2D Convolutional Neural Network on FPGA for High-Resolution TPC Images

Lukas Arnold*, Luca Carloni, Giuseppe Di Guglielmo, Yeon-jae Jwa+ and Georgia Karagiorgi

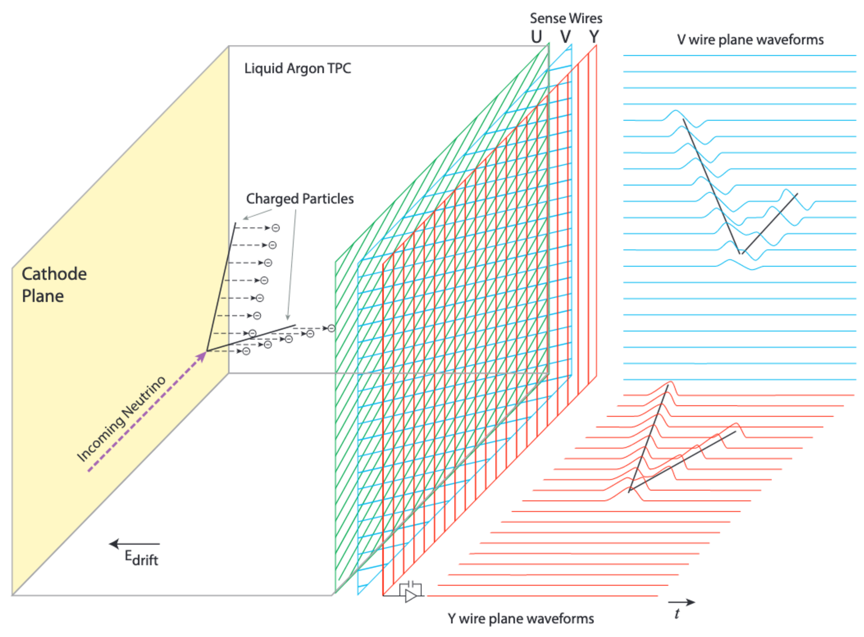
Nevis Laboratories, Department of Physics & Department of Computer Science, Columbia University

This material is based upon work supported in part by the National Science Foundation.

NEVIS LABORATORIES COLUMBIA UNIVERSITY

Intoduction: TPCs

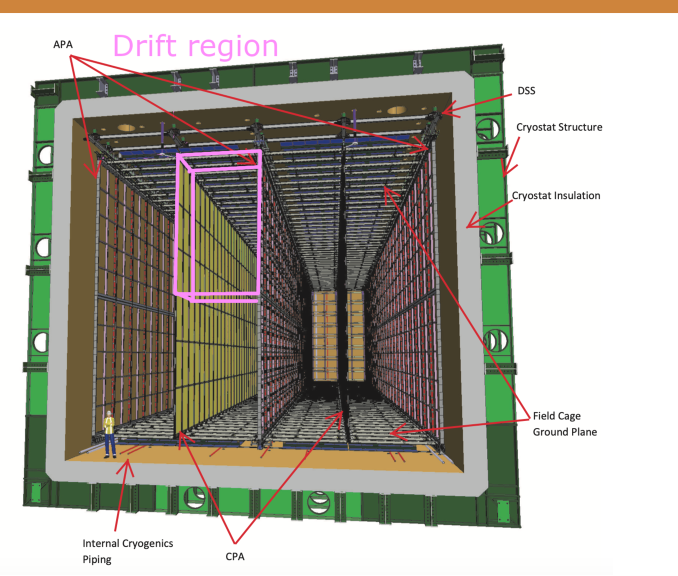
- Liquid Argon Time Projection Chambers (LArTPCs) are a type of particle detector, significant for neutrino physics.
- Working principle for neutrino detection (see *figure* [1]):
  - ▷ Basic structure of LArTPC consist of an anode and cathode plane, generating an electrical field within an area filled with liquid argon
  - ▷ Neutrino interacts with liquid argon, creating charged particles
  - ▷ Due to electrical field, ionization charge trail left behind by the charged particle drifts towards anode plane
  - ▷ Differently oriented wires catch induced signals from charged particles
  - ▷ Using these and drifttime, tracks of charged particles and neutrino can be recontructed



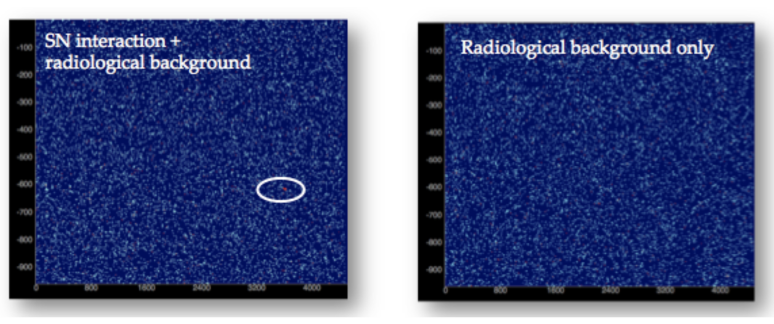
- Current experiments:
  - ▷ ArgoNeuT, MicroBooNE, ProtoDUNE, ICARUS
- Experiments being built or proposed:
  - ▷ **DUNE Far Detector** (to be built in South Dakota), SBND, GRAMS

DUNE Far Detector

- Four modules (horizontal drift and vertical drift)
- up to **200** cells (drift region) per module (see *figure* [2])
- **150** APAs (Anode Plane Assemblies)
- Each APA reads **2560** channels, including **2** × **480** collection channels



Supernova detection at DUNE



- Low-energy events, such as Supernova bursts or proton decay events, are hard to distinguish from noise
  - ▷ practically 100% efficiency needed
  - ▷ maximum 1 False Positive for Supernova burst per month
  - ▷ overall data reduction factor: $\mathbf{10^4}$

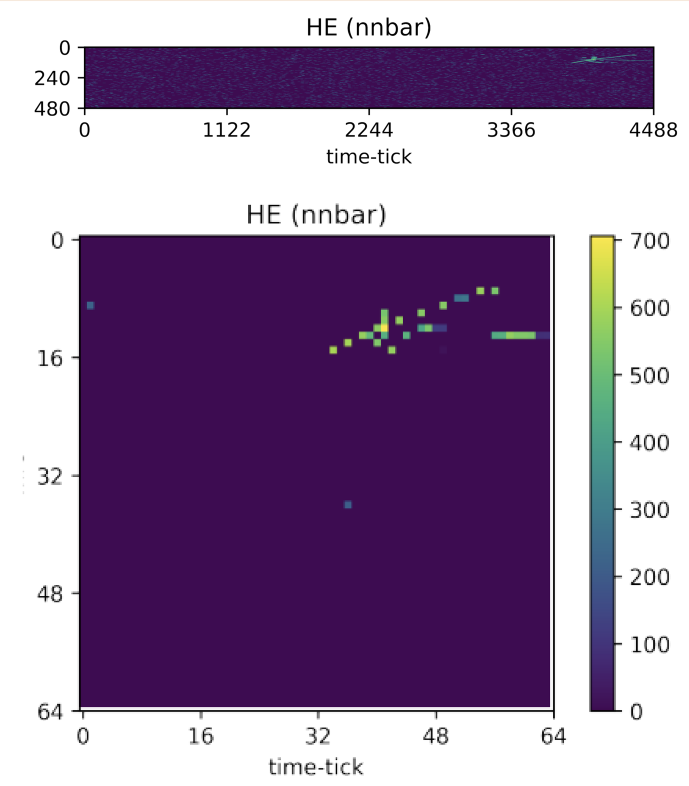
→ deployment of sophisticated data selection algorithm, such as Convolutional Neural Network (CNN). **Can we use CNN for data selection?**

Front-end & Data Acquisition

- DUNE input data
  - ▷ Data is read out continuously, and can be arranged into frames
  - ▷ Data is buffered ($\approx \mathbf{10}$s) and selected data is sent for further processing
  - ▷ 480 channels per collection plane × $\mathbf{2.25}$ms drift length / $\mathbf{2}$MHz sample frequency $=$ $\mathbf{4488}$samples (at $\mathbf{12}$bit ADC resolution)
  - → $\mathbf{11.5}$Gbps per collection plane (one side of APA)
- Functionality of FPGA board: receive data, generate *trigger primitives*, **select data**, route data to CPU host
- target FPGA board for DUNE:
  - ▷ ~75 **FELIX** boards [3] per module
  - ▷ Xilinx Kintex Ultrascale FPGA XCKU115 (1.3M FF, 5.5k DSP)
- current demonstration board:
  - ▷ **Alveo U250** acceleration card
  - ▷ Xilinx Ultrascale FPGA XCU250 (approx. double the FPGA resources of FELIX)

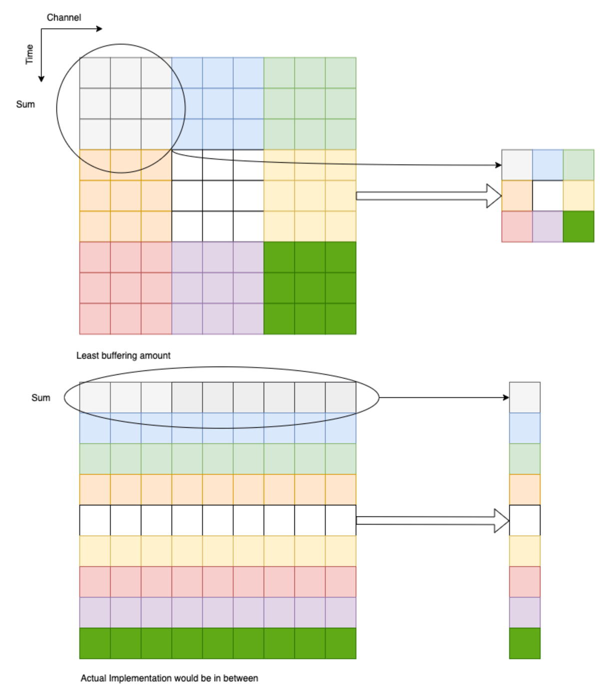
Data flow on FPGA

- Input image: **480** × **4488** (example on right with high-energy event [4])
  - ▷ Throughput: **200** images each $\mathbf{2.25}$ms
- Preprocessing: Downsizing to **64** × **64** pixel image
  - ▷ Denoising (zero suppression)
  - ▷ Different algorithms evaluated (see below)
- CNN
  - ▷ Classification of image into classes {background (NB), high-energy event (HE), low-energy event (LB)}



Preprocessing

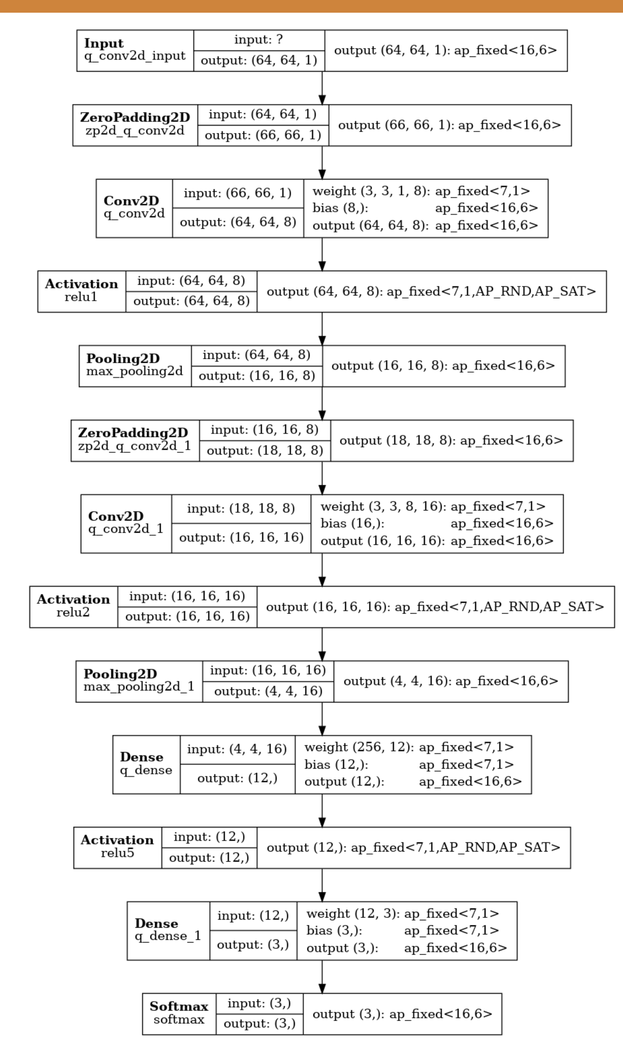
- Different algorithms considered:
  - ▷ Denoising + ROI identification + Cropping ([4], $\geq$ $\mathbf{500}$ms)
  - ▷ Denoising + Cropping around maximum (significant information loss, $\leq$ $\mathbf{1}$ms)
  - ▷ **Denoising + Summing** (chosen, $\mathbf{20 \dots 60}$ms), 5% LUT usage
- Figure: *Summing across multiple samples (upper image, requires large buffer space), and across channels only (lower image).*



Convolutional Neural Network

- Evaluation and emulation using HLS4ML [5]
- Application of fixed-point quantization
- Very good accuracy results, no significant drop for quantization (see table and setup image)

Original	NB	LE	HE
true_NB	99.6%	0.4%	0%
true_LE	3.8%	94.0%	2.2%
true_HE	3.2%	5.4%	91.4%
total accuracy		95.2%	
Quantized	NB	LE	HE
true_NB	99.7%	0.3%	0%
true_LE	3.9%	94.7%	1.4%
true_HE	3.3%	6.4%	90.4%
total accuracy		95.2%	



Implentation results

- Implementation of CNN (without preprocessing) onto Alveo FPGA made in collaboration with HLS4ML group [6].
- CNN Latency: $\mathbf{27.7}\mu$s (emulation: $\mathbf{23.4}\mu$s)

	Block RAM	**DSP Units**	**Flip Flops**	**Look-up tables**
Kernel only	337kbit	2106	142k	139k
Total	391kbit	2106	148k	141k
Available	76Mbit	5.5k	1.3M	600k

Conclusion

- CNN evaluated using *HLS4ML*
- Preprocessing implemented & tested on Alveo
- CNN implemented on Alveo (testing ongoing)
- Next steps
  - ▷ Test bench for implemented version with data from DUNE prototype
  - ▷ Pre-processing and CNN exist as separate Alveo designs → integration
  - ▷ (Long-term) Transfer implemention from Alveo to FELIX
- **CNN for DUNE Data Selection seems to be a viable option**

References: [1] MicroBooNE Collaboration, Design and Construction of the MicroBooNE Detector , JINST 12 (2017) 02, [2] DUNE Collaboration, Far Detector Technical Design Report Vol. III, JINST 15 (2020) 08, [3] G. Karagiorgi, pres., Liquid Argon TPC Trigger Development with SBND, DPF 2019, [4] Y. Jwa, G. Di Guglielmo, L. Carloni, G. Karagiorgi, in: NYSDS 2019, [5] Y. Jwa, G. Di Guglielmo, L. Arnold, L. Carloni, G. Karagiorgi, Accelerating Deep Neural Networks for Real-time Data Selection for High-resolution Imaging Particle, arXiv 2201.05638, [6] J. Duarte, P. Harris, S. Hauck et al, Fast inference of deep neural networks in FPGAs for particle physics, JINST 13 (2018) 07.

*lukas.arnold@nevis.columbia.edu +yj2429@nevis.columbia.edu