Del Algoritmo al Hardware:
Aprendizaje Automático en Sistemas Embebidos

From Algorithm to Hardware: Machine Learning in Embedded Systems

1 al 11 de Abril, 2025. Universidad Nacional de Mar del Plata - Mar del Plata - Argentina.

Part 1 - hls4ml: Bridging Machine Learning and FPGAs for Ultra-Fast Inference.

Part 2 - Workflow for Deep Neural Network Deployment On Embedded Architectures

Romina Soledad Molina, Ph.D.
MLab-STI, ICTP

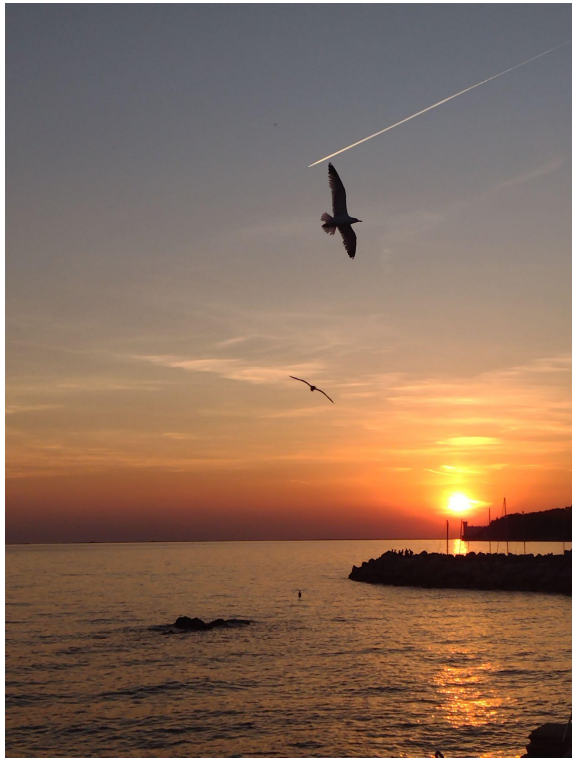
Universidad Nacional de Mar del Plata

Fundación Williams

ICTP The Abdus Salam International Centre for Theoretical Physics

Outline

- Summing up the previous content.
- High-Level Synthesis for Machine Learning.
- hls4ml workflow
- Workflow for Deep Neural Network Deployment On Embedded Architectures.
- Demo: Using hls4ml.



- Part 1 -
hls4ml: Bridging Machine Learning and FPGAs for Ultra-Fast Inference.

Summing up the previous content

- **Rise of real-time ML**
  - ML is now used in latency-critical domains: HEP, robotics, autonomous systems, IoT.

Summing up the previous content

- **Rise of real-time ML**
  - ML is now used in latency-critical domains: HEP, robotics, autonomous systems, IoT.

- **The need for fast and efficient inference**
  - Low-latency, low-power inference.

Summing up the previous content

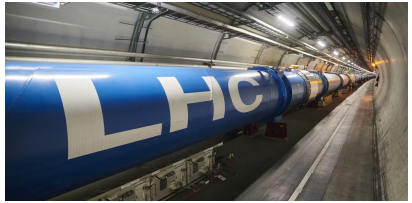
- **Rise of real-time ML**
  - ML is now used in latency-critical domains: HEP, robotics, autonomous systems, IoT.
- **The need for fast and efficient inference**
  - Low-latency, low-power inference
- **Why FPGAs?**
  - Highly parallel and reconfigurable, tuned for latency and energy efficiency.

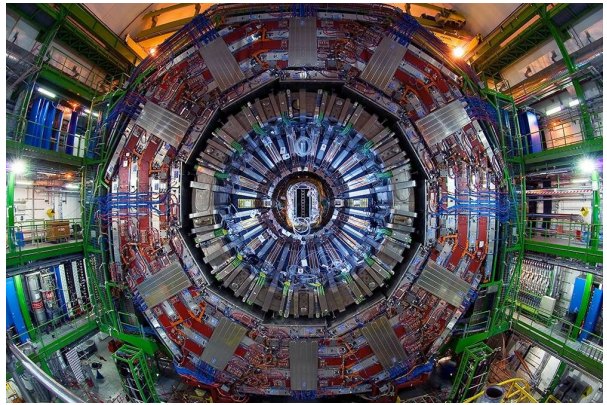
Summing up the previous content

- **Rise of real-time ML**
  - ML is now used in latency-critical domains: HEP, robotics, autonomous systems, IoT.
- **The need for fast and efficient inference**
  - Low-latency, low-power inference.
- **Why FPGAs?**
  - Highly parallel and reconfigurable, tuned for latency and energy efficiency.
- **The challenge**
  - ML models are software-native.
  - Hardware mapping is complex and manual.
  - Needs automation and abstraction.

hls4ml

"**The inspiration for the creation of the hls4ml** package stems from the **high energy physics community at the CERN Large Hadron Collider (LHC)**. While machine learning has already been proven to be extremely useful in analysis of data from detectors at the LHC, it is typically performed in an "offline" environment after the data is taken and agglomerated." [hls4ml]

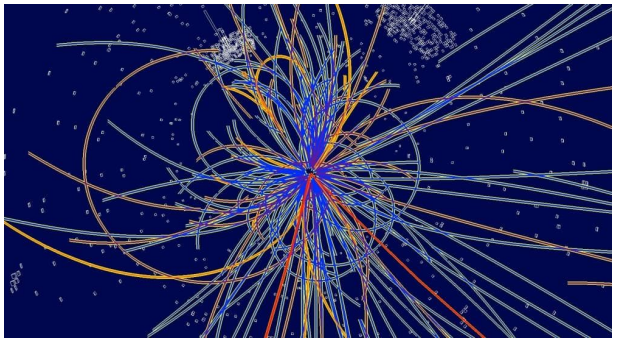




[hls4ml] https://fastmachinelearning.org/hls4ml/

hls4ml

"However, **one of the largest problems at detectors on the LHC is that collisions, or "events", generate too much data for everything to be saved**. As such, filters called **"triggers" are used to determine whether a given event should be kept. Using FPGAs allows for significantly lower latency so machine learning algorithms can essentially be run "live" at the detector level for event selection**. As a result, more events with potential signs of new physics can be preserved for analysis." [hls4ml]



[hls4ml] https://fastmachinelearning.org/hls4ml/

Fast inference of deep neural networks in FPGAs for particle physics

J. Duarte,[a] S. Han,[b] P. Harris,[b] S. Jindariani,[a] E. Kreinar,[c] B. Kreis,[a] J. Ngadiuba,[d] M. Pierini,[d] R. Rivera,[a] N. Tran[a,1] and Z. Wu[e]

[a] *Fermi National Accelerator Laboratory, Batavia, IL 60510, U.S.A.*
[b] *Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.*
[c] *HawkEye360, Herndon, VA 20170, U.S.A.*
[d] *CERN, CH-1211 Geneva 23, Switzerland*
[e] *University of Illinois at Chicago, Chicago, IL 60607, U.S.A.*

E-mail: hls4ml.help@gmail.com

Abstract: Recent results at the Large Hadron Collider (LHC) have pointed to enhanced physics capabilities through the improvement of the real-time event processing techniques. Machine learning methods are ubiquitous and have proven to be very powerful in LHC physics, and particle physics as a whole. However, exploration of the use of such techniques in low-latency, low-power FPGA (Field Programmable Gate Array) hardware has only just begun. FPGA-based trigger and data acquisition systems have extremely low, sub-microsecond latency requirements that are unique to particle physics. We present a case study for neural network inference in FPGAs focusing on a classifier for jet substructure which would enable, among many other physics scenarios, searches for new dark sector particles and novel measurements of the Higgs boson. While we focus on a specific example, the lessons are far-reaching. A companion compiler package for this work is developed based on High-Level Synthesis (HLS) called `hls4ml` to build machine learning models in FPGAs. The use of HLS increases accessibility across a broad user community and allows for a drastic

Duarte, J., Han, S., Harris, P., Jindariani, S., Kreinar, E., Kreis, B., ... & Wu, Z. (2018). Fast inference of deep neural networks in FPGAs for particle physics. *Journal of instrumentation*, *13*(07), P07027.

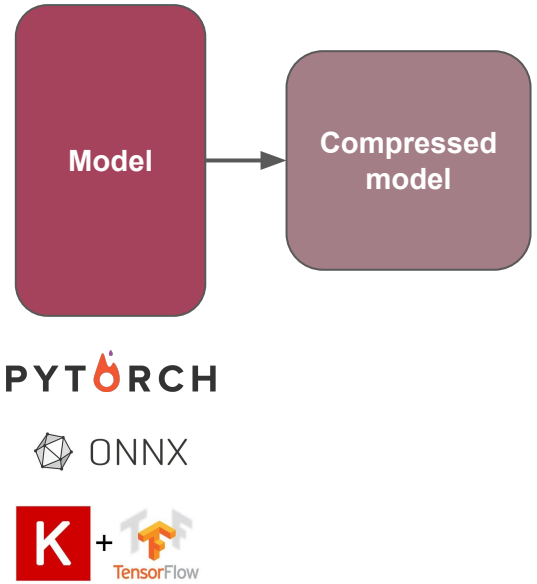
- Binary & Ternary neural networks: [2020 Mach. Learn.: Sci. Technol]
  - Compressed weights for low resource inference
- Boosted Decision Trees: [JINST 15 P05026 (2020)]
  - Low latency for Decision Tree ensembles
- GarNet / GravNet: [arXiv: 2008.03601]
  - Distance weighted graph neural networks suitable for sparse/irregular point-cloud data
- Quantization aware training QKeras + support in hls4ml: [arXiv: 2006.10159]
- Convolutional neural networks: Mach. Learn.: Sci. Technol. 2 045015 (2021)

- Python package.
- It enables the transformation of neural network models into firmware deployable on FPGAs for efficient inference.
  - https://fastmachinelearning.org/hls4ml/
  - https://github.com/hls-fpga-machine-learning/hls4ml
  - pip install hls4ml
- Extremely configurable: precision, resource vs latency/throughput tradeoff.
- Easy to use.

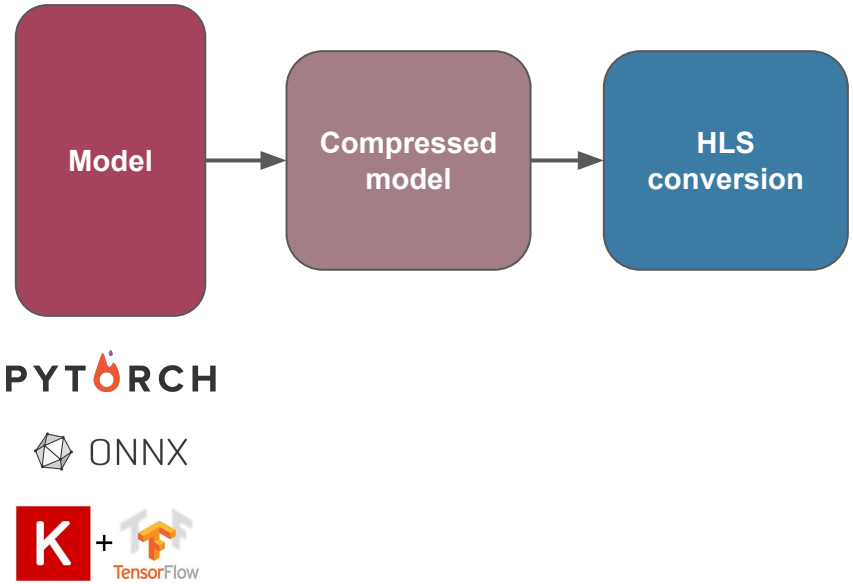
Duarte, J., Han, S., Harris, P., Jindariani, S., Kreinar, E., Kreis, B., ... & Wu, Z. (2018). Fast inference of deep neural networks in FPGAs for particle physics. *Journal of instrumentation*, *13*(07), P07027.

hls4ml workflow

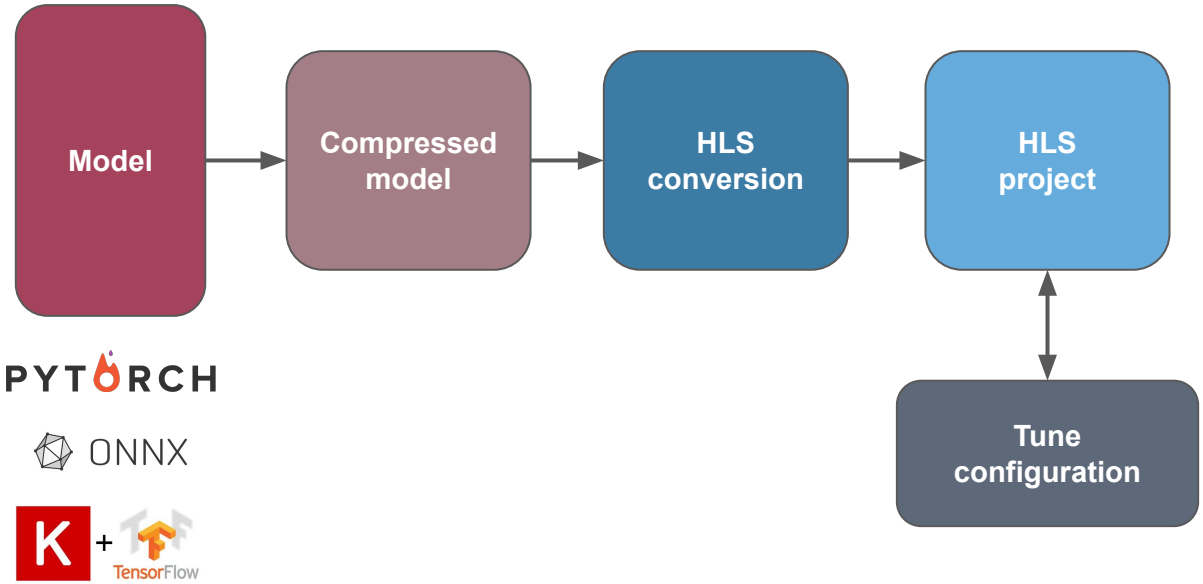
hls4ml workflow



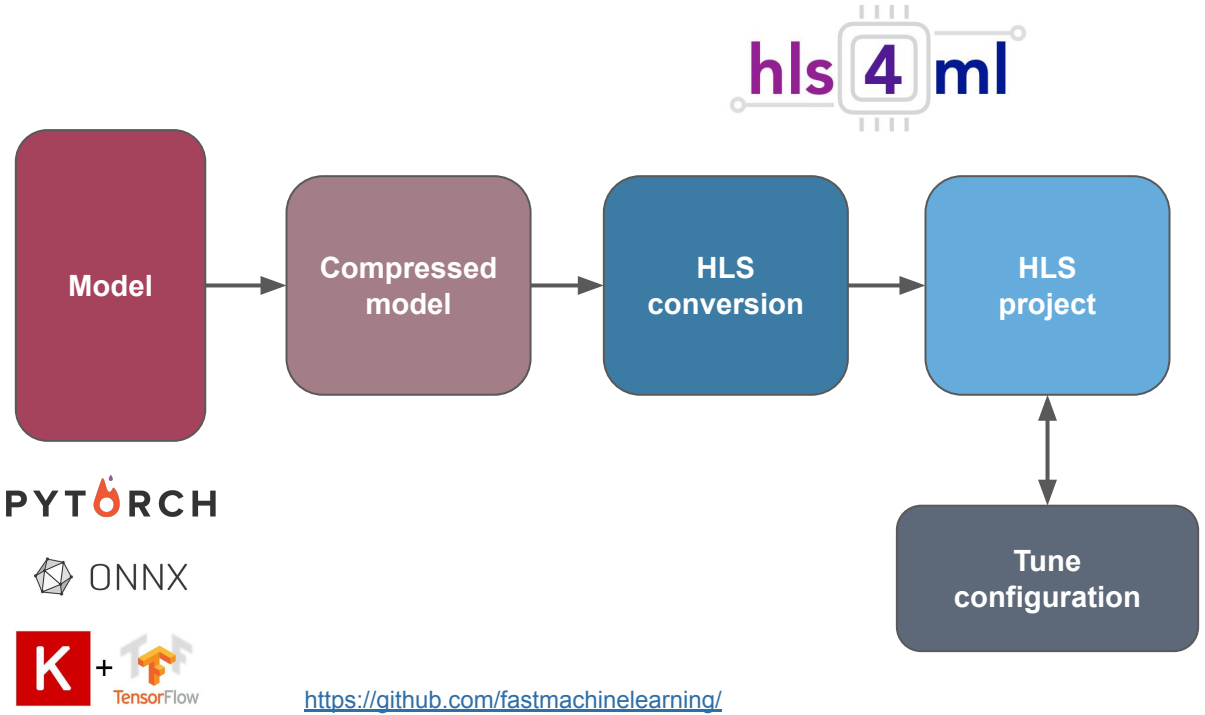
hls4ml workflow



hls4ml workflow

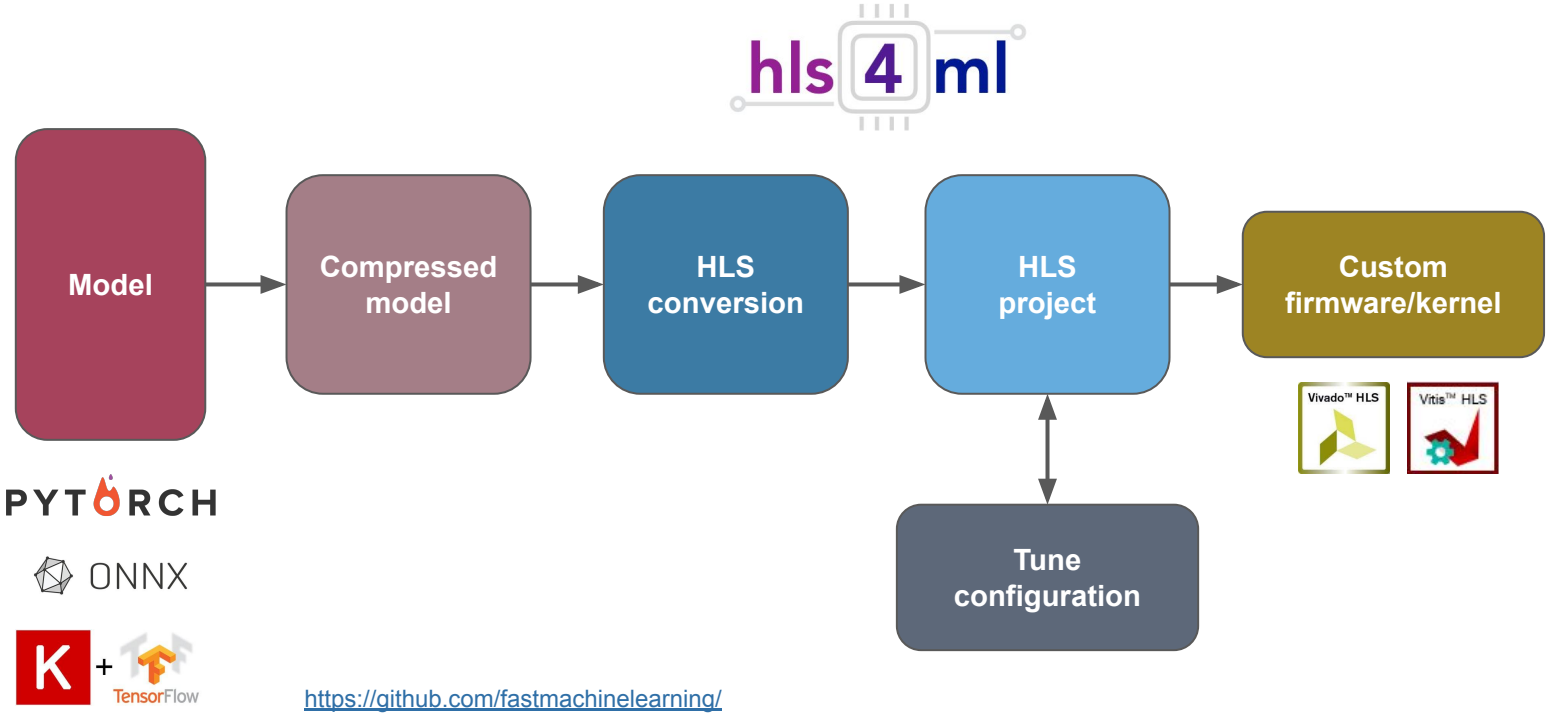


hls4ml workflow



https://github.com/fastmachinelearning/

hls4ml workflow



https://github.com/fastmachinelearning/

High-Level Synthesis for Machine Learning

ML framework support:

- (Q)Keras
- PyTorch
- (Q)ONNX

https://fastmachinelearning.org/hls4ml/

High-Level Synthesis for Machine Learning

ML framework support:

- (Q)Keras
- PyTorch
- (Q)ONNX

Neural networks architectures:

- Fully Connected NN
- Convolutional NN
- Recurrent NN
- Graph NN

https://fastmachinelearning.org/hls4ml/

High-Level Synthesis for Machine Learning

ML framework support:

- (Q)Keras
- PyTorch
- (Q)ONNX

Neural networks architectures:

- Fully Connected NN
- Convolutional NN
- Recurrent NN
- Graph NN

HLS backends:

- Vivado HLS
- Intel HLS
- Vitis HLS
- Catapult HLS
- oneAPI (experimental)

https://fastmachinelearning.org/hls4ml/

High-Level Synthesis for Machine Learning

hls4ml is tested on the following platforms:

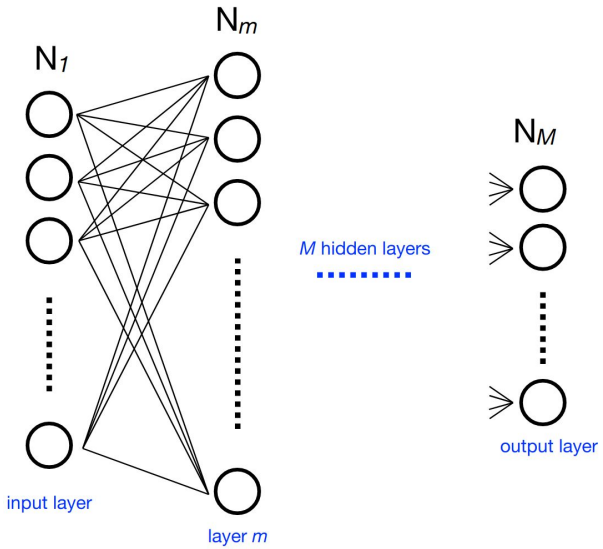
Vivado HLS versions 2018.2 to 2020.1
Intel HLS versions 20.1 to 21.4. Versions > 21.4 have not been tested.
Vitis HLS versions 2022.2 to 2024.1. Versions <= 2022.1 are known not to work.
Catapult HLS versions 2024.1_1 to 2024.2
oneAPI versions 2024.1 to 2025.0

https://fastmachinelearning.org/hls4ml/

hls4ml

How does it work?

With hls4ml, each layer of output values is calculated independently in sequence, using pipelining to speed up the process by accepting new inputs after an initiation interval. The activations, if nontrivial, are precomputed.



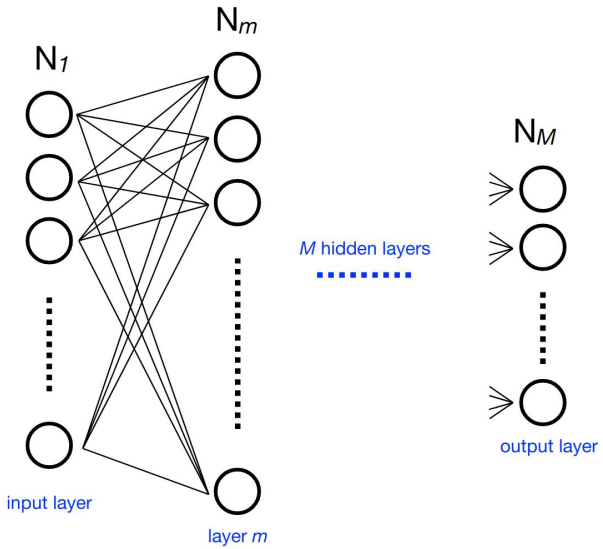
[hls4ml] https://fastmachinelearning.org/hls4ml/

hls4ml

How does it work?

With hls4ml, each layer of output values is calculated independently in sequence, using pipelining to speed up the process by accepting new inputs after an initiation interval. The activations, if nontrivial, are precomputed.

Simplifying the input network must be done before using hls4ml to generate HLS code, for optimal compression to provide a sizable speedup.



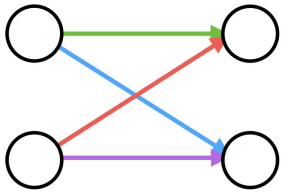
[hls4ml] https://fastmachinelearning.org/hls4ml/

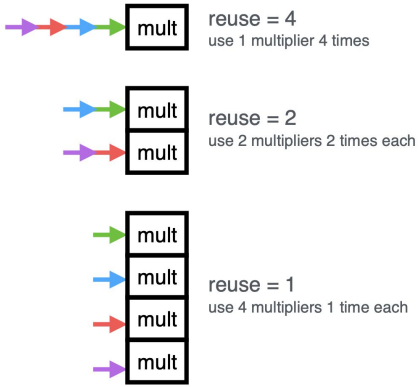
Trade-off between latency and FPGA resource usage determined by the parallelization of the calculations in each layer.

hls4ml

Trade-off between latency and FPGA resource usage determined by the parallelization of the calculations in each layer.

Reuse factor: number of times a multiplier is used to do a computation.



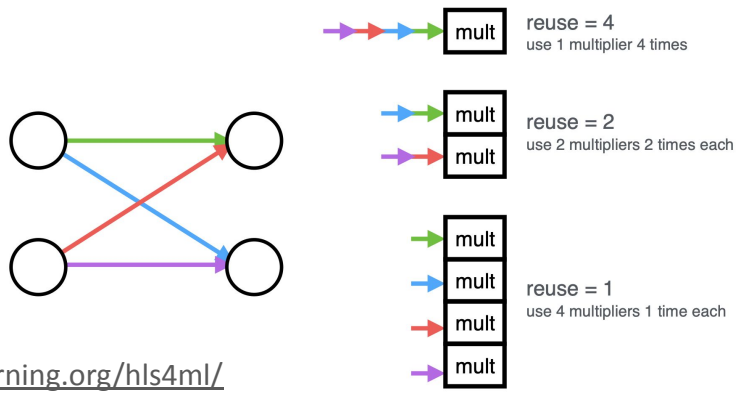


https://fastmachinelearning.org/hls4ml/

hls4ml

Trade-off between latency and FPGA resource usage determined by the parallelization of the calculations in each layer.

Reuse factor: number of times a multiplier is used to do a computation.



Fewer resources,
Lower throughput,
Higher latency

More resources,
Higher throughput,
Lower latency

https://fastmachinelearning.org/hls4ml/

I/O types: supports multiple styles for handling data transfer to/from the network and between layers.

https://fastmachinelearning.org/hls4ml/

I/O types: supports multiple styles for handling data transfer to/from the network and between layers.

io_parallel

io_stream

https://fastmachinelearning.org/hls4ml/

I/O types: supports multiple styles for handling data transfer to/from the network and between layers.

io_parallel

Data is passed in **parallel** between the layers.

This style allows for **maximum parallelism** and is well suited for MLP networks and small CNNs which aim for lowest latency.

https://fastmachinelearning.org/hls4ml/

hls4ml

I/O types: supports multiple styles for handling data transfer to/from the network and between layers.

io_stream

Data is passed **one "pixel" at a time**.

Each pixel is an **array of channels**, which are always sent in parallel. This method for sending data between layers is recommended for larger CNN and RNN networks.

With this IO type, each layer is connected with the subsequent layer through **first-in first-out (FIFO) buffers.**

The implementation of the FIFO buffers contribute to the overall resource utilization of the design, impacting in particular the **BRAM or LUT utilization**.

https://fastmachinelearning.org/hls4ml/

Strategy: the implementation of core matrix-vector multiplication routine, which can be **latency-oriented, resource-saving oriented, or specialized**.

Different strategies will have an **impact on overall latency and resource consumption** of each layer and users are advised to choose based on their design goals.

https://fastmachinelearning.org/hls4ml/

Strategy: the implementation of core matrix-vector multiplication routine, which can be **latency-oriented, resource-saving oriented, or specialized**.

Different strategies will have an **impact on overall latency and resource consumption** of each layer and users are advised to choose based on their design goals.

If one layer would have >**4096 elements**, we should set **['Strategy'] = 'Resource'** for that layer, or **increase the reuse factor** by hand.

4096 elements is related to the maximum size of an array to be partitioned. This value might change, it should be checked with the hardware synthesis tool.

https://fastmachinelearning.org/hls4ml/

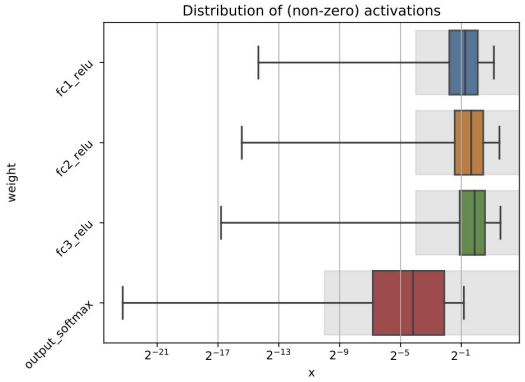
Activations - Implementation parameter

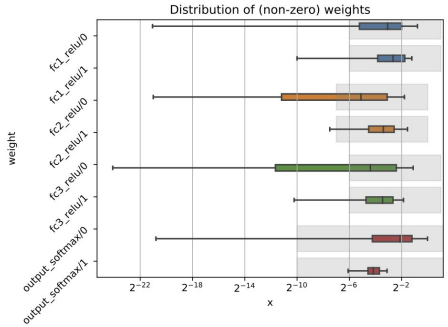
- **latency:** Good latency, high resource usage. **It does not work well if there are many output classes**.
- **stable:** Slower but with better accuracy, useful in scenarios where **higher accuracy is needed**.
- **legacy:** An older implementation with poor accuracy, but good performance. Usually the latency implementation is preferred.
- **argmax:** If you don't care about normalized outputs and only care about which one has the highest value, using argmax saves a lot of resources. This sets the highest value to 1, the others to 0.

https://fastmachinelearning.org/hls4ml/

Profiling

- The tools in **hls4ml.model.profiling** can help to choose the right precision for the model.
- **hls4ml.model.profiling.numerical** method with three objects: **a Keras model object**, **test data**, and an **HLSModel.**





https://fastmachinelearning.org/hls4ml/
Image from ttps://fastmachinelearning.org/hls4ml/api/PROFILING.html

Trace

- When we start using **customised precision** throughout the model, it can be useful to collect the output from each layer. We enable this trace collection by setting **Trace = True** for each layer whose output we want to collect.

```
for layer in config['LayerName'].keys():
        config['LayerName'][layer]['Trace'] = True
```

https://fastmachinelearning.org/hls4ml/
https://fastmachinelearning.org/hls4ml-tutorial/part2_advanced_config.html

High-Level Synthesis for Machine Learning

Python integration


```
1 from tensorflow.keras.models import load_model
2 from sklearn.metrics import accuracy_score
3 model = load_model('model_keras_MLP.h5')
4 model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
fc1 (Dense)                  (None, 60)                3900
_________________________________________________________________
relu1 (Activation)           (None, 60)                0
_________________________________________________________________
fc0 (Dense)                  (None, 40)                2440
_________________________________________________________________
relu0 (Activation)           (None, 40)                0
_________________________________________________________________
fc2 (Dense)                  (None, 30)                1230
_________________________________________________________________
relu2 (Activation)           (None, 30)                0
_________________________________________________________________
fc3 (Dense)                  (None, 10)                310
```

High-Level Synthesis for Machine Learning

Python integration


```
import hls4ml
import plotting

config = hls4ml.utils.config_from_keras_model(teacherMLP, granularity='model')
config['Model'] = {'Precision' : 'ap_fixed<24,16>', 'ReuseFactor' : 1, 'Strategy' : 'Latency'}
print("-----------------------------------")
plotting.print_dict(config)
print("-----------------------------------")
hls_model = hls4ml.converters.convert_from_keras_model(teacherMLP,
                                                       hls_config=config,
                                                       output_dir='model_3/hls_model'
                                                      )
hls_model.compile()
```

High-Level Synthesis for Machine Learning

Python integration


```
import hls4ml
import plotting
hls4ml.model.optimizer.OutputRoundingSaturationMode.layers = ['Activation']
hls4ml.model.optimizer.OutputRoundingSaturationMode.rounding_mode = 'AP_RND'
hls4ml.model.optimizer.OutputRoundingSaturationMode.saturation_mode = 'AP_SAT'

config = hls4ml.utils.config_from_keras_model(model, granularity='name')

config['Model'] = {'Precision' : 'ap_fixed<17,16>', 'ReuseFactor' : 1, 'Strategy' : 'Latency'}

config['LayerName']['fc1']['Precision']['weight'] = 'ap_fixed<9, 1>'

config['LayerName']['softmax']['Precision'] = 'ap_fixed<32,15>'

print("-----------------------------------")
plotting.print_dict(config)
print("-----------------------------------")
hls_model = hls4ml.converters.convert_from_keras_model(model,
                                                       hls_config=config,
                                                       output_dir='model_3/MLP_student_smr3765'
                                                      )

hls_model.compile()
```

High-Level Synthesis for Machine Learning

QKeras for quantization-aware training

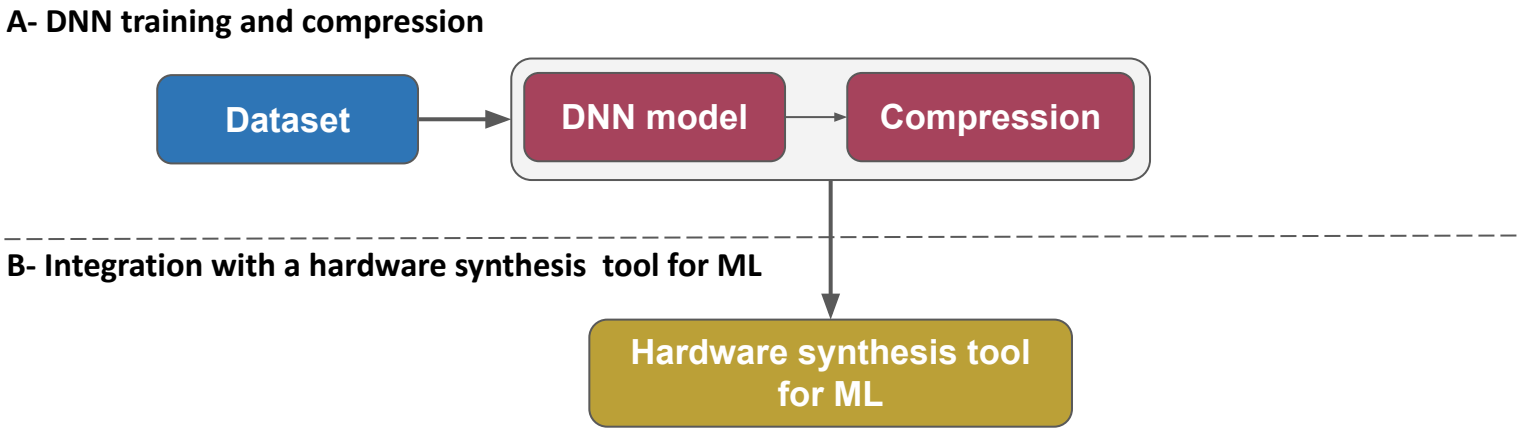

```
# MLP architecture
# Create the student QKERAS
studentQ_MLP = keras.Sequential(
    [
        Input(shape=(30,)),
        QDense(20, name='fc1',
                kernel_quantizer=quantized_bits(9,1,alpha=1), bias_quantizer=quantized_bits(23,15,alpha=1)),
        QActivation(activation=quantized_relu(16,15), name='relu1'),
        QDense(10, name='fc2',
                kernel_quantizer=quantized_bits(9,1,alpha=1), bias_quantizer=quantized_bits(23,15,alpha=1)),
        QActivation(activation=quantized_relu(16,15), name='relu2'),
        QDense(10, name='fc6',
                kernel_quantizer=quantized_bits(9,1,alpha=1), bias_quantizer=quantized_bits(23,15,alpha=1)),
        QActivation(activation=quantized_relu(16,15), name='relu3'),

        QDense(4, name='output',
                kernel_quantizer=quantized_bits(32,15,alpha=1), bias_quantizer=quantized_bits(32,15,alpha=1)),
        Activation(activation='softmax', name='softmax')


    ],
    name="student",
)

print_qstats(studentQ_MLP)
```

- Part 2 -
Workflow to efficiently compress and deploy DNN on SoC/FPGA

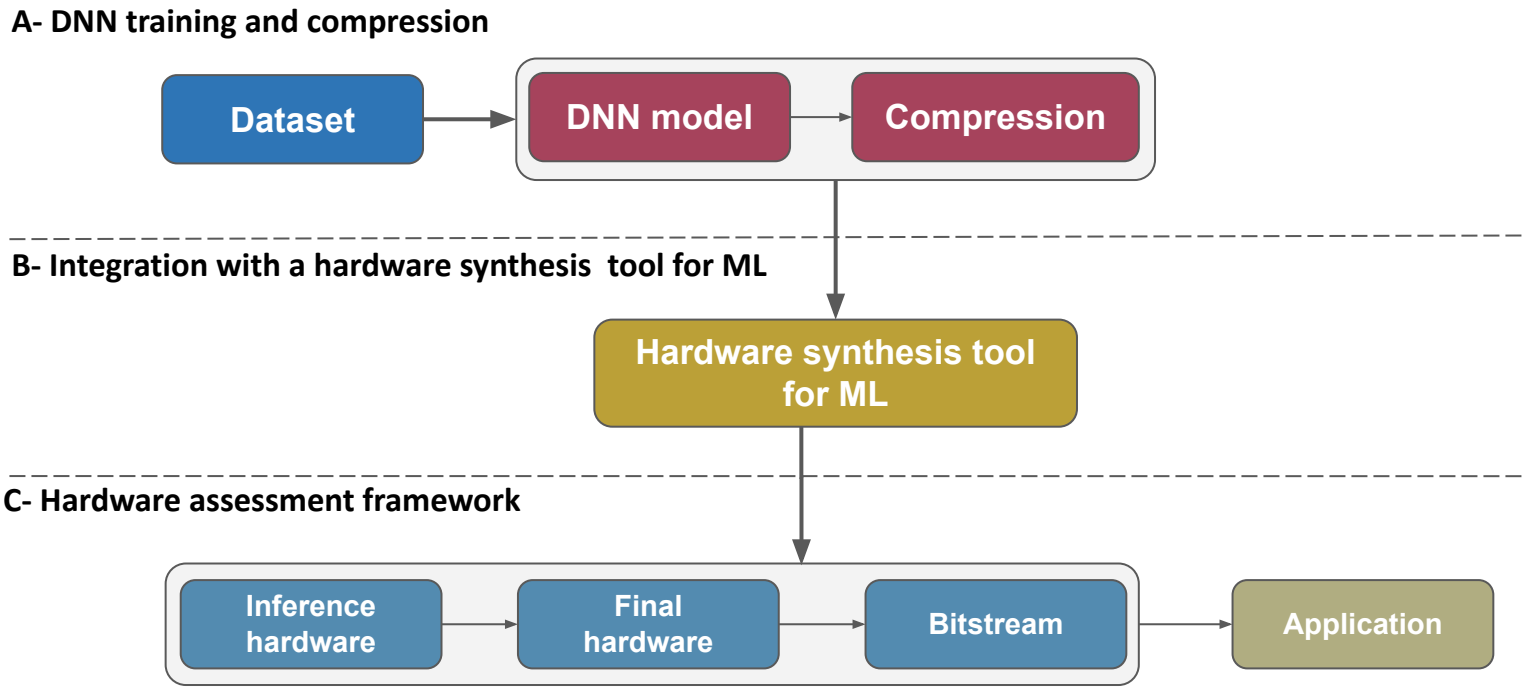
End-to-end workflow

A- DNN training and compression

End-to-end workflow



End-to-end workflow

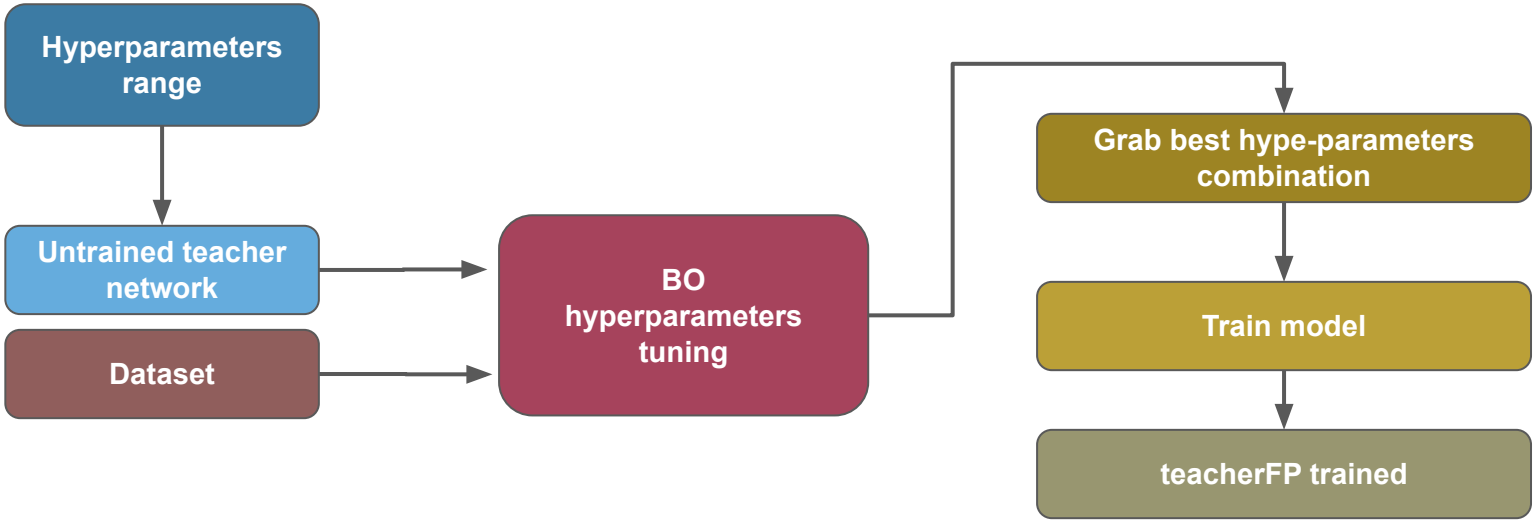


Available at https://github.com/RomiSolMolina/workflowCompressionML

A. DNN training and compression

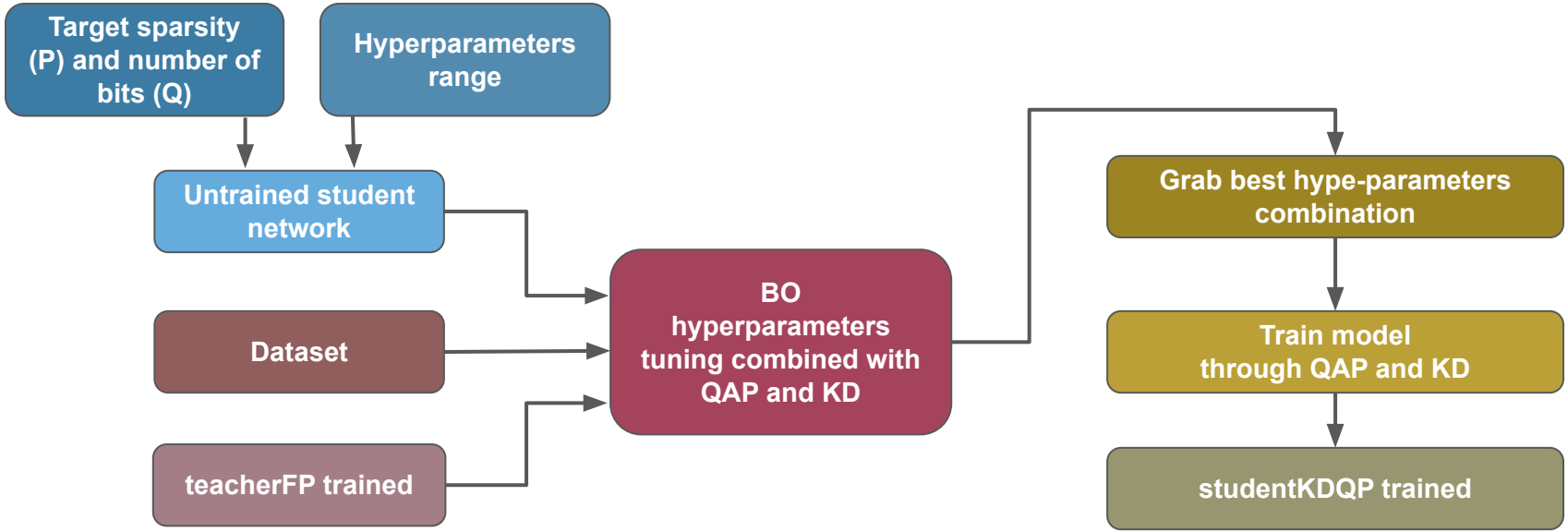
DNN training and compression

Stage 1 - Teacher training



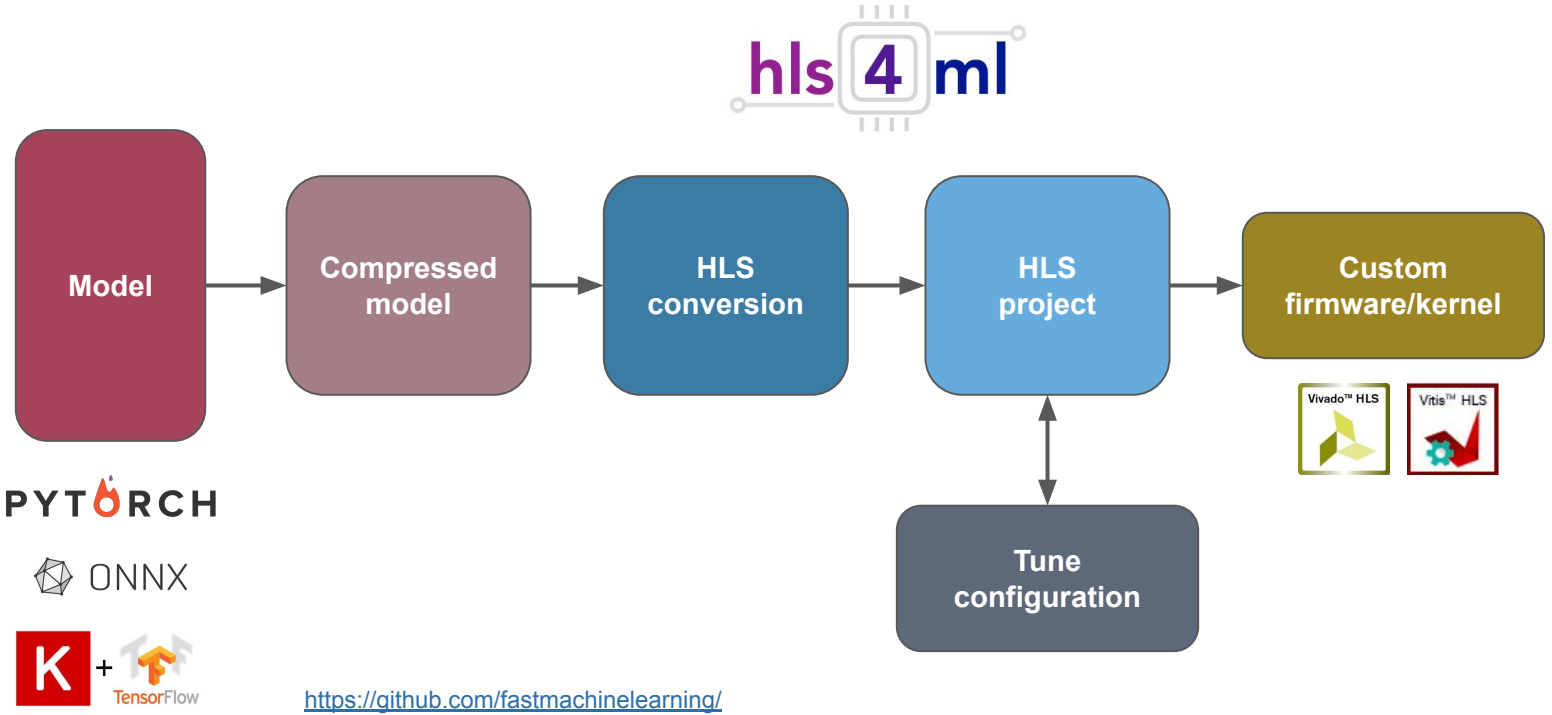
DNN training and compression

Stage 2 - Student training



B. Integration with a hardware synthesis tool for ML

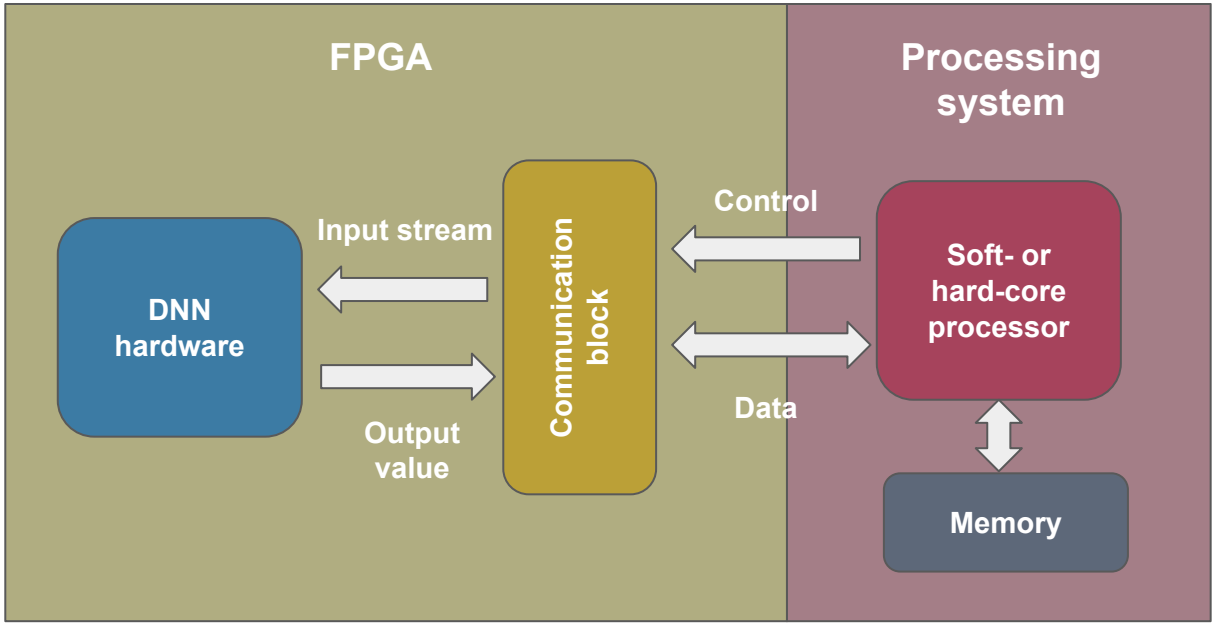
Integration with a hardware synthesis tool for ML



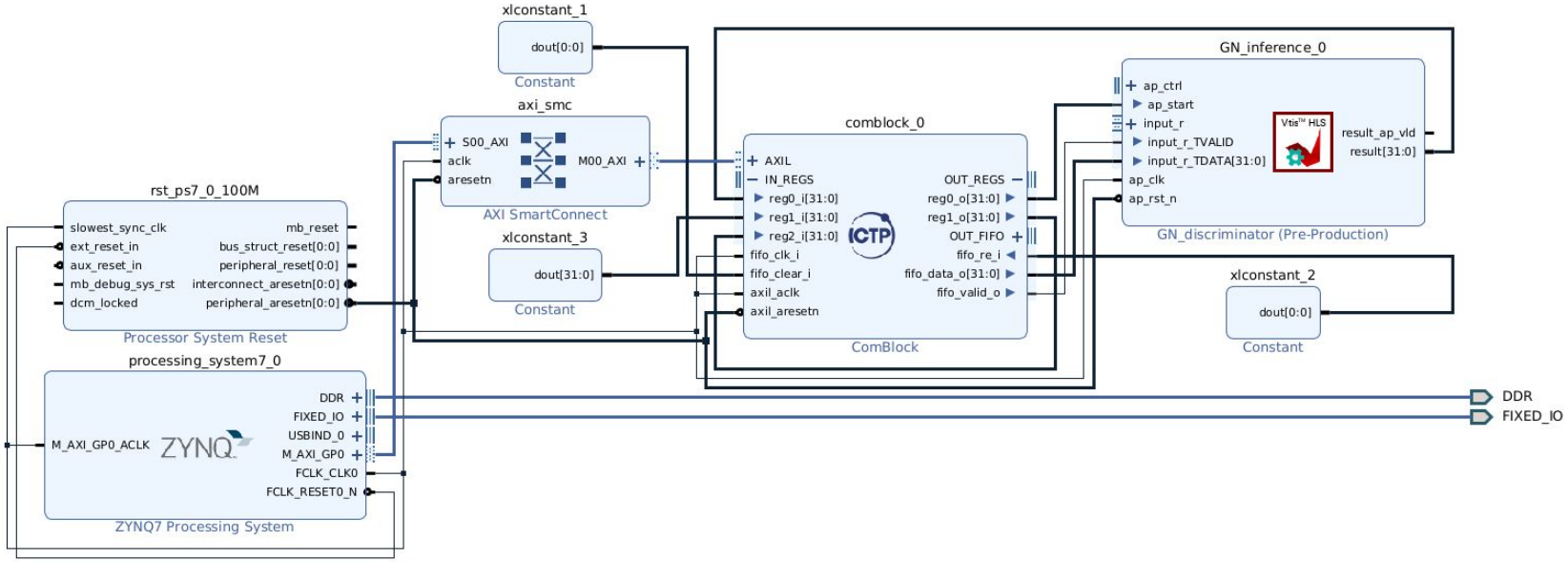
https://github.com/fastmachinelearning/

C. Hardware assessment framework

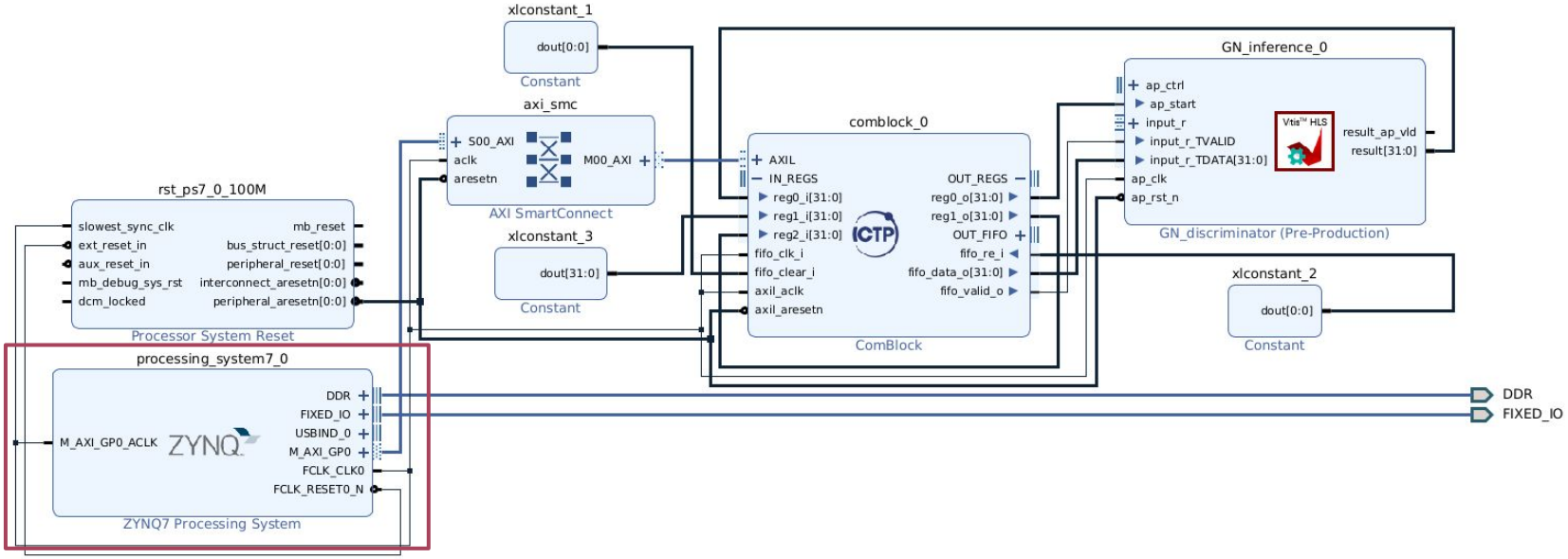
Hardware assessment framework



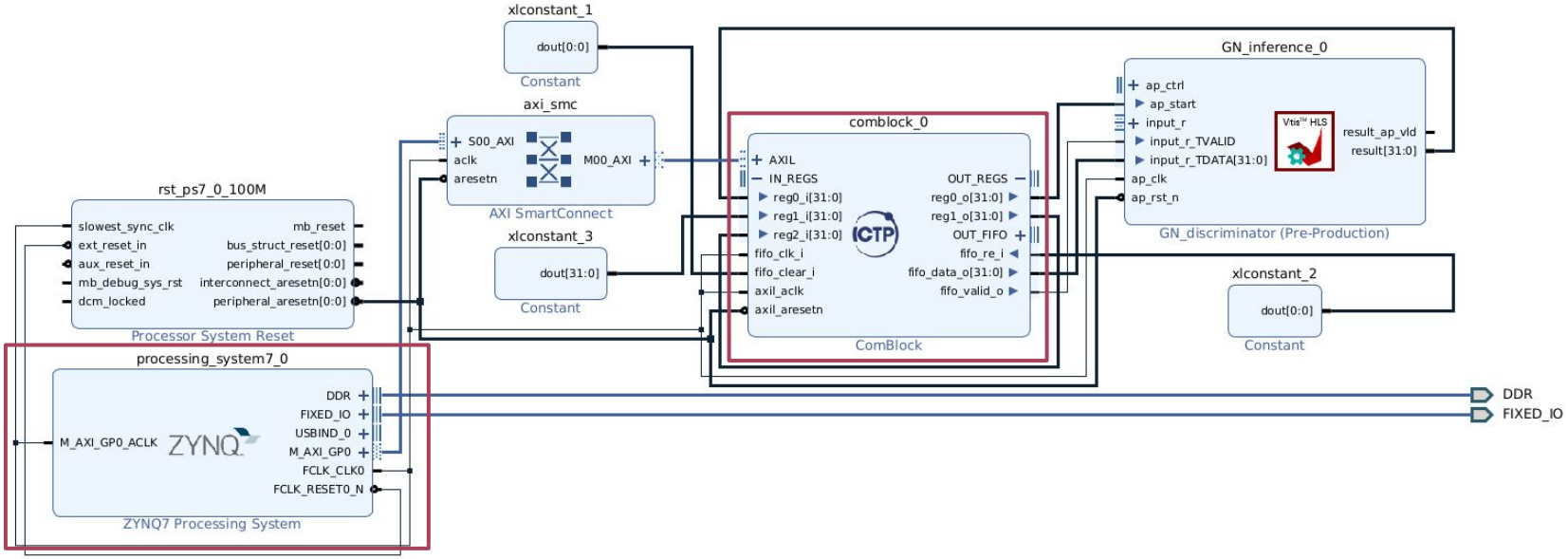
Hardware assessment framework



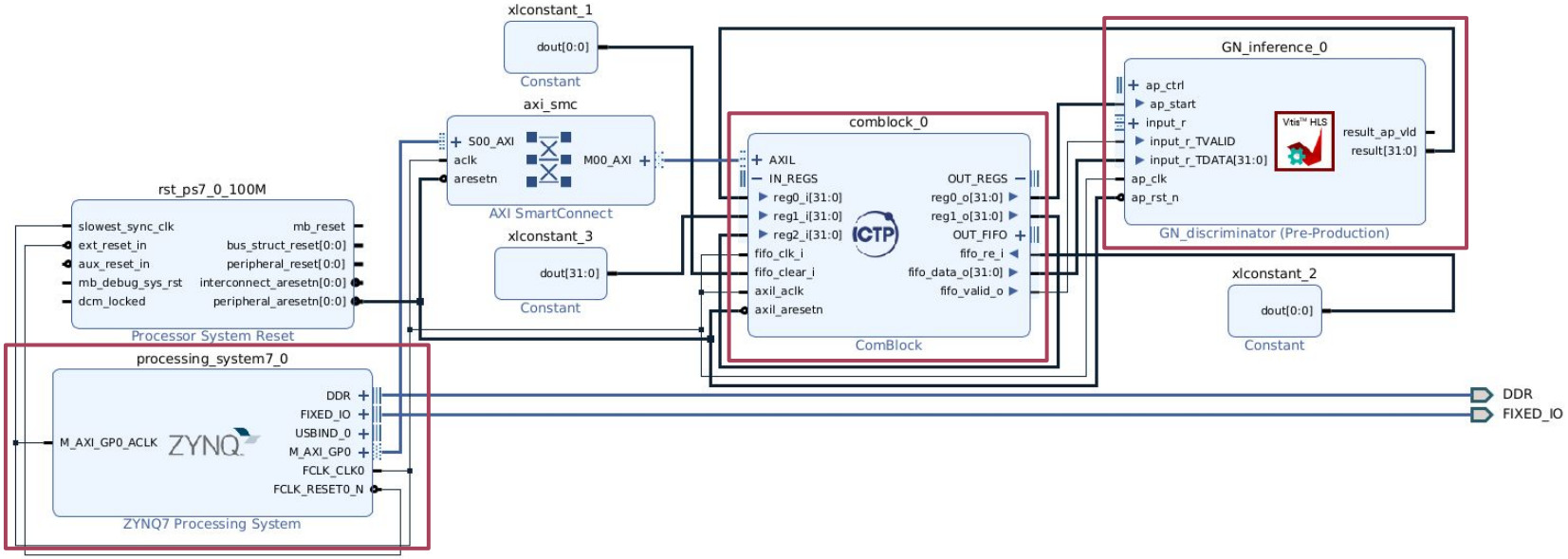
Hardware assessment framework



Hardware assessment framework



Hardware assessment framework

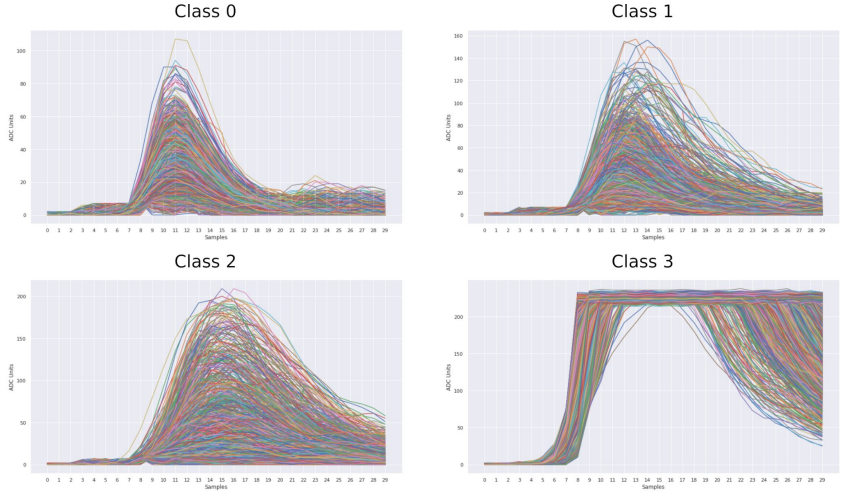


Some results…

Results

Applications

- **1D-MLP** is focused on 1-D signals: a pulse shape discriminator (PSD) based on a Multi-Layer Perceptron (MLP), to be used for event recognition in cosmic rays studies.



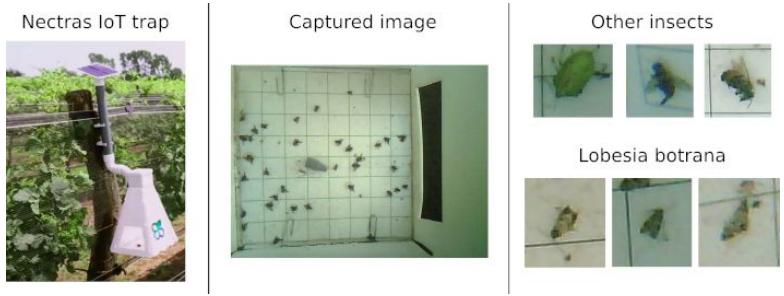
Molina, R. S., Morales, I. R., Crespo, M. L., Costa, V. G., Carrato, S., & Ramponi, G. (2023). An end-to-end workflow to efficiently compress and deploy DNN classifiers on SoC/FPGA. *IEEE Embedded Systems Letters*, *16*(3), 255-258.

Results

Applications

- Application in the field of object classification in 2D-images; its aim is moth classification in the context of pest detection. The solution **2D-CNN** is based on ad-hoc CNN trained with a dataset obtained from in-field traps through an IoT system. This application was further developed in (2D-VGG16) using a larger pre-trained teacher network (VGG16), and a public dataset
-

Ad-hoc CNN



2D-VGG16

Molina, R. S., Morales, I. R., Crespo, M. L., Costa, V. G., Carrato, S., & Ramponi, G. (2023). An end-to-end workflow to efficiently compress and deploy DNN classifiers on SoC/FPGA. *IEEE Embedded Systems Letters*, *16*(3), 255-258.

Results

TABLE I

EVALUATION. THE ACRONYMS IN THE TABLE ARE P: PARAMETERS, CR: COMPRESSION RATIO, NL: NUMBER OF LAYERS (CONSIDERING ONLY CONVOLUTIONAL AND FULLY-CONNECTED LAYERS), OA: OVERALL ACCURACY, SP: SPARSITY, L: LATENCY WITHOUT DATA TRANSFER.

	Teacher model		Compressed student models						Hardware implementation of the compressed student models					
Application	**P**	**OA[%]**	**P**	**CR**	**NL**	**OA[%]**	**Bits**	**SP[%]**	**FPGA**	**BRAM [%]**	**DSP[%]**	**FF[%]**	**LUT[%]**	**L [clk]**
1D-MLP	16,352	99.7	529	30.91	1	98.96	8	20.00	Artix-7	0.00	14.00	2.75	26.27	10
2D-CNN	723,499	99.19	3,699	195.59	11	94.11	8	50.00	ZCU102	5.70	10.90	6.00	18.79	17,411
2D-VGG16	14,818,695	98.87	3,207	4,898.50	16	96.67	8	60.00	ZCU102	10.84	17.54	6.02	15.40	18,005

Molina, R. S., Morales, I. R., Crespo, M. L., Costa, V. G., Carrato, S., & Ramponi, G. (2023). An end-to-end workflow to efficiently compress and deploy DNN classifiers on SoC/FPGA. *IEEE Embedded Systems Letters*, *16*(3), 255-258.

Results

TABLE I

EVALUATION. THE ACRONYMS IN THE TABLE ARE P: PARAMETERS, CR: COMPRESSION RATIO, NL: NUMBER OF LAYERS (CONSIDERING ONLY CONVOLUTIONAL AND FULLY-CONNECTED LAYERS), OA: OVERALL ACCURACY, SP: SPARSITY, L: LATENCY WITHOUT DATA TRANSFER.

	Teacher model		Compressed student models						Hardware implementation of the compressed student models					
Application	**P**	**OA[%]**	**P**	**CR**	**NL**	**OA[%]**	**Bits**	**SP[%]**	**FPGA**	**BRAM [%]**	**DSP[%]**	**FF[%]**	**LUT[%]**	**L [clk]**
1D-MLP	16,352	99.7	529	30.91	1	98.96	8	20.00	Artix-7	0.00	14.00	2.75	26.27	10
2D-CNN	723,499	99.19	3,699	195.59	11	94.11	8	50.00	ZCU102	5.70	10.90	6.00	18.79	17,411
2D-VGG16	14,818,695	98.87	3,207	4,898.50	16	96.67	8	60.00	ZCU102	10.84	17.54	6.02	15.40	18,005

Molina, R. S., Morales, I. R., Crespo, M. L., Costa, V. G., Carrato, S., & Ramponi, G. (2023). An end-to-end workflow to efficiently compress and deploy DNN classifiers on SoC/FPGA. *IEEE Embedded Systems Letters*, *16*(3), 255-258.

Results

TABLE I

EVALUATION. THE ACRONYMS IN THE TABLE ARE P: PARAMETERS, CR: COMPRESSION RATIO, NL: NUMBER OF LAYERS (CONSIDERING ONLY CONVOLUTIONAL AND FULLY-CONNECTED LAYERS), OA: OVERALL ACCURACY, SP: SPARSITY, L: LATENCY WITHOUT DATA TRANSFER.

	Teacher model		Compressed student models						Hardware implementation of the compressed student models					
Application	**P**	**OA[%]**	**P**	**CR**	**NL**	**OA[%]**	**Bits**	**SP[%]**	**FPGA**	**BRAM [%]**	**DSP[%]**	**FF[%]**	**LUT[%]**	**L [clk]**
1D-MLP	16,352	99.7	529	30.91	1	98.96	8	20.00	Artix-7	0.00	14.00	2.75	26.27	10
2D-CNN	723,499	99.19	3,699	195.59	11	94.11	8	50.00	ZCU102	5.70	10.90	6.00	18.79	17,411
2D-VGG16	14,818,695	98.87	3,207	4,898.50	16	96.67	8	60.00	ZCU102	10.84	17.54	6.02	15.40	18,005

Molina, R. S., Morales, I. R., Crespo, M. L., Costa, V. G., Carrato, S., & Ramponi, G. (2023). An end-to-end workflow to efficiently compress and deploy DNN classifiers on SoC/FPGA. *IEEE Embedded Systems Letters*, *16*(3), 255-258.

- Demo -

hls4ml: Bridging Machine Learning and FPGAs for Ultra-Fast Inference

Thank you!

Romina Soledad Molina, Ph.D.
MLab-STI, ICTP

Mar del Plata, Argentina - 2025 -