Large CNN for HLS4ML and Deepcalo

ChiJui Chen[1]
on behalf of the

DeepCalo Team

Bo-Cheng Lai[1], Yan-Lun Huang[1], Lin-Chi Yang[1], Shih-Chieh Hsu[2], Scott Hauck[2],

Ziang (Dennis) Yin[2], Alex Schuy[2], Philip Harris[3], Dylan Rankin[4]

[1]NYCU, [2]UW, [3]MIT, [4]UPenn

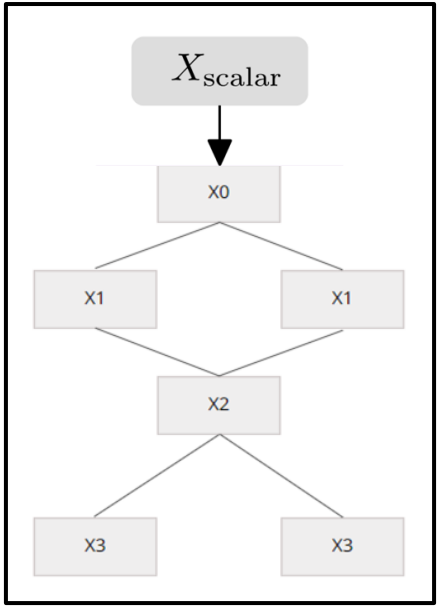
Fast Machine Learning for Science Workshop
10, 3, 2022

DeepCalo: Energy Regression Task of e- and γ

陽明交大 NYCU

ATLAS

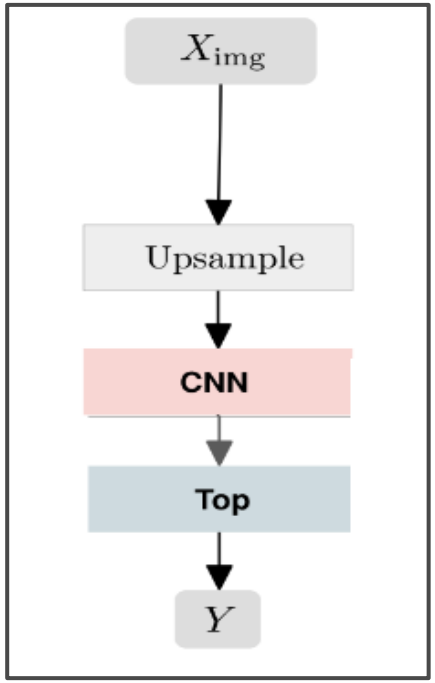
Boosted Decision Tree



IQR_{75} improvement: 0%

DeepCalo

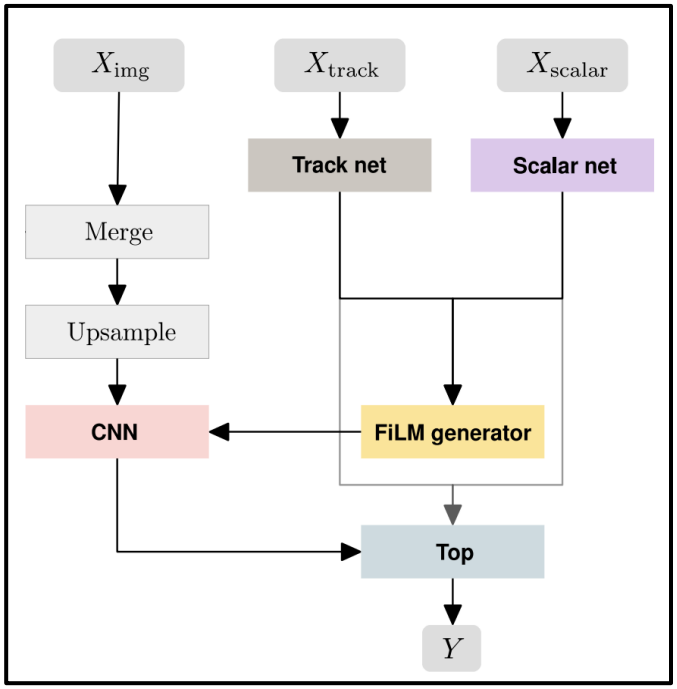
Image-only model



15%

1.8M parameters

Full-variables model

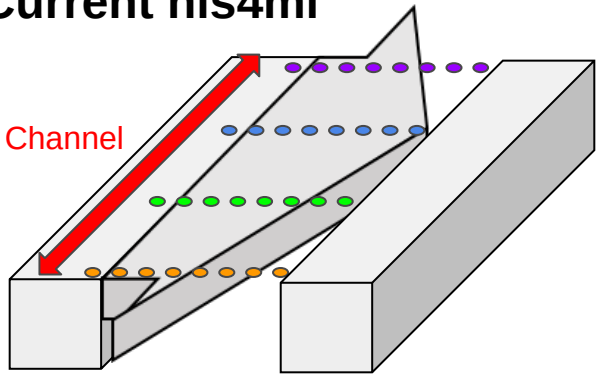


31%

3.6M parameters

Challenges

Current hls4ml

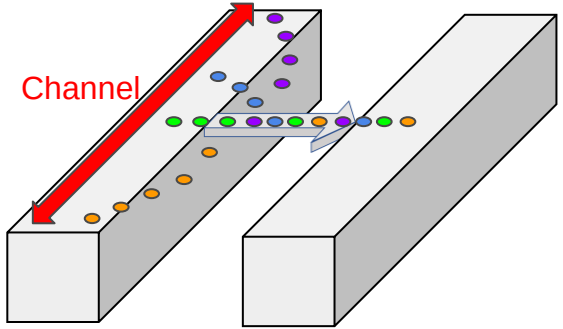


Streams of Struct

Struct { ap_fixed[N] } data_t
hls::stream<data_t>

Synthesize
Large Channel
Layers

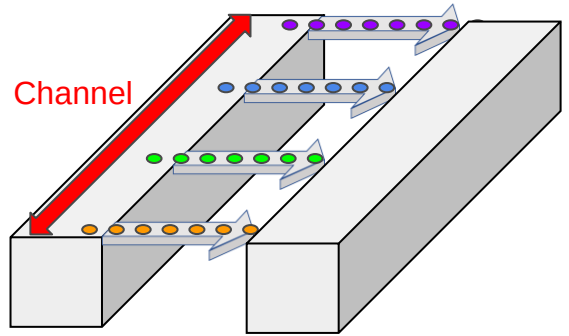
Modified hls4ml



Single-value Streams

hls::stream<ap_fixed>

Channel larger than 64

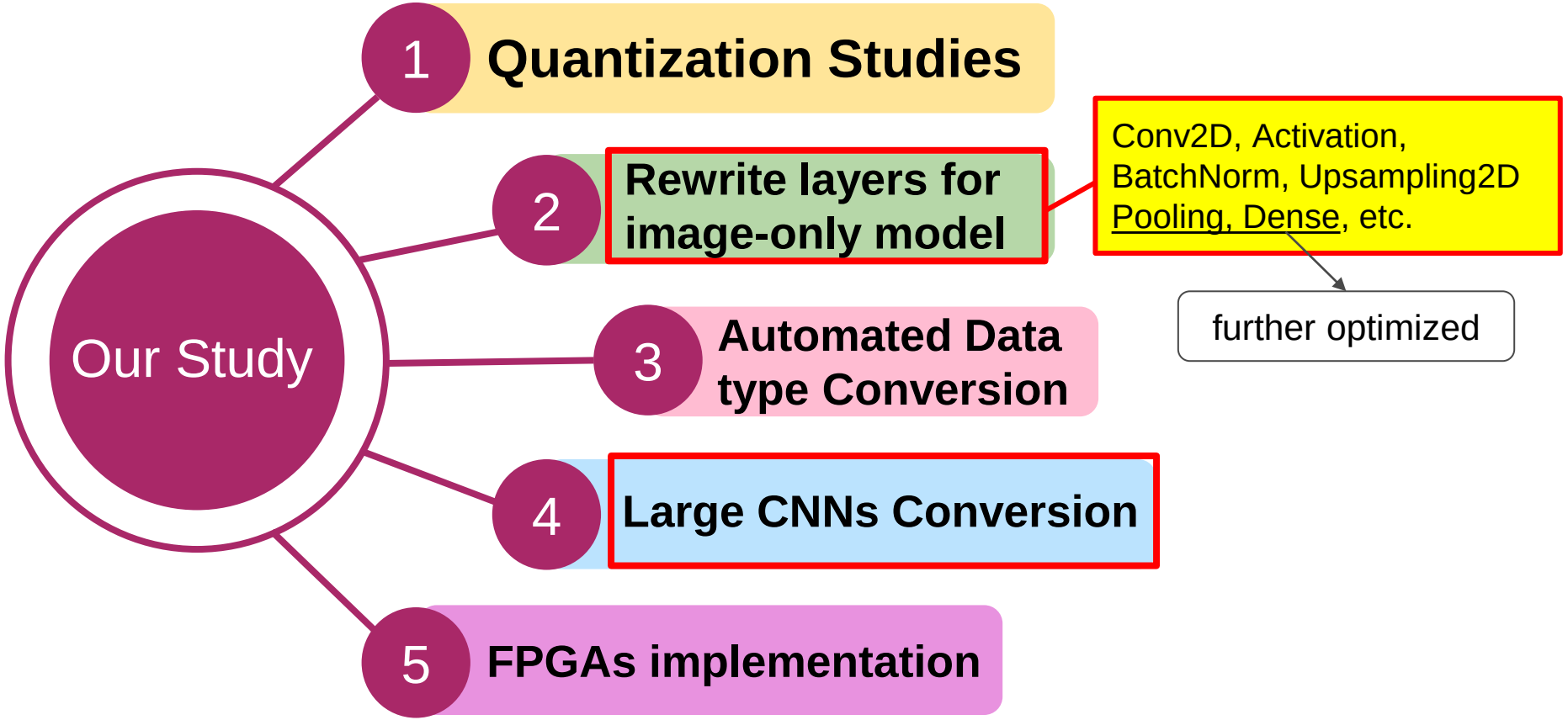


Array of Single-value Streams

hls::stream<ap_fixed>[N]

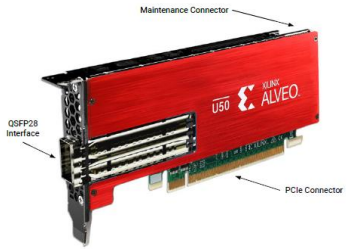
Channel less than 64

Design Flow of Our Study



Results from V-Synthesis

- FPGA: Alveo U50
- Model: image-only
  With 2 kinds of data types
- Precision: ap_fixed<16,6>
- Cycle: 5ns



	hls4ml model (all single)	Resource utilization(%)	hls4ml model (array + single)	Resource utilization(%)
BRAM	662	49.26 %	662	49.26 %
DSP	3,524	59.21 %	3,524	59.21 %
FF	202,011	11.59 %	235,714	13.52 %
LUT	194,744	22.34 %	342,007	39.24 %
URAM	115	17.97 %	115	17.97 %
FPGA time	0.9ms		0.6ms	

Summary

Our current progress:

- Expanded the hls4ml library for supporting **automated large CNNs conversion**.
- Applied the design to the **image-only model** and successfully synthesized and tested it on FPGAs.

Next-steps

- Develop tools to convert the full-model and load it on FPGAs.
- Further quantization study for the DeepCalo models.

Thank you for listening!

References

- ATLAS Collaboration et al. *“The ATLAS Experiment at the CERN Large Hadron Collider.”* In: Journal of Instrumentation3.08 (Aug. 2008), S08003–S08003. issn: 1748-0221. Doi: 10.1088/1748-0221/3/08/S08003.
- Frederik G. Faye: *Energy reconstruction of electrons and photons using convolutional neural networks*, Master’s thesis (cand.scient.), © October 2019
- Alveo U50 Data Center Accelerator Card : https://www.mouser.com/pdfDocs/u50-UserGuide.pdf
- P. Harris and D. Rankin: Deploy DeepCalo on Alveo [slide]
- Kelvin Lin: Convolutional Layer Implementations for hls4ml library [talk] [thesis]
- Github branch: HLS4ML with single stream