

Converting High-Performance and Low-Latency SNNs through Explicit Modelling of Residual Error in ANNs

Zhipeng Huang^{1*}, Jianhao Ding^{2*}, Zhiyu Pan², Haoran Li³, Ying Fang^{1†},
Zhaofei Yu² and Jian K. Liu⁴

¹Fujian Normal University

²Peking University

³Xidian University

⁴University of Birmingham
qsx20221313@student.fjnu.edu.cn

Abstract

Spiking neural networks (SNNs) have garnered interest due to their energy efficiency and superior effectiveness on neuromorphic chips compared with traditional artificial neural networks (ANNs). One of the mainstream approaches to implementing deep SNNs is the ANN-SNN conversion, which integrates the efficient training strategy of ANNs with the energy-saving potential and fast inference capability of SNNs. However, under extreme low-latency conditions, the existing conversion theory suggests that the problem of misrepresentation of residual membrane potentials in SNNs, i.e., the inability of IF neurons with a reset-by-subtraction mechanism to respond to residual membrane potentials beyond the range from resting potential to threshold, leads to a performance gap in the converted SNNs compared to the original ANNs. This severely limits the possibility of practical application of SNNs on delay-sensitive edge devices. Existing conversion methods addressing this problem usually involve modifying the state of the conversion spiking neurons. However, these methods do not consider their adaptability and compatibility with neuromorphic chips. We propose a new approach based on explicit modeling of residual errors as additive noise. The noise is incorporated into the activation function of the source ANN, which effectively reduces the residual error. Our experiments on the CIFAR10/100 dataset verify that our approach exceeds the prevailing ANN-SNN conversion methods and directly trained SNNs concerning accuracy and the required time steps. Overall, our method provides new ideas for improving SNN performance under ultra-low-latency conditions and is expected to promote practical neuromorphic hardware applications for further development.

*these authors contributed equally

†corresponding author

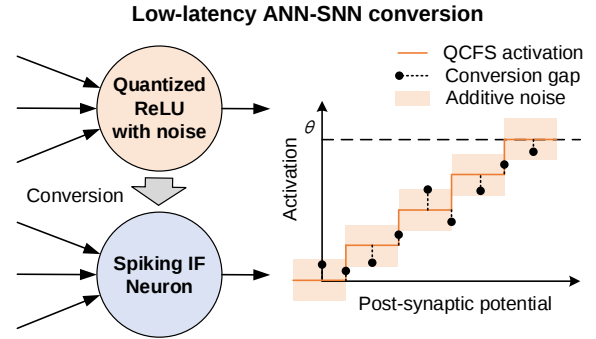


Figure 1: Diagram of our proposed conversion method. State-of-the-art low-latency conversion methods requiring training ANN with quantized activation still bring about a conversion gap in activation. We propose to incorporate additive noise into ANN activation, aiming to compensate for the activation gap.

1 Introduction

Spiking neural networks (SNNs), as third-generation artificial neural networks [40] have attracted a lot of attention from researchers, mainly due to their significant advantages deploying on neuromorphic hardware: low power consumption and high efficiency [32; 33]. Compared with second-generation artificial neural networks (ANNs) full of float-point multiplication computations, SNNs obtain activation of high sparsity, in which only discrete spikes representing 0 and 1 can do the multiplication-free computation. These attributes enable efficient neuromorphic designs [34; 37], which make SNNs a promising candidate for real-time application [38; 13].

Nevertheless, non-differentiable spike information transmission makes training high-performance SNNs not a simple task compared with ANN training. The current major SNN training methods include supervised surrogate backpropagation [35; 6], ANN-SNN conversion [10; 11; 12], unsupervised spike-timing-dependent plasticity learning [36; 39] as well as some hybrid or online training methods [8; 7; 9; 16]. For deep SNNs, unsupervised learning methods perform poorly on large-scale datasets. Supervised surrogate backpropagation requires expansion along the temporal axis

utilizing surrogate functions [6], which requires huge GPU resources. The current progress of online training methods breaks the temporal dependence between spikes of neighboring time steps, yet requires inference multiple time steps and still performs worse among supervised methods. Therefore, in the supervised training domain, ANN-SNN conversion bypasses the non-differentiable problem by training an ANN counterpart and is still promising for large-scale networks and datasets. The key concentration is on reducing the latency that the network needs. For traditional conversion methods, a long inference time is required to match the firing rate of the SNN with the activation value of the ANN [19]. However, this increases the overhead on neuromorphic chips. Addressing the problem of conversion errors at low latency has become an important challenge in current research. Attempts to lower latency by proposing hybrid methods mixing conversion and surrogate training still fail to decrease the training cost.

The challenge of low-latency ANN-SNN conversion arises from conversion errors, which have been identified by previous studies [29; 1], resulting in a performance gap under low-latency conditions. To eliminate these errors, Bu et al. [1] employed quantized QCFS activation in the source ANN from the assumption of bounded residual membrane potentials. They didn't explicitly optimize the so-called "unevenness error" in their work, thus the performance of QCFS is acceptable at 8-time steps but poor at ultra-low time steps (e.g., $T \leq 4$). Efforts to further eliminate the conversion errors with low latency include shifting the initial membrane potential of spiking neurons on the fly [2] and BPTT fine-tuning after conversion [29]. These methods flaw as they either need modifications to neuromorphic hardware or run extra surrogate training on GPUs. By contrast, eliminating conversion errors in ANN training is ideal as it introduces no extra cost. Thus, we propose to further explicitly model residual conversion errors and incorporate them into ANN training. The main contributions of this paper are summarised below:

- We find that the conversion loss for low-latency SNN primarily stems from residual errors between quantized ANNs and converted SNNs. We experimentally observed the distribution of residual errors in each layer with low time steps and find that the variance of the residual error is significant while the mean is rather small.
- Based on the observation, we propose to explicitly model the residual error as a Gaussian noise with a zero mean and integrate the noise into the quantized activation of the source ANN during training, aiming to compensate for the gap between the source ANN and the converted SNN.
- Due to the difficulty of predicting and estimating residual errors after conversion, we induct the noise intensity from a small validation set and apply a layer-wise error-compensating strategy while training.
- We demonstrate the effectiveness of our method on CIFAR10/100. Under low-latency conditions, our method outperforms previous state-of-the-art methods and shows improvement in performance with low time

steps. For example, we attain an outstanding top-1 accuracy of 93.72% on CIFAR-10 with just 2 time steps.

2 Related Work

The core idea of ANN-SNN conversion is to combine the computational efficiency of ANNs and the biological rationality of SNNs to accomplish specific AI tasks. This approach avoids the huge computational consumption associated with directly training SNNs [48] and is a feasible solution for training deep SNNs. Cao et al. [10] found the equivalence between ReLU activation and the fire rate of the IF neuron, laying the foundation for ANN-SNN conversion based on spike firing rate. Subsequently, various weight-threshold balance methods have been popular in reducing the inference delay (denoted as T) to typically up to hundreds [11; 12; 19; 17; 21], yet the delay is still unacceptable for large-scale architectures. To further improve the performance, some works have introduced methods to modify neurons, such as burst spikes [22] and symbolic signed neurons [23; 24]. However, they destroy the binary property of spiking neurons, making it difficult to apply to neuromorphic chips.

Recent advances in ANN-SNN conversion call for low-latency conversion, where conversion error reduction becomes the focus of research. The conversion error is the gap between the ANN activation and the firing rate of SNN. Deng et al. [18] proposed ThreshRelu and found that the conversion error can be optimized layer by layer. Later, quantization clip-floor-shift (QCFS) activation proposed by [1] simulates the characteristics of SNNs as much as possible and has already achieved good performance under low-latency conditions ($T \leq 8$). However, there is still a performance gap between ANN and SNN when the latency is lower. The fundamental reason is that the so-called "unevenness error" pointed out by the authors has not been effectively solved. This is precisely an important challenge. To resolve this error, Hao et al. proposed SRP [47] and OffsetSpikes methods [2]. However, these methods include restricting the converted SNN to release and move the initial membrane potential, increasing the overhead in deployment. Wang et al. [29] considered modifying the initial membrane potentials and fine-tuning them by surrogate training in the converted SNN to solve the unevenness error. While these methods improve performance under low latency compared to previous methods, they do not consider the usability of the methods in neuromorphic hardware chips. The goal of ANN-SNN conversion is to take advantage of GPU training of ANNs to create a well-trained, low-latency SNN model without introducing complex operations. The paper aims to further examine the conversion error and reduce the performance gap between ANNs and SNNs at low latency by modeling the error in the activation function of the source ANN. This approach will help improve the performance of SNNs on neuromorphic hardware chips.

3 Preliminaries

3.1 Neuron Model

ANN neurons. For traditional ANNs, the input tensor is fed into the ANN and processed layer by layer through weighted summation and the continuous nonlinear activation function

$f(\cdot)$ to produce output activation. The forward propagation for neurons in layer l in an ANN can be expressed as:

$$\mathbf{a}^l = f(\mathbf{W}^l \mathbf{a}^{l-1}), \quad l = 1, 2, \dots, M \quad (1)$$

where $\mathbf{a}^{l-1}$ and $\mathbf{a}^l$ are the pre-activation and post-activation vectors of the l -th layer, $\mathbf{W}^l$ is the weight matrix, and $f(\cdot)$ is usually set to the ReLU activation function.

SNN neurons. Unlike ANNs, SNNs introduce a temporal dimension and employ non-differentiable spike activation. For deep SNNs, the inputs are usually repeated over T time steps during forward propagation for the purpose of better performance, which can produce the final mean output as logits [14; 15]. We deployed the integrate-and-fire (IF) neuron model [27] as reported in previous studies of ANN-SNN conversion [10; 11]. The overall discrete kinetics of the IF neuron can be expressed as follows:

$$\mathbf{v}^l(t) = \mathbf{v}^l(t-1) + \mathbf{I}^l(t) - \mathbf{s}^l(t-1)\theta^l, \quad (2)$$

$$\mathbf{I}^l(t) = \mathbf{W}^l \mathbf{s}^{l-1}(t)\theta^{l-1} = \mathbf{W}^l \mathbf{x}^{l-1}(t). \quad (3)$$

Here, $\mathbf{v}^l(t)$ and $\mathbf{I}^l(t)$ denote the membrane potential and the input current of layer l at time step t , respectively. $\mathbf{W}^l$ is the synaptic weight, and θ^l is the firing threshold. $\mathbf{s}^l(t)$ indicates whether the spike is triggered at time step t . For the i th neuron, the neuron fires a spike if the potential after charging $u_i^l(t) = v_i^l(t-1) + I_i^l(t)$ exceeds the firing threshold θ^l . The subscript i denotes the i th element of the vector unless otherwise specified. This firing rule can be described by the following equation:

$$s_i^l(t) = H(u_i^l(t) - \theta^l) = \begin{cases} 1, & u_i^l(t) \geq \theta^l \\ 0, & u_i^l(t) < \theta^l \end{cases} \quad (4)$$

where $H(\cdot)$ is the Heaviside step function. $\mathbf{x}^{l-1}(t) = \mathbf{s}^{l-1}(t)\theta^{l-1}$ denotes the post-synaptic potential of neurons in layer $l-1$ as introduced by Bu et al. [1]. In addition, to minimize information loss during inference, neurons in SNN employ a reset-by-subtraction mechanism in Eq. 2, that is, once a spike is triggered, the membrane potential after the spike needs to be subtracted by the firing threshold θ^l .

3.2 ANN-SNN Conversion

The idea behind ANN-SNN conversion is to create a mapping from spiking neurons' postsynaptic potential in SNN to ReLU activation in ANN. Denote that $\phi^l(T) = \frac{\sum_{t=1}^T \mathbf{x}^l(t)}{T}$ is the average post-synaptic potential during T time steps. The average post-synaptic potential of neurons in the neighboring layers is related as follows:

$$\phi^l(T) = \mathbf{W}^l \phi^{l-1}(T) - \frac{\mathbf{v}^l(T) - \mathbf{v}^l(0)}{T}. \quad (5)$$

The derivation of Eq. 5 can be found in the work of Bu et al. [1]. It is obvious that $\phi^l(T)$ is in the range of $[0, \theta^l]$ and only takes discrete values due to $\mathbf{x}^l(t) = \mathbf{s}^l(t)\theta^l$. Since $\phi^l(T) > 0$, one can build the equivalent mapping from ANN activation $\mathbf{a}^l$ to $\phi^l(T)$ by constraining $\mathbf{a}^l$ into discrete values and setting a specific upper bound for $\mathbf{a}^l$. If $T \rightarrow \infty$,

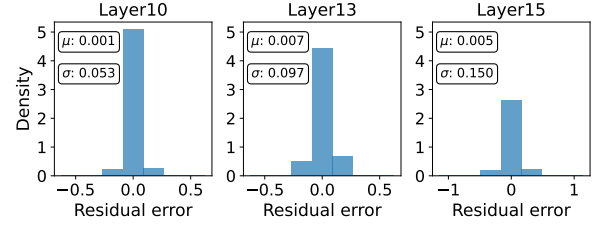


Figure 2: The distribution of the residual error of ANN output and the average postsynaptic potential for SNN. We calculate the residual error of some activation layers during the training of VGG16 on the CIFAR10 dataset.

$(\mathbf{v}^l(0) - \mathbf{v}^l(T))/T \rightarrow 0$, in this case, lossless ANN-SNN conversion can be achieved. Nevertheless, a large value of T significantly increases the cost of the application of efficient SNN hardware, which calls for low-latency conversion. When T is small, $(\mathbf{v}^l(0) - \mathbf{v}^l(T))/T$ is not approaching 0, which indicates the existence of conversion errors.

To address the low-latency conversion, Bu et al. [1] proposed a concept based on the finite residual membrane potential assumption $\mathbf{v}(T) \in [0, \theta^l]$ and derived the QCFS activation function as a substitute for the commonly used ReLU activation function in source ANNs:

$$\mathbf{a}^l = \text{QCFS}(\mathbf{z}^l) = \lambda^l \text{clip}\left(\frac{1}{L} \left\lfloor \frac{\mathbf{z}^l L}{\lambda^l} + 0.5 \right\rfloor, 0, 1\right), \quad (6)$$

where L indicates the activation quantization step, $\mathbf{z}^l = \mathbf{W}^l \mathbf{a}^{l-1}$. λ^l is the trainable threshold for layer l in ANN. When one performs a conversion, λ^l should be copied as θ^l in SNN. Such that given the same average input $\mathbf{z}^l$ as ANN, when $T = L$, $\mathbf{v}^l(0)$ is set to $0.5 \cdot \theta^l$, and $\mathbf{v}^l(T) \in [0, \theta^l]$, the converted average postsynaptic potential can be expressed as:

$$\phi^l(T) = \theta^l \text{clip}\left(\frac{1}{T} \left\lfloor \frac{\mathbf{z}^l T + \mathbf{v}^l(0)}{\theta^l} \right\rfloor, 0, 1\right). \quad (7)$$

It is proven that such a source ANN activation function can more accurately approximate the SNN's activation function and eliminate quantization error in expectation. Other work, such as [24; 3] also shared a similar quantization framework. After applying quantization to ANN, the low-latency performance has significantly improved. However, this work still incurs unresolved conversion errors, which we will brief in the following section.

4 Residual Error in Low-Latency Conversion

In this section, we discuss three potential errors that occurred in the process of converting ANN to SNN, i.e., clipping error, quantization error, and residual error, which contribute to the performance gap between source ANNs and target SNNs. Moreover, we examine the distribution of the residual error.

4.1 Clipping and Quantization error

Clipping and quantization errors bear a certain resemblance as they are both due to the difference between $\mathbf{a}^l$ and $\phi^l(T)$, where $\mathbf{a}^l$ is the ANN quantized activation of layer l and $\phi^l(T)$

is the average postsynaptic potential from layer l in converted SNN. Clipping error arises from the difference in value ranges. In Eq. 6, if $\hat{\lambda}^l$ is the actual maximum value of output a^l and larger than θ^l , then the unbounded activation larger than θ^l cannot be exactly expressed after conversion. This may cause clipping errors.

Quantization error arises from differences in distribution. Output a^l in ANN is continuous values spread all over the range of $[0, a_{max}^l]$, while $\phi^l(T)$ can be distributed only on discrete values like $\frac{\theta^l}{T}, \frac{2\theta^l}{T}, \frac{3\theta^l}{T}, \dots$, thus they cannot be perfectly matched.

Reducing clipping and quantization errors can be achieved by modifying the activation functions of the source ANN to quantized ones [25]. Besides, adding the trainable thresholds is proven to be effective as it can reduce the clipping error to zero by directly mapping the trainable upper bound of the activation to the threshold of the SNN [21].

4.2 Residual error

Previous studies have noticed one error beyond clipping and quantization errors, which usually attribute to neuronal differences. These errors are used to be considered as transient dynamics [12], temporal jitter of spike trains [28], or unevenness error [1]. We deem that these studies do not accurately reflect the cause of the error. Here, we refer to this type of error as residual error. Firstly, the primary reason leading to this error is that IF neurons with reset-by-subtraction mechanisms fail to respond to residual membrane potentials outside the range of $[0, \theta]$. Besides, when the quantization parameter L mismatches the inference time step T , the average postsynaptic membrane potential also mismatches the activation.

Residual error seriously affects the performance of SNNs under low-latency conditions. Here, we would like to give the form of residual error denoted as $g^l(T)$ for layer l . IF neurons in layer l receive weighted input $\hat{z}^l = \mathbf{W}^l \phi^{l-1}(T)$, and their initial membrane potential is denoted as $v^l(0)$. Then we can reformulate Eq. 5 into:

$$\phi^l(T) = \theta^l \text{clip} \left(\frac{1}{T} \left\lfloor \frac{\hat{z}^l T + v^l(0)}{\theta^l} \right\rfloor, 0, 1 \right) + g^l(T). \quad (8)$$

Eq. 8 now is not related to the source ANN and $g^l(T)$ is not properly modeled. For QCFS conversion, when $T = L$, $v^l(0) = 0.5 \cdot \theta^l$, and $v^l(T) \in [0, \theta^l]$, a^l in Eq. 6 matches $\phi^l(T)$ in Eq. 7. And we have $\phi^l(T) = \phi^l(T) + g^l(T)$. However, these ideal conditions are too harsh; in fact, we can only obtain:

$$\phi^l(T) = \phi^l(T) + g^l(T), \quad (9)$$

$$= a^l + g^l(T), \quad (10)$$

where $g^l(T)$ absorbs the original residual error $g^l(T)$ and other errors caused by $v^l(T) \notin [0, \theta^l]$ or $T \neq L$. According to Eq. 10, we can observe the distribution of $g^l(T)$ based on the source ANN activation and converted SNN postsynaptic potential. Specifically, we train an ANN with QCFS activation on CIFAR-100 for VGG-16, and convert it to an SNN by fixing $T = L$. Please refer to Fig. 2. We find that the standard deviation of the distribution of $g^l(T)$ is large while the

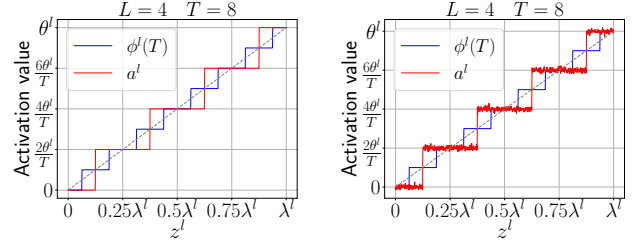


Figure 3: Comparison of SNN output $\phi^l(T)$ and ANN output a^l with the same input z^l . The figure shows two activation functions for source ANNs: quantization clip-floor-shift (QCFS) activation (left) and our proposed Noisy Quantized activation with residual error noise modeling (right).

mean is comparably small (close to 0). The distribution is almost symmetric around 0. This suggests that the distribution of $g^l(T)$ will be highly dispersed around 0.

5 Methods

In this section, we come up with an explicit modeling method and improve activation functions by adding Gaussian noise to lessen the error between ANN and converted SNN. We manage to improve the low-latency conversion performance by adding a fixed noise for all layers. Finally, we propose a layer-wise error-compensating strategy to set the noise intensity more accurately for each activation layer.

5.1 Explicit Modeling of Residual Error

Since the distribution of $g^l(T)$ is almost symmetric around 0 and has a significant standard deviation, we consider using $\delta^l \cdot \mathbf{G}$ as an approximate alternative to $g^l(T)$, where G_i is the i th item in $\mathbf{G}$, $G_i \sim N(0, 1)$, and $\delta^l \cdot G_i$ is sampled from an i.i.d. Gaussian distribution for each neuron i in layer l .

$$g_i^l(T) \approx \delta^l \cdot G_i, \quad (11)$$

where $g_i^l(T)$ is the i th item in $g^l(T)$. In this case, $\delta^l \cdot G_i \sim N(0, \delta^{l^2})$ for each neuron i . With such an approximate, by combining Eq. 9 and Eq. 11, we update a more accurate estimation expression for ϕ^l in the SNN:

$$\phi^l(T) \approx \phi^l(T) + \delta^l \cdot \mathbf{G}. \quad (12)$$

We consider the additional Gaussian noise as compensation for the residual error. According to Eq. 9, we propose to introduce an quantized activation function with a parameterized noise model of the residual error to train the ANN:

$$a^l = \text{NQ}(z^l) = \lambda^l \text{clip} \left(\frac{1}{L} \left\lfloor \frac{z^l L}{\lambda^l} + 0.5 \right\rfloor, 0, 1 \right) + \delta^l \cdot \mathbf{G} \quad (13)$$

We name this activation Noisy Quantized activation (NQ). In Eq. 13, considering that the effect of quantization error cannot be offset in the case of $T \neq L$, we adopt a treatment similar to that of [1], i.e., introducing a shift term of 0.5 in the activation function. Note that Eq. 13 degenerates to a QCFS activation function when $\delta^l = 0$.

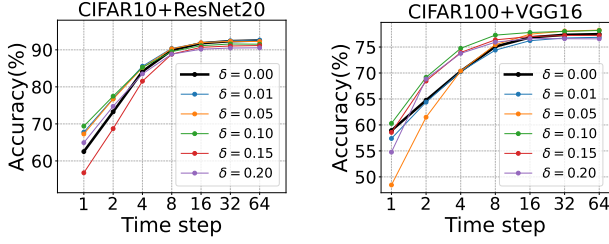


Figure 4: The effect of noise intensity. Adding a certain amount of noise to the activation function benefits inference performance at short time steps.

Based on the NQ activation and the average postsynaptic potential of the converted SNN activation in Eq. 5, we can derive the conversion error ϵ^l between the ANN and the SNN:

$$\epsilon^l = \mathbf{W}^l \phi^{l-1}(T) - \frac{v^l(T) - v^l(0)}{T} - \text{NQ}(z^l). \quad (14)$$

With the definition of conversion error above, we show Theorem 1, which proves that under some conditions, the expectation of the conversion error is zero.

Theorem 1. *Given an ANN using our proposed NQ activation in Eq. 13, the trained ANN is converted to an SNN with IF neurons with the same weights and satisfying $\theta^l = \lambda^l$ for each layer. Assume that $v^l(0) = \frac{\theta^l}{2}$, $v^l(T) \in [0, \theta^l]$ where $v^l(T)$ is the membrane potential at time T . Then we have:*

$$\forall T, L, \delta^l \quad \mathbb{E}_z(\tilde{\epsilon}) = \mathbf{0}. \quad (15)$$

The proof of Theorem 1 given in the Appendix. Theorem 1 shows that for any δ^l , even if $L \neq T$, the additional introduction of $\delta^l \mathbf{G}$ in the NQ activation function modeled is complementary to the quantization and clipping error. $\delta^l \mathbf{G}$ does not affect the expected value of the conversion error. These nice properties ensure the feasibility of our high-performance converted SNN at ultra-low inference time steps.

5.2 Minimizing the Gap Between ANN and SNN

Determining the appropriate Gaussian noise intensity δ^l in real training is a critical task. Initially, we consider manually setting a fixed noise intensity for all activation layers. To evaluate the effect of adding noise to the activation on the performance of ANN-SNN conversion, we conduct experiments for VGG16 on the CIFAR10 dataset and ResNet20 on the CIFAR100 dataset. The experimental results are shown in Fig. 4. Specifically, we add a fixed and shared zero-mean Gaussian distribution to all activation layers during the training of the ANN, where $\delta^l = \delta$ for each layer. We try different settings of the noise intensity δ . The experimental results in Fig. 4 show that different settings of the noise intensity will directly affect the conversion performance. In the case of adding less noise, the performance of the converted SNN is improved for all time steps, but there is still room for further improvement. However, when the noise intensity is too large, we observe that the converted SNN fails to converge to a higher accuracy. We believe that this is due to the fact that the noise introduces randomness that affects the accuracy of

the ANN and further affects the accuracy of the SNN. To obtain a fixed noise setting that is most beneficial for ANN-SNN conversion usually requires a lot of experimentation and tuning for different datasets and network structures. In addition, according to our observation on residual error across different layers in Fig. 2, it can be seen that the error distributions across layers are not consistent. Therefore, personalizing the noise setting for each activation layer would be a more reasonable way.

We propose a hierarchical error compensation strategy to optimize the ANN-SNN conversion during ANN training. Since it is difficult to predict and estimate the residual error before conversion, we infer the converted SNN on a small validation set and decide the noise intensity for each layer. Specifically, before training, we split a small portion of the training dataset as the validation dataset. For the first epoch, we choose to train using NQ by setting $\delta^l = 0$ for all l . After training one epoch of the training set, we successively process the validation set with the training ANN with NQ activation and the corresponding SNN. During the validation process, we first calculate the output of each activation layer of the ANN and then the average postsynaptic potential of the corresponding SNN neuron layer at $T = \tau$. After obtaining vectors from ANN and SNN, we calculate the standard deviation of the difference between the two vectors. Subsequently, we set the noise intensity δ^l of each activation layer to the standard deviation obtained, respectively. That is, in each epoch, we adjust the noise intensity δ^l according to the statistics obtained from the previous epoch on the validation set. Such a procedure ensures that the additive noise is independent within the different activation layers, allowing a more precise compensation of the effects caused by residual error.

In addition, we find that the ANN accuracy and SNN accuracy on the testing set are not always optimal at the same epoch during the training process due to the introduction of the additive noise. We record the accuracy of VGG16 architecture of ANN and ANN on the CIFAR10 dataset and the CIFAR100 dataset during training and display the accuracy curve in Fig. 5. As can be seen from the results in Fig. 5, if we save the model with the best ANN performance during training, we will miss the best epoch when SNN have the best performance. Therefore, during training, we choose to evaluate the converted SNN model with T time steps rather than the model that is optimal for the ANN.

6 Experiments

In this section, we use the image classification task to evaluate the effectiveness and performance of our proposed method on the CIFAR-10 and CIFAR-100 datasets. We use the widely adopted VGG-16, ResNet-18, and ResNet-20 network structures as source ANNs for conversion.

6.1 Experimental Setting

For ANN training, we use the SGD training optimizer and a cosine decay scheduler to tune the learning rate. The initial learning rate is set to 0.1 for CIFAR-10 and 0.05 for CIFAR-100. Besides, we set the weight decay to 5×10^{-4} for both

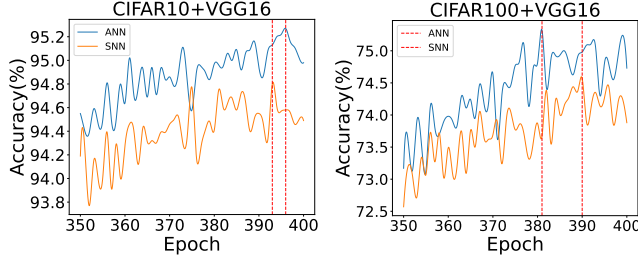


Figure 5: Accuracy curves of ANN and SNN on the testing dataset during the training process

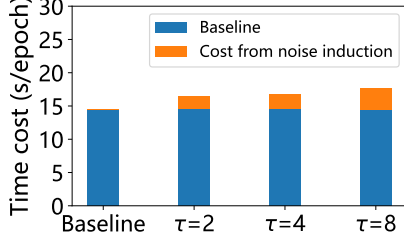


Figure 6: Comparison of training overhead of ANN

datasets. We train all models for 400 epochs. Our setting of the quantization step L is consistent with Bu et al. [1]. When training on the CIFAR-10 dataset, L is set to 4. When training VGG-16 and ResNet-18 on the CIFAR-100 dataset, L is set to 4. When training ResNet-20 on the CIFAR-100 dataset, L is set to 8. When applying the layer-wise strategy, we suggest that the noise-induction time step τ be consistent with L . Please refer to the Appendix for a detailed experimental setting.

6.2 Comparison With the State-of-the-Art Conversion Methods

We compare our method with the state-of-the-art ANN-SNN conversion methods, including RMP [17], RTS [18], RNL [20], OPI [26], SNNC-AP [25], TCL [21], QCFS [1] and SNM [23]. Table 1 shows the performance of our proposed method. Our model almost outperforms all the other conversion methods when $2 \leq T \leq 8$. On the CIFAR-10 dataset, for VGG-16, we achieve an accuracy of 94.80%, very close to the accuracy of its ANN counterpart (95.21%) with only 4 time steps. For ResNet-18, the accuracy of the converted SNN is 18.28% higher than the current SOTA QCFS SNN at $T = 2$ and 4.94% higher than the QCFS SNN at $T = 4$. Besides, on CIFAR-100, the proposed method achieves 60.93% top-1 accuracy at 8 time steps for ResNet-20, which is 37.84% and 5.56% higher than OPI and QCFS, respectively. For VGG-16, we achieve 74.57% top-1 accuracy when $T = 4$, which is 4.95% higher than QCFS (69.62%) at the same time step and 14.08% higher than OPI at $T = 8$ (60.49%).

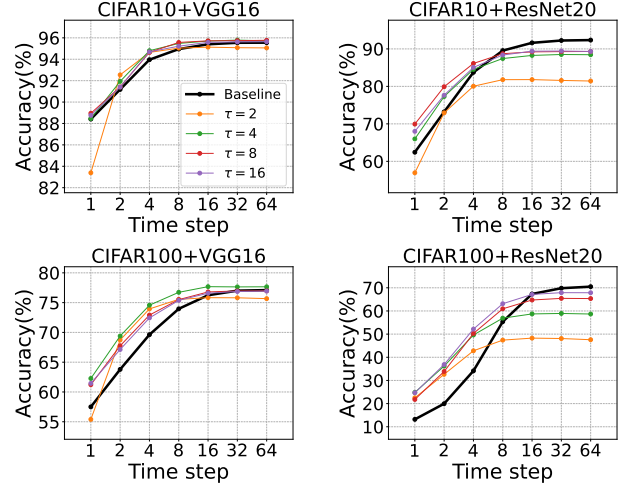


Figure 7: Effect of noise-induction time step τ

6.3 Comparison With Other SNN Training Methods

Table 2 presents the results between our proposed method and recent advanced SNN training methods on the CIFAR10/100 dataset, involving STBP-tdBN [14], TET [45], TEBN [46], Diet-SNN [30], Dual-Phase [29], RecDis-SNN [31] under the low latency condition. The accuracy achieved by our method when $T = 4$ exceeds the accuracy of other SNN training methods using the same or even longer time step. It is worth emphasizing that our proposed method does not need to consume as much memory and computational resources as other SNN training methods to achieve high performance with very low inference latency. In addition, our method does not require any other additional operations or optimizations on the converted SNNs. This advantage enhances the usefulness of ANN-SNN conversion algorithms in neuromorphic chips.

6.4 Effect of Noise-Induction Time Step

We further explored the effect of the noise-induction time step τ . Fig. 7 shows the accuracy of VGG16 and ResNet20 on CIFAR-10 and CIFAR-100 for different τ values. The settings of L are the same as those in the experimental setup. When $\tau = L$, the noise intensity δ mainly represents the effect of residual error during the validation process. When $\tau \neq L$, δ represent the effect of both quantization error from $\tau \neq L$ and residual error from $\tau = L$. We find that when $\tau = L$, for VGG16, SNN achieves better performance than the baseline at almost all time steps. For ResNet20, when $\tau = L$, the accuracy of SNN outperforms baseline at shorter time steps ($T \leq \tau$). However, when $T > \tau$, the accuracy of SNN falls short of baseline. In short, the setting of $\tau = L$ can estimate the residual error more accurately, thus improving the performance of SNN with low latency ($T \leq L$). The training process may primarily eliminate the performance gap of low-latency conversion, potentially introducing more noise and affecting ANN and thus SNN performance when $T \leq \tau$. Notably, this degradation problem can be solved by setting a larger τ . The results in Fig. 7 (purple lines) show that in the

Table 1: Comparison with existing state-of-the-art ANN-SNN conversion methods

Method	ANN	Architecture	T=1	T=2	T=4	T=8	T=16	T=32	T=64
CIFAR-10 Dataset									
SNM [23]	94.09%	VGG16	—	—	—	—	—	93.43%	94.07%
RMP [17]	93.96%		—	—	—	—	—	60.30%	90.35%
RTS [18]	95.72%		—	—	—	—	—	76.24%	90.64%
TCL [21]	94.57%		—	—	—	—	—	93.64%	94.26%
RNL [20]	92.82%		—	—	—	—	57.90%	85.40%	91.15%
SNNC-AP [25]	95.72%		—	—	—	—	—	93.71%	95.14%
OPI [26]	94.57%		—	—	—	90.96%	93.38%	94.20%	94.45%
QCFS [1]	95.52%		88.41%	91.18%	93.96%	94.95%	95.40%	95.54%	95.55%
Ours	95.21%		88.46%	91.93%	94.80%	95.48%	95.70%	95.79%	95.72%
RTS [18]	92.32%	ResNet18	—	—	—	—	92.41%	93.30%	93.55%
SNM [18]	95.39%		—	—	—	—	—	94.03%	94.03%
SNNC-AP [25]	95.46%		—	—	—	—	—	94.78%	95.30%
OPI [26]	96.04%		—	—	—	75.44%	90.43%	94.82%	95.92%
QCFS [1]	96.04%		—	75.44%	90.43%	94.82%	95.92%	96.08%	96.06%
Ours	95.52%		89.64%	93.72%	95.37%	96.21%	96.38%	96.31%	96.36%
OPI [26]	92.74%	ResNet20	—	—	—	66.24%	87.22%	91.88%	92.57%
QCFS [1]	91.77%		62.43%	73.20%	83.75%	89.55%	91.62%	92.24%	92.35%
Ours	85.18%		65.99%	77.30%	84.54%	87.43%	88.26%	88.51%	88.45%
CIFAR-100 Dataset									
SNM [18]	74.13%	VGG16	—	—	—	—	—	71.80%	73.69%
RTS [18]	77.89%		—	—	—	—	65.94%	69.80%	70.35%
TCL [21]	76.32%		—	—	—	—	—	52.30%	71.17%
SNNC-AP [25]	77.89%		—	—	—	—	—	73.55%	76.64%
OPI [26]	76.31%		—	—	—	60.49%	70.72%	74.82%	75.97%
QCFS [1]	76.28%		—	63.79%	69.62%	73.96%	76.24%	77.01%	77.10%
Ours	74.86%		62.27%	69.39%	74.57%	76.73%	77.68%	77.64%	77.66%
RTS [18]	67.08%	ResNet18	—	—	—	—	63.73%	68.40%	69.27%
SNNC-AP [25]	77.16%		—	—	—	—	—	76.32%	77.29%
QCFS [1]	78.80%		—	70.79%	75.67%	78.48%	79.48%	79.62%	79.54%
Ours	76.66%		62.35%	70.86%	76.81%	78.85%	79.44%	79.62%	79.46%
RMP [17]	68.72%	ResNet20	—	—	—	—	—	27.64%	46.91%
OPI [26]	70.43%		—	—	—	23.09%	52.34%	67.18%	69.96%
QCFS [1]	69.94%		—	19.96%	34.14%	55.37%	67.33%	69.82%	70.49%
Ours	62.34%		21.78%	33.87%	50.28%	60.93%	64.73%	65.43%	65.35%

Table 2: Comparison with other SNN training methods

Method	Type	Arch	Acc	T
CIFAR-10 Dataset				
STBP-tdBN	BPTT	ResNet19	93.16	6
TET	BPTT	ResNet19	94.50	6
TEBN	BPTT	ResNet19	94.71	6
Ours	ANN2SNN	ResNet18	95.37	4
CIFAR-100 Dataset				
Diet-SNN	Hybrid	VGG16	69.67	5
Dual-Phase	Hybrid	VGG16	70.08	4
RecDis-SNN	BPTT	VGG16	69.88	5
Ours	ANN2SNN	VGG16	74.57	4

case of $\tau \geq L$, our method maintains a good generalization ability.

6.5 Training Overhead

To assess the possible computational overhead associated with the inclusion of the validation process, we calculate the average training time per epoch of the CIFAR-100 dataset during the training of VGG16 and display the results in Fig. 6. The noise induction operation introduced to automatically ad-

just the noise intensity does not lead to a significant additional computational overhead, which shows that our training method is efficient and has potential practical applications.

7 Conclusions

This paper improves the low-latency ANN-SNN conversion by explicitly modeling an additive noise representing the effect of residual error in source ANNs, allowing the output of quantized activation to be closer to the average postsynaptic potential of spiking neurons. This approach maintains the expectation of conversion error, resulting in high accuracy and ultra-low latency. The method also proposes to adjust the noise intensity with low overhead, further reducing the performance gap between ANNs and SNNs under low-latency conditions. We evaluate the proposed method on the CIFAR10/100 dataset and showed that our method improves performance under low-latency conditions. We believe that this work will facilitate the further development of neuromorphic hardware applications.

References

- [1] T. Bu, W. Fang, J. Ding, P. Dai, Z. Yu, and T. Huang, "Optimal ANN-SNN conversion for high-accuracy and ultra-low-latency spiking neural networks," in *International Conference on Learning Representations*, 2022.
- [2] Z. Hao, J. Ding, T. Bu, T. Huang, and Z. Yu, "Bridging the gap between ANNs and SNNs by calibrating offset spikes," in *International Conference on Learning Representations*, 2023.
- [3] H. Jiang, S. Anumasa, G. De Masi, H. Xiong, and B. Gu, "A unified optimization framework of ANN-SNN conversion: towards optimal mapping from activation values to firing rates," in *Proceedings of the 38th International Conference on Machine Learning*, 2023, pp. 14 945–14 974.
- [4] P. O'Connor, E. Gavves, M. Reisser, and M. Welling, "Temporally efficient deep learning with spikes," in *International Conference on Learning Representations*, 2018.
- [5] Y. Wu, L. Deng, G. Li, J. Zhu, and L. Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers in Neuroscience*, vol. 12, p. 331, 2018.
- [6] F. Zenke and T. P. Vogels, "The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks," *Neural computation*, vol. 33, no. 4, pp. 899–925, 2021.
- [7] A. Tavanaei and A. Maida, "BP-STDP: Approximating backpropagation using spike timing dependent plasticity," *Neurocomputing*, vol. 330, pp. 39–47, 2019.
- [8] C. Lee, P. Panda, G. Srinivasan, and K. Roy, "Training deep spiking convolutional neural networks with STDP-based unsupervised pre-training followed by supervised fine-tuning," *Frontiers in Neuroscience*, vol. 12, p. 435, 2018.
- [9] N. Rathi, G. Srinivasan, P. Panda, and K. Roy, "Enabling deep spiking neural networks with hybrid conversion and spike timing dependent backpropagation," in *International Conference on Learning Representations*, 2020.
- [10] Y. Cao, Y. Chen, and D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition," *International Journal of Computer Vision*, vol. 113, pp. 54–66, 2015.
- [11] P. U. Diehl, D. Neil, J. Binas, M. Cook, S.-C. Liu, and M. Pfeiffer, "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *International Joint Conference on Neural Networks*. IEEE, 2015, pp. 1–8.
- [12] B. Rueckauer, I.-A. Lungu, Y. Hu, M. Pfeiffer, and S.-C. Liu, "Conversion of continuous-valued deep networks to efficient event-driven networks for image classification," *Frontiers in Neuroscience*, vol. 11, p. 682, 2017.
- [13] R. Massa, A. Marchisio, M. Martina, and M. Shafique, "An efficient spiking neural network for recognizing gestures with a dvs camera on the loihi neuromorphic processor," in *International Joint Conference on Neural Networks*. IEEE, 2020, pp. 1–9.
- [14] H. Zheng, Y. Wu, L. Deng, Y. Hu, and G. Li, "Going deeper with directly-trained larger spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 11 062–11 070.
- [15] Q. Xu, Y. Li, J. Shen, J. K. Liu, H. Tang, and G. Pan, "Constructing deep spiking neural networks from artificial neural networks with knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7886–7895.
- [16] M. Xiao, Q. Meng, Z. Zhang, D. He, and Z. Lin, "On-line training through time for spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 717–20 730, 2022.
- [17] B. Han, G. Srinivasan, and K. Roy, "RMP-SNN: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2020, pp. 13 558–13 567.
- [18] S. Deng and S. Gu, "Optimal conversion of conventional artificial neural networks to spiking neural networks," in *International Conference on Learning Representations*, 2021.
- [19] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: VGG and residual architectures," *Frontiers in neuroscience*, vol. 13, p. 95, 2019.
- [20] J. Ding, Z. Yu, Y. Tian, and T. Huang, "Optimal ANN-SNN conversion for fast and accurate inference in deep spiking neural networks," in *International Joint Conference on Artificial Intelligence*, 2021.
- [21] N.-D. Ho and I.-J. Chang, "TCL: an ANN-to-SNN conversion with trainable clipping layers," in *2021 58th ACM/IEEE Design Automation Conference*. IEEE, 2021, pp. 793–798.
- [22] Y. Li and Y. Zeng, "Efficient and accurate conversion of spiking neural network with burst spikes," in *International Joint Conference on Artificial Intelligence*, 2022.
- [23] Y. Wang, M. Zhang, Y. Chen, and H. Qu, "Signed neuron with memory: Towards simple, accurate and high-efficient ANN-SNN conversion," in *International Joint Conference on Artificial Intelligence*, 2022.
- [24] C. Li, L. Ma, and S. Furber, "Quantization framework for fast spiking neural networks," *Frontiers in Neuroscience*, vol. 16, p. 918793, 2022.
- [25] Y. Li, S. Deng, X. Dong, R. Gong, and S. Gu, "A free lunch from ANN: Towards efficient, accurate spiking neural networks calibration," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 6316–6325.

- [26] T. Bu, J. Ding, Z. Yu, and T. Huang, "Optimized potential initialization for low-latency spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 11–20.
- [27] W. Gerstner and W. M. Kistler, *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.
- [28] J. Wu, Y. Chua, M. Zhang, G. Li, H. Li, and K. C. Tan, "A tandem learning rule for effective training and rapid inference of deep spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [29] Z. Wang, Y. Zhang, S. Lian, X. Cui, R. Yan, and H. Tang, "Toward high-accuracy and low-latency spiking neural networks with two-stage optimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [30] N. Rathi and K. Roy, "DIET-SNN: A low-latency spiking neural network with direct input encoding and leakage and threshold optimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [31] Y. Guo, X. Tong, Y. Chen, L. Zhang, X. Liu, Z. Ma, and X. Huang, "RecDis-SNN: Rectifying membrane potential distribution for directly training spiking neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2022, pp. 326–335.
- [32] E. Z. Farsa, A. Ahmadi, M. A. Maleki, M. Gholami, and H. N. Rad, "A low-cost high-speed neuromorphic hardware based on spiking neural network," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 9, pp. 1582–1586, 2019.
- [33] Y. Liu, Y. Chen, W. Ye, and Y. Gui, "FPGA-NHAP: A general FPGA-based neuromorphic hardware acceleration platform with high speed and low power," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2553–2566, 2022.
- [34] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He *et al.*, "Towards artificial general intelligence with hybrid tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, 2019.
- [35] Y. Wu, L. Deng, G. Li, J. Zhu, Y. Xie, and L. Shi, "Direct training for spiking neural networks: Faster, larger, better," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1311–1318.
- [36] S. R. Kheradpisheh, M. Ganjtabesh, S. J. Thorpe, and T. Masquelier, "STDP-based spiking deep convolutional neural networks for object recognition," *Neural Networks*, vol. 99, pp. 56–67, 2018.
- [37] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
- [38] S. Kim, S. Park, B. Na, and S. Yoon, "Spiking-YOLO: spiking neural network for energy-efficient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- [39] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers in Computational Neuroscience*, vol. 9, p. 99, 2015.
- [40] W. Maass, "Networks of spiking neurons: the third generation of neural network models," *Neural networks*, vol. 10, no. 9, pp. 1659–1671, 1997.
- [41] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade: Second Edition*. Springer, 2012, pp. 421–436.
- [42] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2017.
- [43] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with Cutout," *arXiv preprint arXiv:1708.04552*, 2017.
- [44] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 113–123.
- [45] S. Deng, Y. Li, S. Zhang, and S. Gu, "Temporal efficient training of spiking neural network via gradient re-weighting," in *International Conference on Learning Representations*, 2022.
- [46] C. Duan, J. Ding, S. Chen, Z. Yu, and T. Huang, "Temporal effective batch normalization in spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 377–34 390, 2022.
- [47] Z. Hao, T. Bu, J. Ding, T. Huang, and Z. Yu, "Reducing ANN-SNN conversion error through residual membrane potential," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, pp. 11–21.
- [48] W. Zhang and P. Li, "Temporal spike sequence learning via backpropagation for deep spiking neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 022–12 033, 2020.
- [49] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.
- [50] N.-D. Ho and I.-J. Chang, "Mice: an ann-to-snn conversion technique to enable high accuracy and low latency," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2023.
- [51] B. Wang, J. Cao, J. Chen, S. Feng, and Y. Wang, "A new ann-snn conversion method with high accuracy, low latency and good robustness," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, 2023, pp. 3067–3075.