Muon Trigger using Deep Neural Networks accelerated by FPGAs

Seulgi Kim, Jason Lee, Inkyu Park,
Youngwan Son, Ian James Watson, Seungjin Yang

서울시립대학교
UNIVERSITY OF SEOUL

ICHEP 2020 - 29th Jul 2020

Introduction

❑ Motivation

- Trigger decision can be treated as image segmentation problem in Computer Vision, and Convolutional Neural Network(CNN) is a powerful tool in segmentation problem.

- Our plan is to increase trigger accuracy with CNN based segmentation, while reducing latency by FPGA to enable online use.

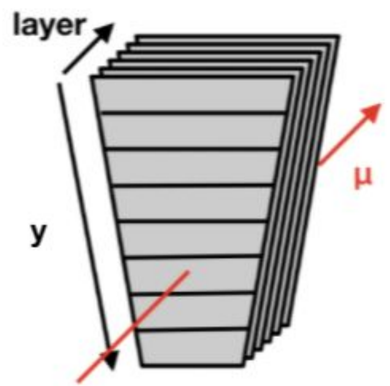

Fig1. Detector we aim for

❑ Dataset Info.

- We aim for generic muon detectors. In this study, the chamber has 6 layers, with layer consisting of 384 strips in the x axis and segmented in 8 partitions in the y axis.

- A chamber image made up of 6*8*384 voxels, and each voxel is binary (0 or 1). When the current flows on a strip, a voxel turned 1 from 0.

- Muons are ranging from p_T 5~100 GeV.

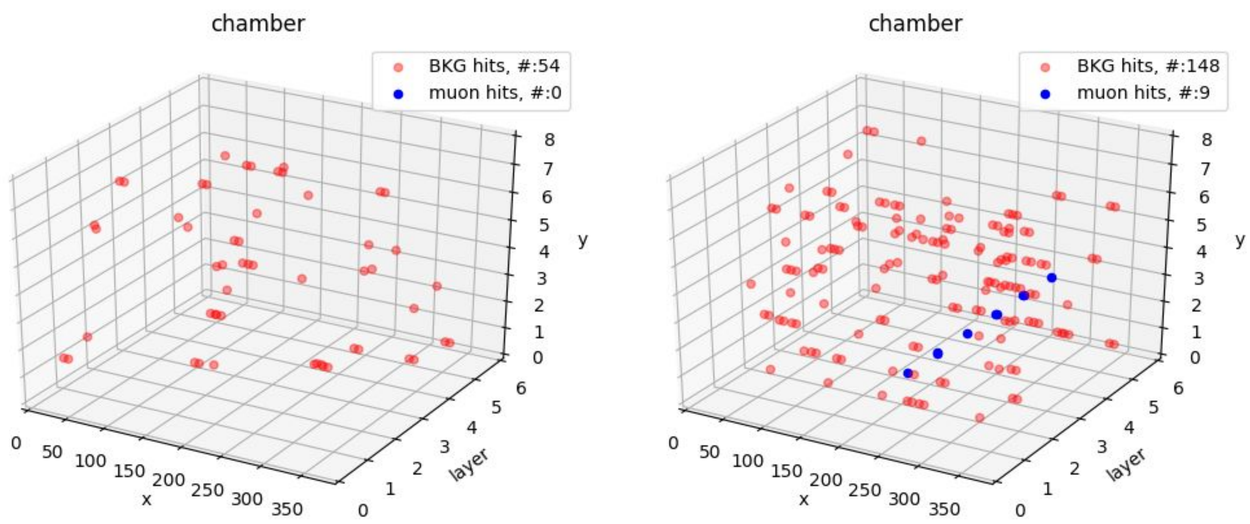

Fig2. Visualization of chambers with hits

Method

❑ Training CNN

- Convolution kernels extract the visual features of muon hits so that model output close to ground truth.
- Loss - weighted binary cross entropy loss

$$\mathrm{WCE}\left(p,\hat{p}\right) = -\left(\beta p \log\left(\hat{p}\right) + (1-p)\log\left(1-\hat{p}\right)\right)$$

- Keras with tensorflow 1.14 for the backend.

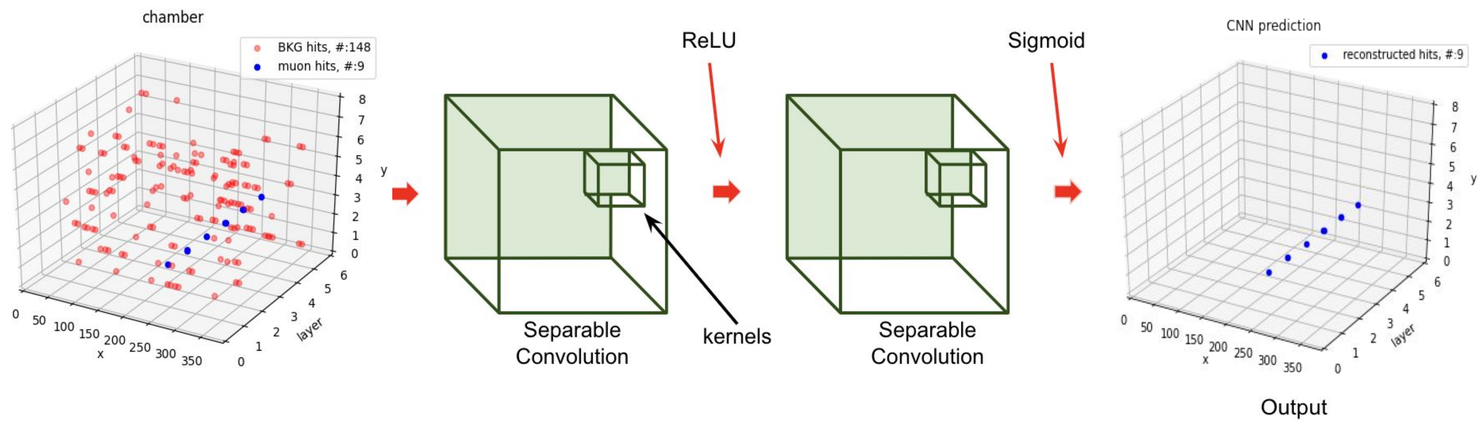

Fig3. CNN based track segment reconstruction

❑ Deploying on FPGA

- Quantization for smaller model size and faster computation.

Quantization

- Changing float 32→int 8 for lighter model

Layer optimization

- Deleting batch normalization layer and sigmoid layer
  - Batch normalization layer is used for training, not inferencing.
  - Sigmoid layer can be replaced by some appropriate threshold.

Compilation & Deploying

- To deploy the CNN in FPGA, we used the compilation tool provided by Vitis-AI (Xilinx AI framework).

Results

❑ Efficiency & Fake rate

- About 300,000 images are used for test.
- If True Positive Rate (TP/P) of a image is over 0.6, we define it as a “matched”.
- In efficiency plot, 0~5GeV is empty, so the number of samples in first bin is half of others.

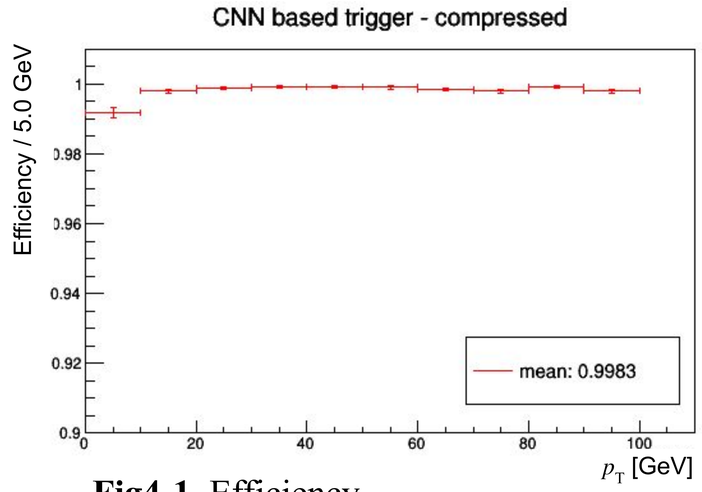

Fig4-1. Efficiency

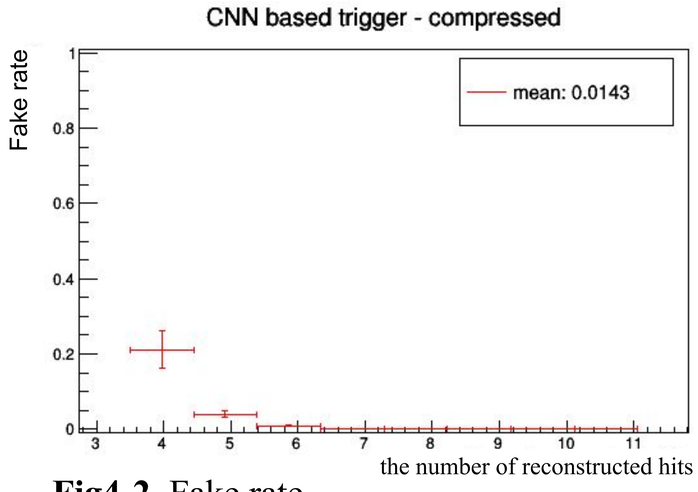

Fig4-2. Fake rate

※ Definitions of efficiency and fake rate are on postsers.

❑ Throughput

- Throughput is the rate at which something is processed.
- On Fig5, throughput was measured in processed images per second (FPS), and the buffer sizes are 1, 2, 4, ⋯ , 2048 from left.
- Only 1~6 buffer size is allowed for FPGA now.

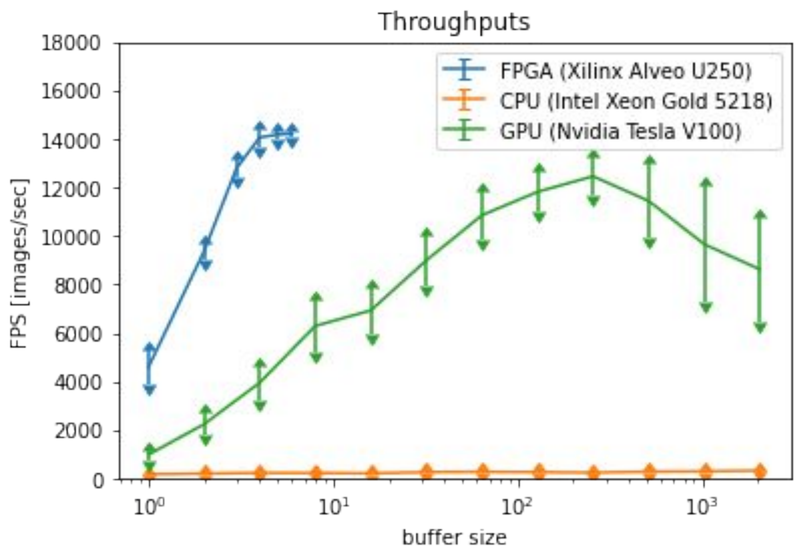

Fig5. Throughput according to processors