ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746865

EVENT-BASED MULTIMODAL SPIKING NEURAL NETWORK WITH ATTENTION MECHANISM

Qianhui Liu[1,2]*, Dong Xing*[1]*, Lang Feng*[1]*, Huajin Tang*[1,3]*, Gang Pan*[1,3*]

[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[3]Zhejiang Lab, Hangzhou, China
{qianhuiliu, dongxing, langfeng, htang, gpan}@zju.edu.cn

ABSTRACT

Human brain can effectively integrate visual and auditory information. Dynamic Vision Sensor (DVS) and Dynamic Audio Sensor (DAS) are event-based sensors imitating the mechanism of human retina and cochlea. Since the sensors record the visual and auditory input as asynchronous discrete events, they are inherently suitable to cooperate with the spiking neural network (SNN). Existing works of SNNs for processing events mainly focus on unimodality, however, audiovisual multimodal SNNs are still limited. In this paper, we propose an end-to-end event-based multimodal spiking neural network. The network consists of visual and auditory unimodal subnetworks and a novel attention-based cross-modal subnetwork for fusion. The attention mechanism measures the significance of each modality and allocates the weights to two modalities. We evaluate our proposed multimodal network on an event-based audiovisual joint dataset (MNIST-DVS and N-TIDIGITS datasets). Experimental results show the performance improvement of this multimodal network and the effectiveness of our proposed attention mechanism.

Index Terms— spiking neural networks, multimodal learning, dynamic vision sensors, dynamic audio sensors

1. INTRODUCTION

Humans perceive the environment through multiple modalities, such as vision, hearing and so on. By effectively integrating these modalities, human brain can better understand external information [1]. For example, vision information may be affected under poor lighting, but the audio is not hampered, whereas the audio may be attenuated by a background noise but the vision is not [2]. Thus audiovisual multimodality can overcome and compensate for the weaknesses of unimodality.

Dynamic Vision Sensor (DVS) [3] and Dynamic Audio Sensor (DAS) [4] are neuromorphic devices imitating the mechanisms of human retina and cochlea and have been explored in many works [5, 6, 7, 8]. Each pixel in DVS individually emits events when it monitors sufficient light changes in receptive field. DAS produces events from delta changes in the rectified channel output using an asynchronous delta modulation scheme [5]. The output of each sensor is a stream of events collected from each pixel (channel), forming an asynchronous and sparse representation of the scene.

These event-based representations are inherently suitable to cooperate with the spiking neural network (SNN) since SNN also has the event-based property. SNN uses discrete spikes to transmit information between units, which mimics the behavior of human brain [9]. The strength in processing spatio-temporal information [10, 11, 12] also makes it an ideal choice for visual or auditory events. Existing research has utilized SNN to deal with unimodal events [13, 14, 15, 16, 17], however, works of SNN on event-based audiovisual multimodal information processing are still limited.

This paper proposes an event-based multimodal SNN. We utilize a convolutional SNN for visual events from DVS and a recurrent SNN for audio events from DAS. The outputs of these two unimodal subnetworks are concatenated and fed to the attention-based cross-modal subnetwork. The proposed attention mechanism automatically measures the significance of each modality and dynamically allocates the weights to two modalities. The overall network is trained in an end-to-end manner to optimize the subnetworks towards the common goal. We evaluate our proposed multimodal network on an event-based audiovisual joint digit classification dataset (MNIST-DVS and N-TIDIGITS datasets). The results show the performance improvement of this multimodal network and the effectiveness of the proposed attention mechanism.

2. METHOD

In this section, we introduce the proposed multimodal spiking neural network in detail. The framework of this network

*Corresponding Author.

This work is supported by the Natural Science Foundation of China (No. 61925603, U1909202), the Key Research and Development Program of Zhejiang Province in China (2020C03004), Key Realm R&D Program of Guangzhou (202007030005), Zhejiang Lab and A*STAR under its RIE2020 Advanced Manufacturing and Engineering Domain (AME) Programmatic Grant (Grant No. A1687b0033, Project Title: Spiking Neural Networks).

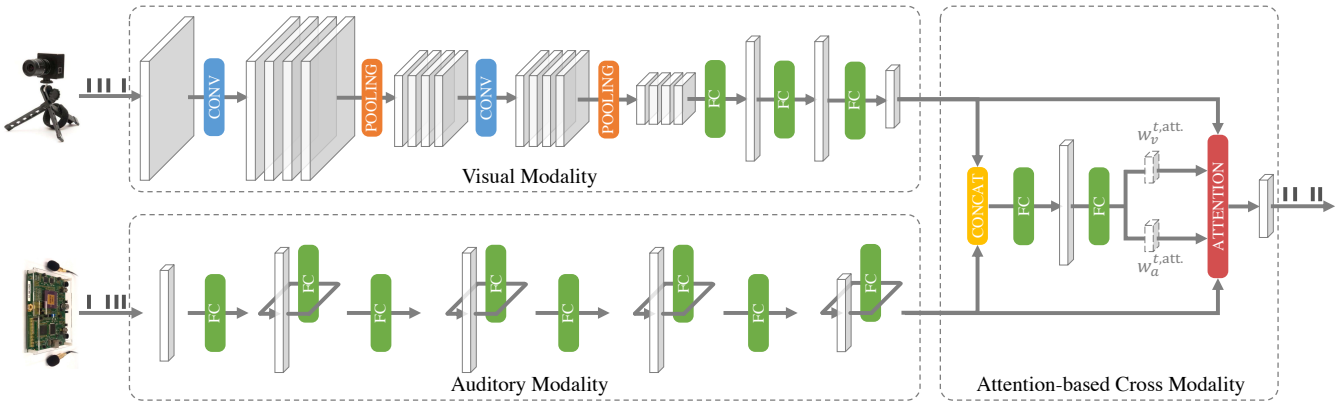

Fig. 1. The framework of the proposed multimodal spiking neural network. The network consists of three subnetworks, i.e., visual modality, auditory modality and attention-based cross modality.

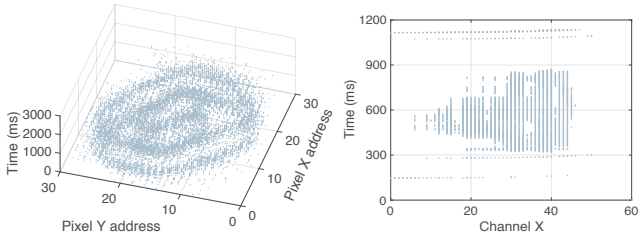

Fig. 2. Visualization of visual and auditory event streams representing the digit '0' in MNIST-DVS dataset (left) and N-TIDIGITS dataset (right).

is shown in Fig. 1. We first present the visual modal subnetwork and auditory modal subnetwork. Then we introduce our proposed attention-based cross-modal subnetwork. Finally, we elaborate the end-to-end training scheme of the overall multimodal network.

2.1. Visual Modality

The events from DVS carry the information of timestamp (the time when the event was emitted) and address (the position of the corresponding pixel in the sensor). Fig. 2 shows a visualization of the visual event stream representing the digit '0' in MNIST-DVS dataset [3]. The events are first encoded in *Input* layer to a compatible format for the following processing. This layer can be seen as a neural map, comprised of spiking neurons that have no internal dynamics and emit spikes when receiving the corresponding events.

The events are then sent to the convolutional SNN. In this paper, we adopt the iterative Leaky Integrate-and-Fire (LIF) model [10] to describe the neural dynamics since it has biological reality and is suitable for the error back propagation algorithm. Specifically, the dynamics of iterative LIF model can be described as:

$$u_i^{t,n} = u_i^{t-1,n} f(o_i^{t-1,n}) + \mathcal{F}(o_i^{t,n-1}) \tag{1}$$

$$o_i^{t,n} = g(u_i^{t,n}) \tag{2}$$

$$f(x) = \begin{cases} \tau, & x = 0 \\ 0, & x = 1 \end{cases} \tag{3}$$

$$g(x) = \begin{cases} 1, & x \geq V_{th} \\ 0, & x < V_{th} \end{cases} \tag{4}$$

where the superscript t denotes the timestep and n denotes the n-th layer. $o_i^{t,n}$ is the output spike at timestamp t of i-th neuron in n-th layer. $u_i^{t,n}$ is the membrane potential. The forget gate $f(\cdot)$ controls the leaky extent of the membrane potential and the output gate $g(\cdot)$ generates a spike activity when the input is above the threshold V_{th}. τ denotes the potential decay constant. $\mathcal{F}(\mathbf{o}^{t,n}) = \mathbf{w}^n \mathbf{o}^{t,n} + \mathbf{b}^n$ integrates the input to form the stimulus of the current neuron. In visual modal subnetwork, the instantiation of $\mathcal{F}$ involves convolutional operation and fully connected (FC) operation. We also employ the average pooling for convolutional output spikes in order to control the network size. Finally, for a C-class categorization task, the output layer of visual unimodal network, audio unimodal network and overall network all employ C neurons and each of them represents one particular class.

2.2. Auditory Modality

The events from DAS carry the information of timestamp and channel. Fig. 2 shows a visualization of the audio event stream representing 'zero' in N-TIGIDITS dataset [6]. The events from DAS are also first sent to *Input* layer to a compatible format for the following processing.

To explore the temporal features contained in auditory events, recurrent layers are adopted in this auditory unimodal subnetwork. In the recurrent layer, $\mathcal{F}$ in neural dynamics is instantiated as a FC operation, which not only receives spikes from the previous layer but also receives their own membrane potentials at the previous timestep. With this setting, the temporal effect of audio information can be strengthened.

2.3. Attention-based Cross Modality

This section dynamically fuses two modalities with the proposed attention mechanism to determine the final classification result. Information provided in a single modality is limited and vulnerable. Multimodal information can compensate for the weakness of unimodality, providing a more comprehensive and accurate understanding. In order to effectively integrate these two modalities, we need to allocate the attention to each modality appropriately. This is naturally supported by the human brain, as experimental results reveal that human can selectively activate different brain regions to deal with information from different modalities [18].

To this end, we propose an attention mechanism which is described with the following equations:

$$\begin{aligned}
&u_i^{t,n} = u_i^{t-1,n} f(o_i^{t-1,n}) + \mathcal{F}_{\mathrm{FC}}([o_{v,i}^{t,l(v)}, o_{a,i}^{t,l(a)}]) \\
&o_i^{t,n} = g(u_i^{t,n}) \\
&\cdots \\
&[w_v^{t,\mathrm{att.}}, w_a^{t,\mathrm{att.}}] = \mathcal{F}_{\mathrm{FC}}(o_i^{t,l(m)-1}) \\
&u_i^{t,l(m)} = u_i^{t-1,l(m)} f(o_i^{t-1,l(m)}) + w_v^{t,\mathrm{att.}} o_{v,i}^{t,l(v)} + w_a^{t,\mathrm{att.}} o_{a,i}^{t,l(a)} \\
&o_i^{t,l(m)} = g(u_i^{t,l(m)})
\end{aligned} \tag{5}$$

Specifically, we first concatenate the output spikes $o_{v,i}^{t,l(v)}$ from visual modality and $o_{a,i}^{t,l(a)}$ from auditory modality, and then feed them to FC layers (one FC layer used in our experiment). Next, the cross-modal subnetwork learns to estimate two weights $w_v^{t,\mathrm{att.}}$ and $w_a^{t,\mathrm{att.}}$ for visual modality and auditory modality respectively. The weights represent the degree to which the multimodal network allocates its attention on each modality. The output spikes from two modalities are weighted by attention and fed to the output layer of overall multimodal network for a final result. In this procedure, the attention mechanism automatically compares and judges the significance of each modality, and dynamically allocates higher weights on the one with more confidence. Therefore, the overall multimodal network becomes more flexible and can compensate for partial information of a single modality.

2.4. Overall Training

We now present the end-to-end training scheme of overall multimodal network. The loss function L is formed by three components:

$$L = L_v + L_a + L_m \tag{6}$$

where L_v, L_a and L_m denote the loss of visual modality, auditory modality and overall multimodal network, respectively. The loss function measures the mean square error between the averaged voting results and label vector Y within a given time window T:

$$L_m = \|\mathbf{Y} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{o}^{t,l(m)}\|_2^2 \tag{7}$$

where $\mathbf{o}^{t,l(m)}$ denotes the output vector of overall multimodal network. L_v and L_a are calculated with a similar equation. The only difference is that they replace the output vector $\mathbf{o}^{t,l(m)}$ by the results of visual modality $\mathbf{o}_v^{t,l(v)}$ and auditory modality $\mathbf{o}_a^{t,l(a)}$ to enhance the ability of each single modality.

Spikes in iterative LIF model not only propagate layer-by-layer in spatial domain, but also affect the neural states in temporal domain. Therefore, gradient-based training should consider the derivatives in these two domains. We here adopt Spatio-Temporal Backpropagation (STBP) [10] to train our network. The derivative of loss function L_m with respect to u and o in the n-th layer at time step t can be computed by:

$$\begin{aligned}
\frac{\partial L_m}{\partial o_i^{t,n}} &= \sum_j \frac{\partial L_m}{\partial u_j^{t,n+1}} \frac{\partial u_j^{t,n+1}}{\partial u_i^{t,n}} + \frac{\partial L_m}{\partial u_i^{t+1,n}} \frac{\partial u_i^{t+1,n}}{\partial o_i^{t,n}} \\
\frac{\partial L_m}{\partial u_i^{t,n}} &= \frac{\partial L_m}{\partial o_i^{t,n}} \frac{\partial o_i^{t,n}}{\partial u_i^{t,n}} + \frac{\partial L_m}{\partial u_i^{t+1,n}} \frac{\partial u_i^{t+1,n}}{\partial u_i^{t,n}}
\end{aligned} \tag{8}$$

The derivations of L_v and L_a are the same. Due to the non-differentiable property of spike activities, $\frac{\partial o^t}{\partial u^t}$ does not exist. We take rectangular function $h(u)$ to approximate the derivative of spike activity [10]:

$$h(u) = \frac{1}{a} sign(|u - V_{th}| < \frac{a}{2}) \tag{9}$$

where the width parameter a determines the shape of $h(u)$.

3. EXPERIMENTAL RESULTS

3.1. Datasets

The joint digit classification dataset consists of

MNIST-DVS dataset [3]**:** it is obtained with a DVS128 sensor by recording 10,000 original handwritten images in MNIST. The images were displayed on an LCD monitor moving with slow motion. The full length of each recording is about 2000 ms and the spatial resolution is 28×28.

N-TIDIGITS dataset [6]**:** it consists of output spikes of a 64-channel CochleaAMS1b sensor in response to audio waveforms from the original Tidigits dataset. It has 4,950 samples for 11 categories including 'oh', 'zero', and the digits '1' to '9'. The full length of each recording is about 2000 ms.

In order to maintain the category consistency between two datasets, we use 10 kinds of digits ('zero', '1'-'9'). Therefore, the joint digit classification dataset has 10,000 pairs of inputs (4500 audio samples are repeated).

3.2. Experimental Settings

The first 50% samples of each dataset are used for training and the remaining ones are used for testing. Each experiment is repeated five times to obtain the mean and standard deviation of the accuracy. The time window T is set as 64 timesteps. The learning rate is set as 1e-3. The derivative

Methods	Structure	Modality	Dataset	Mean	Std	Best
SPA [19]	3136-100	Unimodal	MNIST-DVS	91.58	0.46	92.24
SLAYER [20]	32C3-AP2-64C3-AP2-512-10[i]	Unimodal	MNIST-DVS	93.03	0.17	93.16
STCA [21]	32C3-AP2-64C3-AP2-512-10	Unimodal	MNIST-DVS	94.21	0.56	94.82
HM2-BP [22]	250-250-11	Unimodal	N-TIDIGITS	-	-	89.69
GRN [6]	2× G200-100-11[ii]	Unimodal	N-TIDIGITS	-	-	90.90
Phased-LSTM [6]	2× 250L-11[iii]	Unimodal	N-TIDIGITS	-	-	91.25
ST-RSBP [14]	400-R400-400-11[iv]	Unimodal	N-TIDIGITS	93.63	0.27	93.90
STBP [10]	512R-512R-512R-11R	Unimodal	N-TIDIGITS	89.19	0.66	89.90
V-Network	32C3-AP2-32C3-AP2-128-10	Unimodal	MNIST-DVS	96.33	0.20	96.64
A-Network	512R-512R-512R-10R	Unimodal	N-TIDIGITS	90.66	1.10	91.74
This Work	32C3-AP2-32C3-AP2-128-10 512R-512R-512R-10R -256-10	Multimodal	MNIST-DVS + N-TIDIGITS	98.95	0.17	99.10

[i]C represents convolutional layer and AP represents average pooling layer; [ii]G represents GRN layer; [iii]L represents LSTM layer; [iv]R represents recurrent layer.

Table 1. Comparison of our multimodal network with other unimodal methods.

approximation parameter a is set as 1.0. The initial membrane potential and threshold of neurons are set as 0 and 0.5. The decay constant τ of neurons in visual, auditory and cross modality is set as 0.2, 0.1 and 0.9.

3.3. Comparison of Multimodal and Unimodal Networks

We evaluate the performance of the proposed multimodal network on the joint digit classification dataset. As shown in Table 1, our multimodal network achieves a classification accuracy of 98.95%, which is higher than other unimodal SNN methods. Meanwhile, the standard deviation we achieved remains one of the lowest among all methods, which reflects the stability of our multimodal network. As an ablation study, we have also compared our multimodal network with two single modalities that are also trained by STBP. We can notice that our multimodal network surpasses visual unimodality (V-Network) by 2.62% and auditory unimodality (A-Network) by 8.29%. This indicates that our proposed multimodal framework can effectively integrate two unimodal information and make more accurate decisions.

3.4. Effectiveness of Attention Mechanism

To further investigate the performance improvement, we conduct an experiment to compare our network with a variant having no attention mechanism. This variant integrates two unimodal subnetworks by directing connecting them to two consecutive FC layers. Fig. 3 shows the performance of two multimodal networks on testing set. We can observe that the multimodal network with attention has a higher performance than that without attention. It is worth noting that the performance of network without attention is slightly lower than visual modality. Since the performance of auditory modality is inferior to that of visual modality, if we simply concatenate two single modalities, the audio input will affect the overall performance. Instead, our proposed attention mechanism effectively allocates the weights to two modalities and obtains a superior performance over single modality.

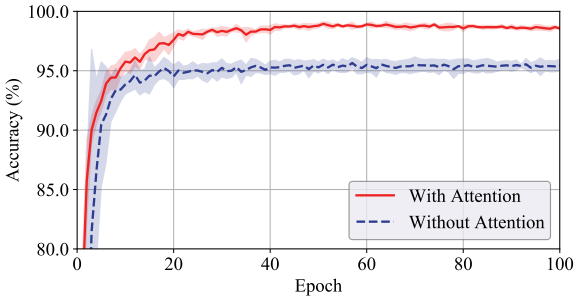

Fig. 3. Performance comparison between cross modality with and without the proposed attention.

4. CONCLUSION

This paper presents an end-to-end framework of event-based multimodal SNN with an attention mechanism. We utilize a convolutional SNN for visual modality from DVS and a recurrent SNN for audio modality from DAS. To mitigate the defect brought by single modality, we propose an attention mechanism in cross modality to dynamically allocate the weights to two modalities. We evaluate our multimodal network on a joint digit classification dataset. The experimental result shows that our work surpasses methods with unimodality. We have also demonstrated the effectiveness of our attention mechanism with an ablation study.

5. REFERENCES

[1] Malu Zhang, Xiaoling Luo, Yi Chen, Jibin Wu, Ammar Belatreche, Zihan Pan, Hong Qu, and Haizhou Li, "An efficient threshold-driven aggregate-label learning algorithm for multimodal information processing," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 3, pp. 592–602, 2020.

[2] Nitin Rathi and Kaushik Roy, "Stdp-based unsupervised multimodal learning with cross-modal processing in spiking neural network," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 5, no. 1, pp. 143–153, 2021.

[3] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck, "A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[4] Shih-Chii Liu, Andre van Schaik, Bradley A Minch, and Tobi Delbruck, "Asynchronous binaural spatial audition sensor with 2 × 64 × 4 channel output," *IEEE Trans. Biomed. Circuits Syst.*, vol. 8, no. 4, pp. 453–464, 2014.

[5] Xiaoya Li, Daniel Neil, Tobi Delbruck, and Shih-Chii Liu, "Lip reading deep network exploiting multi-modal spiking visual and auditory sensors," in *ISCAS*, 2019, pp. 1–5.

[6] Jithendar Anumula, Daniel Neil, Tobi Delbruck, and Shih-Chii Liu, "Feature representations for neuromorphic audio spike streams," *Frontiers in Neuroscience*, vol. 12, pp. 23, 2018.

[7] Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, and Gang Pan, "Indoor lighting estimation using an event camera," in *CVPR*, 2021, pp. 14760–14770.

[8] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong, "Object tracking by jointly exploiting frame and event domain," in *ICCV*, 2021, pp. 13043–13052.

[9] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda, "Towards spike-based machine intelligence with neuromorphic computing," *Nature*, vol. 575, no. 7784, pp. 607–617, 2019.

[10] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi, "Spatio-temporal backpropagation for training high-performance spiking neural networks," *Frontiers in Neuroscience*, vol. 12, pp. 331, 2018.

[11] Malu Zhang, Hong Qu, Ammar Belatreche, Yi Chen, and Zhang Yi, "A highly effective and robust membrane potential-driven supervised learning method for spiking neurons," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 1, pp. 123–137, 2019.

[12] Malu Zhang, Jiadong Wang, Burin Amornpaisannon, Zhixuan Zhang, VPK Miriyala, Ammar Belatreche, Hong Qu, Jibin Wu, Yansong Chua, Trevor E Carlson, et al., "Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, 2021.

[13] Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman, "HFirst: A temporal approach to object recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2028–2040, 2015.

[14] Wenrui Zhang and Peng Li, "Spike-train level backpropagation for training deep recurrent spiking neural networks," in *NeurIPS*, 2019, pp. 7800–7811.

[15] Rong Xiao, Huajin Tang, Yuhao Ma, Rui Yan, and Garrick Orchard, "An event-driven categorization model for AER image sensors using multispike encoding and learning," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 9, pp. 3649–3657, 2020.

[16] Qianhui Liu, Gang Pan, Haibo Ruan, Dong Xing, Qi Xu, and Huajin Tang, "Unsupervised AER object recognition based on multiscale spatio-temporal features and spiking neurons," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 31, no. 12, pp. 5300–5311, 2020.

[17] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan, "Event-based action recognition using motion information and spiking neural networks.," in *IJCAI*, 2021, pp. 1743–1749.

[18] Ryuta Kawashima, Satoshi Imaizumi, Koichi Mori, Ken Okada, Ryoi Goto, Shigeru Kiritani, Akira Ogawa, and Hiroshi Fukuda, "Selective visual and auditory attention toward utterancesa pet study," *Neuroimage*, vol. 10, no. 2, pp. 209–215, 1999.

[19] Qianhui Liu, Haibo Ruan, Dong Xing, Huajin Tang, and Gang Pan, "Effective AER object classification using segmented probability-maximization learning in spiking neural networks.," in *AAAI*, 2020, pp. 1308–1315.

[20] Sumit Bam Shrestha and Garrick Orchard, "SLAYER: Spike layer error reassignment in time," in *NeurIPS*, 2018, pp. 1412–1421.

[21] Pengjie Gu, Rong Xiao, Gang Pan, and Huajin Tang, "STCA: Spatio-temporal credit assignment with delayed feedback in deep spiking neural networks," in *IJCAI*, 2019, pp. 1366–1372.

[22] Yingyezhe Jin, Wenrui Zhang, and Peng Li, "Hybrid macro/micro level backpropagation for training deep spiking neural networks," in *NeurIPS*, 2018, pp. 7005–7015.