Deep Learning Models for Particle Identification and Energy Regression in CMS HGCAL L1 Trigger

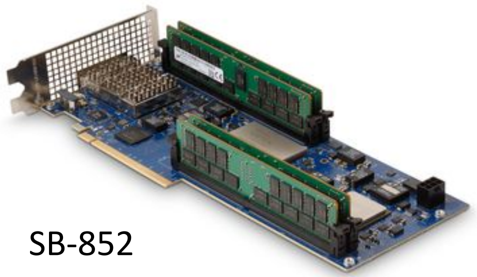
*Using **Micron's** FPGA based inference engines and*
FWDNXT firmware + software for compiling models and running the inference

Anwesha Bhattacharya, Emilio Meschi, Thomas James, Dejan Golubovic

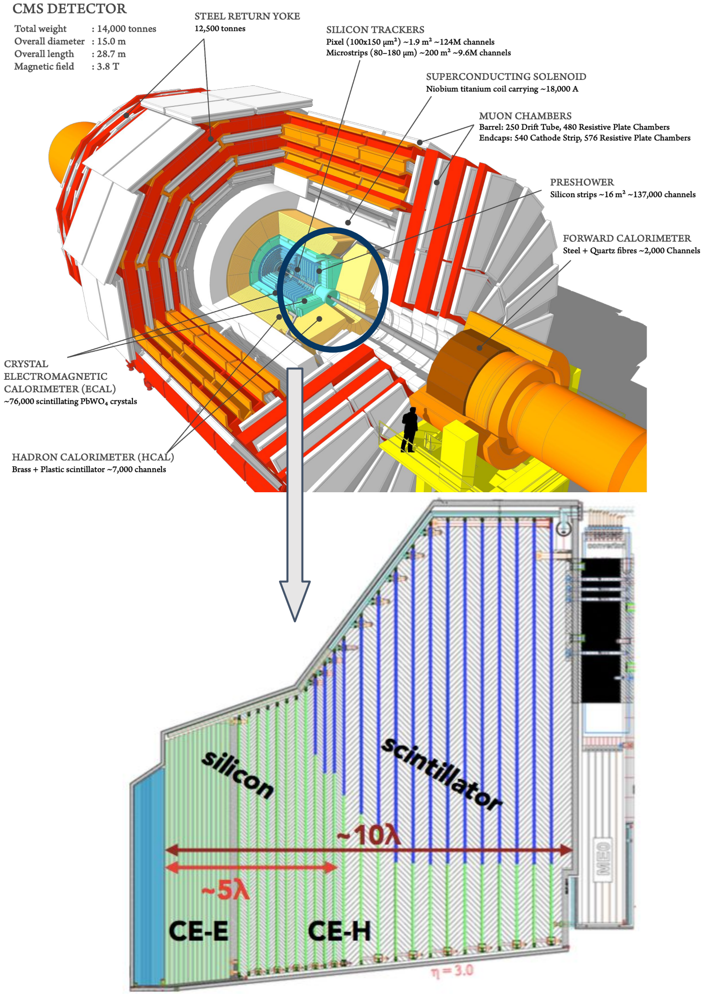
13th August, 2019

Introduction

- Classify **electrons** and **photons** and estimate the energy of particles for the L1 trigger
    - Unsupervised learning for clustering
    - Supervised learning for classification and regression
    - Implement the models in software and Micron hardware

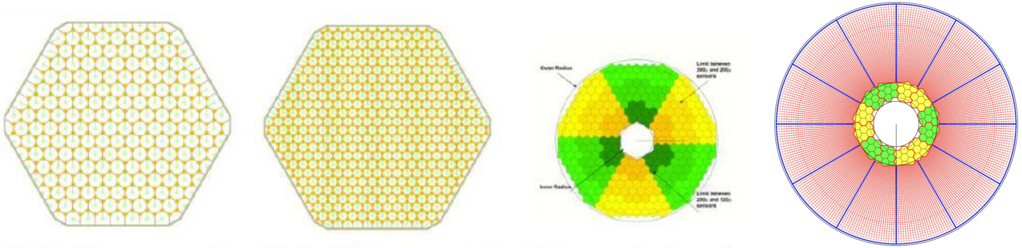


SB-852

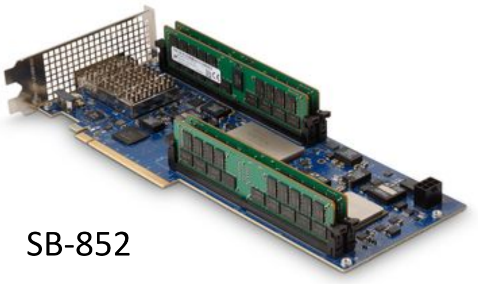


Introduction

- Classify **electrons** and **photons** and estimate the energy of particles for the L1 trigger
  - Unsupervised learning for clustering
  - Supervised learning for classification and regression
  - Implement the models in software and Micron hardware



(a) Big Silicon sensors, 1.18 cm^2 (b) Small Silicon sensors, 0.52 cm^2 (c) Layout of a layer where only silicon sensors are present (d) Layout of wafers and Scintillator tiles in a layer where both are present: the 22nd layer of CE-H

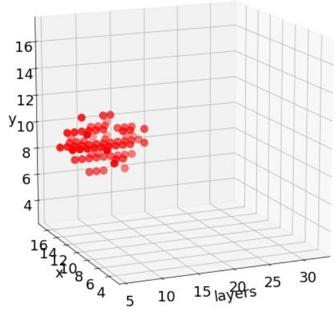


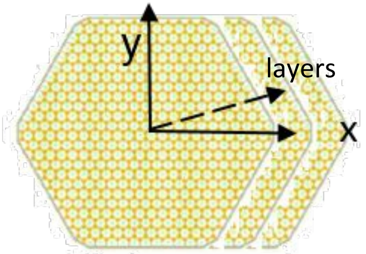
SB-852

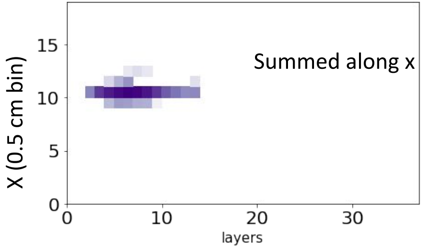
- DL implementations
  - Python -> keras/tensorflow
- Hardware & Software connecting DL models and hardware- by Micron
  - SB-852 FPGA board
  - Drivers for hardware
  - Compiler to translate DL models to FPGA bitstream - proprietary
  - Python wrapper around compiler - open source

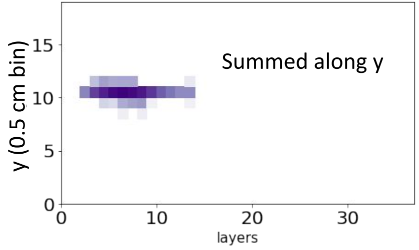
Dataset

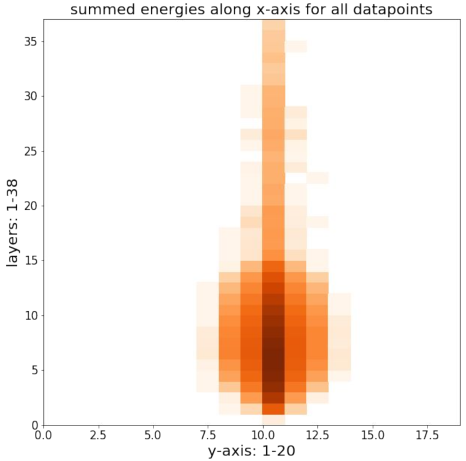
- Simulated single particle events with no pileup. Interpreted as 3D images of energy deposits in the calorimeter.
- 100k photons, e^+ and e^- generated from the interaction region. 40% photons decay to e^+-e^- pairs.
- Constrained topological clustering is used to extract the clusters
- CNN model trained on this dataset for classification and regression



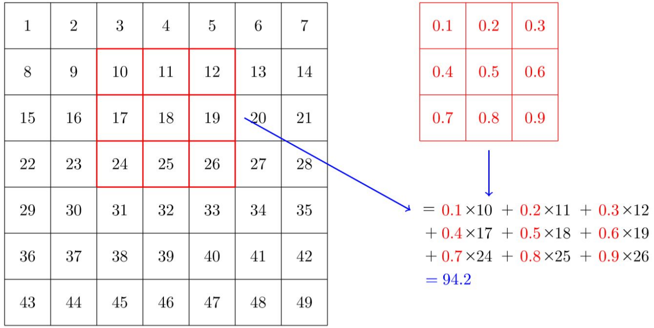




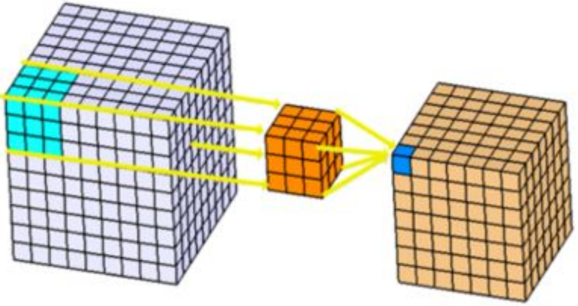




DL Model Architecture

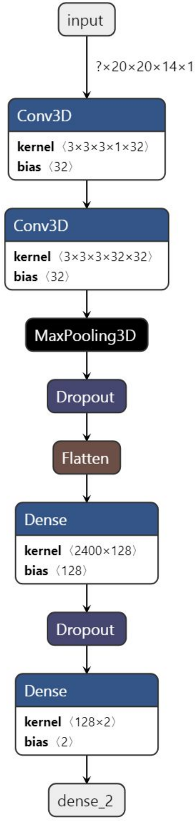


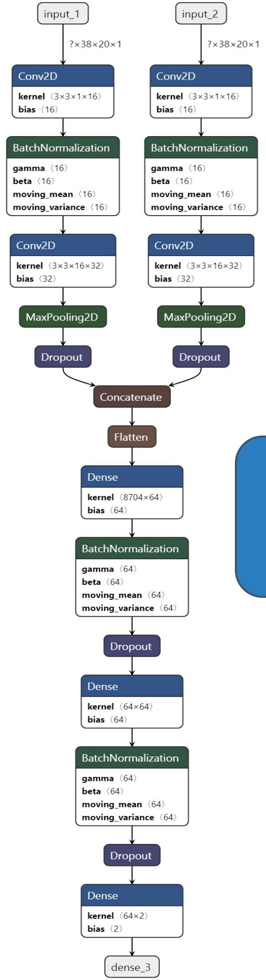
2D CNN



3D CNN

Single input 3D CNN





Multi-input 2D CNN

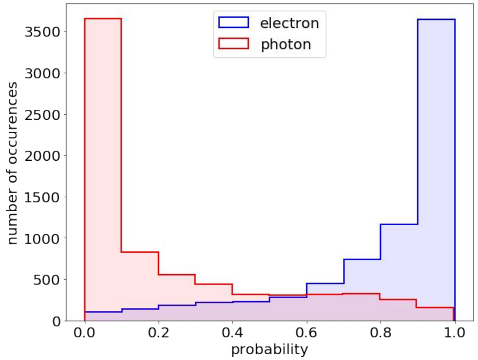
Training -> 860us/image 55us/image

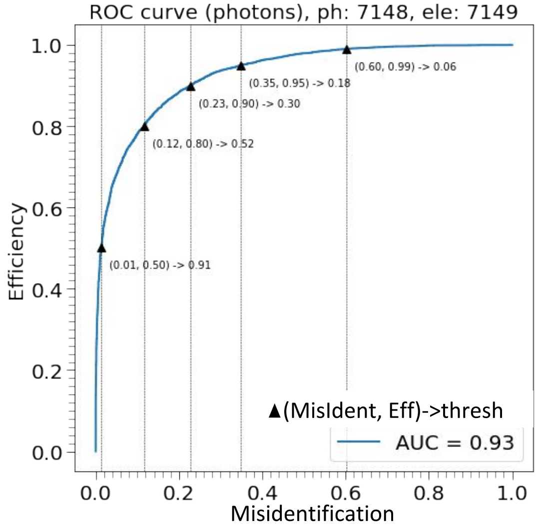
CERN openlab

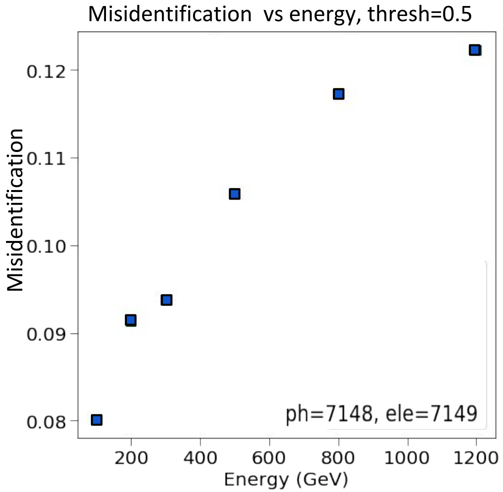
Classification Model

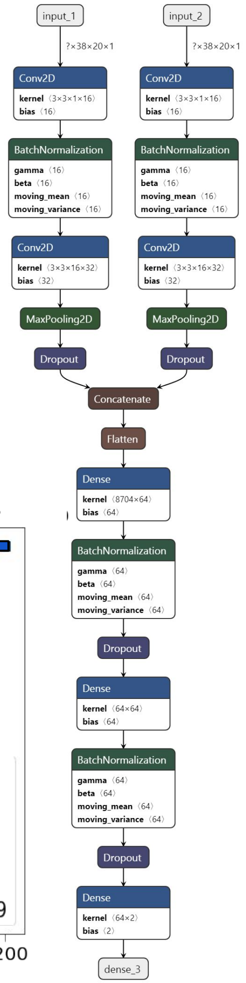
2D 2-input CNN model that takes into account both x and y channels.

	Pred Photons	Pred e^+/e^-
True photons	0.82	0.18
True e^+/e^-	0.13	0.87





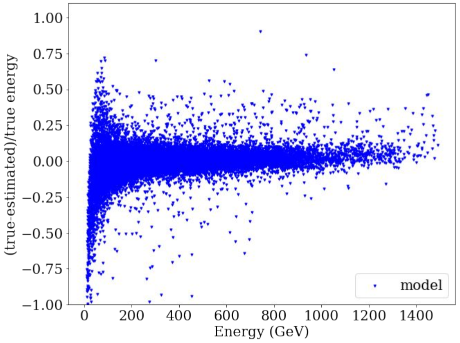


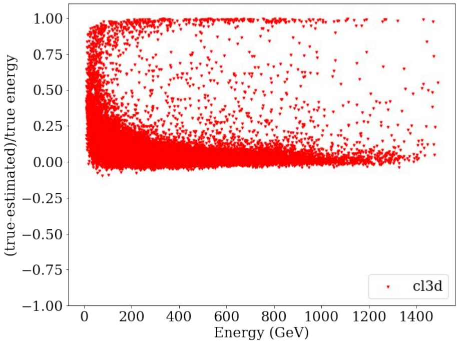


CERN openlab

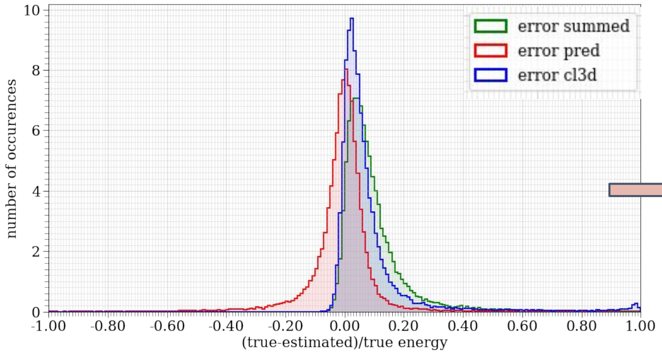
Regression Model

Energy error from cl3d* and model pred vs true energy

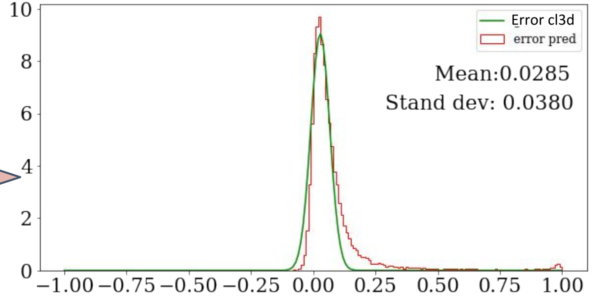




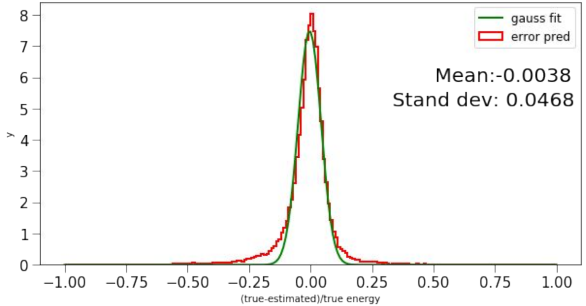
Relative error on energy estimate from the model compared to cl3d* and summed trigger cells



Gaussian fit for cl3d



Gaussian fit for the model



- 46k - train samples
- 20k - validation
- 28k - test
- 95k- total

cl3d = Classical density-based clustering of each detector layer and then sum

Summary

- Trained classification and regression models for CMS HGCAL L1 trigger.
- Both 2D 2-input and Conv3D models compile and run on the micron board using FWDNXT SDK.
- Promising results for particle identification. Energy regression needs further tuning.
- A test comparison shows good results when using single inputs. Multiple input mode gives results inconsistent with the python implementation- to be followed up with Micron/FWDNXT.

Next Steps

- Compare performance of Conv3D and 2D 2-input models for the board
- Compare classification results with BDT results from HGCAL backend group.
- Optimize and tune the model for the board.
- Merge regression and classification model into one.
- Include pile-up events.
- Explore regression models for cluster position.

Useful Links

- Micron hardware
  - https://www.micron.com/products/advanced-solutions/advanced-computing-solutions
- FWDNXT SDK
  - https://github.com/FWDNXT/SDK
- The Phase-2 Upgrade of the CMS L1 Trigger Interim TDR
  - https://cds.cern.ch/record/2283192?ln=en
- The Phase-2 Upgrade of the CMS Endcap Calorimeter
  - https://cds.cern.ch/record/2293646

Thank you

QUESTIONS?

96anwesha@gmail.com

Specifications - sb852

- Xilinx Virtex Ultrascale+ UV7P or UV9P FPGA
- 2GB Hybrid Memory Cube
- Two full-width (x16) links with 15 Gb/s transceivers
- Up to 120 GB/s HMC bandwidth
- Up to 30 GB/s (RX and TX combined) via each full-width (x16) link
- 64GB DDR4 SODIMM (standard configuration); upgradeable to 512GB of high-performance memory
- 2 QSFP transceiver connectors
- PCIe x16 Gen3 to the host
- SDAccel (OpenCL™) support
- Easy design framework with simple FPGA bitstream loading from host
- Complete suite of analytics tools

Confusion matrix for 3D CNN

	Pred photons	Pred e^+/e^-
True photons	0.95	0.05
True e^+/e^-	0.31	0.69