Object Recognition From Event Cameras Using Spiking Neural Networks

Diego Palma[1,2], Shihao Zou[2] and Li Cheng[2]

[1] Research Intern from University of Engineering and Technology (UTEC)

[2] Vision and Learning Lab – University of Alberta

Introduction

Event Cameras

- Event cameras are bio-inspired vision sensors whose pixels work asynchronously, reporting brightness changes as they occur (Figure 1).
- Advantages of event cameras:
  - Very high temporal resolution and low latency (~μs).
  - High dynamic range (twice that of common cameras).
  - Low power consumption.

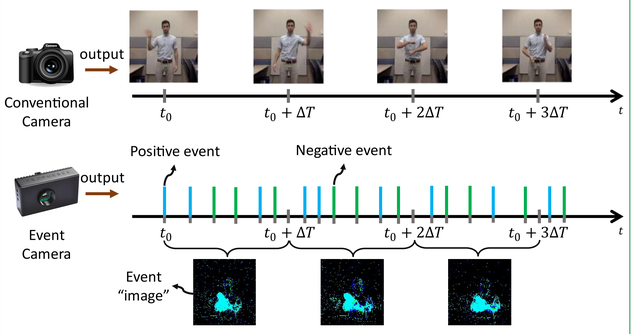

Figure 1. The upper part shows the output of a common camera, which is a 2D static image at a fixed frequency rate. The bottom part shows the event camera output that corresponds to asynchronous events. For visualization purposes, events are grouped in a period to visualize an event "image". Images were taken from DVS Gesture dataset [1]

Objective

Task: Object Recognition

Event cameras are suitable for object recognition, i.e, classifying objects based on event data (Figure 2).

Problem: State-of-the-art techniques can not be directly applied to events, due to their asynchronous nature and the lack of pixel intensity information.

Objective: The present work aims to implement different Spiking Neural Network (SNN) architectures, fed by event-based data, to perform object recognition on two different datasets: DVS128 Gesture and CIFAR 10-DVS, in order to obtain the highest accuracy as possible.

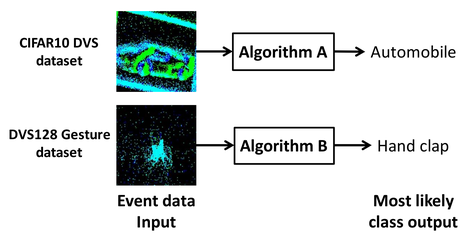

Figure 2. Object recognition task. First, event data is the input of the algorithm. Then, this should classify which class it belongs to by returning the most likely class output.

Methodology

Overview

- An overview of the methodology used is shown in Figure 3.
- Event data is captured by an event camera. In this case, we use datasets that already have all the event data stored in specific classes (See Table 1).
- Event data are the inputs of the Spiking Neural Networks (SNN) which are composed of the following blocks: spiking encoder (SE), spike-element-wise block (SEW block) [3] and fully connected layers (FC).

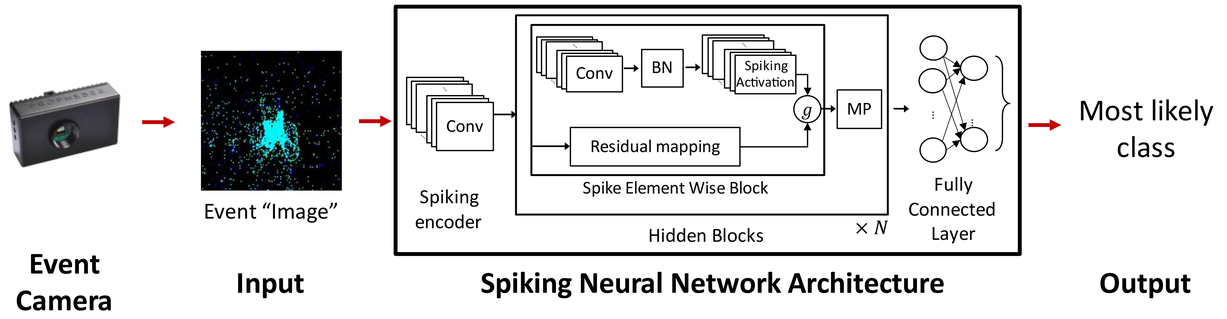

Figure 3. General overview of the methodology used to recognize objects based on event data.

Table 1. Characteristics of the two datasets used as input data.

Dataset	N° classes	Possible classes
DVS128 Gesture [1]	11	arm roll, hand clap, left hand clockwise, left hand counterclockwise, left-hand wave, right-hand wave, right-hand clockwise, right-hand counterclockwise, air drums, air guitar and other gesture.
CIFAR 10 DVS [2]	10	airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck.

Why SNNs?

- Sparsity of event signals results in the efficiency of the SNNs (Figure 4), especially for hardware implementation.

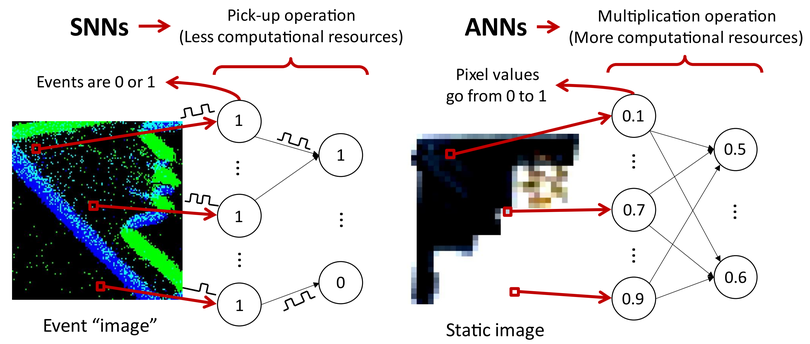

Figure 4. The nature of events reduces the operations to simply pick-up operations. Then, SNNs result more efficients than ANNs.

Configurations of SNNs

- Based on this methodology, different configurations of the SNN architecture were proven on each dataset. We took different hyperparameter values regarding the training and the components of the architecture:
  - ❑ **Training:** Batch size, epochs, learning rate and simulating time-steps.
  - ❑ **Architecture:** Number of SEW blocks (N) and the number of channels of the spiking encoder.

Results

- Table 2 shows the accuracy for the most successful architectures[1].
- Figure 5 and 6 show visual examples of the SNN outputs for both datasets, where green and red boxes indicate a true positive and false positive prediction, respectively.

Table 2. Accuracy of the successful architectures for each dataset.

Dataset	Architecture	Accuracy (%)
DVS128 Gesture	SE(32)-{SEW(7,ADD)-MP(2)}x7-FC(11)	97.22
	SE(64)-{SEW(7,AND)-MP(2)}x7-FC(11)	94.19
CIFAR 10 DVS	SE(64,128)-{SEW(7,ADD)-MP(2)}x7-FC(10)	70.40
	Same architecture, but LR=0.001 and T=16	68.12

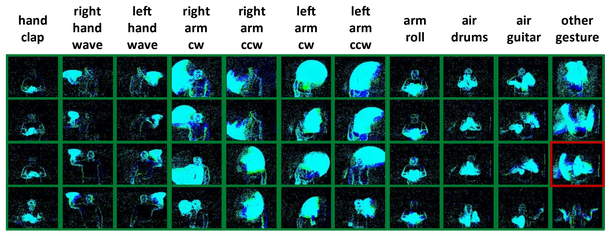

Figure 5. Examples of SNN predictions for DVS128 Gesture.

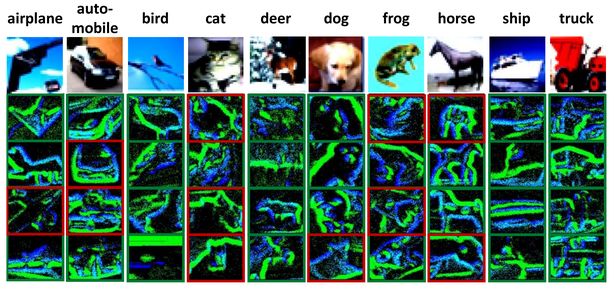

Figure 6. Examples of SNN predictions for CIFAR10 DVS.

[1] *These are partial results because the project is currently in progress, seeking for improve the results and find novel architectures.*

Conclusions

The present work implemented different SNN architectures to perform object recognition based on event data for the DVS128 Gesture and CIFAR10 dataset. On the first dataset a maximum accuracy of 97.22% was obtained, while on the other one was 70.4 %.

Impact

Object recognition based on event cameras have a large potential for robotics in challenging scenarios for standard cameras, such as high speed and high dynamic range. Implementing this technology would increase time response and precision of their real-time applications.

References

[1] A. Amir et al., "A Low Power, Fully Event-Based Gesture Recognition System," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 7388-7397, doi: 10.1109/CVPR.2017.781.

[2] Li H, Liu H, Ji X, Li G and Shi L (2017) CIFAR10-DVS: An Event-Stream Dataset for Object Classification. Front. Neurosci. 11:309. doi: 10.3389/fnins.2017.00309

[3] Fang, W., Yu, Z., Chen, Y., Huang, T., Masquelier, T., & Tian, Y. (2021). Deep Residual Learning in Spiking Neural Networks. Neural Information Processing Systems.