Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration

Yingfu Xu[1], Kevin Shidqi[1], Gert-Jan van Schaik[1], Refik Bilgic[2], Alexandra Dobrita[1], Shenqi Wang[1], Roy Meijer[1], Prithvish Nembhani[1], Cina Arjmand[1], Pietro Martinello[1], Anteneh Gebregiorgis[3], Said Hamdioui[3], Paul Detterer[1], Stefano Traferro[1], Mario Konijnenburg[1], Kanishkan Vadivel[1], Manolis Sifalakis[1], Guangzhi Tang[4], and Amirreza Yousefzadeh[5]

[1] imec the Netherlands, Eindhoven, Netherlands
[2] imec, Leuven, Belgium
[3] TU Delft, Delft, Netherlands
[4] Maastricht University, Maastricht, Netherlands
guangzhi.tang@maastrichtuniversity.nl
[5] University of Twente, Enschede, Netherlands

Abstract. Modern AI systems are prohibitively unsustainable. Inspired by our brains, neuromorphic computing promises low-latency and energy-efficient neural network processing. Yet, current neuromorphic solutions still struggle to rival conventional deep learning accelerators' performance and area efficiency in practical applications. In this encore abstract, we present our published work [8] on explorations of optimizing sparse event-based neural network inference on SENECA, a scalable and flexible neuromorphic architecture. We introduce the event-driven depth-first convolution to increase area efficiency and latency in convolutional neural networks (CNNs) on the neuromorphic processor. We benchmarked our optimized solution on sensor fusion, digit recognition, and high-resolution object detection tasks, and showed significant improvements in energy, latency, and area, compared with other state-of-the-art large-scale neuromorphic processors. To extend our published results, we performed energy-efficient event-based optical flow prediction using our proposed methods on the neuromorphic processor. The extension study shows that sparsely activated artificial neural networks can achieve the same level of efficiency as spiking neural networks.

Keywords: Energy Efficiency · Neuromorphic · Neural Networks

1 Event-driven Depth-first Convolution

Compared to standard convolutional neural networks, event-driven convolution in neuromorphic computing processes sparse events from the previous layer one by one in their order of arrival and accumulates them incrementally, directly into the neural states of the corresponding fanned-out postsynaptic neurons.

However, this process requires maintaining high-dimensional neural states of convolutional layers in memory, which is impractical for the limited size of the on-chip memory, if the output tensor has a high dimension. To overcome this challenge, we propose the event-driven depth-first convolution.

The depth-first inference [7, 4] is a scheduling method in neural network inference that prioritizes the network's layer (depth) dimension by consuming activations right after their generation. We present event-driven depth-first convolution in our published paper [8]. The input events within a time step are assumed to be sorted in spatial order from the top-left corner of the (X, Y) plane to the bottom-right corner. Under this assumption, a neuron will receive all of its input events in a pre-defined order. Accordingly, its neural state updates will be concluded earlier than those of spatially lower-ranked neurons. As a result, it can fire immediately after its last neuron state update without needing to wait to process all the input events. After the event-generation process of a neuron, the memory for its neuron state can be released. Therefore, each layer only needs to buffer a small portion of neural states that are incomplete/partially summed (the amount of required memory increases with the kernel size).

To characterize and quantify the improvements, we carry out experiments in two classification tasks: gesture recognition [1] and handwritten digit classification [2], and one high-resolution object detection task [5] using the energy-efficient event-based camera. Compared with other state-of-the-art large-scale neuromorphic processors, our proposed optimizations result in a 6× to 300× improvement in energy efficiency, a 3× to 15× improvement in latency, and a 3× to 100× improvement in area efficiency.

2 Sparse Event-based Optical Flow Prediction

Estimating optical flow using an event camera is robust to motion blur and varying illumination thanks to the event stream that captures pixel brightness changes asynchronously in high dynamic ranges. Spiking neural networks (SNNs) for event-based optical flow are claimed to be computationally more efficient than their deep artificial neural networks (ANNs) counterparts [3, 6], but a fair comparison is missing in the literature.

To extend our published methods in [8], we propose an event-based optical flow solution based on activation sparsification on the SENECA neuromorphic processor using event-driven depth-first convolution. Therefore, the implementation can exploit the sparsity in ANN activations and SNN spikes to accelerate the inference of both types of neural networks. The ANN and the SNN for comparison have similar low activation/spike density (5%) thanks to our novel sparsification-aware training on a modified FireNet architecture [3]. In the hardware-in-loop experiments designed to deduce the average time and energy consumption, the SNN consumes $0.9mJ$ and the ANN consumes $1.2mJ$ per event frame prediction on average. As a result, ANN can achieve the same level of efficiency as SNN on the neuromorphic processor while maintaining state-of-art prediction accuracy.

References

1. Ceolini, E., Frenkel, C., Shrestha, S.B., Taverni, G., Khacef, L., Payvand, M., Donati, E.: Hand-gesture recognition based on emg and event-based camera sensor fusion: A benchmark in neuromorphic computing. Frontiers in neuroscience **14**, 637 (2020)
2. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
3. Hagenaars, J., Paredes-Vallés, F., De Croon, G.: Self-supervised learning of event-based optical flow with spiking neural networks. Advances in Neural Information Processing Systems **34**, 7167–7179 (2021)
4. Mei, L., Goetschalckx, K., Symons, A., Verhelst, M.: Defines: Enabling fast exploration of the depth-first scheduling space for dnn accelerators through analytical modeling. In: 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA). pp. 570–583. IEEE (2023)
5. Perot, E., De Tournemire, P., Nitti, D., Masci, J., Sironi, A.: Learning to detect objects with a 1 megapixel event camera. Advances in Neural Information Processing Systems **33**, 16639–16652 (2020)
6. Schnider, Y., Woźniak, S., Gehrig, M., Lecomte, J., Von Arnim, A., Benini, L., Scaramuzza, D., Pantazi, A.: Neuromorphic optical flow and real-time implementation with event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4129–4138 (2023)
7. Waeijen, L., Sioutas, S., Peemen, M., Lindwer, M., Corporaal, H.: Convfusion: A model for layer fusion in convolutional neural networks. IEEE Access **9**, 168245–168267 (2021)
8. Xu, Y., Shidqi, K., van Schaik, G.J., Bilgic, R., Dobrita, A., Wang, S., Meijer, R., Nembhani, P., Arjmand, C., Martinello, P., et al.: Optimizing event-based neural networks on digital neuromorphic architecture: a comprehensive design space exploration. Frontiers in Neuroscience **18**, 1335422 (2024)