Live Demonstration: Convolutional neural network driven by dynamic vision sensor playing RoShamBo

Iulia-Alexandra Lungu, Federico Corradi, Tobi Delbruck
Institute of Neuroinformatics, University of Zurich and ETH Zurich, Zurich, Switzerland

***Abstract*—This demonstration presents a convolutional neural network (CNN) playing "RoShamBo" ("rock-paper-scissors") against human opponents in real time. The network is driven by dynamic and active-pixel vision sensor (DAVIS) events, acquired by accumulating events into fixed event-number frames.**

Associated tracks: 8.4: Neuromorphic and Event-Based Systems

I. DEMONSTRATION

CNNs have become widely used for visual detection and recognition tasks in recent years. Such networks are usually driven by images captured using conventional frame-based cameras. However, frame-based approaches are generally power-hungry at high frame rate. The DAVIS is a neuromorphic camera that outputs static pixel sensor (APS) image frames (not used here) along with dynamic vision sensor (DVS) temporal contrast events [1]. The DVS address-events report brightness (log intensity) changes in the scene and have a sub-millisecond latency. The advantage of using such a sensor over conventional cameras is its low power consumption when there are no changes in the scene, as well as its rapid output.

Here we followed the approach of [2], by using the DVS sensor to drive a CNN playing the game of RoShamBo against a human. CNN input images are generated at a variable, data-driven rate between 1-200 Hz by accumulating asynchronous DVS address-events into 64x64 pixel 2D histograms of a constant total number of events. By using a conventional 5-layer CNN, our approach makes use of existing knowledge in the field of deep learning and is compatible with current accelerator technologies.

Training data consisted of labeled continuous DAVIS240C (inilabs.com) recordings of a person showing a single symbol, obtained using jAER [3], a software developed to process DAVIS data. The resulting recordings were converted to AVI movies, cut into frames and compiled in an LMDB database. 15 participants were each recorded for (3 symbols) x (2 hands) x (2 minutes), amounting to around 1.3 million images for each of 4 classification categories: rock, paper, scissors and background. Input normalization comprises rectifying all DVS events to positive 'on' events with a 200-event maximum grayscale bin value and mapping the image pixel values to a 0-1 range by performing a 3-sigma normalization. These methods are applied during training as well as at inference time. Two data augmentation mechanisms were used. Firstly, all the original recordings were sampled using 4 different numbers of accumulated events: 0.5k, 1k, 2k and 4k. Secondly, the resulting images were randomly mirrored. The final training dataset, consisting of 5M images, is available on request.

The final CNN architecture consists of 5 convolutional layers, each using the ReLU activation function and followed by 2x2 max pooling. Kernels are square with dimension 5, 3, 3, 3, 1 respectively. The network has 114k parameters, needing 18 MOp to classify one image. 40 training epochs were performed in Caffe [4], requiring about 5h on an Nvidia GTX980 GPU. Robust accuracy was achieved by collecting sufficient data and systematic architecture exploration. Accuracy of 99.3% was achieved for the 10% test set.

II. DEMONSTRATION SETUP AND READINESS

This demo has been shown internally, at NIPS 2016, and in several seminars and is convincing. The setup (Fig. 1) uses a DAVIS240C camera and an Arduino board connected to a laptop. CNN inference in jAER requires about 70ms on a laptop. The Arduino board receives the CNN output and drives a robotic hand and an LED box.

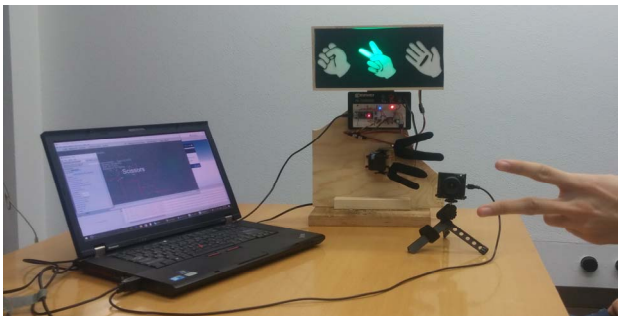

Fig. 1. RoShamBo demonstration setup.

III. VISITOR EXPERIENCE

Visitors flash one of the three symbols traditionally played in the RoShamBo game: rock, paper or scissors. The LED box lights up the symbol which is currently being played, while the robotic hand demonstrates the winning, opposite action. The reaction is fast enough that visitors believe they are being beaten. Visitors learn advantages of combining event sensors with CNNs.

References

[1] P. Lichtsteiner, C. Posch, and T. Delbrück, "A 128 x 128 120dB 15us Latency Asynchronous Temporal Contrast Vision Sensor," *IEEE J Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[2] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbruck, "Steering a Predator Robot using a Mixed Frame/Event-Driven Convolutional Neural Network," in *2016 IEEE Conf. on Event Based Control Communication and Signal Processing (EBCCSP 2016)*, Krakow, Poland, 2016, p. in press [Online]. Available: https://www.researchgate.net/publication/303614947_Steering_a_Predator_Robot_using_a_Mixed_FrameEvent-Driven_Convolutional_Neural_Network

[3] "jAER Open Source Project," *jAER Open Source Project*, 23-Mar-2007. [Online]. Available: http://jaerproject.org. [Accessed: 23-May-2016]

[4] "Caffe | Deep Learning Framework." [Online]. Available: http://caffe.berkeleyvision.org/. [Accessed: 01-Feb-2016]

978-1-4673-6853-7/17/$31.00 ©2017 IEEE