Contribution ID: **28** Type: **Parallel session talk**

SparsePixels: Efficient Convolution for Sparse Data on FPGAs

Thursday 9 October 2025 15:00 (20 minutes)

Inference of standard convolutional neural networks (CNNs) on FPGAs often incurs high latency and long initiation intervals due to the nested loops required to slide filters across the full input, especially when the input dimensions are large. However, in some datasets, meaningful signals may occupy only a small fraction of the input, say sometimes just a few percent of the total pixels or even less. In such cases, most computations are wasted on regions containing no useful information. In this work, we introduce SparsePixels, a framework for efficient convolution for sparsely populated input data on FPGAs operating under tight resource and low-latency constraints. Our approach implements a special class of CNNs where only active pixels (non-zero or above a threshold) are retained and processed at runtime, while the inactive ones are discarded and ignored. We show that our framework can achieve performance comparable to standard CNNs in some target datasets while significantly reducing both latency and resource usage on FPGAs. This also demonstrates its potential for efficient readout in next-generation detectors, where inputs are massive but signals could be sparse. Custom kernels for training and the HLS implementation are developed to support sparse convolution operations.

Authors: TSOI, Ho Fung (University of Pennsylvania (US)); RANKIN, Dylan Sheldon (University of Pennsylvania (US)); LONCAR, Vladimir (CERN); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: TSOI, Ho Fung (University of Pennsylvania (US))

Session Classification: SHARED SESSION

Track Classification: RDC 5 Trigger & DAQ