FUDAN UNIVERSITY

Grips

GLOBAL RESEARCH IMMERSION PROGRAM FOR YOUNG SCIENTISTS

Leveraging Dynamic Vision Sensors for Speck Event Recording to Develop Spiking Neural Networks in AIOT Home Appliances

Author: Chixia Ye (cye9@Sheffield.ac.uk)
Department of Automatic Control & System Engineering, The University of Sheffield (United Kingdom)

1. Abstract

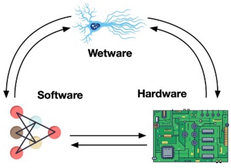
In the pursuit of advancing AIOT (Artificial Intelligence of Things) for modern home appliances, this paper explores the application of Dynamic Vision Sensors (DVS) to capture human activities and facilitate intelligent applications. Unlike conventional cameras, DVS offers significant advantages, including lower power consumption and enhanced privacy protection by recording only the outlines of human figures. By leveraging DVS technology, the development of Spiking Neural Networks (SNN) capable of interpreting speck events aims to create more efficient and privacy-conscious smart environments.

The applications extend beyond typical household tasks, encompassing scenarios such as protecting elderly individuals living alone and ensuring the safety of children. This research underscores the potential of DVS in transforming everyday home activities, making them more intelligent and context-aware, while simultaneously addressing energy efficiency and privacy concerns. The findings highlight the effectiveness of DVS in AIOT implementations and provide a foundation for future advancements in smart home technologies.

2. Dynamic Vision Sensor (DVS)

In the research, the advanced capabilities of the Synsense Speck2fDevKit are utilized to create datasets for neural network training. This device is a state-of-the-art Dynamic Vision Sensor (DVS) that integrates an asynchronous neuromorphic dynamic vision processor (DYNAPTMCNN) and a Dynamic Vision Sensor (DVS), also known as an Event Camera, Dynamic Event-based Sensor (DES), or Event-based Vision Sensor (EVS). This device features a large-scale spiking convolutional neural network (sCNN) architecture, configurable with up to 320,000 spiking neurons, and an internally integrated 128x128 resolution DVS, facilitating real-time and efficient dynamic vision input.

The device's event-driven nature provides high-speed signals through a sparse data stream, significantly reducing the amount of data that needs to be processed. This is particularly beneficial in the context of creating datasets for neural network training, where low latency and high efficiency are crucial. The integrated sCNN computing architecture optimizes highly sparse computations and minimizes latency to just 3.36μs per incoming event, resulting in an extremely low-latency visual processing pipeline. This efficient data handling capability makes the device ideal for generating precise and comprehensive datasets needed for training sophisticated neural networks.



By combining both the sensor and processing unit on a single chip, the device reduces production costs and simplifies the data collection process. The end-to-end nature of this System on Chip (SoC) facilitates the creation of detailed and accurate datasets, which are essential for training high-performance neural networks. This makes the device a highly effective tool for advancing research in developing and refining neural network models.

2. Neural network construction: SNN

In this study, a Spike Neural Network (SNN) is utilized to process and classify event-based data from a Dynamic Vision Sensor (DVS). The DVS inherently captures data with temporal information, making it a natural fit for SNNs, which are designed to handle such spiking dynamics.

The SNN model is based on Leaky Integrate-and-Fire (LIF) neurons, described by the differential equation:

$$C\frac{dV_m(t)}{dt} = -g_L(V_m(t) - E_L) + I(t)$$

where C is the membrane capacitance, $V_{m(t)}$ is the membrane potential, g_L is the leak conductance, E_L is the leak reversal potential, and $I(t)$ is the input current.

In this implementation, the LIF model is simplified while maintaining its core dynamics. The membrane potential update is given by

$$u_{t+1} = \tau \cdot u_t \cdot (1 - o_t) + W \cdot o_t$$

, and spike generation is described by

$$o_{t+1} = H(u_{t+1} - V_{th})$$

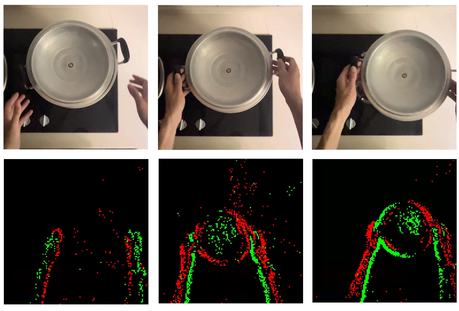
where H is the Heaviside step function and V_{th} is the spiking threshold.

The SNN architecture comprises convolutional layers followed by fully connected layers to handle spatiotemporal data from the DVS. It includes three convolutional layers (`conv1`, `conv2`, `conv3`) and pooling layers (`pool1`, `pool2`, `pool3`), with fully connected layers (`fc1`, `fc2`) for classification. The network processes timesteps to accumulate spike outputs, effectively leveraging the temporal resolution of event-based data while maintaining low power consumption.

4. Datasets Recording – Speck Event

To ensure that the model training and subsequent inference are optimally suited to the capabilities of the Dynamic Vision Sensor (DVS), I opted to create a custom dataset. The dataset is divided into six categories, each representing a specific kitchen-related activity: opening the refrigerator door, pulling a drawer, opening the washing machine, removing a pot, opening a pot lid, and closing a pot lid. For each category, I recorded 80 training samples and 20 testing samples, with each sample having a duration of 5 seconds and a frame rate of 1s per frame.

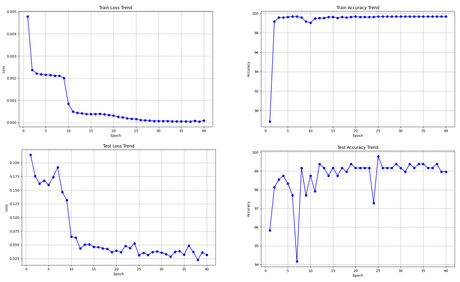
The DVS outputs a stream of events in the format <X, Y, P, T>, where X and Y represent the spatial coordinates, P represents the polarity of the brightness change, and T denotes the timestamp of the event. This event-based output captures dynamic changes in the visual scene by detecting variations in natural light intensity, thus providing bio-inspired, spiking neural signals.



From the above images, it is evident that the Dynamic Vision Sensor (DVS) camera captures motion trajectories with high temporal resolution, generating event-driven data. The first row consists of RGB images taken at three consecutive time points, illustrating the process of moving a pot away from a stove. The second row displays the corresponding DVS-rendered data, showcasing motion events at the same time points. The DVS camera records the sequence of motion events from the hands approaching the pot to its removal, with green and red dots indicating the polarity of the events (increase or decrease). Compared to traditional RGB image capture, the DVS camera offers significant advantages: it provides microsecond-level temporal resolution for accurately recording fast movements, operates in an event-driven manner to reduce data redundancy and improve processing efficiency, and consumes less power due to its selective event recording.

4. Model training - STBP

In this study, I employed the Spatio-Temporal Backpropagation (STBP) model, developed by the Center for Brain-Inspired Computing Research at Tsinghua University, due to its superior accuracy and performance. The STBP model is specifically designed for spiking neural networks (SNNs), optimizing the backpropagation process in both spatial and temporal dimensions, which is crucial for handling the dynamic nature of DVS data. This approach allows for efficient and accurate gesture recognition, leveraging the event-driven characteristics of dynamic vision sensors. The training conditions were meticulously set as follows: a batch size of 10 for training and 5 for testing, a total of 40 training epochs, an initial learning rate of 0.0001, and an SGD momentum of 0.5. CUDA was utilized to accelerate the training process, with a random seed of 1 to ensure reproducibility. Logs were recorded every 10 batches.



The final training results demonstrate the model's exceptional performance, with an average loss of 0.0021 and 100% accuracy on the training set, and an average loss of 0.0320 and 99% accuracy on the test set. These results indicate that the model has effectively learned and generalized well, making it highly suitable for practical applications in action recognition using DVS data.

5. Results and Future

In summary, the STBP model has shown excellent compatibility with DVS data, which is highly suitable for embedding in household appliances for action detection due to its excellent data processing capabilities and low power consumption. However, due to the limited project timeframe, there was not enough time to conduct real-time recognition testing and development, so real-time inference data was not obtained. Nonetheless, it is believed that this is feasible and can be achieved in future work.

Reference List:
1. Gallego, Guillermo, et al. "Event-Based Vision: A Survey." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, Jan. 2020, pp. 1–24, https://doi.org/10.1109/tpami.2020.3008413. Accessed 19 July 2024.
2. Kulkarni, Shruti, et al. "Learning and Real-Time Classification of Hand-Written Digits with Spiking Neural Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 1, 1 Dec. 2017, https://doi.org/10.1109/icecs.2017.8292015. Accessed 20 July 2023.
3. Wu, Yujie, et al. "Spatio-Temporal Backpropagation for Training High-Performance Spiking Neural Networks." Frontiers in Neuroscience, vol. 12, 23 May 2018, https://doi.org/10.3389/fnins.2018.00331.
4. Xing, Yannan, et al. "A New Spiking Convolutional Recurrent Neural Network (SCRNN) with Applications to Event-Based Hand Gesture Recognition." Frontiers in Neuroscience, vol. 14, 17 Nov. 2020, https://doi.org/10.3389/fnins.2020.590164. Accessed 24 July 2024.