A Free Lunch From ANN: Towards Efficient, Accurate Spiking Neural Networks Calibration

Yuhang Li [*1,2] **Shikuang Deng** [*1] **Xin Dong** [3] **Ruihao Gong** [4] **Shi Gu** [1]

Abstract

Spiking Neural Network (SNN) has been recognized as one of the next generation of neural networks. Conventionally, SNN can be converted from a pre-trained ANN by only replacing the ReLU activation to spike activation while keeping the parameters intact. Perhaps surprisingly, in this work we show that a proper way to *calibrate* the parameters during the conversion of ANN to SNN can bring significant improvements. We introduce SNN Calibration, a cheap but extraordinarily effective method by leveraging the knowledge within a pre-trained Artificial Neural Network (ANN). Starting by analyzing the conversion error and its propagation through layers theoretically, we propose the calibration algorithm that can correct the error layer-by-layer. The calibration only takes a handful number of training data and several minutes to finish. Moreover, our calibration algorithm can produce SNN with state-of-the-art architecture on the large-scale ImageNet dataset, including MobileNet and RegNet. Extensive experiments demonstrate the effectiveness and efficiency of our algorithm. For example, our advanced pipeline can increase up to 69% top-1 accuracy when converting MobileNet on ImageNet compared to baselines. Codes are released at a GitHub repo.

1. Introduction

Spiking neural networks (SNNs) are based on the spiking neural behavior in biological neurons (Hodgkin & Huxley, 1952; Izhikevich, 2003). Each neuron in SNNs elicits a spike when its accumulated membrane potential exceeds a threshold, otherwise, it would stay inactive in the current time step. Compared with ANNs, the activation values in SNNs are binarized (i.e., neuromorphic computing (Roy et al., 2019b)), thus resulting in an advantage of energy efficiency for SNNs. Existing works reveal that on specialized hardware, SNNs can save energy by orders of magnitude compared with ANNs (Roy et al., 2019a; Deng et al., 2020). Another vital attribute of SNN is its ability to make inferences in a spatial-temporal paradigm. Specifically, the forwarding pass in SNN is repeated for T steps to get the final result, where the final result is the expectation of the ultimate layer's output across T steps. This allows the flexibility of adjusting T to balance between the latency and accuracy of SNNs for different application scenarios.

Features	Training	Conversion	Calibration
Accuracy($T \leq 100$)	High	Low	High
Scalability	Tiny	Large	Large
Training Speed	Slow	Fast	Fast
# Required Data	Full-set	1000	1000
Inference Speed	Fast	Slow	Fast

Table 1. Features comparison between SNN direct training, ANN-SNN conversion and our SNN calibration.

Conventionally, there are two distinct routes to obtain a functional SNN: (1) training SNN from scratch (Shrestha & Orchard, 2018; Kheradpisheh et al., 2018), and (2) converting a pretrained ANN to SNN (Cao et al., 2015; Diehl et al., 2015). For training from scratch, it is hard to adopt gradient-based optimization methods because of the non-differentiability of the binary activation function in SNNs (Neftci et al., 2019). Although several approaches like surrogate gradients (Wu et al., 2018; Shrestha & Orchard, 2018) and synaptic plasticity (Kheradpisheh et al., 2018) are proposed to mitigate this problem, training SNN from scratch still lacks the scalability to obtain an effective SNN on the ImageNet dataset. Another notable problem is the tremendous resources required to complete the training process. The binary acceleration cannot be employed in GPU training since no CUDA instructions support this kind of computation. As a result, training an SNN may require $T\times$ more time than ANN training.

Besides directly training SNNs from scratch, another family

[*]Equal contribution [1]University of Electronic Science and Technology of China [2]Yale University [3]Harvard University [4]SenseTime Research. Correspondence to: Yuhang Li <yuhang.li@yale.edu>, Shi Gu <gus@uestc.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

of approaches is converting a pretrained ANN into SNN (Diehl et al., 2016; Rueckauer et al., 2017; Sengupta et al., 2018). The conversion process demands less computation and memory than training from scratch. Although some progress in SNN conversion is made, such as threshold balancing (Diehl et al., 2015; Sengupta et al., 2018), weight normalization (Rueckauer et al., 2017), and soft-reset mechanism (Rueckauer et al., 2017; Han et al., 2020), all of them fail to convert ANN with BN layers in low latency time steps (≤ 256), which may significantly increase the latency especially for resource-limited devices. We think the simple *copy-paste* of parameters without any dedicated calibration on SNN will inevitably result in activation mismatch.

In this work, we aim to obtain an SNN in extremely low latency (less than 256 time steps) and in extremely low cost. We choose to utilize a pre-trained ANN and convert it to SNN. Unlike previous conversion work which simply **transplants the weights parameters** to the SNN, in this work we show that the **activation transplating** is much more important. In order to accomplish this, we propose SNN calibration, a new technology family by calibrating the parameters in SNN to match the activation after conversion, and thus significantly narrow the gap in activation distribution between the source ANN and calibrated SNN. We summarize the comparison between our calibration method and the existing conversion & training methods in Table 1. The novel contributions of the paper are threefold:

- We formulate the conversion equation and divide the conventional conversion error into flooring error and clipping error. And then we analyze the error propagation through layers.
- We propose layer-wise calibration algorithm to adjust the network parameters including weights, bias, and initial potential to diminish the conversion error. To accommodate different user requirements, we provide Light Pipeline and Advanced Pipeline to balance accuracy and practical utility.
- We verify our algorithms on large-scale datasets like ImageNet (Deng et al., 2009). In addition to ResNets and VGG networks in previous work, we test a lightweight model MobileNet (Howard et al., 2017) and a large model RegNetX-4GF (Radosavovic et al., 2020) (79.4% top-1 accuracy) for the first time in ANN-to-SNN conversion. Our method can increase up to 69% accuracy in Spiking MobileNet conversion with 256 time steps.

2. Related Work

For training-based SNN, there are several supervised learning algorithms divided into (1) synaptic plasticity and (2) surrogate gradient. Synaptic plasticity methods are based on time-sensitivity and update the connection weight via the two neurons' firing time interval (Kheradpisheh et al., 2018; Iyer & Chua, 2020; LI & LI, 2019). They are more suitable for the neuromorphic image (Amir et al., 2017) or rate coding from static images. On the other hand, surrogate gradient (spiking-based backpropagation) methods use a soft relaxed function to replace the hard step function and train SNN like RNN (Wu et al., 2018; Shrestha & Orchard, 2018). They suffer from the computationally expensive and slow during the training process on complex network architecture(Rathi et al., 2019).

Unlike training from scratch, ANN-to-SNN conversion methods, such as data-based normalization (Diehl et al., 2015; Rueckauer et al., 2016) or threshold balancing (Diehl et al., 2015; 2016), adapt to more complex situations (Tavanaei et al., 2019). The major bottleneck of these methods is how to balance accuracy and inference latency as they require more than 2k time steps to get accurate results. Recently, many methods have been proposed to reduce the conversion loss and simulation length. The soft-reset also called the reset-by-subtraction mechanism, is the most common technique to address the potential reset's information loss (Rueckauer et al., 2016; Han & Roy, 2020). Our IF neuron model also adopts this strategy. Rueckauer et al. (2017) suggest using percentile threshold, which avoids picking the outlier in the activation distribution. Spike-Norm (Sengupta et al., 2018) tests architectures like VGG-16 and ResNet-20. In this work, we further extend the source architecture to MobileNet and RegNet. RMP (Han et al., 2020) and TSC (Han & Roy, 2020) achieves near-to-origin accuracy by adjusting the threshold according to the input and output spike frequency. Deng & Gu (2021) decompose the conversion loss into each layer and reduce it via shifting bias. Low latency converted SNN is an on-going research challenge since it still requires a considerable amount of simulation length. At the same time, most SNN conversion work does not address the BN layers in low latency settings.

3. Preliminaries

Neuron Model for ANN. Considering the ℓ-th fully-connected layer or convolutional layer in the ANNs, its forwarding process can be formulated as,

$$\mathbf{x}^{(\ell+1)} = h(\mathbf{z}^{(\ell)}) = h(\mathbf{W}^{(\ell)}\mathbf{x}^{(\ell)}), 1 \leq \ell \leq n, \quad (1)$$

where $\mathbf{x}^{(\ell)}$, $\mathbf{W}^{(\ell)}$ denote the input activation and weight parameters in that layer respectively, and $h(\cdot)$ is the ReLU activation function. One can optionally train a bias parameter $\mathbf{b}^{(\ell)}$ and add it to pre-activation.

Neuron Model for SNN. Here we use the Integrate-and-Fire (IF) neuron model (Liu & Wang, 2001; Barbi et al., 2003). In specific, suppose at time step t the spiking neurons in layer ℓ receive its binary input $\mathbf{s}^{(\ell)}(t) \in \{0, V_{th}^{(\ell-1)}\}$, the

neuron will update its temporary membrane potential by,

$$\mathbf{v}_{temp}^{(\ell)}(t+1) = \mathbf{v}^{(\ell)}(t) + \mathbf{W}^{(\ell)}\mathbf{s}^{(\ell)}(t), \tag{2}$$

where $\mathbf{v}^{(\ell)}(t)$ denotes the membrane potential at time step t, and $\mathbf{v}_{temp}^{(\ell)}(t+1)$ denotes the intermediate variable that would be used to determine the update from $\mathbf{v}^{(\ell)}(t)$ to $\mathbf{v}^{(\ell)}(t+1)$. If this temporary potential exceeds a pre-defined threshold $V_{th}^{(\ell)}$, it would produce a spike output $\mathbf{s}^{(\ell+1)}(t)$ with the value of $V_{th}^{(\ell)}$. Otherwise, it would release no spikes, i.e. $\mathbf{s}^{(\ell+1)}(t) = 0$. The membrane potential at the next time step $t+1$ would then be updated by *soft-reset* mechanism, also known as *reset-by-subtraction*. Mathematically, we describe the updating rule as

$$\mathbf{v}^{(\ell)}(t+1) = \mathbf{v}_{temp}^{(\ell)}(t+1) - \mathbf{s}^{(\ell+1)}(t), \tag{3}$$

$$\mathbf{s}^{(\ell+1)}(t) = \begin{cases} V_{th}^{(\ell)} & \text{if } \mathbf{v}_{temp}^{(\ell)}(t+1) \geq V_{th}^{(\ell)} \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Note that V_{th}^{ℓ} can be distinct in each layer. Thus, we cannot represent the spike in the whole network with binary signals. This problem can be avoided by utilizing a weight normalization technique to convert the $\{0, V_{th}^{(\ell-1)}\}$ spike to $\{0, 1\}$ spike in every layers, given by:

$$\mathbf{W}^{(\ell)} \leftarrow \frac{V_{th}^{(\ell-1)}}{V_{th}^{(\ell)}}, \quad V_{th}^{(\ell)} \leftarrow 1. \tag{5}$$

Recursively applying the above euqalization, we can use 0,1 spike to represent the intermediate activation for each layer. For the rest of the paper, we shall continue using the notation of $\{0, V_{th}^{(\ell-1)}\}$ spike for simplicity.

As for the input to the first layer and the output of the last layer, we do not employ any spiking mechanism. We use the first layer to direct encode the static image to temporal dynamic spikes, this can prevent the undesired information loss of the Poission encoding. For the last layer output, we only integrate the pre-synaptic input and does not firing any spikes. This is because the output can be either positive or negative, yet Eq. (4) can only convert the ReLU activation.

Converting ANN to SNN Compared with ANN, SNN employs binary activation (i.e. spikes) at each layer. To compensate the loss in representation capacity, researchers introduce the time dimension to SNN by repeating the forwarding pass T times to get final results. Ideally, the converted SNN is expected to have approximately the same input-output function mapping as the original ANN, i.e.,

$$\mathbf{x}^{(\ell)} \approx \bar{\mathbf{s}}^{(\ell)} = \frac{1}{T}\sum_{t=0}^{T}\mathbf{s}^{(\ell)}(t). \tag{6}$$

In practice, the above approximation only holds when T grows to 1k or even higher. However, high T would lead to large inference latency thus damage SNN's practical utility.

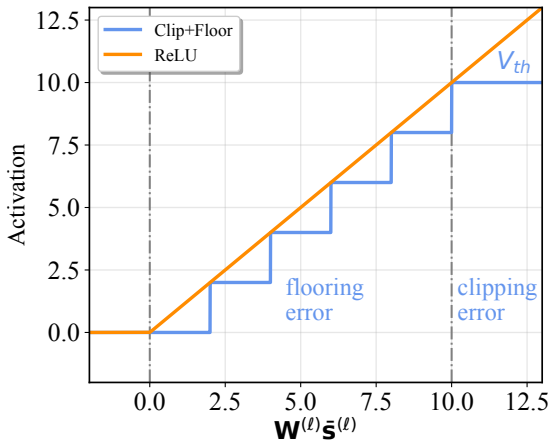

Figure 1. The conversion error between the ReLU activation used in ANN and the output spike in SNN ($V_{th} = 10, T = 5$) contains flooring error and clipping error.

4. Methodology

4.1. Dividing the Conversion Loss

We first use the derivation in Deng & Gu (2021) to deduce the relationship between $\bar{\mathbf{s}}^{(\ell)}$ and $\bar{\mathbf{s}}^{(\ell+1)}$. Suppose the initial membrane potential $\mathbf{v}^{(\ell)}(0) = \mathbf{0}$. Substitute Eq. (2) into Eq. (3) and sum over T, then we get

$$\mathbf{v}^{(\ell)}(T) = \mathbf{W}^{(\ell)}\left(\sum_{t=0}^{T}\mathbf{s}^{(\ell)}(t)\right) - \sum_{t=0}^{T}\mathbf{s}^{(\ell+1)}(t). \tag{7}$$

Since at each time step, the output can be either 0 or $V_{th}^{(\ell)}$, the accumulated output $\sum_{t=0}^{T}\mathbf{s}^{(\ell+1)}(t)$ can be written to $mV_{th}^{(\ell)}$ where $m \in \{0, 1, \ldots, T\}$ denotes the total number of spikes. Note that we assume the terminal membrane potential $\mathbf{v}^{(\ell)}(T)$ lies within the range $[0, V_{th}^{(\ell)})$. Therefore, according to Eq. (7), we have

$$T\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)} - V_{th}^{(\ell)} < mV_{th}^{(\ell)} \leq T\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)}. \tag{8}$$

where $\bar{\mathbf{s}}^{(\ell)}$ is defined in Eq. (6). Then, we can use floor operation and clip operation to determine the m:

$$m = \text{clip}\left(\left\lfloor \frac{T}{V_{th}^{(\ell)}}\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)} \right\rfloor, 0, T\right). \tag{9}$$

Here the clip function sets the upper bound T and lower bound 0. Floor function $\lfloor x \rfloor$ returns the greatest integer that less than or equal to x. Given this formula, we can calculate the expected output spike:

$$\begin{aligned} \bar{\mathbf{s}}^{(\ell+1)} &= \text{clipfloor}(\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)}, T, V_{th}^{(\ell)}) \\ &= \frac{V_{th}^{(\ell)}}{T}\text{clip}\left(\left\lfloor \frac{T}{V_{th}^{(\ell)}}\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)} \right\rfloor, 0, T\right) \end{aligned} \tag{10}$$

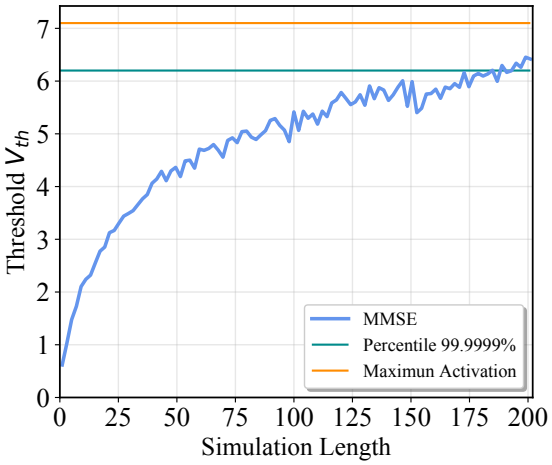

Figure 2. Comparison of the threshold determined by three different approaches. Our MMSE threshold will vary with time steps.

According to Eq. (10), the conversion loss (difference between $\mathbf{x}^{(\ell+1)}$ and $\bar{\mathbf{s}}^{(\ell+1)}$) comes from two aspects, namely the *flooring error* and the *clipping error*.

In Fig. 1, we further indicate that $V_{th}^{(\ell)}$ is crucial for conversion loss because it affects both the flooring and the clipping errors. Increasing $V_{th}^{(\ell)}$ leads to lower clipping error but higher flooring error. Previous work (Diehl et al., 2015; 2016) sets $V_{th}^{(\ell)}$ to the maximum pre-activations across samples in ANN to eliminate the clipping error. However, the maximum pre-activations are usually outliers. Given this insight, the outliers may tremendously increase the flooring error. As a result, they have to use a very large T (for example, 2000) to decrease the flooring error.

4.2. Adaptive Threshold by MMSE

In an effort to better balance flooring error and the clipping error, we use Minimization of Mean Squared Error (MMSE) to obtain the threshold $V_{th}^{(\ell)}$ under different simulation length T. Here we adopt the similar layer-wise optimization problem in (Deng & Gu, 2021), which is formulated by

$$\min_{V_{th}} \left(\text{clipfloor}(\mathbf{x}^{(\ell+1)}, T, V_{th}^{(\ell)}) - \text{ReLU}(\mathbf{x}^{(\ell+1)})\right)^2 \tag{11}$$

Note that the above problem is not guaranteed to be convex, and there is no closed-form solution to this minimization. We hereby sample several batches of training images and use grid search to determine the final result of $V_{th}^{(\ell)}$. Specifically, we linearly sample N grids between $[0, \max(\mathbf{x}^{(\ell+1)})]$, and find the grid that has lowest MSE. We set $N = 100$ and find this option is precise enough to obtain a good solution. Fig. 2 shows the dynamics of our proposed method.

It is worthwhile to note that $V_{th}^{(\ell)}$ does not monotonically increase along with T, because the flooring error may be decreased by slightly increment the threshold. We can further apply MMSE threshold channel-wisely to further decrease the MSE error, as did in (Kim et al., 2019).

4.3. Layer-wise Calibration

Besides adaptive threshold, we further reduce the conversion error by calibrating the parameters of SNN. We first analyze how conversion errors accumulate through layers, and then present a set of layer-wise algorithms to calibrate different types of SNN parameters, including bias, weights and initial membrane potential.

As aforementioned, the output layer of our SNN only accumulate the pre-synaptic inputs through time, that is to say, the final output of SNN is arithmetic sum of output from each time step $\sum_{t=1}^{T} \mathbf{W}^{(n)}\mathbf{s}^{(n)}(t)$. Note that this modification doesn't introduce additional overhead because network output won't participate in further operation. With this modified last layer, our object becomes to minimize the difference between the input to the last layer, i.e. $\bar{\mathbf{s}}^{(n)}$ and $\mathbf{x}^{(n)}$.

Lemma 4.1. *Denote the Frobenius norm as* $||\cdot||$*, the conversion error in the last layer in given by*

$$||\mathbf{e}^{(n)}|| = ||\mathbf{x}^{(n)} - \bar{\mathbf{s}}^{(n)}|| \leq ||\sum_{\ell=1}^{n} Err^{(\ell)} \prod_{k=\ell}^{n} \mathbf{W}^{(k)}||, \tag{12}$$

where $Err^{(\ell)} = \text{clipfloor}(\bar{\mathbf{s}}^{(\ell)}) - \text{ReLU}(\bar{\mathbf{s}}^{(\ell)})$.

We provide the detailed derivation of above lemma in Appendix A. The above lemma indicates that the conversion errors in former layers have a cumulative effect on the subsequent layers. In addition, the conversion error in the ultimate layer is upper bounded by weighted linear combination of layer-wise error. Based on this observation, we develop a set of greedy layer-wise calibration algorithm to correct the conversion error in each layer progressively.

We introduce Light Pipeline and Advanced Pipeline, which can be chose by users according to their memory and computation budgets for layer-wise calibration in practice. The light pipeline achieves fast calibration with less memory and computation consumption by only adjusting bias in SNNs. With a little effort, the light pipeline can outperform state-of-the-art methods by a large margin. We also propose an Advanced Pipeline that achieves best results by calibrating the weights as well as the initial membrane potential in a fine-grained way.

Light Pipeline Light Pipeline only contains Bias Calibration (BC). In order to calibrate the bias parameters, we first

define a reduced mean function:

$$\mu_c(\mathbf{x}) = \frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h}\mathbf{x}_{c,i,j} \tag{13}$$

where w, h are the width and height of the feature-map, and $\mu_c(\mathbf{x})$ computes the spatial mean of the feature-map in each channel c. We notice that the spatial mean of conversion error $\mathbf{e}^{(\ell)} = \mathbf{x}^{(\ell)} - \bar{\mathbf{s}}^{(\ell)}$ can be written by

$$\mu_c(\mathbf{x}^{(\ell)}) = \mu_c(\bar{\mathbf{s}}^{(\ell)}) + \mu_c(\mathbf{e}^{(\ell)}). \tag{14}$$

To ensure the mean output of SNN is equal to the mean output of ANN, we can add the expected conversion error into the bias term as $\mathbf{b}_c^{(\ell)} \leftarrow \mathbf{b}_c^{(\ell)} + \mu_c(\mathbf{e}^{(\ell+1)})$. In practice, we only sample one batch training images and compute the reduced mean to calibrate the bias.

Advanced Pipeline Calibrating the bias only corrects partial error. We need a more fine-grained calibration method. To this end, we propose advanced pipeline which consists of Potential Calibration (PC) and Weights Calibration (WC).

Now consider a non-zero initial membrane potential $\mathbf{v}^{(\ell)}(0)$, we can rewrite Eq. (10) to

$$\tilde{\bar{\mathbf{s}}}^{(\ell+1)} = \frac{V_{th}^{(\ell)}}{T}\text{clip}\left(\left\lfloor \frac{T}{V_{th}^{(\ell)}}\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)} + \frac{\mathbf{v}^{(\ell)}(0)}{V_{th}^{(\ell)}}\right\rfloor, 0, T\right), \tag{15}$$

where $\tilde{\bar{\mathbf{s}}}^{(\ell+1)}$ is the calibrated expected output with non-zero initialization of membrane potential. To obtain a fast calibration for the initial membrane potential, here we make an approximation that:

$$\begin{aligned}\tilde{\bar{\mathbf{s}}}^{(\ell+1)} &\approx \frac{V_{th}^{(\ell)}}{T}\text{clip}\left(\left\lfloor \frac{T}{V_{th}^{(\ell)}}\mathbf{W}^{(\ell)}\bar{\mathbf{s}}^{(\ell)}\right\rfloor, 0, T\right) + \frac{\mathbf{v}^{(\ell)}(0)}{T} \\ &= \bar{\mathbf{s}}^{(\ell+1)} + \mathbf{v}^{(\ell)}(0)/T\end{aligned} \tag{16}$$

Similar to BC, $\mathbf{v}^{(\ell)}(0)/T$ can correct the output distribution of SNN. We can directly set $\mathbf{v}^{(\ell)}(0)$ to $T \times \mathbf{e}^{(\ell+1)}$ to calibrate the initial potential. Note that Potential Calibration does not need to compute spatial mean.

In advanced pipeline, we also introduce Weights Calibration to correct the conversion error in each layer. In WC, the formulation is given by:

$$\min_{\mathbf{W}^{(\ell)}} ||\mathbf{e}^{(\ell+1)}||^2. \tag{17}$$

Here we optimize the whole weights tensor in SNN layer-by-layer and can reduce the conversion error even further. For practical implementation, we will first store input samples in ANN $\mathbf{x}^{(\ell)}$ and input spikes samples in SNN of each time step $\mathbf{s}^{(\ell)}(t)$ and then compute the expected spike input by $\bar{\mathbf{s}}^{(\ell)} = \sum_{t=1}^{T}\mathbf{s}^{(\ell)}(t)$. To further compute the gradient of clipfloor function, we apply the StraightThrough Estimator (Bengio et al., 2013) of the floor operation, i.e.

$$\frac{\partial\lfloor x\rfloor}{\partial x} = 1. \tag{18}$$

To this end, we can use regular training methods like stochastic gradient descent for calibrating the weights. When conducting calibration for weights, the optimization process is very efficient compared to other direct training methods. This is because we first store the expected input from previous layers, and we do not have to perform T times convolution like direct training methods. The major bottleneck of WC is storing the input of SNN. For example, if we set $T = 1024$, then we will do 1024 times forwarding pass for one batch and accumulate them to get the final expected results.

Algorithm 1 Overall Algorithms

input Pretrained ANN; simulation length T
  Fold BN Layers into Conv Layers (cf. Eq. (19))
  Replace AvgPooling Layers to depthwise Conv Layers
  for all $i = 1, 2, \ldots, N$-th layers in the ANN **do**
    Collect input data $\mathbf{x}^{(i)}$, output data $\mathbf{x}^{(i+1)}$ in one batch
    Get MMSE threshold $V_{th}^{(i)}$ using grid search
    Get SNN output $\bar{\mathbf{s}}^{(i+1)}$
    Compute Error term $\mathbf{e}^{(i+1)} = \mathbf{x}^{(i+1)} - \bar{\mathbf{s}}^{(i+1)}$
    if *Light Pipeline* **then**
      Calibrate bias term $\mathbf{b}^{(i)} \leftarrow \mathbf{b}^{(i)} + \mu(\mathbf{e}^{(i+1)})$
    else
      Calibrate Potential $\mathbf{v}^{(i)}(0) \leftarrow T \times \mathbf{e}^{(i+1)}$
      Optimize weights to minimize $||\mathbf{e}^{(\ell+1)}||^2$ via stochastic gradient descent
    end if
  end for
output Converted SNN model

4.4. Average Pooling Layers

Most SNN works do not include the Max Pooling layers since finding the maximum activation neuron ahead of time is impractical, i.e. we cannot determine the maximum neuron $\bar{\mathbf{s}}$ when we only observe $\mathbf{s}(1)$. Therefore, they use Average Pooling Layers to downsample the feature-maps and do not convert them in SNN. However, we argue that Average Pooling Layers will produce non-binary information. For example, a 2×2 kernel AvgPool layer can output 4 possible values $[0, 0.25, 0.5, 1]$ with spike inputs. To make SNN run on corresponding hardware, we convert the AvgPool layer by treating the AvgPool layer as a convolutional layer with specific values. So we can convert the AvgPool layer just like other convolutional layers. More details are included in Appendix.

4.5. Converting BN layers

There is no corresponding module in SNN for Batch Normalization (BN) layers. Rueckauer et al. (2017) propose to absorb the BN parameters to the weight and bias, which can be represented by:

$$\mathbf{W} \leftarrow \mathbf{W}\frac{\gamma}{\sigma}, \quad \mathbf{b} \leftarrow \beta + (\mathbf{b} - \mu)\frac{\gamma}{\sigma}, \tag{19}$$

where μ, σ are the running mean and standard deviation, and γ, β are the transformation parameters in the BN layer.

5. Experiments

To demonstrate the effectiveness and the efficiency of the proposed algorithm, we conduct experiments on CIFAR (Krizhevsky et al.) and ImageNet (Deng et al., 2009) datasets with extremely low simulation length (say $T \leq 256$). In Sec. 5.2, we study the impact of the approximations and design choices made in Sec. 4. In Sec. 5.3, we compare our methods to other methods.

5.1. Implementation Details

For all ANN with BN layers, we fold the BN layer before conversion. We do not convert input images to binary spikes because generating binary spikes requires time and degrades the accuracy. We also do not convert network output to spikes as explained in Sec. 4.3. To correct the bias and membrane potential, we sample one batch of unlabeled data (128 training images). To estimate the MMSE threshold and calibrate weights, we use 1024 training images. In our experiments, we apply the bias shift as described in Deng & Gu (2021). We use Stochastic Gradient Descent with 0.9 momentum to optimize weights in WC, followed by a cosine learning rate decay (Loshchilov & Hutter, 2016). The learning rate for WC is set to 10^{-5}, and no L2 regularization is imposed. We optimize the weights in each layer with 5000 iterations. We will analyze the time and space complexity of our algorithm in the next section. Note that the training details of ANN are included in the Appendix.

5.2. Ablation Study

In this section, we verify the design choices of our proposed adaptive threshold and layer-wise calibration. In all ablation experiments, we test *VGG-16 and ResNet-20* (Sengupta et al., 2018; Han et al., 2020) *on CIFAR100*. We also conduct variance studies by running 5 times with different random seeds and report the mean and standard deviation of the (top1) accuracy on the validation set.

Effect of MMSE Threshold in Conversion We study the effect of choosing different threshold V_{th}. In Table 2, we show that maximum activation has the lowest effect because of the under-fire problem in the initial stage. Our MMSE threshold achieves better results than maximum activation (Diehl et al., 2015) and percentile (Rueckauer et al., 2017) threshold when $T = 16$. As an example, our method is 15.7% higher in accuracy than percentile when converting VGG-16. In ResNet-20, better threshold can significantly improve the accuracy of conversion. Finally, we apply the channel-wise MMSE threshold and further boost the accuracy from 43.5% to 47.4% in VGG-16 and from 72.1% to 73.5% in ResNet-20.

Method	VGG-16 (77.89)		ResNet-20 (77.16)	
	T=16	T=32	T=16	T=32
Maximum Act	2.38±0.17	5.01±1.23	25.67±4.67	51.27±3.82
Percentile 99.9%	3.73±0.38	42.11±0.37	55.00±0.47	71.94±0.19
MMSE (Ours)	**19.42**±0.81	43.53±0.72	58.38±0.62	72.13±0.18
MMSE* (Ours)	17.50±2.46	**47.40**±2.71	**63.55**±0.60	**73.57**±0.06

Table 2. (Top-1) Accuracy comparison on different threshold determination methods for ANN with BN layers. * denotes channel-wise threshold.

Method	VGG-16 (77.89)		ResNet-20 (77.16)	
	T=16	T=32	T=16	T=32
MaxAct + BC	24.61±2.19	32.60±2.61	55.17±5.06	68.78±2.48
Percentile + BC	40.56±1.29	66.87±0.78	69.60±0.14	75.26±0.20
MMSE + BC	44.95±0.52	67.61±0.72	70.78±0.15	75.53±0.15
MMSE* + BC	**52.96**±1.91	**69.19**±0.75	**72.33**±0.13	**75.94**±0.23

Table 3. Light pipeline combines BC with different thresholds for ANN with BN layers and consistently improves accuracy. * denotes channel-wise threshold.

Light Pipeline: Combining Bias Calibration Next, we verify the effect of the proposed Bias Calibration by applying it to different threshold methods. Results are summarized in Table 3, where we can find BC can *consistently improve the accuracy of converted SNN by simply tuning the bias parameters.* For example, BC can boost 22% accuracy in VGG-16 using percentile threshold when $T = 16$. On our MMSE threshold, the Bias Calibration can increase up to 35% accuracy. We should emphasize that BC is cheap and only requires tiny memory space to store the bias term for different T. Therefore our light pipeline is flexible to make trade-off between accuracy and latency.

Advanced Pipeline: Potential and Weights Calibration Our advanced pipeline contains Potential and Weights Calibration that will alter the ANN's parameters to adapt better in spiking configuration. We validate the effect of them on our MMSE threshold mode in Table 5. We can find that Potential Calibration can substantially improve the accuracy of SNN. As an example, the MMSE + BC on VGG-16 only has 44.95% accuracy. However, with PC, we can uplift the accuracy to 59.52%. We also find WC is slightly more stable since the variance of the results is lower.

Method	Use BN	Convert AP	ANN Acc.	$T = 32$	$T = 64$	$T = 128$	$T = 256$	$T \geq 2048$
ResNet-34 (He et al., 2016) ImageNet								
Spike-Norm (Sengupta et al., 2018)	✗	✗	70.69	-	-	-	-	65.47
Hybrid Train (Rathi et al., 2019)	✗	✗	70.20	-	-	-	61.48	65.10
RMP (Han et al., 2020)	✗	✗	70.64	-	-	-	55.65	69.89
TSC (Han & Roy, 2020)	✗	✗	70.64	-	-	-	55.65	69.93
Opt. (Deng & Gu, 2021)*	✗	✓	70.95	33.01	59.52	67.54	70.06	70.98
Ours (Light Pipeline)	✗	✓	70.95	**62.34**	**68.38**	**70.15**	**70.75**	70.97
Opt. (Deng & Gu, 2021)*	✓	✓	75.66	0.09	0.12	3.19	47.11	75.08
Ours (Light Pipeline)	✓	✓	75.66	**50.21**	**63.66**	**68.89**	**72.12**	75.44
Ours (Advanced Pipeline)	✓	✓	75.66	**64.54**	**71.12**	**73.45**	**74.61**	75.45
VGG-16 (Simonyan & Zisserman, 2014) ImageNet								
Spike-Norm (Sengupta et al., 2018)	✗	✗	70.52	-	-	-	-	69.96
Hybrid Train (Rathi et al., 2019)	✗	✗	69.35	-	-	-	62.73	65.19
RMP (Han et al., 2020)	✗	✗	73.49	-	-	-	48.32	73.09
TSC (Han & Roy, 2020)	✗	✗	73.49	-	-	-	69.71	73.46
Opt. (Deng & Gu, 2021)*	✗	✓	72.40	54.92	66.51	69.94	71.35	72.09
Ours (Light Pipeline)	✗	✓	72.40	**69.30**	**71.12**	**71.85**	**72.20**	72.29
Opt. (Deng & Gu, 2021)*	✓	✓	75.36	0.114	0.118	0.122	1.81	73.88
Ours (Light Pipeline)	✓	✓	75.36	**24.88**	**56.77**	**70.49**	**73.66**	75.15
Ours (Advanced Pipeline)	✓	✓	75.36	**63.64**	**70.69**	**73.32**	**74.23**	75.32
MobileNet (Howard et al., 2017) ImageNet								
Opt. (Deng & Gu, 2021)*	✓	✓	73.40	0.110	0.104	0.100	0.964	68.21
Ours (Light Pipeline)	✓	✓	73.40	0.254	**12.62**	**53.91**	**65.86**	**72.19**
Ours (Advanced Pipeline)	✓	✓	73.40	**37.43**	**56.26**	**65.40**	**69.02**	**72.38**
RegNetX-4GF (Radosavovic et al., 2020) ImageNet								
Opt. (Deng & Gu, 2021)*	✓	✓	**80.02**	0.218	3.542	48.60	71.22	78.33
Ours (Light Pipeline)	✓	✓	**80.02**	**35.63**	**65.28**	**74.37**	**77.33**	**79.15**
Ours (Advanced Pipeline)	✓	✓	**80.02**	**55.70**	**70.96**	**75.78**	**77.50**	**79.21**

Table 4. Comparison of our algorithm with other existing SNN conversion works. *Use BN* means use BN layers to optimize ANN, *Convert AP* means use depthwise convolutional layers to replace Average Pooling layers. * denotes self-implementation results.

Method	VGG-16 (77.89)		ResNet-20 (77.16)	
	T=16	T=32	T=16	T=32
MMSE + PC	59.52±1.06	70.62±0.47	73.37±0.28	76.21±0.06
MMSE* + PC	65.29±0.86	73.53±0.27	**74.22**±0.25	**76.68**±0.12
MMSE + PC + WC	65.02±0.33	73.51±0.23	73.36±0.28	76.32 ±0.13
MMSE* + PC + WC	**67.14**±0.88	**74.52**±0.36	74.02±0.20	76.52±0.16

Table 5. Advanced pipeline use Potential and Weights Calibration to optimize SNN. * denotes channel-wise threshold.

5.3. Comparison to Previous Work

In this section, we compare our proposed algorithm with other existing work. We first test ImageNet models[1]. Here we choose the widely adopted ResNet-34 (He et al., 2016) and VGG-16 (Simonyan & Zisserman, 2014) in the existing literature. Note that we test ANNs both with and without BN layers. We additionally verify our algorithm on **MobileNet** (Howard et al., 2017). To our best knowledge, this is the first work that studies Spiking MobileNet conversion.

Results can be found in Table 9. For both ResNet-34 and VGG-16 without BN, our light pipeline is within 1% accuracy loss when $T = 128$. On models with BN layers, our method can substantially improve the conversion loss. In particular, the light pipeline can improve 50.1%, and the advanced pipeline can improve 64.4% accuracy in ResNet-34 with BN Conversion when $T = 32$. Baseline methods still produce a large accuracy gap even on VGG-16 without BN layers and AvgPool Conversion. While our light pipeline reaches 73.66 (less than 2% accuracy drop) when $T = 256$. The superiority of our algorithm is also reflected in Spiking MobileNet conversion, where the baseline method (Deng & Gu, 2021) crashed when $T \leq 256$. To reach acceptable accuracy of Spiking MobileNet, the baseline method has to increase T to 2048. However, our advanced pipeline can achieve higher accuracy while reducing $8\times$ simulation length (69.02 when $T = 256$). Finally, we test our algorithm on a large ANN, RegNetX-4GF (Radosavovic et al., 2020) which achieves 79.4% top-1 accuracy. Our light pipeline reaches 73% accuracy when $T = 128$ and our advanced pipeline reaches 75.8% accuracy when $T = 128$.

[1]We include the comparison on CIFAR dataset in Appendix D.

Model	BC	PC	WC
VGG-16	0.098±0.003	0.106±0.017	4.70±0.037
MobileNet	2.29±0.005	2.19±0.01	27.6±1.62

Table 6. Conversion time (minutes) for different calibration algorithm. We set $T = 64$ and test VGG-16 on CIFAR100 and MobileNet on ImageNet.

#Samples	32	64	128	256
VGG-16	64.50±1.19	65.12±1.02	66.04±0.72	66.20±0.58
ResNet-20	75.35±0.16	75.41±0.18	75.43±0.09	75.41±0.08

Table 7. Comparison of the accuracy using different number of data samples for bias correction.

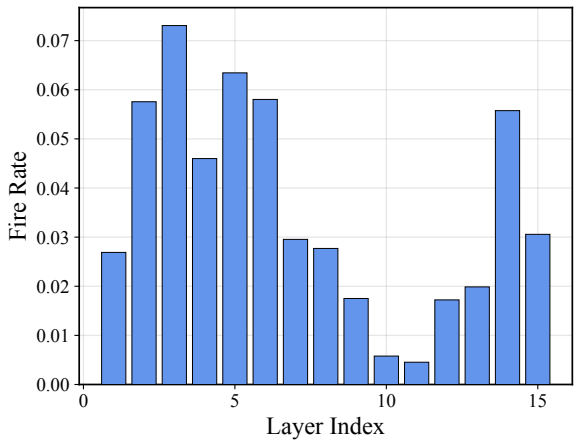

Figure 3. Firing rate visualization of VGG-16.

5.4. Complexity Study

Time Complexity During run-time, our converted SNN will not produce additional inference time. However, converting SNN using light or advanced pipeline may require time and computing resources. The time needed for each calibration is described in the table below. All experiments were tested on a single NVIDIA GTX 1080TI with 5 runs. In Table 6, we can see that the Bias Calibration and Potential Calibration only takes limited time on CIFAR100 and ImageNet. Using Weights Calibration is much expensive than the other two methods. For example, calibrate a MobileNet on ImageNet may take 30 minutes using the advanced pipeline. We should emphasize that our advanced pipeline is still much cheaper than Hybrid Train (Rathi et al., 2019), which requires 20 epochs of end-to-end training (hundreds of GPU hours).

Space Complexity In this section, we report the memory requirements for each calibration algorithm. Since our method will calibrate a new set of parameters for different T, therefore it is necessary to study the model size if we want to deploy SNNs under different T. Specifically, calibrating the bias of ResNet-34 on ImageNet only requires 0.3653MB memory. However, calibrating the weights and potential requires 83.25MB and 18.76MB, respectively. Thus, our proposed light pipeline is both computational and memory cheap and is optimal for flexible SNN conversion. In contrast, the advanced pipeline (PC and WC) requires much more memory space. One may optionally only apply PC to lower down the memory footprint of ResNet-34. Interestingly, some tiny structures like MobileNet share less weights memory (12.21MB) but higher activation memory (19.81MB).

Data Sample Complexity We study the robustness of the our algorithm by increasing the size of calibration dataset. Here we test Bias Calibration in Table 7 on ResNet-20 and VGG-16 ($T = 32$). By increasing the number of samples for calibration, the accuracy will also increase. However, we can see that in ResNet-20 the effect of samples is trivial. While in VGG-16, increasing the number of samples from 32 to 256 can increase 1.7% mean accuracy. We also find that more samples lead to a stable calibration result. Therefore, we recommend using at least 128 images for calibration. In our experiments, the same trend is also observed in other calibration algorithms.

5.5. Efficiency and Sparsity

In this section we visualize the sparsity of our calibrated SNN. We choose the spike VGG-16 on the ImageNet dataset, with $T = 64$. We only leverage light pipeline (bias calibration) and record the mean firing ratio across the whole validation dataset, which also corresponds to sparsity of the activation. The firing ratio is demonstrated in Fig. 3, where we can find the maximum firing ratio is under 0.08, and the minimum firing ratio can be 0.025. To quantitaively compute the energy saving, we use the energy-estimation euqation in Rathi & Roy (2020). For addition, we measure it by $0.9J$ per operation; for multiplication, we measure it by $4.6J$ per operation. On the event-driven neuromorphic hardware, a non-firing neuron will not cost any energy. Based on this rule, our calibrated spiking VGG-16 only costs 69.36% energy of ANN's consumption.

Conclusion

In this work, we analyze the composition of conversion error and its cumulative effect. To reduce the gap between ANN activation and SNN activation, we propose adaptive threshold to determine the threshold in different time steps. We also introduce the layer-wise calibration, which significantly improves the performance of SNN compared with other *simple-copy* methods. Layer-wise calibration is easy to use and only requires a few training images. Our method estab-

lishes new state-of-the-art performance for SNN conversion. It can successfully convert challenging architectures like MobileNet and RegNetX-4GF with a low latency (less than 256 time steps) for the first time. Even when converting the ANN with Batch Normalization layers, our method can preserve high classification accuracy.

Acknowledgement

This work is supported by NSFC 61876032. We greatly thank annoymous reviewers for their kind suggestions to this work.

References

Amir, A., Taba, B., Berg, D., Melano, T., McKinstry, J., Di Nolfo, C., Nayak, T., Andreopoulos, A., Garreau, G., Mendoza, M., et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7243–7252, 2017.

Barbi, M., Chillemi, S., Di Garbo, A., and Reale, L. Stochastic resonance in a sinusoidally forced lif model with noisy threshold. *Biosystems*, 71(1-2):23–28, 2003.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Cao, Y., Chen, Y., and Khosla, D. Spiking deep convolutional neural networks for energy-efficient object recognition. *International Journal of Computer Vision*, 113(1): 54−−66, 2015.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Deng, L., Wu, Y., Hu, X., Liang, L., Ding, Y., Li, G., Zhao, G., Li, P., and Xie, Y. Rethinking the performance comparison between snns and anns. *Neural Networks*, 121: 294 – 307, 2020.

Deng, S. and Gu, S. Optimal conversion of conventional artificial neural networks to spiking neural networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=FZ1oTwcXchK.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Diehl, P. U., Neil, D., Binas, J., Cook, M., and Liu, S. C. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015.

Diehl, P. U., Zarrella, G., Cassidy, A., Pedroni, B. U., and Neftci, E. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–8. IEEE, 2016.

Han, B. and Roy, K. Deep spiking neural network: Energy efficiency through time based coding. In *European Conference on Computer Vision*, 2020.

Han, B., Srinivasan, G., and Roy, K. 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*, pp. 13558–13567, 2020.

He, K., Zhang, X., Ren, S., and Jian, S. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

Hodgkin, A. L. and Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117 (4):500–544, 1952.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Iyer, L. R. and Chua, Y. Classifying neuromorphic datasets with tempotron and spike timing dependent plasticity. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Izhikevich, E. M. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. Stdp-based spiking deep convolutional

neural networks for object recognition. *Neural Networks*, 99:56–67, 2018.

Kim, S., Park, S., Na, B., and Yoon, S. Spiking-yolo: Spiking neural network for energy-efficient object detection. *arXiv preprint arXiv:1903.06530*, 2019.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

LI, S.-L. and LI, J.-P. Research on learning algorithm of spiking neural network. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pp. 45–48. IEEE, 2019.

Liu, Y.-H. and Wang, X.-J. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. *Journal of computational neuroscience*, 10(1):25–45, 2001.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Neftci, E. O., Mostafa, H., and Zenke, F. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10428–10436, 2020.

Rathi, N. and Roy, K. Diet-snn: Direct input encoding with leakage and threshold optimization in deep spiking neural networks. *arXiv preprint arXiv:2008.03658*, 2020.

Rathi, N., Srinivasan, G., Panda, P., and Roy, K. Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation. In *International Conference on Learning Representations*, 2019.

Roy, D., Chakraborty, I., and Roy, K. Scaling deep spiking neural networks with binary stochastic activations. In *2019 IEEE International Conference on Cognitive Computing (ICCC)*, pp. 50–58. IEEE, 2019a.

Roy, K., Jaiswal, A., and Panda, P. Towards spike-based machine intelligence with neuromorphic computing. *Nature*, 575(7784):607–617, 2019b.

Rueckauer, B., Lungu, I.-A., Hu, Y., and Pfeiffer, M. Theory and tools for the conversion of analog to spiking convolutional neural networks. *arXiv: Statistics/Machine Learning*, (1612.04052):0–0, 2016.

Rueckauer, B., Lungu, I.-A., Hu, Y., Pfeiffer, M., and Liu, S.-C. Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. *Frontiers in neuroscience*, 11:682, 2017.

Sengupta, A., Ye, Y., Wang, R., Liu, C., and Roy, K. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in Neuroence*, 13, 2018.

Shrestha, S. B. and Orchard, G. Slayer: Spike layer error reassignment in time. In *Advances in Neural Information Processing Systems*, pp. 1412–1421, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, 2019.

Wu, Y., Deng, L., Li, G., Zhu, J., and Shi, L. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12: 331, 2018.