Spiking-Diffusion: Vector Quantized Discrete Diffusion Model with Spiking Neural Networks

1st Mingxuan Liu
Tsinghua University
Beijing, China
arktis@qq.com

2nd Jie Gan
Beijing Smartchip Microelectronics Technology Co., Ltd.
Beijing, China
ganjie@sgchip.sgcc.com.cn

3rd Rui Wen
Tsinghua University
Beijing, China
1766649947@qq.com

4th Tao Li
Tsinghua University
Beijing, China
qh_litao717@mail.tsinghua.edu.cn

5th Yongli Chen
Beijing Smartchip Microelectronics Technology Co., Ltd.
Beijing, China
chenyongli@sgchip.sgcc.com.cn

6th Hong Chen*
Tsinghua University
Beijing, China
hongchen@tsinghua.edu.cn

***Abstract*—Spiking neural networks (SNNs) have tremendous potential for energy-efficient neuromorphic chips due to their binary and event-driven architecture. SNNs have been primarily used in classification tasks, but limited exploration on image generation tasks. To fill the gap, we propose a Spiking-Diffusion model, which is based on the vector quantized discrete diffusion model. First, we develop a vector quantized variational autoencoder with SNNs (VQ-SVAE) to learn a discrete latent space for images. In VQ-SVAE, image features are encoded using both the spike firing rate and postsynaptic potential, and an adaptive spike generator is designed to restore embedding features in the form of spike trains. Next, we perform absorbing state diffusion in the discrete latent space and construct a spiking diffusion image decoder (SDID) with SNNs to denoise the image. Our work is the first to build the diffusion model entirely from SNN layers. Experimental results on MNIST, FMNIST, KMNIST, Letters, and Cifar10 demonstrate that Spiking-Diffusion outperforms the existing SNN-based generation model. We achieve FIDs of 37.50, 91.98, 59.23, 67.41, and 120.5 on the above datasets respectively, with reductions of 58.60%, 18.75%, 64.51%, 29.75%, and 44.88% in FIDs compared with the state-of-art work. Our code will be available at https://github.com/Arktis2022/Spiking-Diffusion.**

***Index Terms*—Spiking neural networks, Diffusion model, Image generation, Bionic Learning**

I. Introduction

Spiking neural networks (SNNs) are the third generation neural networks with biological plausibility that encode and transmit information in form of spikes by mimicking the dynamics of neurons in the brain. Compared with Artificial neural networks (ANNs), the event-driven nature of SNNs allows for significant reduction in energy consumption when running on neuromorphic chips [1].

SNNs with deep learning techniques have shown promising results on simple tasks such as image classification, image segmentation, and optical flow estimation[2]. However, the scope of their utility in complex tasks, particularly image generation, is still confined. Spiking-GAN [3] is the first fully SNN-based GAN that uses time-to-first-spike (TTFS) encoding to generate MNIST images through adversarial training. Despite its significant progress, the quality of the generated images still falls short of ANN-based GAN. Another model, fully spiking variational autoencoder (FSVAE) [4], is a VAE model constructed entirely from SNN layers. This model produces comparable results to those of ANN-based VAEs, but its performance is still constrained by the inherent limitations of VAEs, namely the weak generative capacity due to the overly simplified decoder and compressed latent space [5]. Thus, more exploration and development are needed to harness the full potential of SNNs in image generation tasks.

In this paper, we propose a vector quantized discrete diffusion model with spiking neural networks (called Spiking-Diffusion), which is the first diffusion model using only SNN layers. The pipeline of proposed Spiking-Diffusion contains two stages. First, an image is transformed into a discrete matrix through proposed vector quantized spiking variational autoencoder (VQ-SVAE). However, creating VQ-VAEs [6] in SNNs brings two challenges. The one is converting spike sequences into dense features suitable for storage in the codebook, because storing spike sequences directly within the codebook will consume too much memory. To overcome this challenge, we combine spike frequency rate (SFR) and postsynaptic potential (PSP) to model the spike sequences and introduce the learnable operator k to optimize the weights of PSP and SFR. The other challenge is losslessly converting embedded features to spiking input for SNN decoder, because Poisson coding incurs information loss. To address this challenge, we design an adaptive spike generator (ASG) before the decoder and train it by using the same dictionary learning algorithm as traditional VQ-VAE. In the second stage, we utilize a spiking diffusion image decoder (SDID) to fit the prior discrete latent codes. In the forward process, we select a Markov transition matrix with an absorbing state [7] to gradually add masks to the discrete matrix, and adopt an SNN denoising network SDID to recover the masks and achieve inverse process parameterization. Experiments on MNIST [8], FMNIST [9], KMNIST [10], Letters [11], and Cifar10 [12] show that Spiking-Diffusion outperforms the SOTA SNN-based generative model.

II. Methodology

A. SNNs Learning Algorithms

In this study, we adopt he SNNs learning algorithms in [13]. Spike neurons in SNNs emulates the behavior of biological neurons by generating discrete pulses, or spikes, in response to input stimuli. The LIF neuron model [14] is used, which is described by the following dynamic equation:

$$H[t] = V[t-1] + \frac{1}{\tau}(X[t] - (V[t-1] - V_{reset})) \tag{1}$$

$$S[t] = \Theta(H[t] - V_{th}) \tag{2}$$

$$V[t] = H[t](1 - S[t]) + V_{reset}S[t] \tag{3}$$

where τ represents the membrane time constant, $X[t]$ denotes the synaptic input current at time step t, $H[t]$ represents the membrane potential of a neuron after charging but before firing a spike, and $S[t]$ denotes the spike generated by a LIF neuron when its membrane potential exceeds the discharge threshold V_{th}. $\Theta(v)$ is the Heaviside step function, which is equal to 1 when $v \geq 0$ and 0 otherwise. $V[t]$ stands for the membrane potential after a spike event, which is reset to the V_{reset} using a hard reset [15]. If no spike is generated, it is equal to $H[t]$.

B. VQ-SVAE

Although diffusion models [16] can generate high-quality images, they need to predict the noise added in the forward step. However, SNNs are difficult to fit and process continuous analog signals due to their discrete spike encoding [17]. In order to solve the problem, we propose a VQ-SVAE to learn a discrete latent representation of images.

As shown in Fig. 1, Given a 2D image $x \in \mathbb{R}^{1\times H\times W}$, we first convert it into sequences $x_{1:T_s} \in \mathbb{R}^{T_s\times 1\times H\times W}$ using direct input encoding [18], and then the $x_{1:T_s}$ is encoded as a spike sequence $z^E_{1:T_s} \in \{0,1\}^{T_s\times C\times H'\times W'}$ after the SNN encoder. In order to prevent the codebook from consuming too much memory, we utilize spike firing rate (SFR) and postsynaptic potential (PSP) to model the spike sequence. SFR has been proven in neurobiology to have the ability to represent information. For example, auditory nerves use SFR to encode steady-state vowels [19].

$$\mathrm{SFR}(z^E_{1:T_s}) = \frac{1}{T_s}\sum_{t=1}^{T_s} z^E_t \tag{4}$$

PSP simulates the response of postsynaptic neurons to the action potential (AP) sequence. Similarly, studies have shown that PSP is related to experience-dependent plasticity in the mammalian nervous system [20]. We use the following formula to update the PSP [21]:

$$\mathrm{PSP}(z^E_{\leq t_s}) = (1 - \frac{1}{\tau_{syn}}) \times \mathrm{PSP}(z^E_{\leq t_{s-1}}) + \frac{1}{\tau_{syn}} \times z^E_t \tag{5}$$

where τ_{syn} is the synaptic time constant and $\mathrm{PSP}(z^E_{\leq 0})$ is set to 0. We use a trainable operator k to allocate the weights of the PSP and SFR, so the features used for quantized encoding $z^E \in \mathbb{R}^{C\times H'\times W'}$ can be obtained from the following formula:

$$z^E = k \times \mathrm{SFR}(z^E_{1:T_s}) + (1-k) \times \mathrm{PSP}(z^E_{\leq T_s}) \tag{6}$$

After obtaining z^E, we construct a codebook $\mathbb{Z} \in \mathbb{R}^{K\times C}$, with $z_k \in \mathbb{Z}$ and $z^E_{i,j} \in z^E$. The quantized encoding of z^E, denoted as a discrete matrix h_0, can then be derived through nearest neighbor search. According to h_0, each spatial feature $z^E_{i,j}$ is mapped to its nearest codebook entry z_k by indexing in $\mathbb{Z}$, thereby obtaining the embedding feature z^Q:

$$h_0 = Q(z^E, \mathbb{Z}) = \mathrm{argmin}_k ||z^E_{i,j} - z_k||_2 \tag{7}$$

$$z_Q = \mathrm{Index}(\mathbb{Z}, h_0) \tag{8}$$

However, the embedding feature z^Q incorporates the SFR and PSP features of spike sequences. Consequently, it cannot be Poisson encoded [21] as input to the SNN decoder without loss of information. To solve the problem, we use a SNN layer to construct a trainable adaptive spike generator (ASG) that learns to restore z^Q to the original spike sequence $z^Q_{1:T_s}$. At this stage, all parameters in the network can be trained end-to-end with the following loss function:

$$\begin{aligned} \mathcal{L}_{\mathrm{VQ-SVAE}} &= ||x - \hat{x}||_2 + ||z^E - z^Q||_2^2 \\ &+ (\sum_{t=1}^{T} ||\mathrm{PSP}(sg[z^E_{\leq t_s}]) - \mathrm{PSP}(z^Q_{\leq t_s})||^2 \\ &+ \beta\sum_{t=1}^{T} ||\mathrm{PSP}(sg[z^Q_{\leq t_s}]) - \mathrm{PSP}(z^E_{\leq t_s})||^2) \end{aligned} \tag{9}$$

In the formula, $sg[\cdot]$ denotes the stop gradient operation. Notably, compared with VQ-VAE [6], VQ-SVAE imposes an additional constraint on training the ASG, which corresponds to the third term above. Moreover, the straight-through estimator propagates the gradients from the spike sequences $z^Q_{1:T_s}$ to $z^E_{1:T_s}$ during backpropagation, rather than propagating the gradients of the embedded features z^Q. Since spike sequences exhibit sparsity, directly computing the mean squared error (MSE) loss between sequences is not the optimal distance metric, therefore, we set the third term in Eq. (9) to maximum mean discrepancy (MMD), whose kernel function is PSP.

C. Absorbing State Diffusion

The training dataset is encoded into discrete matrices h_0 by the VQ-SVAE to train the spiking diffusion image decoder (SDID). For the forward diffusion process $q(h_{t_d}|h_{t_d-1})$, we use a transition matrix with an absorbing state [7] to convert all the elements in the discrete matrix into a mask after a fixed number of T_d time steps. In detail, for each discrete random variable element $h_{t_d}(i,j)$ in h_{t_d} with K categories, denoting

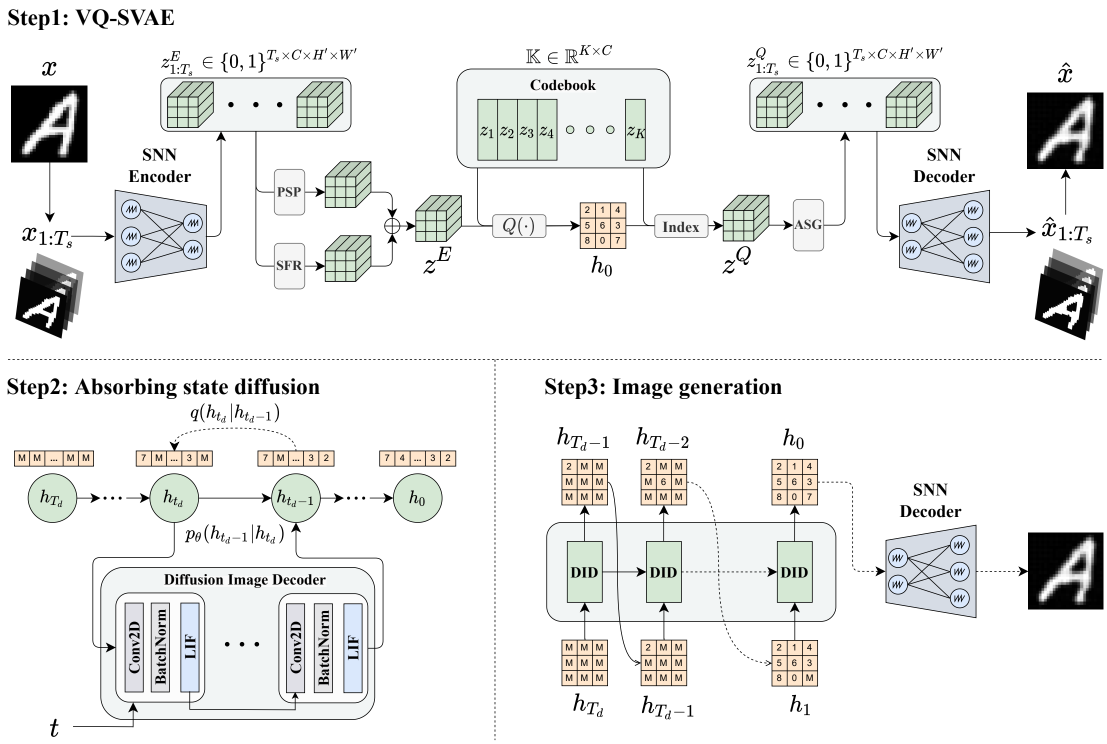

Fig. 1. The training process of Spiking-Diffusion consists of two stages. Step 1: Compress the images into discrete variables through VQ-SVAE. Step 2: The spiking diffusion image decoder (SDID) models the discrete latent space by reversing the forward diffusion process, which gradually adds masks in the discrete matrix through a fixed Markov chain. Step 3: During the test process, the SDID lifts the masks through an autoregressive process to obtain a discrete matrix with the target distribution.

the one-hot version of h with the row vector $\boldsymbol{h}$, we can define the forward process as:

$$q(\boldsymbol{h}_{t_d}|\boldsymbol{h}_{t_d-1}) = \mathrm{Cat}(\boldsymbol{h}_{t_d}; \boldsymbol{p} = \boldsymbol{h}_{t_d-1}\boldsymbol{Q}_{t_d}) \tag{10}$$

where $\mathrm{Cat}(\boldsymbol{h}; \boldsymbol{p})$ is a categorical distribution parameterized by $\boldsymbol{p}$, and $[\boldsymbol{Q}_{t_d}]_{ij} = q(h_{t_d} = j|h_{t_d-1} = i)$ is the transition matrix of the forward Markov chain. We adopt absorbing state diffusion which is compatible with SNNs. Its transition matrix $Q_{t_d} \in \mathbb{R}^{(K+1)\times(K+1)}$ can be expressed as follows:

$$Q_{t_d} = \begin{bmatrix} 1-\gamma_{t_d} & 0 & \cdots & \gamma_{t_d} \\ 0 & 1-\gamma_{t_d} & \cdots & \gamma_{t_d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \tag{11}$$

$$\gamma_{t_d} = \frac{1}{T_d - t_d + 1} \tag{12}$$

The Eq. (11) indicates that at each timestep, each regular token has probability γ_{t_d} of being replaced by the [MASK] token, and probability $1-\gamma_{t_d}$ of remaining unchanged, but the [MASK] token remains unchanged. In our work, we set the [MASK] token as K, as a result, $\boldsymbol{Q}_{t_d} \in \mathbb{R}^{(K+1)\times(K+1)}$ and each token possesses $K+1$ states. Therefore, the transition matrix with the absorbing state is non-zero only on the diagonal and the last column, so the sparsity of transition matrix will help to reduce the computational cost of the forward process.

The reverse process is defined as a Markov chain parameterized by θ, which gradually denoises from the data distribution $p_\theta(\boldsymbol{h}_{0:T_d}) = p(\boldsymbol{h}_{T_d})\prod_{t_d=1}^{T_d} p_\theta(\boldsymbol{h}_{t_d-1}|\boldsymbol{h}_{t_d})$ and is optimized by the evidence lower bound (ELBO). With t_d^{th} term the loss can be written as:

$$\mathcal{L}_{t_d} = D_{\mathrm{KL}}(q(\boldsymbol{h}_{t_d-1}|\boldsymbol{h}_0)||p_\theta(\boldsymbol{h}_{t_d-1}|\boldsymbol{h}_{t_d})) \tag{13}$$

To reduce the randomness of training, SDID $S(h_{t_d}, t_d)$ is employed to predict $p_\theta(\boldsymbol{h}_0|\boldsymbol{h}_{t_d})$ rather than learn $p_\theta(\boldsymbol{h}_{t_d-1}|\boldsymbol{h}_{t_d})$ directly. As a result, variational bound reduces to:

$$\mathbb{E}_{q(\boldsymbol{h}_0)}\left[\sum_{t_d=1}^{T_d}\frac{1}{t_d}\mathbb{E}_{q(\boldsymbol{h}_{t_d}|\boldsymbol{h}_0)}\Big[\sum_{\boldsymbol{h}_{t_d}(i,j)=m}\log p_\theta(\boldsymbol{h}_0(i,j)|\boldsymbol{h}_{t_d})\Big]\right] \tag{14}$$

III. Experiments

A. Datasets

We test our model on five datasets: The MNIST [8] consists of 60,000 handwritten digit images divided into 10 classes (digits 0 to 9); FMNIST [9] contains 60,000 images of 10 different classes of clothing; KMNIST [10] consists of 60,000 images of Japanese characters, with each image belonging to one of 10 classes; Letters [11] provides 145,000 images of 26 Latin letters; for Cifar10[12], 50,000 images are used for training and 10,000 images for evaluation.

TABLE I
PERFORMANCE COMPARISONS OF SPIKING-DIFFUSION(OURS) AND FSVAE ON SPIKINGJELLY [13].

Dataset	Model	MSE ↓	SSIM ↓	KID ↓	FID ↓
MNIST	FSVAE	0.023	0.219	0.054	90.57
	Ours	**0.006**	**0.077**	**0.018**	**37.50**
FMNIST	FSVAE	0.023	0.378	0.070	113.2
	Ours	**0.011**	**0.233**	**0.055**	**91.98**
KMNIST	FSVAE	0.054	0.452	0.272	166.9
	Ours	**0.014**	**0.151**	**0.068**	**59.23**
Letters	FSVAE	0.015	0.180	0.140	95.96
	Ours	**0.005**	**0.078**	**0.075**	**67.41**
Cifar10	FSVAE	0.024	0.689	0.181	218.6
	Ours	**0.009**	**0.399**	**0.121**	**120.5**

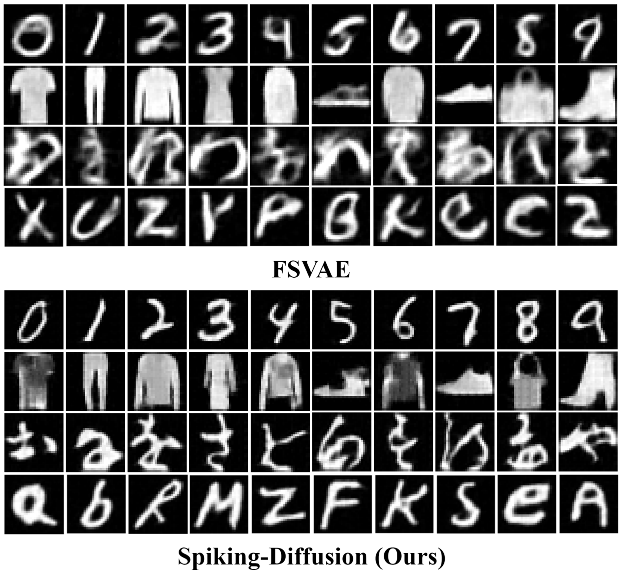

Fig. 2. Generated images of FSVAE and our Spiking-Diffusion (From top to down: MNIST, FMNIST, KMNIST, and Letters).

B. Implementation details

We make a comprehensive comparison between Spiking-Diffusion and FSVAE [4], which is the state-of-the-art SNN-based generative model generating images with quality equal to or better than ANN-based models. We test the FSVAE and Spiking-Diffusion on all datasets with the same experimental setup. The LIF neuron model's surrogate gradient function is $g(x) = \frac{1}{\pi}\arctan(\frac{1}{\pi}\alpha x) + \frac{1}{2}$, and its derivative is $g'(x) = \frac{\alpha}{2(1+(\frac{\pi}{2}\alpha x)^2)}$, where α represents the slope parameter. For all neurons, $\alpha = 2$, $V_{reset} = 0$, and $V_{th} = 1$. The Optimizer is AdamW with $\beta = (0.9, 0.999)$ and weight decay 0.001. The training lasts for 100 epochs. In addition, the original FSVAE used the STBP-tdBN [22] framework to train SNNs, while we adopt SpikingJelly [13] in our work.

C. Evaluation Metrics and Results

Evaluation Metrics. We evaluate the reconstruction ability of the model by comparing the input and output images using two metrics: mean squared error (MSE) loss and structural similarity (SSIM) loss. To assess the quality of sampled images, we use two commonly used scores: Kernel Inception Distance (KID) and Fréchet inception distance (FID).

Results. Table I shows that our Spiking-Diffusion model outperforms FSVAE in reconstruction and generation with the SpikingJelly framework. Spiking-Diffusion yields lower MSE and SSIM losses of reconstruction quality, and lower FID and KID scores of generated image quality. Figure 2 shows the examples of generated images on the four datasets, from which we find that with the same SNN encoder and decoder structures, the images generated by FSVAE are blurrier, while the images generated by our Spiking-Diffusion are clearer with obvious boundaries. The is because that only the Bernoulli distribution is used in FSVAE to approximate the true posterior distribution, which cannot fully capture the rich semantic information in the image. In contrast, the Spiking-Diffusion can approximate the complex posterior distribution through nearest neighbor search and generate images by seeing context.

IV. CONCLUSION

In this work, we present Spiking-Diffusion, the first implementation of diffusion models in SNNs. Our approach is based on VQ-DDM, which has two stages. First, we train the VQ-SVAE model to achieve image reconstruction, by using both PSP and SFR to model spiking features and constructing a codebook for image discretization. Secondly, we construct a discrete diffusion model in the discrete feature domain of images, and using absorbing diffusion state in Spiking-Diffusion is proven to generate highe quality images. The experimental results demonstrate that Spiking-Diffusion currently outperforms other SNN-based generative models. Future work will focus on how to train larger scale SNN generative models.

ACKNOWLEDGMENTS

This work is supported by the National Science and Technology Major Project from Minister of Science and Technology, China (Grant No.2018AAA0103100), and the Laboratory Open Fund of Beijing Smartchip Microelectronics Technology Co., Ltd. (No. SGSC0000SJQT2207792) (Mingxuan Liu, Jie Gan, and Rui Wen contributed equally to this work).

REFERENCES

[1] Jilin Zhang, Dexuan Huo, Jian Zhang, Chunqi Qian, Qi Liu, Liyang Pan, Zhihua Wang, Ning Qiao, Kea-Tiong Tang, and Hong Chen, "22.6 anp-i: A 28nm 1.5 pj/sop asynchronous spiking neural network processor enabling sub-o. 1 μj/sample on-chip learning for edge-ai applications," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 21–23.

[2] João D. Nunes, Marcelo Carvalho, Diogo Carneiro, and Jaime S. Cardoso, "Spiking neural networks: A survey," *IEEE Access*, vol. 10, pp. 60738–60764, 2022.

[3] Vineet Kotariya and Udayan Ganguly, "Spiking-gan: A spiking generative adversarial network using time-to-first-spike coding," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–7.

[4] Hiromichi Kamata, Yusuke Mukuta, and Tatsuya Harada, "Fully spiking variational autoencoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 7059–7067.

[5] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015.

[6] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[7] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg, "Structured denoising diffusion models in discrete state-spaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17981–17993, 2021.

[8] Li Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE signal processing magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[9] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[10] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha, "Deep learning for classical japanese literature," *arXiv preprint arXiv:1812.01718*, 2018.

[11] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[12] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[13] Wei Fang, Yanqi Chen, Jianhao Ding, Ding Chen, Zhaofei Yu, Huihui Zhou, Timothée Masquelier, Yonghong Tian, and other contributors, "Spikingjelly," https://github.com/fangwei123456/spikingjelly, 2020, Accessed: 2023-04-18.

[14] RB Stein and Alan Lloyd Hodgkin, "The frequency of nerve action potentials generated by applied currents," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 167, no. 1006, pp. 64–86, 1967.

[15] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2661–2671.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

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

[18] Youngeun Kim, Hyoungseob Park, Abhishek Moitra, Abhiroop Bhattacharjee, Yeshwanth Venkatesha, and Priyadarshini Panda, "Rate coding or direct coding: Which one is better for accurate, robust, and energy-efficient spiking neural networks?," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 71–75.

[19] Murray B Sachs and Eric D Young, "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 470–479, 1979.

[20] Uma R Karmarkar and Yang Dan, "Experience-dependent plasticity in adult visual cortex," *Neuron*, vol. 52, no. 4, pp. 577–585, 2006.

[21] Friedemann Zenke and Surya Ganguli, "Superspike: Supervised learning in multilayer spiking neural networks," *Neural computation*, vol. 30, no. 6, pp. 1514–1541, 2018.

[22] Hanle Zheng, Yujie Wu, Lei Deng, Yifan Hu, and Guoqi Li, "Going deeper with directly-trained larger spiking neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 11062–11070.