta(\cdot)$ is the step function, which satisfies $\Theta(x) = 0$ when $x < 0$, otherwise $\Theta(x) = 1$.

The SNN expressed in Equations 1-2 mimics natural neural networks more closely than a traditional ANN. In this way, we represent a neuron as the parallel combination of a "leaky" resistor and a capacitor. The second term of the r.h.s. of Equation 1 is used as external current input to charge up the capacitor to update the potential u_t. If the neuron emits a spike $z_t = 1$ at timestep t, the capacitor discharges to a resting potential (which we fix at zero throughout this paper) by using the first term in Equation 1.

An SNN is built by hooking together many of these simple "neurons", so that the output of a neuron can be the input of another. We let $u_i^{t,n}$ and $z_i^{t,n}$ denote the membrane and output of neuron i in layer n at timestep t. The network has parameters $\mathbf{W} = \{\mathbf{W}^1, \ldots, \mathbf{W}^{N-1}\}$, where $\mathbf{W}_{ij}^n$ denote the parameter associated with the connection between neuron j in layer n, and neuron i in layer $n+1$. We also let $l(n)$ denote the number of neurons in layer n and let N be the number of layers in our network. Therefore, for layer $n \in \{2, \ldots, N\}$, we write $\mathbf{u}^{t,n} = (u_1^{t,n}, \ldots, u_{l(n)}^{t,n})^\top$ and $\mathbf{z}^{t,n} = (z_1^{t,n}, \ldots, z_{l(n)}^{t,n})^\top$ to denote the membrane and output vector of neurons in layer n at timestep t. For $n = 1$, we will use $\mathbf{z}^{t,1} = \mathbf{x}^t$ to denote the input vector. Thus, the expression of an SNN is given by:

$$\mathbf{u}^{t,n} = \tau \mathbf{u}^{t-1,n} \odot (1 - \mathbf{z}^{t-1,n}) + \mathbf{W}^{n-1}\mathbf{z}^{t,n-1}, \tag{3}$$

$$\mathbf{z}^{t,n} = \Theta(\mathbf{u}^{t,n} - V_{\text{th}}). \tag{4}$$

From Equations 3-4, the spike signals not only propagate through the layer-by-layer spatial domain, but also affect the neuronal states through the temporal domain. Therefore, it considers both the spatial and temporal directions during the error backpropagation, *i.e.*, spatio-temporal backpropagation (STBP) [25, 26], which significantly improves the network accuracy. During backpropagation, because the activity function $\Theta(\cdot)$ is non-differentiable, it is common to use the rectangular function to approximate the corresponding derivative.

Given the expressions above, we can easily solve a standard SNN classification problem by training a classifier $f\colon \mathbb{R}^{P\times T} \mapsto \{1, \ldots, N\}$ on a given dataset $\{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(K)}, y^{(K)})\}$ that contains K training samples, of which each instance $\mathbf{x}^{(i)} \in \mathbb{R}^{P\times T}$ has an observed label $y^{(i)} \in \{1, \ldots, l(N)\}$. P is the number of input entries and T denotes the length of spike train. To train the SNN, we define the following loss function L for a single training example $(\mathbf{x}, y)$:

$$L = \left(y - \frac{1}{T}\sum_t^T \mathbf{M}\mathbf{z}^{t,N}\right)^2 \tag{5}$$

where $\mathbf{z}^{t,N}$ denotes the voting vector of the last layer N at time step t, $\mathbf{M}$ denotes a constant voting vector connecting neurons in the output layer to a specific class. Thus, we can use STBP to propagate the gradients $\frac{\partial L}{\partial o_i^{t,n+1}}$ from the $(n+1)$-th layer and $\frac{\partial L}{\partial o_i^{t+1,n}}$ from time step $t+1$ as follows:

$$\frac{\partial L}{\partial o_i^{t,n}} = \sum_{j=1}^{l(n+1)} \frac{\partial L}{\partial o_j^{t,n+1}} \frac{\partial o_j^{t,n+1}}{\partial o_i^{t,n}} + \frac{\partial L}{\partial o_i^{t+1,n}} \frac{\partial o_i^{t+1,n}}{\partial o_i^{t,n}} \tag{6}$$

$$\frac{\partial L}{\partial u_i^{t,n}} = \frac{\partial L}{\partial o_i^{t,n}} \frac{\partial o_i^{t,n}}{\partial u_i^{t,n}} + \frac{\partial L}{\partial o_i^{t+1,n}} \frac{\partial o_i^{t+1,n}}{\partial u_i^{t,n}} \tag{7}$$

4 METHODOLOGY

In this work, we propose a sparsification procedure for deep SNNs that accelerates both training and inference while improving the their generalization capabilities through regularization.

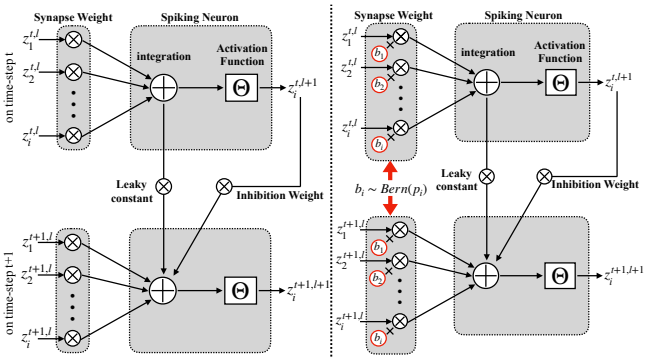

Figure 3: Difference between traditional and sparse SNNs.

4.1 Sparsity regularization and optimization

To build a sparse structure, we consider a re-parametrization of $\mathbf{W}^n_{ij}$, inspired by [13]:

$$\mathbf{W}^n = \tilde{\mathbf{W}}^n \odot \mathbf{b}^n, \quad \mathbf{b}^n_{ij} \in \{0, 1\}, \quad \tilde{\mathbf{W}}^n_{ij} \neq 0 \tag{8}$$

where the $\mathbf{b}^n_{ij}$ correspond to binary "gates" that denote whether the corresponding parameter $\tilde{\mathbf{W}}^n_{ij}$ is utilized or not utilized. $\tilde{\mathbf{W}}^n$ and $\mathbf{b}^n$ is also independent of time t. To simplify the later derivations, we reformulate the minimization of Equation 5 as $L = f(y, \mathbf{x}; \tilde{\mathbf{W}}, \mathbf{b})$.

By letting $p(\mathbf{b}^n_{ij}|\mathbf{\Pi}^n_{ij}) = \text{Bern}(\mathbf{\Pi}^n_{ij})$ be a Bernoulli distribution over each gate $\mathbf{b}^n_{ij}$, we reconsider a sparse network structure as a regularized minimization procedure with a regularization on the number of parameters being used, on average, as follows:

$$L = L_E + L_C, \tag{9}$$

$$L_E = \mathbb{E}_{p(\mathbf{b}|\mathbf{\Pi})}\left[f(y, \mathbf{x}; \tilde{\mathbf{W}}, \mathbf{b})\right], \tag{10}$$

$$L_C = \lambda \sum_{n=1}^{N} \|\mathbf{\Pi}^n\|_1 \tag{11}$$

where L_E denotes the expectation of loss with respect to the Bernoulli distribution of $\mathbf{b}$. Meanwhile, L_C corresponds to the complexity loss that measures the sparsity of the model. Due to the positive nature of each $\mathbf{\Pi}^n_{ij}$, this term also corresponds to the expectation of the amount of gates being "on." Based on [16], the objective described in Equation 9 is a close surrogate to a variational bound involving a spike and slab distribution over the parameters and a fixed coding cost for the parameters when the gates are active. However, the first term in Equation 9 is problematic for $\mathbf{\Pi}$ due to the discrete nature of $\mathbf{b}$, which does not allow for efficient gradient-based optimization. The unbiased gradient estimator in [24] could be employed, however, it suffers from high variance. The straight-through estimator in [3] can also be used in this problem, but it provides biased gradients as it ignores the Heaviside function during gradient evaluation.

In this paper, inspired by [13], we find a simple alternative way to smooth the objective function such that we allow for efficient gradient-based optimization of Equation 9. Let $\mathbf{s}^n_{ij}$ be a continuous random variable with a distribution $q(\mathbf{s}^n_{ij})$ that has parameters $\mathbf{\Phi}^n_{ij}$. We can now let each gate be given by a hard-sigmoid rectifiation of $\mathbf{s}^n_{ij}$ as follows:

$$\mathbf{s}^n_{ij} \sim q\left(\mathbf{s}^n_{ij}|\mathbf{\Phi}^n_{ij}\right), \quad \mathbf{b}^n_{ij} = \min\left(1, \max(0, \mathbf{s}^n_{ij})\right) \tag{12}$$

This allows $\mathbf{b}^n_{ij}$ to be exactly zero. Due to the i.i.d assumption of each $\mathbf{s}^n_{ij}$, we can thus smooth the binary Bernoulli gates by replacing each $\mathbf{b}^n_{ij}$ appearing in the first term of Equation 9 with $\mathbf{s}^n_{ij}$ and the second term with the probability of the variable $\mathbf{s}^n_{ij}$ being positive:

$$L_E = \mathbb{E}_{q(\mathbf{s}|\mathbf{\Phi})}\left[f(y, \mathbf{x}; \tilde{\mathbf{W}}, \mathbf{s})\right], \tag{13}$$

$$L_C = \lambda \sum_{ijn} P(\mathbf{s}^n_{ij} > 0|\mathbf{\Phi}^n_{ij}) \tag{14}$$

Here we similarly have a cost that explicitly penalizes the probability of a gate being different from zero, thus Equations 13-14 act as a close surrogate to the original loss function in Equation 10-11. By following the reparameterization trick [11], we can describe the expression in Equation 13 as an expectation over a parameter-free noise distribution $p(\epsilon)$ and a deterministic and differentiable transformation $g(\cdot)$ of the parameter $\mathbf{\Phi}$ and ϵ. This allows us to make the following Monte Carlo approximation to the intractable expectation over the noise distribution:

$$L_E = \frac{1}{M} \sum_{m=1}^{M} \left[f(y, \mathbf{x}; \tilde{\mathbf{W}}, \mathbf{s}^{(m)})\right], \tag{15}$$

$$\mathbf{s}^{(m)} = \min\left(1, \max(0, g(\mathbf{\Phi}, \epsilon^{(m)}))\right), \epsilon^{(m)} \sim p(\epsilon) \tag{16}$$

Next we provide more details about $g(\cdot)$ in Equations 16.

4.2 The hard concrete distribution

The framework above enables us to employ efficient stochastic gradient-based optimization, while still allowing for exact zeros of the parameters. For the differentiable transformation $g(\cdot)$, we follow [14]: assume that we have a binary concrete random variable s distributed in the interval $(0, 1)$. The parameters of this distribution include $\log \alpha$ and β, where $\log \alpha$ denotes the location and β is referred to as the temperature.

Temperature β controls the degree of approximation. With $\beta = 0$, we recover the original Bernoulli distribution, whereas with $0 < \beta < 1$ we obtain a probability density that concentrates its mass near 0 and 1. Therefore the hard concrete distribution can inherit statistical properties very similar to that of the Bernoulli distribution. We then stretch s to the interval (γ, ς), with $\gamma < 0$ and $\varsigma > 1$. Following [14], we fix $\gamma = -0.1, \varsigma = 1.1,$ and all $\beta = \frac{2}{3}$ throughout this paper. Then we sample b based on the expressions as follows:

$$s = \sigma\left(\left(\log u - \log(1-u) + \log \alpha\right)/\beta\right), \tag{17}$$

$$\bar{s} = s(\varsigma - \gamma) + \gamma, \quad u \sim U(0, 1), \tag{18}$$

$$b = \min(1, \max(0, \bar{s})). \tag{19}$$

Thus, the complexity loss L_C of the objective function in Equation 14 under the hard concrete distribution can be calculated as:

$$L_C = \sum_{ijn} \sigma\left(\log \alpha^n_{ij} - \beta \log \frac{-\gamma}{\varsigma}\right). \tag{20}$$

Algorithm 1 Training code for sparse SNN

Require: : i: Network inputs $\{X^t\}_t^T$; ii: class label Y; iii: parameters and states of convolutional layers ($\{\mathbf{W}^n, \mathbf{b}^l, \mathbf{u}^{0,n}, \mathbf{o}^{0,n}\}_{n=1}^{N_1-1}$); iv: full-connected layers ($\{\mathbf{W}^n, \mathbf{b}^n, \mathbf{u}^{0,n}, \mathbf{o}^{0,n}\}_{n=1}^{N_2-1}$); v: simulation window T; vi: the parameters of the hard-concrete distribution ($\log\alpha^n, \beta, \gamma, \varsigma$); vii: the parameters of iterative LIF ($T, k_\tau, \delta, V_{th}$)

Ensure: : Update network parameters

Forward (inference):

for all $t = 1$ to T **do**
  $\mathbf{b}^n \leftarrow \text{Generate}(\log\alpha^n, \beta, \gamma, \varsigma)$ //Eq. (17)
  $\mathbf{o}^{t,1} \leftarrow \text{EncodingLayer}(X^t)$
  for all $l = 2$ to $N_1 - 1$ **do**
    $(\mathbf{u}^{t,n}, \mathbf{o}^{t,n}) \leftarrow \text{StateUpdate}(\mathbf{W}^{n-1}, \mathbf{b}^{b-1}, \mathbf{u}^{t-1,n}, \mathbf{o}^{t-1,n}, \mathbf{o}^{t,n-1}, \mathbf{x}^{t,n-1})$//Eq. (21,22)
  end for
end for

Loss:

$L \leftarrow \text{ComputeLoss}(\mathbf{Y}, \mathbf{o}^{t,N_2}, \log\alpha)$//Eq. (9)

Backward:

Gradient Initialization: $\frac{\partial L}{\partial \mathbf{o}^{t+1,*}} = 0$
for all $t = T$ to 1 **do**
  $\frac{\partial L}{\partial \mathbf{o}^{t,N_2}} \leftarrow \text{LossGradient}(L, \frac{\partial L}{\partial \mathbf{o}^{t+1,N_2}})$//Eq. (6,7,9)
  for all $l = N_2 - 1$ to 1 **do**
    $(\frac{\partial L}{\partial \mathbf{o}^{t,n}}, \frac{\partial L}{\partial \mathbf{u}^{t,n}}, \frac{\partial L}{\partial \mathbf{W}^n}, \frac{\partial L}{\partial \alpha^n}) \leftarrow \text{BackwardGradient}(\frac{\partial L}{\partial \mathbf{o}^{t,n+1}}, \frac{\partial L}{\partial \mathbf{o}^{t+1,n}}, \mathbf{W}^n, \log\alpha^n)$//Eq. (6,7,9)
  end for
  for all $l = N_1$ to 2 **do**
    $(\frac{\partial L}{\partial \mathbf{o}^{t,n}}, \frac{\partial L}{\partial \mathbf{u}^{t,n}}, \frac{\partial L}{\partial \mathbf{W}^{n-1}}, \frac{\partial L}{\partial \alpha^{n-1}}) \leftarrow \text{BackwardGradient}(\frac{\partial L}{\partial \mathbf{o}^{t,n+1}}, \frac{\partial L}{\partial \mathbf{o}^{t+1,n}}, \mathbf{W}^{n-1}, \log\alpha^{n-1})$//Eq. (6,7,9)
  end for
end for

Given these derivations, we can easily obtain the corresponding iterative state update equations and gradients for sparse deep SNNs.

$$\mathbf{u}_i^{t+1,n+1} = k_\tau \mathbf{u}_i^{t,n+1}\left(1 - \mathbf{o}_i^{t,n+1}\right) + \sum_j^{l(n)} \tilde{\mathbf{W}}_{ij}^n \mathbf{b}_{ij}^n \mathbf{o}_j^{t+1,n} \tag{21}$$

$$\mathbf{o}_i^{t+1,n+1} = \Theta\left(\mathbf{u}_i^{t+1,n+1} - V_{\text{th}}\right) \tag{22}$$

$$\mathbf{b}_{ij}^n = \min\left(1, \max(0, \bar{\mathbf{s}}_{ij}^n)\right), \tag{23}$$

$$\bar{\mathbf{s}}_{ij}^n = \mathbf{s}_{ij}^n(\varsigma - \gamma) + \gamma, \tag{24}$$

$$\mathbf{s}_{ij}^n = \sigma\left(\left(\log u - \log(1-u) + \log\alpha_{ij}^n\right)/\beta\right) \tag{25}$$

We also summarize the overall training process of our proposed sparse SNNs as pseudo-code in Algorithm 1.

5 EMPIRICAL STUDY

To comprehensively validate the effectiveness of our proposed method, we conduct experiments to answer two questions: First, we are interested in computational improvement with very negligible degradation in accuracy. Our work in this paper thus aims to improve the state-of-the-art SNN in this regard. We choose the Spiking CNN (SCNN) [26] as the basic model to which we apply our proposed sparsification procedure on this model and name it sparse SCNN. We then compare the efficiency and accuracy of our Sparse SCNN with SCNN, M-SNN [5], and stochastic SNN [1] on various classification tasks. To better compare our work with them, we follow the same experimental setting as in [26], including the same experimental datasets and the same network structure. Second, we want to explore the generalizability of our proposed model, especially for high dimensional data with very few training samples. We thus test on small training subsets of MNIST and N-MNIST. We validate our sparse deep SNN framework by using the state-of-the-art fully connected and convolutional architectures for deep SNNs [26] on these datasets. To combat randomness in the experiment system, we run all experiments 10 times and report the average results, except when otherwise stated.

5.1 Datasets

We evaluate our sparse SNN models and baselines on various datasets. Using the same datasets as in [26], we test on both static (non-spiking) as well as dynamic (neuromorphic) data.

5.1.1 Static Datasets. MNIST is a popular dataset comprised of a training set with 60,000 samples and a testing set with 10,000 samples of hand-written digits $0-9$. CIFAR-10 is an established computer-vision dataset used for object recognition. It consists of 60,000 32×32 color images containing one of 10 object classes, with 6,000 images per class. Since our method and baselines are spike based learning algorithm, the static images should be converted to spike trains. To this end, we use the Bernoulli sampling conversion from original pixel intensity to the spike trains in this paper. Each normalized pixel is converted to a spike event "1") or no spike event "0") at each time step by using an independent and identically distributed Bernoulli sampling. The probability of generating a spike event is proportional to the normalized value of the entry. Thus, given a certain time window T, the spike events form a spike train. During training, we set T to 12 and 30ms in MNIST and CIFAR-10, respectively.

5.1.2 Dynamic Datasets. Compared to the static datasets, dynamic datasets contain richer temporal features and are therefore more suitable for evaluating SNNs since SNNs can take advantage of the added information. We use the N-MNIST[1] and DVS-Gesture[2] datasets to evaluate the capability of our method on dynamic datasets. The N-MNIST dataset [19] consists of MNIST images converted into a spiking dataset using a Dynamic Vision Sensor (DVS) moving on a pan-tilt unit. Each dataset sample is 300ms long, with a shape of 34×34 pixels, containing both "on" and "off" spikes. The dataset is split into training and test sets following the original split in MNIST of 60,000 training samples and 10,000 testing samples.

The DVS-Gesture dataset [2] contains 1,342 instances of a set of 11 hand and arm gestures, grouped into 122 trials and collected from 29 subjects under 3 different lighting conditions. During each trial, one subject stood against a stationary background and performed all 11 gestures sequentially under the same lighting conditions.

[1] https://www.garrickorchard.com/datasets/n-mnist
[2] https://ibm.ent.box.com/s/3hiq58ww1pbbjrinh367ykfdf60xsfm8

Table 1: Fixed parameter values for the various experiments.

Parameter	Description	Chosen Value (MNIST/CIFAR10/N-MNIST/DVS-Gesture)
T	Time window	30ms,12ms,300ms,1450ms
k_τ	Decay factor	0.1ms,0.3ms,0.2ms,0.2ms
δ	Derivative approximation parameter	1.0,0.5,0.5,0.5
V_{th}	Threshold	0.5
β	Temperature of hard-concrete distribution	2/3
γ, ς	Other parameters of hard-concrete distribution	-0.1, 1.1
λ	The weight factor of sparse regularization	0.001

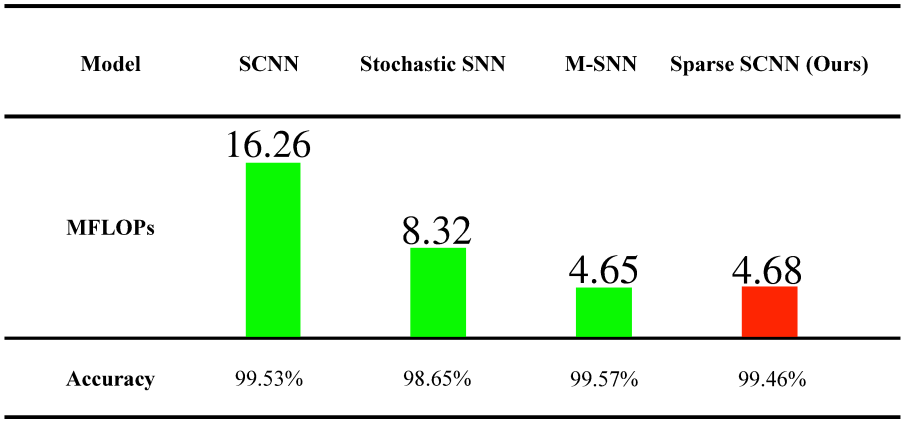

Figure 4: Comparison with SCNN[25], M-SNN[5] and stochastic SNN[1] on MNIST. We show the number of million floating point operations (MFLOPs) after training for each model. These were computed by assuming one FLOP for multiplication and another FLOP for addition.

These gestures are recorded using a DVS128 camera, which is a 28×28-pixel Dynamic Vision Sensor. The problem is to identify the correct action label associated with each action sequence video.

5.2 Network structure

Throughout this section, we use the following notations to describe the deep SNN architecture. Layers are separated by "$-$" and spatial dimensions are separated by "$\times$". A convolution layer is represented by "C" and a pooling layer is represented by "P". For example, "$28 \times 28 - 15C5 - P2 - 10$" represents a 4-layer spiking CNN with 28×28 input, followed by 15 convolution filters that are (5×5), followed by 2×2 pooling layer and finally a dense layer connected to 10 output neurons. Table 2 provides the network structures for experiments. We use the exact same network architecture for our model and baselines for a fair comparison.

5.3 Initialization

In our proposed model, some parameters, such as the model weights and the locations of the hard-concrete distribution, need to be learned while others need to be fixed throughout the optimization. We now discuss our choice for initializing these parameters, which includes the weights, the thresholds and the decay factor for each neuron, the weighting factor for the sparse regularization, and

Table 2: Network structures used for experiments.

Static Dataset	
MNIST	$28\times28 - 15C5 - P2 - 40C5 - P2 - 300 - 10$
CIFAR10	$34\times34\times2 - 32C3 - P2 - 64C3 - P2 - 256 - 10$
Dynamic Dataset	
N-MNIST	$34\times34\times2 - 16C5 - P2 - 32C3 - P2 - 64C3 - 10$
DVS-Gesture	$128\times128\times2 - P4 - 16C5 - P2 - 32C3 - P2 - 512 - 11$

the parameters of the hard-concrete distribution. We divide these parameters into two sets to consider.

First, to better mimic the neural dynamics, we need to control the relative magnitude between the weights and thresholds to avoid too much spiking, which reduces neuronal selectivity. In practice, and as a simplification, we fix the threshold value as a constant for each neuron and only adjust the weights that is responsible for controlling/balancing activity. We initialize all the weight parameters by sampling from the standard uniform distribution followed by normalization.

Second, while sparsifying the network, we follow [14] and set $\gamma = -0.1, \varsigma = 1.1, \beta = \frac{2}{3}$ for the concrete distributions. Meanwhile, we initialize the locations $\log\alpha$ by sampling from a normal distribution

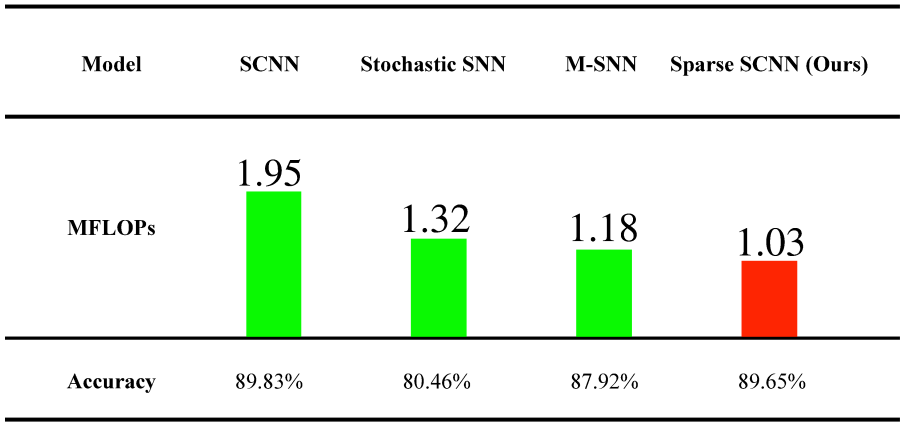

Figure 5: Comparison with SCNN[25], M-SNN[5] and stochastic SNN[1] on CIFAR-10.

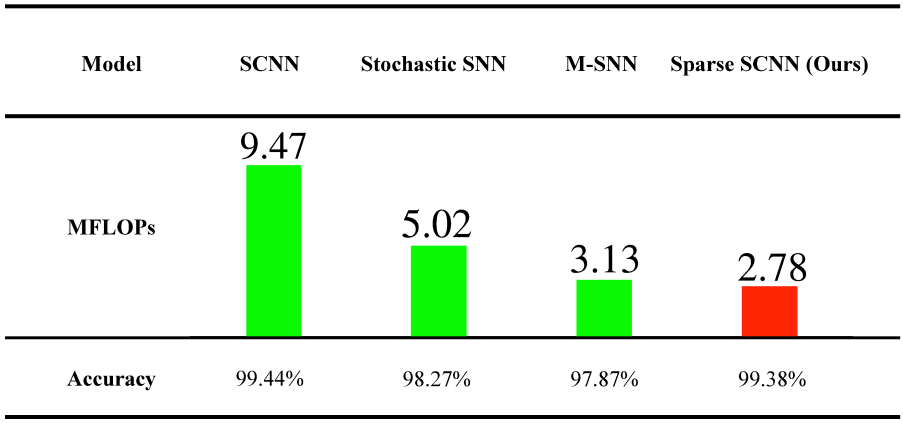

Figure 6: Comparison with SCNN[25], M-SNN[5] and stochastic SNN[1] on N-MNIST.

with a standard deviation of 0.01 and a mean of 1. In practice, we use a single sample of the gate z for each mini-batch of the dataset during the training, even though this can lead to larger variance in the gradients. This way, we show that we can obtain the speedups in optimization with a practical implementation without incurring a significant loss in classification accuracy. A summary of the values of the fixed parameters used is shown in Table 1.

5.4 Evaluation metrics

To evaluate classification performance, we use the standard Accuracy metric. To evaluate the computational efficiency, we count the floating point operations (FLOPs) to measure the potential speedup. FLOPs are computed by assuming one flop for multiplication and one flop for addition.

5.5 Experiment results

In this section, we discuss the experimental results pertaining to each of the two previously-raised research questions separately.

5.5.1 Potential speedup. The tables shown in Figures 4 and 5 compare our proposed sparse deep SNNs with a traditional SNN, M-SNN, and stochastic SNN on the static MNIST and CIFAR-10 datasets, respectively. Even without a complex architecture, the proposed deep SNNs and their competitors still perform well on these datasets. We find that there is only a slight difference in accuracy between our sparse deep SNNs and their competitors (*i.e.*, only between 0.05% and 0.1%). This is a negligible difference. However, as we can observe, there is a significant improvement in the FLOP count between our sparse deep SNNs and the competitors. On CIFAR-10, our sparse network and M-SNN incurs only half the computational cost compared to traditional SNN. On MNIST, this ratio is further reduced to less than 25%, which allows for a potentially significant speedup in inference phase.

Table 3: Comparison on small datasets. Best accuracy highlighted.

Dataset	Method	Accuracy
MNIST	SCNN	69%
	Sparse SCNN (Ours)	**92%**
N-MNIST	SCNN	95%
	Sparse SCNN (Ours)	**97%**

Model	SCNN	Stochastic SNN	M-SNN	Sparse SCNN (Ours)
MFLOPs	5.18	2.61	1.32	1.28
Accuracy	90.53%	77.82%	60.18%	90.15%

Figure 7: Comparison with SCNN[25], M-SNN[5] and stochastic SNN[1] on DVS-Gesture.

The results for the neuromorphic datasets are shown in Figures 6 and 7. We find that both M-SNN and stochastic SNN, which perform well on static datasets, have a significant degradation in accuracy. This demonstrates that the works proposed in [1, 5] are not as suitable for neuromorphic datasets. Meanwhile, by using a sparse network structure, our proposed model incur a slight degradation in accuracy (*i.e.*, decrease between 0.05% and 0.4%) but sparsity can provide a significant speedup – nearly 5x times. Our experimental results show that sparsifying deep SNNs using our proposed framework can greatly speed up training and inference while only incurring a minimal and negligible loss in classification performance. In summary, our proposed model achieves better computational efficiency than previous works when tested on both neuromorphic as well as static datasets and achieves very negligible degradation in accuracy.

5.5.2 Better generalization. To evaluate the ability of the model to generalize, we first compare the performance of our proposed method to SCNN on MNIST as the size of the training set is varied We continuously reduce the training set of MNIST and test on a test set of the same size.

As shown in Figure 8, although both methods achieve very competitive accuracy when the whole training set is used, our sparse deep SNN demonstrates much higher robustness when training set size is decreased. In particular, we observe that the performance of the non-sparse model drops sharply when the training set size is reduced to below 3,000 while the sparse deep SNN's performance remains fairly steady. We conclude that when there are not enough training samples, deep SNNs will easily overfit and even memorize random patterns in the training set. This overfitting can lead to poor generalization. In contrast, by using sparse architecture in deep SNNs, the model shows better generalization even when training samples are limited.

We summarize the results of these experiments, run on MNIST and N-MNIST, in Table 3. During training, we limit the percentage of available training samples to only 1.67% (*i.e.*, only 1,000 samples). As can be observed from the results, all the sparse deep SNNs demonstrate higher accuracy than that of their competitors. This demonstrates that by inducing model sparsity in the architecture, deep SNNs can achieve better generalization in practice.

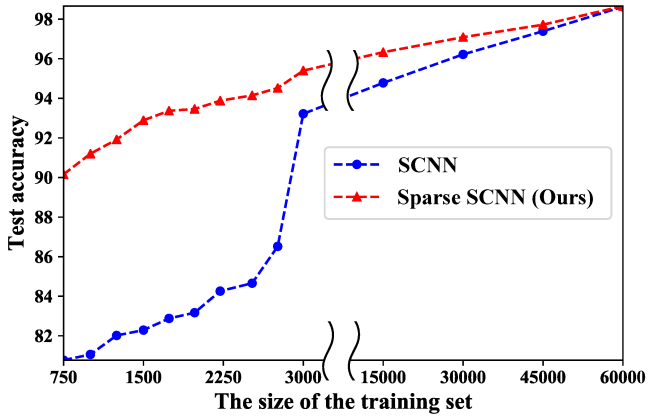

Figure 8: Generalization test on MNIST dataset. The Y-axis denotes the accuracy of each model on the test set (10,000 samples). The X-axis denotes the size of the training set.

5.6 Impact of the weight factor of sparse regularization

In this paper, we achieve sparsity in the network structure of deep SNNs via sparsity regularization, which balances the accuracy with the percent of non-zero weights. Now we quantitatively analyze the impact of this sparsity regularization. We implement a sparse spiking CNN on MNIST, keeping the model configurations the same as our previous experiment on MNIST (28×28-15C5-P2-40C5-P2-300-10).

In Figure 9, the left side shows the level of sparsity at each layer when $\lambda = 0.001$. We use subgraphs of different widths to correspond to the number of weights of each layer in the network, while using the height of blue shaded area to correspond to the sparsity of each layer. The right side shows the overall level of sparsity under different values of λ. The X-axis denotes the value of the sparsity regularization term while the Y-axis denotes the percent of non-zero weights.

The results in Figure 9 show that sparse regularization has a different influence on convolutional layers and fully connected layers. One reason why the the fully connected layers are sparser than the convolutional layers may be due to the difference in nature

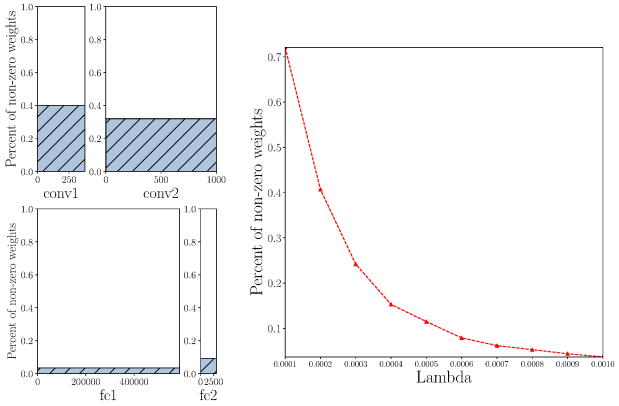

Figure 9: (Left) Level of sparsity with fixed λ=0.001; (Right) Overall sparsity when λ is varied.

of the two types of layers. Convolutional layers apply the same set of weights repeatedly at different positions of the input. On the other hand, each weight in a fully connected layer will only be used once. The parameters in convolutional layers therefore learn general features at possibly multiple locations while each parameter in fully connected layers computes a single feature. As a result, the effect of sparse regularization is more significant in fully connected layers than in convolutional layers.

6 CONCLUSION

In this paper, we aim to design a novel algorithm for high-dimensional spike train classification to be performed on energy-limited smart devices. To this end, we propose a novel sparse architecture for deep SNNs. Our sparsification is achieved by reparametrizing original weights among neurons in the network and then employing a sparsity regularization during optimization. In addition, we also propose an algorithm that can directly train sparse deep SNNs via back-propagation. In empirical study, we choose SCNN as the basic model and apply our proposed sparsification procedure on it. To validate the effectiveness of our proposed method, we compare our model with SCNN, M-SNN and stochastic SNN. Our experimental results on both non-spiking (MNIST and CIFAR-10) and neuromorphic datasets (N-MNIST and DVS-Gesture) show that we can achieve significant speedup with little or no loss in classification accuracy. Furthermore, compared with densely-connected SNNs, we also show through extensive experiments that sparsification can result in better generalizability of the trained model on small-size datasets.

ACKNOWLEDGMENTS

Hang Yin and Xiangnan Kong were supported in part by NSF Grant CNS-1815619. Thomas Hartvigsen was supported in part by the U.S. Dept. of Education Grant P200A150306. Sihong Xie was supported by NSF through grants CNS-1931042, IIS- 2008155, and IIS-1909879.

REFERENCES

[1] Mohammed Alawad, Hong-Jun Yoon, and Georgia Tourassi. Energy efficient stochastic-based deep spiking neural networks for sparse datasets. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 311–318. IEEE, 2017.

[2] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7243–7252, 2017.

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[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.

[5] Ruizhi Chen, Hong Ma, Shaolin Xie, Peng Guo, Pin Li, and Donglin Wang. Fast and efficient deep sparse multi-strength spiking neural networks with dynamic pruning. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[6] 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*, pages 1–8, 2015.

[7] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. In *Advances in neural information processing systems*, pages 3581–3590, 2017.

[8] Yangfan Hu, Huajin Tang, Yueming Wang, and Gang Pan. Spiking deep residual network. *arXiv preprint arXiv:1805.01352*, 2018.

[9] Giacomo Indiveri and Shih-Chii Liu. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 103(8):1379–1397, 2015.

[10] Yingyezhe Jin, Wenrui Zhang, and Peng Li. Hybrid macro/micro level backpropagation for training deep spiking neural networks. In *Advances in neural information processing systems*, pages 7005–7015, 2018.

[11] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583, 2015.

[12] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016.

[13] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.

[14] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[15] Rivu Midya, Zhongrui Wang, Shiva Asapu, Saumil Joshi, Yunning Li, Ye Zhuo, Wenhao Song, Hao Jiang, Navnidhi Upadhay, Mingyi Rao, et al. Artificial neural network (ann) to spiking neural network (snn) converters based on diffusive memristors. *Advanced Electronic Materials*, 5(9):1900060, 2019.

[16] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[17] Srinjoy Mitra, Stefano Fusi, and Giacomo Indiveri. Real-time classification of complex patterns using spike-based learning in neuromorphic vlsi. *IEEE transactions on biomedical circuits and systems*, 3(1):32–42, 2008.

[18] Sharan Narang, Eric Undersander, and Gregory Diamos. Block-sparse recurrent neural networks. *arXiv preprint arXiv:1711.02782*, 2017.

[19] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[20] Michael Pfeiffer and Thomas Pfeil. Deep learning with spiking neurons: opportunities and challenges. *Frontiers in neuroscience*, 12:774, 2018.

[21] Sumit Bam Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. In *Advances in neural information processing systems*, pages 1412–1421, 2018.

[22] Evangelos Stromatias, Miguel Soto, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. An event-driven classifier for spiking neural networks fed with synthetic or dynamic vision sensor data. *Frontiers in neuroscience*, 11:350, 2017.

[23] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, 2019.

[24] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[25] 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.

[26] 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.