Deep Learning with Micron FPGAs on protoDune

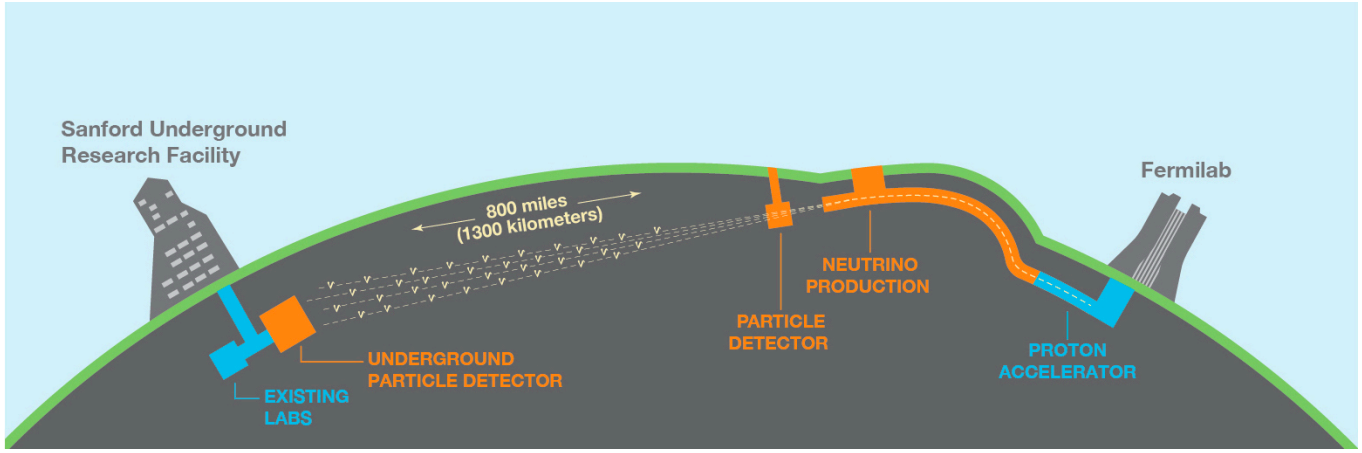
Miroslav Kubu

2022 CERN openlab Technical Workshop

21.03. 2022

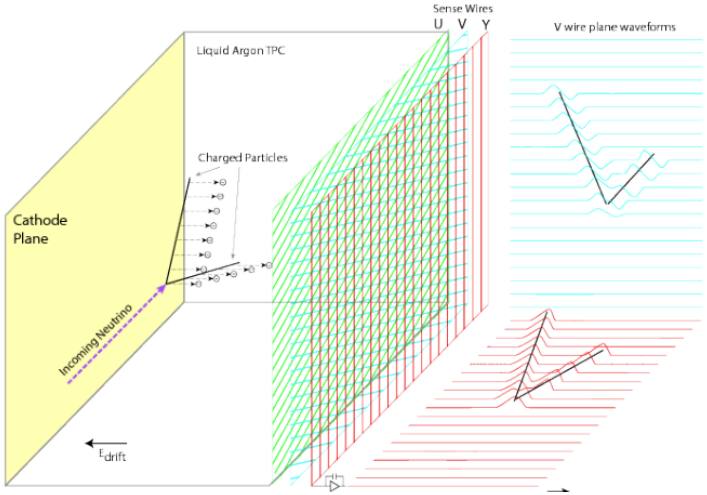
DUNE Experiment

- Deep Underground Neutrino Experiment
- Leading-edge experiment for long-baseline neutrino oscillation studies, neutrino astrophysics and proton decay searches
- Far Detector (FD) is 800 miles from the neutrino beam source
  - Four modules, each with 10,000 ton of liquid argon
- High power muon neutrino beam produced at Fermilab
  - Can switch polarity to produce a muon antineutrino beam.



DUNE Experiment

- ProtoDUNE is the DUNE Far Detector prototype at the CERN Neutrino Platform
  - Consists of 6 APA (Anode Plane Assembly) units, while DUNE Far Detector will consist of 150 APAs
  - 1:1 scaled design components with same technologies
  - Data collected with 2 collection and 1 induction planes
  - Could be reconstructed as 3x 2D image



Real-time data processing

DUNE Data collection in numbers (one single-phase module):

- Sampling rate: 2 MHz
- Trigger window: 3 ms
- Number of pixels to process during the time window: 15,360,000 (2560x6000) per APA
  - 2 induction planes with 800 channels
  - 1 collection plane with 960 channels

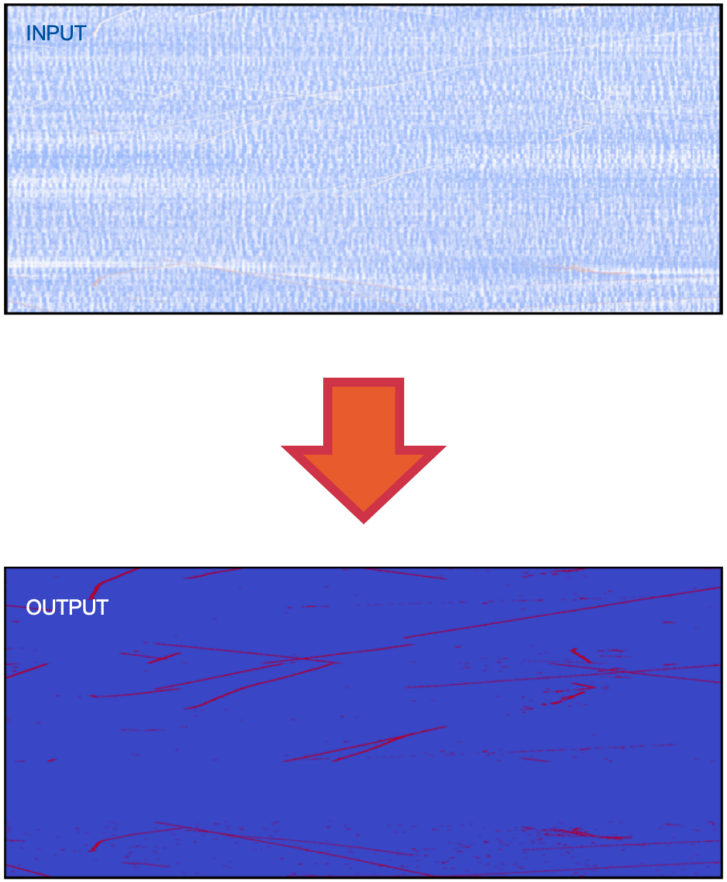
Our goal:

- Reduce the amount of data stored for offline reconstruction
  - Locate where the hits are
- DUNE data could be processed by Deep Learning methods
- Deep Learning accelerators such as Micron FPGAs might enable us to filter the data online using a suitable Deep Learning model

Real-time data processing

Possible approach:

- Perform semantic segmentation for regions of interest (ROI) using the raw inputs
- Target is to find a mask for ROI in the raw data

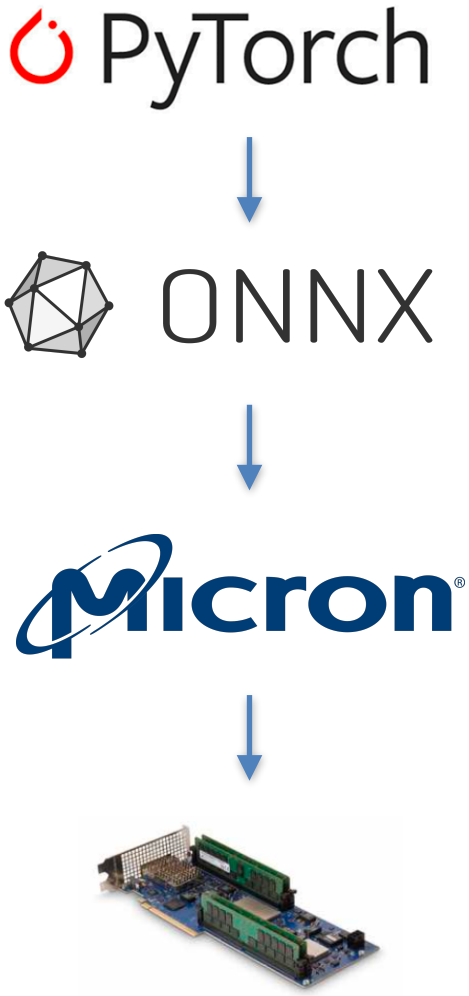


Micron framework

- Direct deployment of neural networks on the inference engine
- Supports majority of the layers used in computer vision
- Any framework that supports export to ONNX could be used

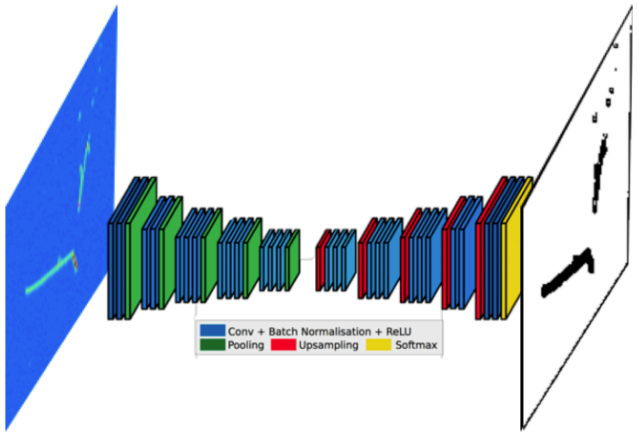
Workflow:

1. Train the network
2. Convert it into ONNX
3. Compile it using the Micron framework
4. Deploy it using the inference engine

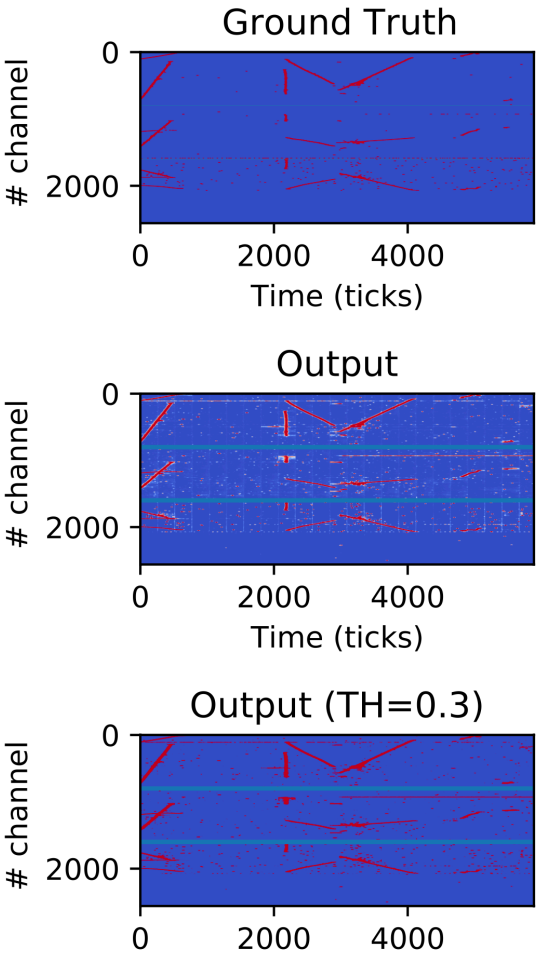


Approach: Linknet

- Pixel-wise semantic segmentation for visual scene understanding
- Promising results in terms of accuracy
- We managed to use a minimalistic version of Linknet on a Micron FPGA



LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation (https://arxiv.org/abs/1707.03718)

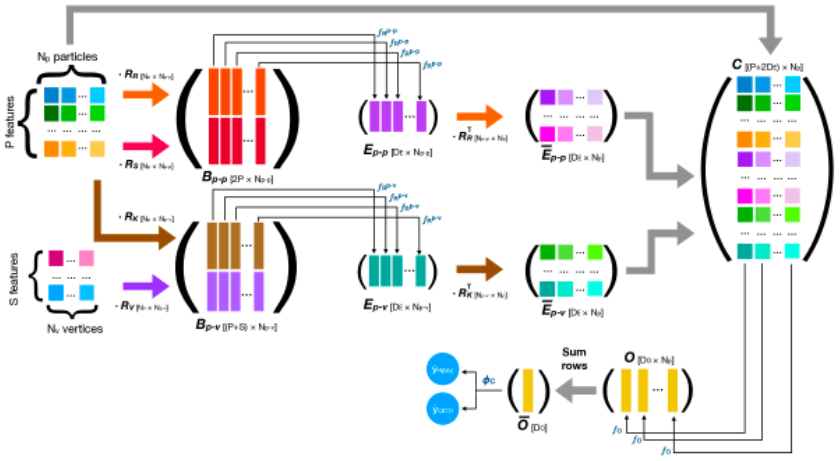


Approach: Linknet

- We aim to find these ROIs in real time
  - To analyze a whole trigger window of 3 ms, we need to run the inference over 15,360,000 pixels (2560 channels x 6000 ms)
- After reducing the network to a minimalistic working version, the best reported results are ~100 ms
- We have large volume of sparse data
  - We need to think about an approach that wouldn't need to process full event input
  - Graph Neural Networks (GNNs) are often used for similar cases

Approach: Interaction network

- Possibilities of compiling Graph Networks with advanced layers on FPGAs are limited
- Interaction network
  - Alternative to common Graph Convolutional layers
  - Series of MLP and matrix multiplications using receiving and sending matrices describing the connections between hits in the image



Interaction networks for the identification of boosted H→bb decays (https://arxiv.org/abs/1909.12285)

- Designed for classification
  - We created a test dataset for shower and track classification
  - Next step would be to find a way how to use the network for semantic segmentation

Approach: Interaction network

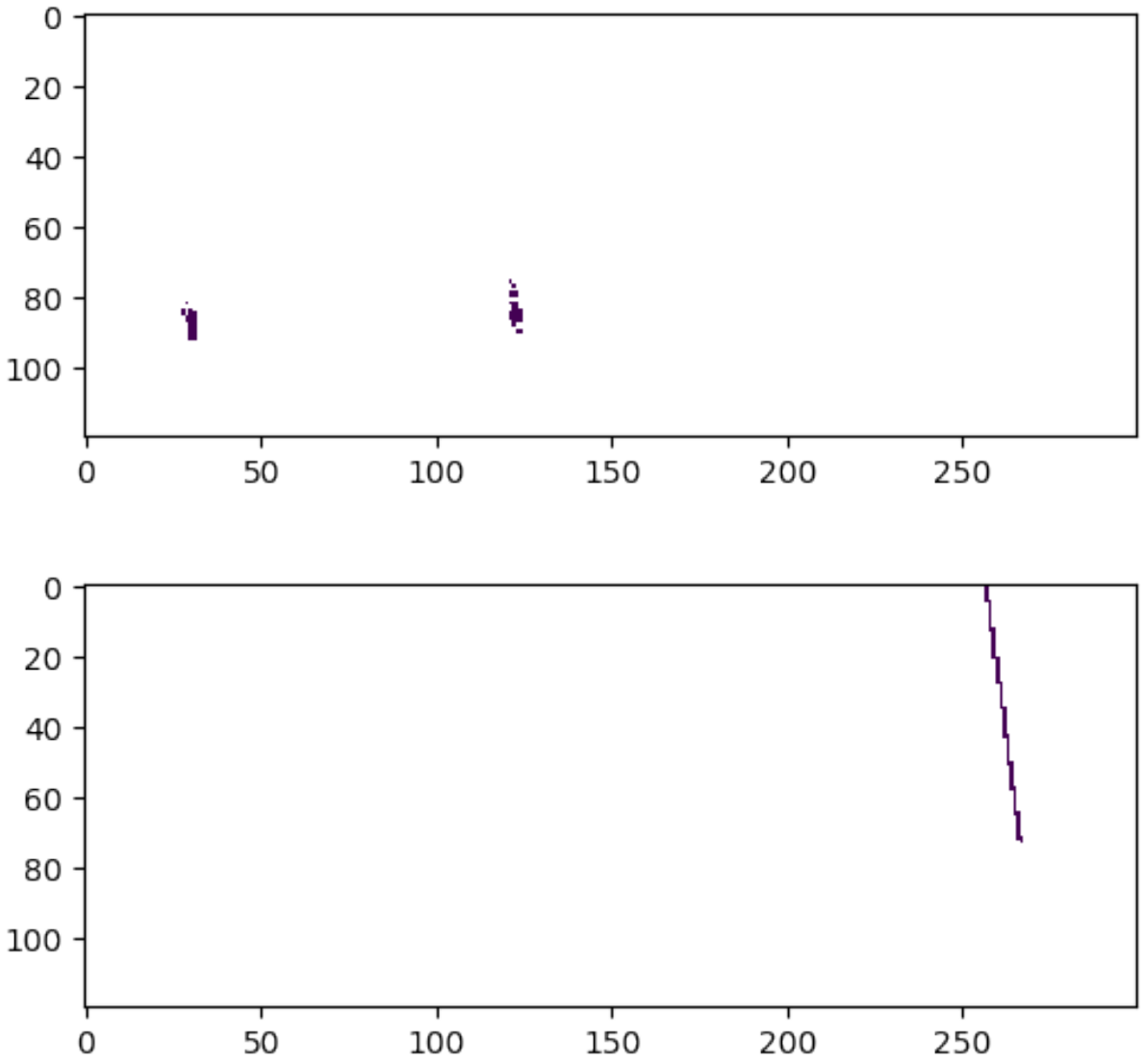
- Graph analogies:
  - Nodes are represented by positions of the hits (pixels)
  - Nodes are connected only when the pixels are close to each other
- Matrices defining the connections between the nodes are too large for the FPGAs (size N x N(N-1))
- After restricting connections on 2 closest pixels, we need to limit to N ~ 200 hits
- Good accuracy:
  - 0.85 ACC for shower
  - 0.95 ACC for track
- Bottleneck: With 200 hits restriction, we can use only a small portion of the event

Summary

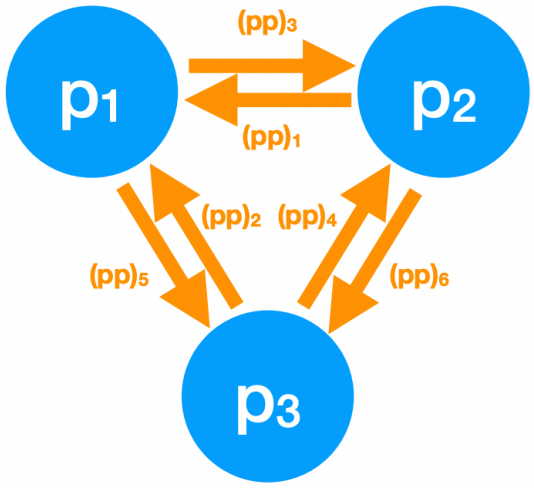
- We are testing fast inference solutions for Deep Learning using the Micron FPGAs
- We aim to use the FPGAs for data filtering on protoDUNE
- The tested semantic segmentation Linknet model works offline, but we need to increase the inference speed for the online filtering
- We tested a graph-inspired Interaction network, but we were not able to use it on a full-scale event
- Plan: we are looking for a way to speed up our Linknet solution to make it fast enough for online filtering

Thank you

Track X Shower binary classification using the event coordinates



Interaction network receiving and sending matrices



$$R_R = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{array}{c} \begin{array}{cccccc} (pp)_1 & (pp)_2 & (pp)_3 & (pp)_4 & (pp)_5 & (pp)_6 \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array},$$

$$R_S = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \end{array} \begin{array}{c} \begin{array}{cccccc} (pp)_1 & (pp)_2 & (pp)_3 & (pp)_4 & (pp)_5 & (pp)_6 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array},$$

Interaction networks for the identification of boosted H→bb decays
(https://arxiv.org/abs/1909.12285)