CNN.tar.gz

Compression for fast CNNs with FPGA

Graziella Russo-Offshell 2021

Simone Francescato-Stefano Giagu-Federica Riti-Graziella Russo-Luigi Sabetta-Federico Tortonesi

Sapienza Università di Roma
ATLAS ROMA1
INFN

The pipeline

The implementation of Deep Neural Networks on FPGAs can lead to several challenges.

The Level-0 trigger logic of ATLAS experiment for HL-LHC will be implemented on FPGAs.

Generation of a toy simulation of the detector and trigger realistic response.

Training of state-of-the-art Deep Neural Network architectures and optimisation of physics performance.

Model compression techniques are adopted.

Implementation of the compressed and simplified models on FPGAs.

Dataset

It is possible to arrange the RPC strips into image-like objects, to be used as input for ML convolutional models particularly suitable for muon tracks recognition.

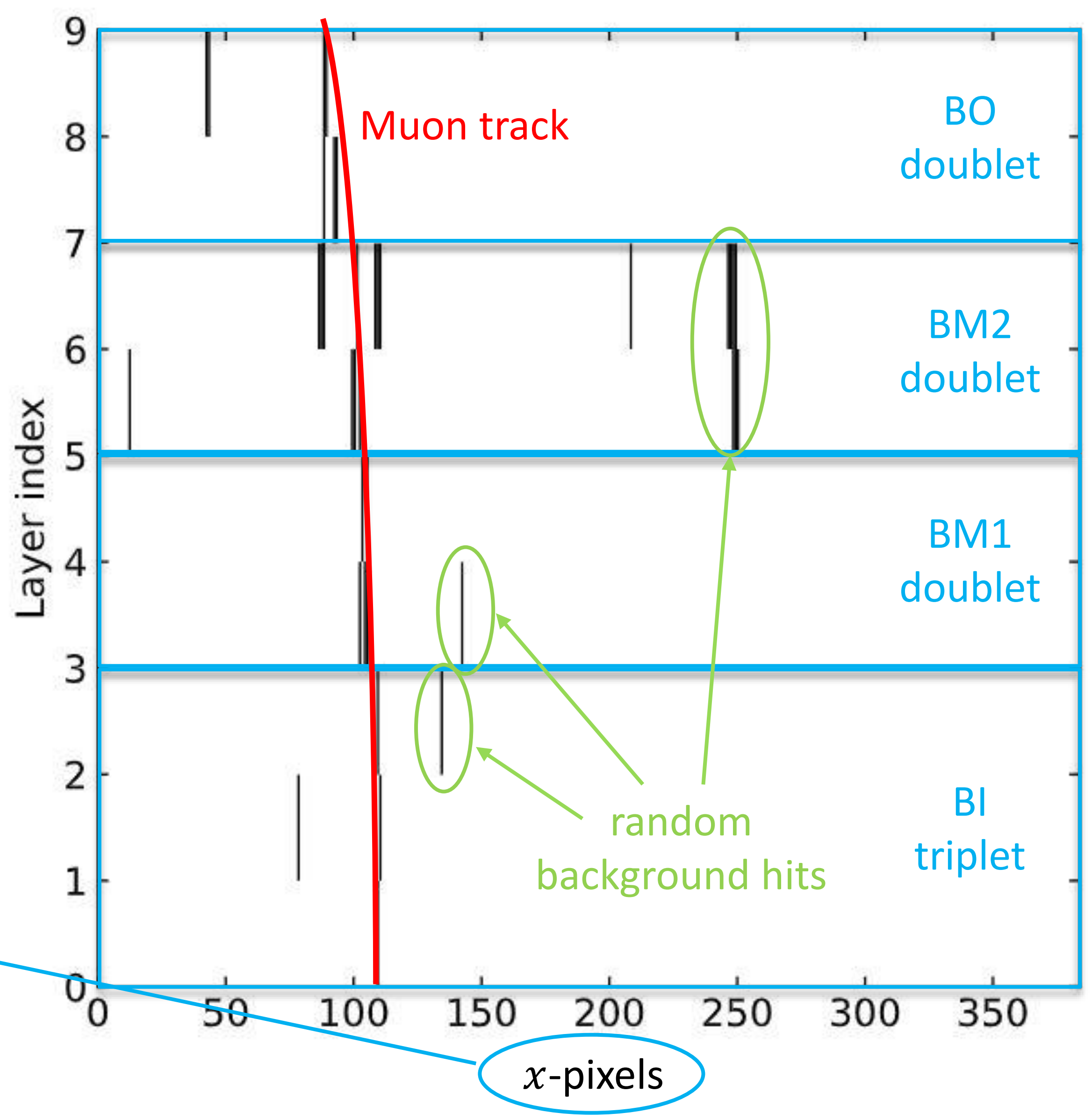
toy-events for ML dataset

Generation of 700k images with:

- single muon tracks ($p_T \in [3{,}20]$GeV) + random hit background;
- random hit background only.

Each vertical bin represents an RPC layer; each horizontal bin linearly maps the pseudorapidity into the pixel x-coordinate ($x \in [1{,}384]$).

Target labels: $[p_T, \eta]$



Model compression and simplification

The extreme experimental conditions impose some constraints on the algorithm architecture and performance:

- Fit within the Virtex UltraScale^{+} 13 FPGA resources; a lower occupancy is always recommended;
- Maximum latency of ~ 400 ns due to high event-rate of the experiment;
- Fake Rate (trigger efficiency of background events) < 2 ‰.

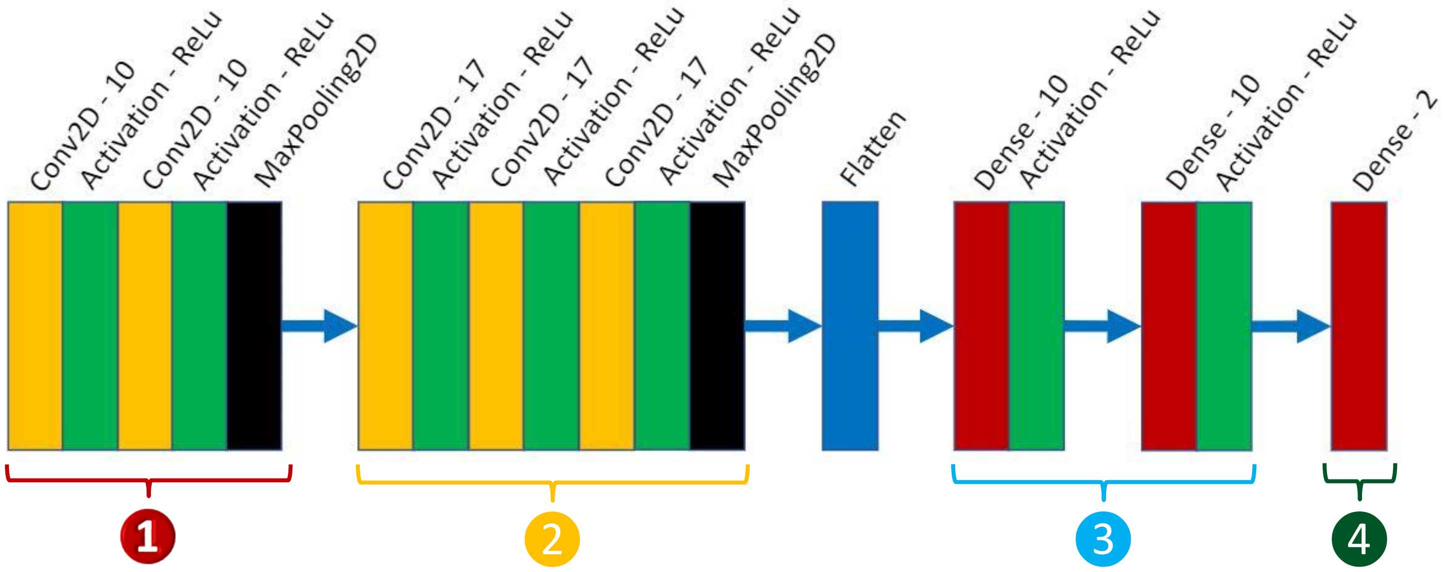
ADOPTED COMPRESSION TECHNIQUES

Compression by Knowledge Distillation

Compression by Quantization

Compression by Fragmentation of the Input

Teacher model



Knowledge Distillation (KD)

A relatively big Teacher model is used to teach the smaller Student model, by knowledge transmission during the Student training phase.

The Teacher model is based on a simplified version of the VGG architecture well suited for the task.

1° convolutional block

- Two *Conv2D* layers with 10 filters each;
- ReLu activation;
- final (1,2) *MaxPooling* layer.

2° convolutional block

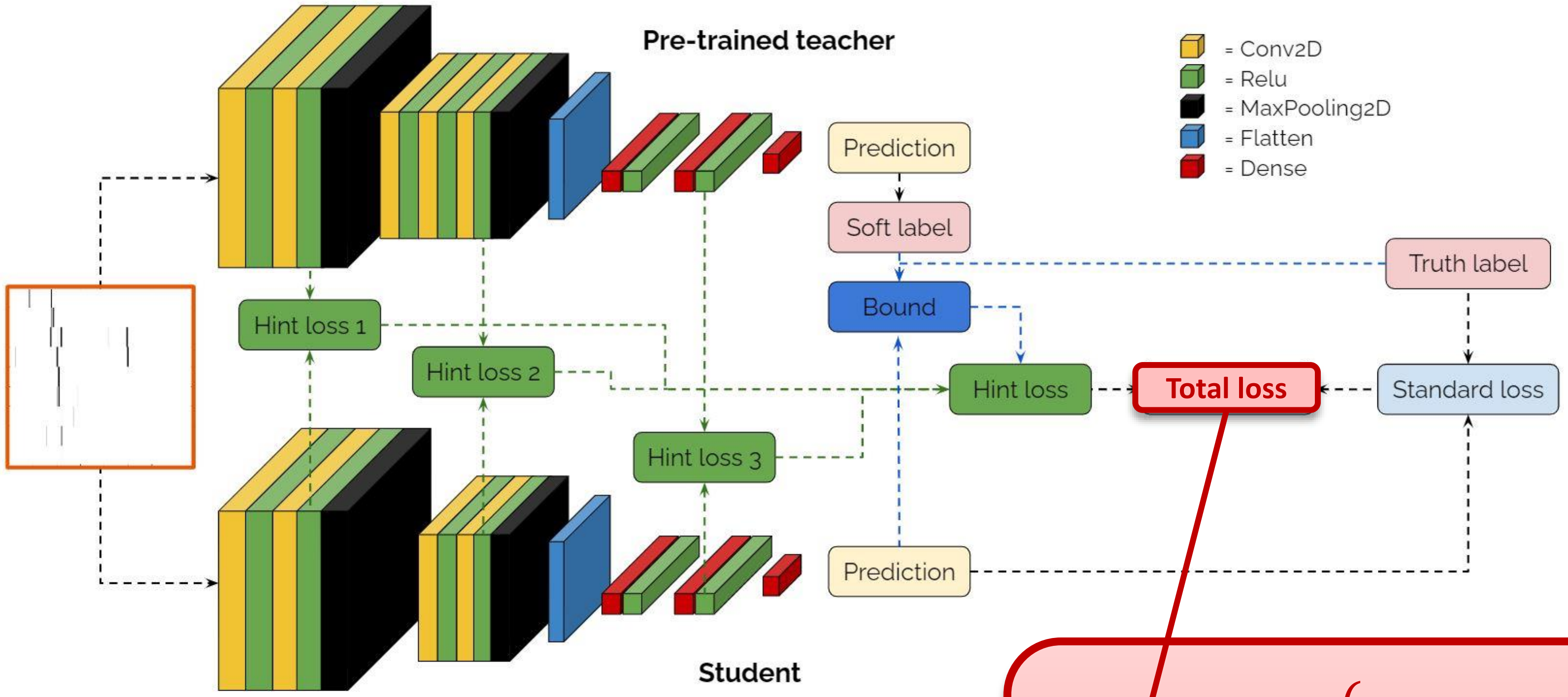
- Three *Conv2D* layers with 17 filters each;
- ReLu activation;
- final (2,2) *MaxPooling* layer.

Dense block

- Two *Dense* layers with 10 neurons each;
- ReLu activation.

Output: $[p_T, \eta]$

Student model training



During the training the Student model learns from the Teacher via the 3 *hints* ($H_i = \left\|A_i - T_i^H\right\|^2$, i = 1,2,3) added to the final loss (L).

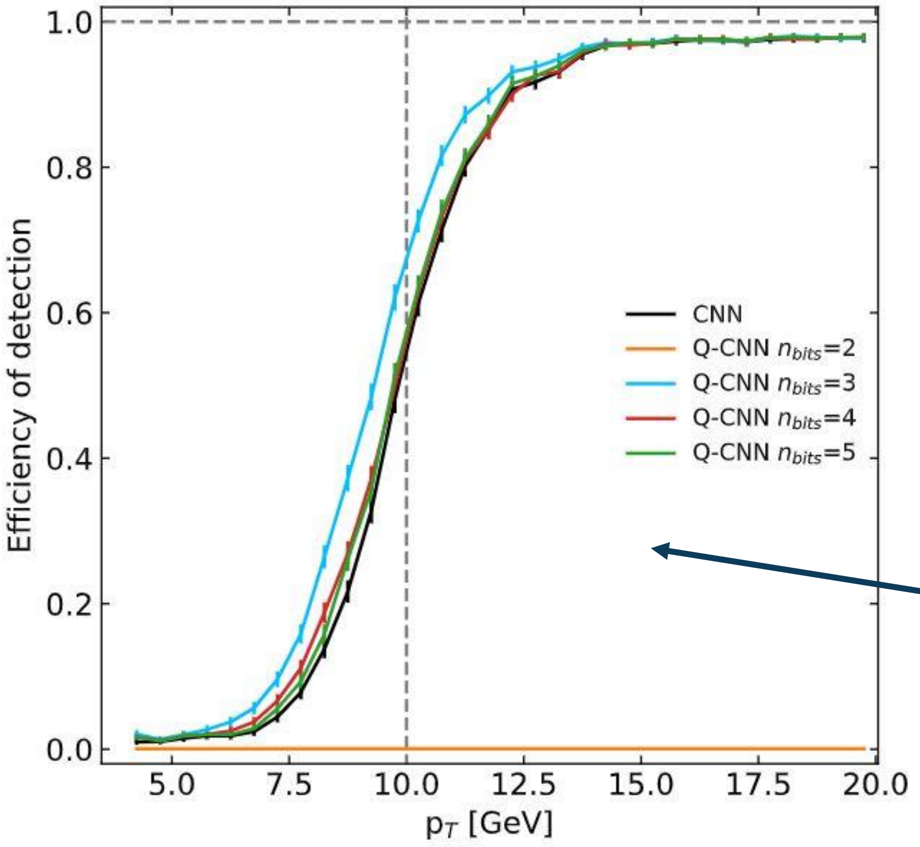
Adaptation layers are introduced in order to match the Teacher and Student intermediate outputs.

y = true labels
y_S = Student predictions
y_T = Teacher predictions
A_i = output of the i^{th} Adaptation layer
T_i^H = output of the i^{th} Teacher layer used for the hint
γ_i = tuned weight for each hint

$$L(y, y_S, y_T) = \begin{cases} \|y - y_S\|^2 + \sum_{i=1}^{3} \gamma_i H_i & \text{if } \|y - y_T\|^2 < \|y - y_S\|^2 \\ \|y - y_S\|^2 & \text{otherwise} \end{cases}$$

Teacher quantization

The model weights and activations (except the output layer) are not trained in the usual 32 or 64 bit precision floating-point aritmethic, but by fixing a lower number of bit $n_{bits} = 2, 3, 4, 5$.



Efficiency curve for 10 GeV p_T threshold

- Each bin corresponds the fraction of events which the algorithm predicts to have a $p_T \geq 10$ GeV;
- at low p_T the curve should be the closest as possible to zero;
- at high p_T it should reach the plateau efficiency of 1.

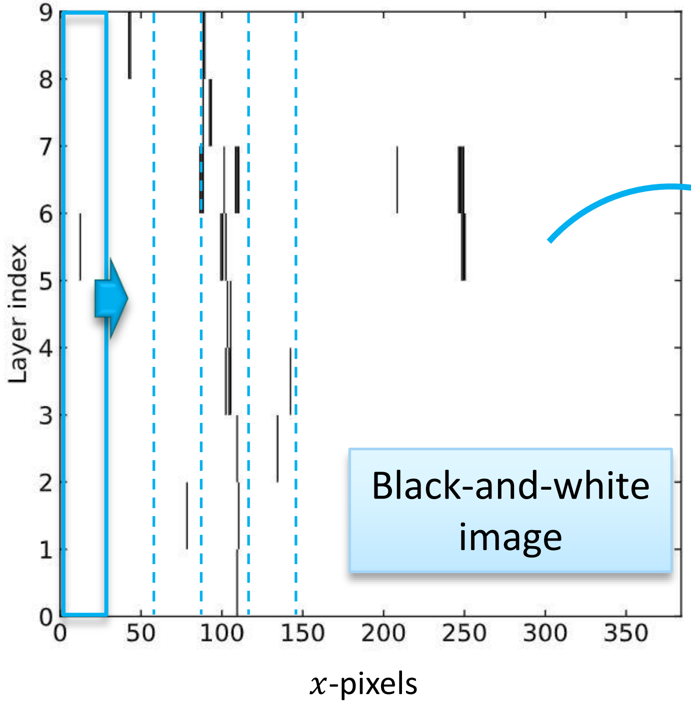
For $n_{bits} = 2$, the quantization is too aggressive and the network is so unstable that it is unable to learn the task.

For $n_{bits} > 2$, the degradation is progressively decreasing.

Compression by Fragmentation of the Input

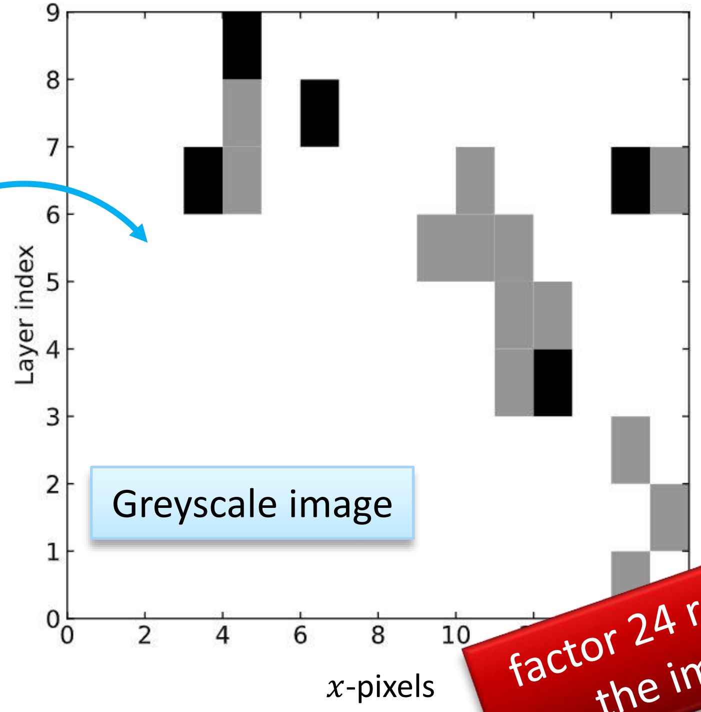
The size of the input images is reduced from $9x384$ to $9x16$ with a specified information-aware technique, in order to reduce the latency with minimal information loss.

Low p_T muon tracks bend within ~ 30 pixel
$\Rightarrow$ 9x32 fixed windows



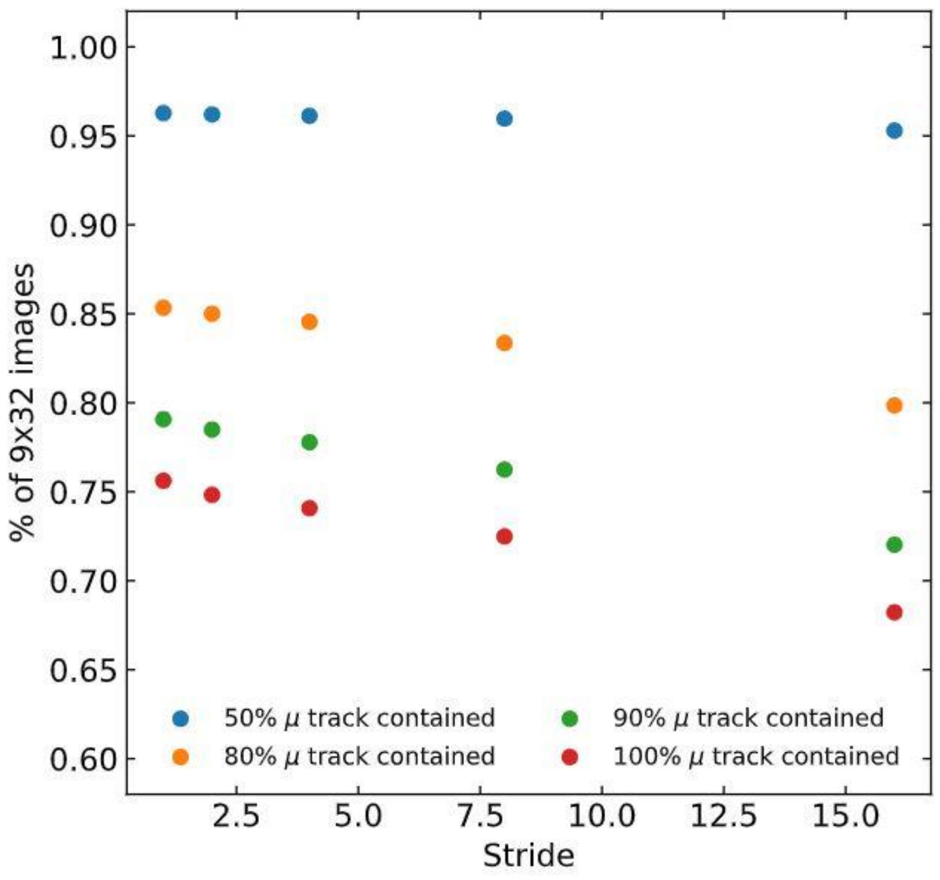
The image with highest number of non-empty pixels is selected

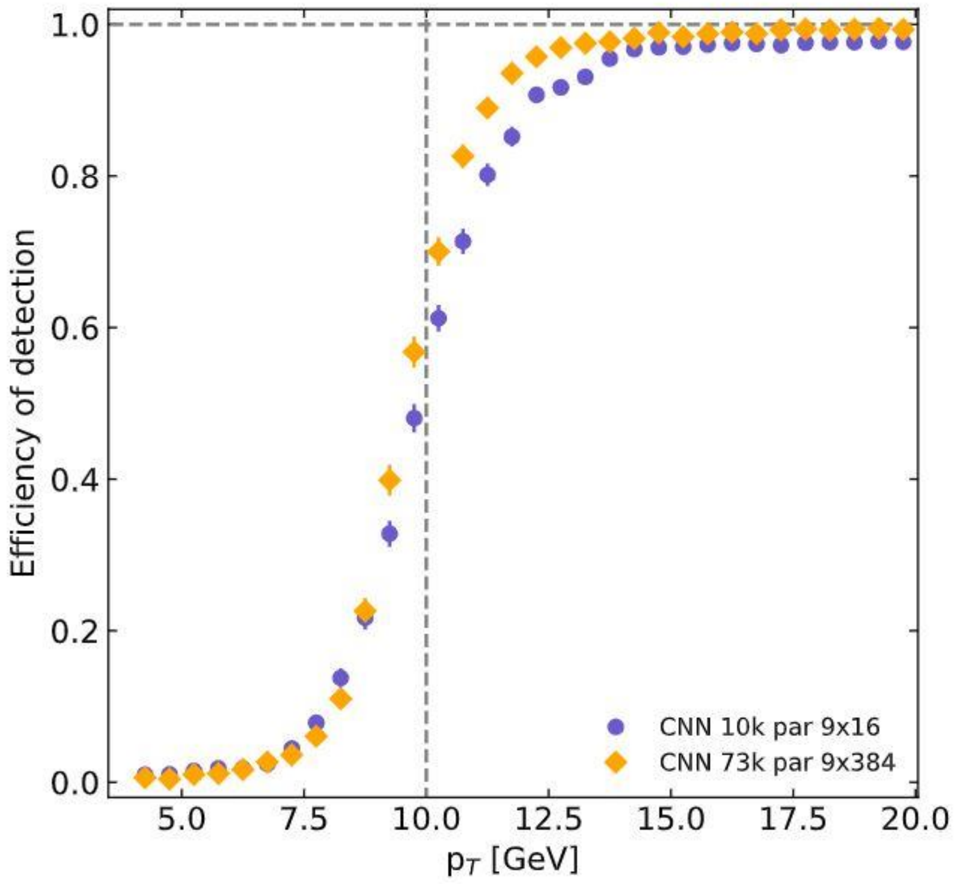
Further image compression through (1,2) AveragePooling $\Rightarrow$ 9x16 final image



factor 24 reduction in the image size

During the image-fragmentation, inevitably part of the muon track gets lost...





... BUT the performance is still good!

The different techniques are applied to the following models and then performance is evaluated:

Fragmentation

Distillation

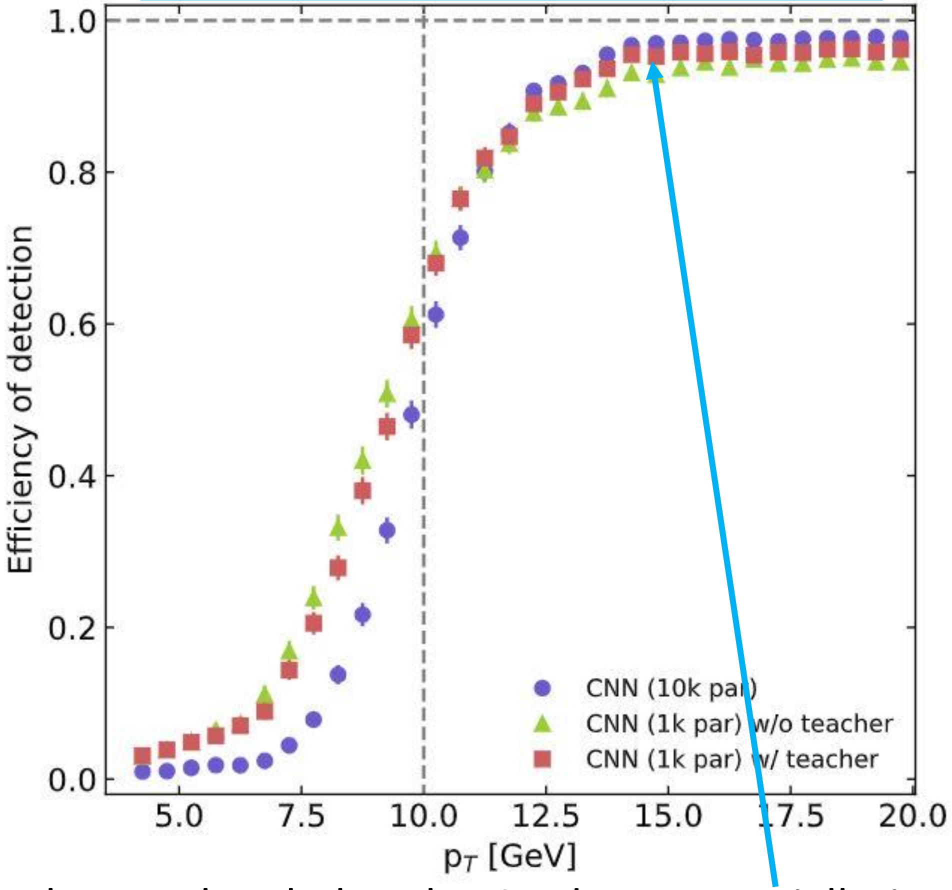
Quantized

not Quantized

Layer type	Teacher 9x384 Output shape	Teacher 9x384 Weights	Teacher 9x16 Output shape	Teacher 9x16 Weights	Student 9x16 Output shape	Student 9x16 Weights
Input	(9, 384, 1)	0	(9, 16, 1)	0	(9, 16, 1)	0
Conv2D	(9, 384, 10)	100	(9, 16, 10)	100	(7, 14, 1)	10
Conv2D	(9, 384, 10)	910	(9, 16, 10)	910	(5, 12, 1)	10
MaxPooling2D	(9, 192, 10)	0	(9, 8, 10)	0		
	Activation: ReLU, padding: same		Activation: ReLU, padding: same		Activation: ReLU, padding: valid	
Conv2D	(9, 192, 17)	1547	(9, 8, 17)	1547	(3, 10, 6)	60
Conv2D	(9, 192, 17)	2618	(9, 8, 17)	2618	(1, 8, 6)	330
Conv2D	(9, 192, 17)	2618	(9, 8, 17)	2618		
MaxPooling2D	(4, 96, 17)	0	(4, 4, 17)	0		
	Activation: ReLU, padding: same		Activation: ReLU, padding: same		Activation: ReLU, padding: valid	
Flatten	6528	0	272	0	48	0
Dense	10	65290	10	2730	10	490
Dense	10	110	10	110	10	110
	Activation: ReLU		Activation: ReLU		Activation: ReLU	
Dense	2	22	2	22	2	22
Model total		73215		10655		732

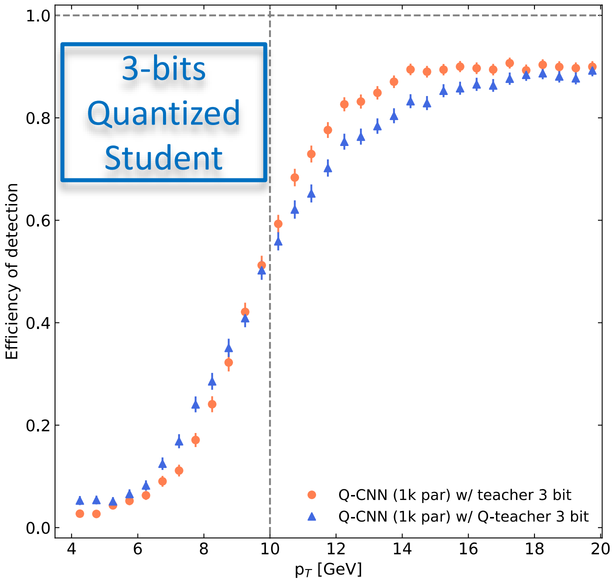
Study of the best KD approach

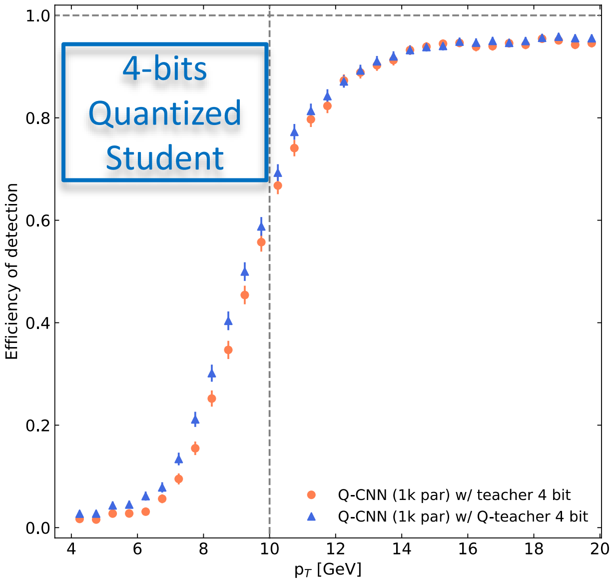
Not Quantized Student and Teacher



The Teacher helps the Student, expecially in reaching a higher plateau.

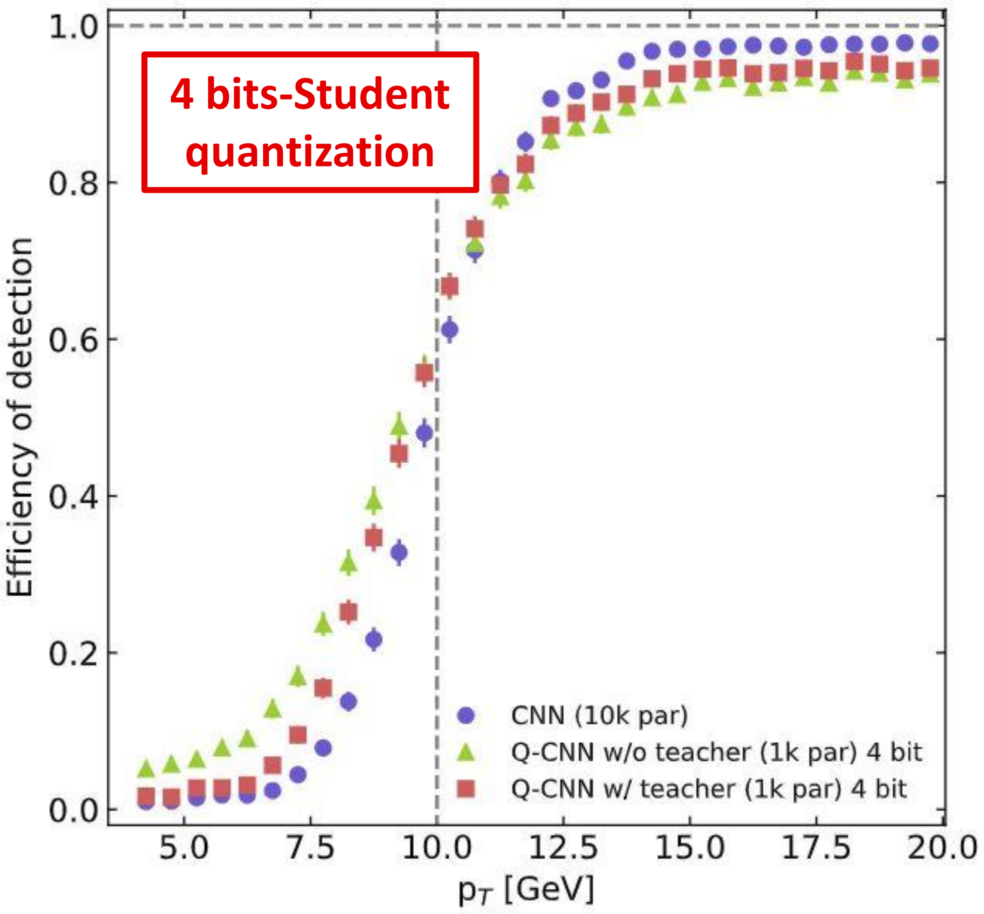
Quantized Student + Teacher → Quantized / not Quantized

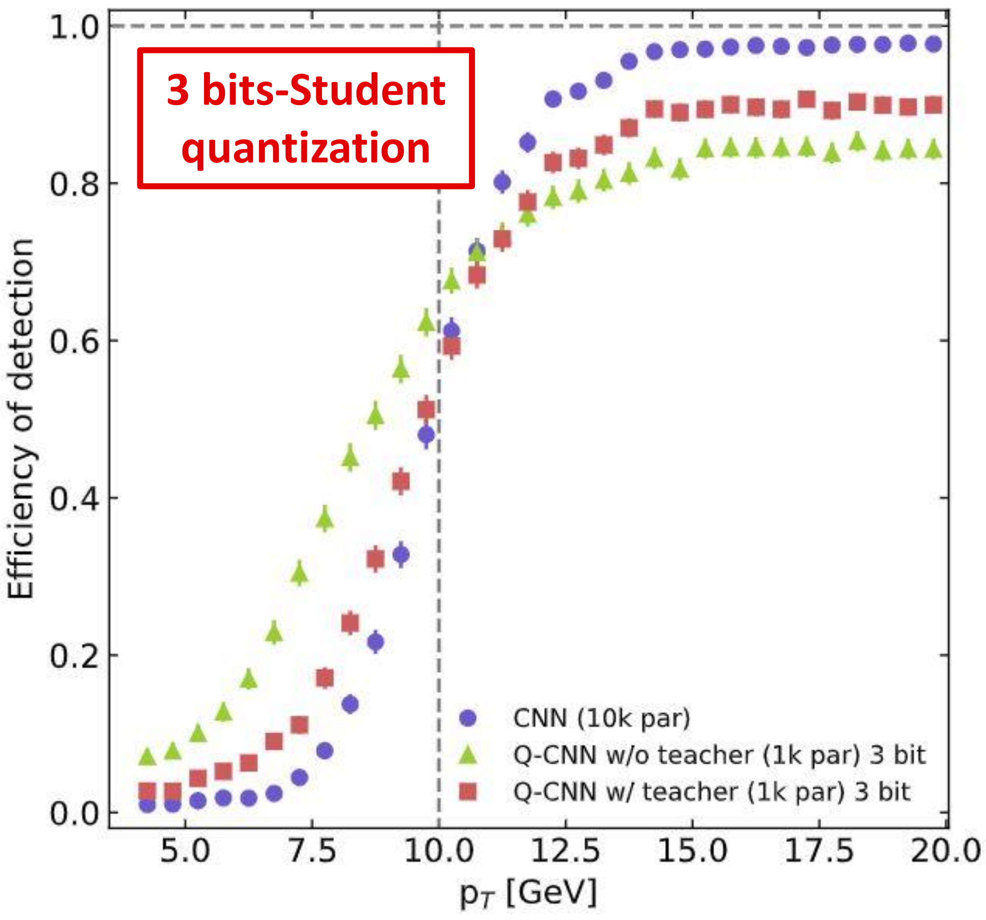




When the Student is quantized, the not-quantized Teacher is more helpful.

Teacher and quantized-Student comparison



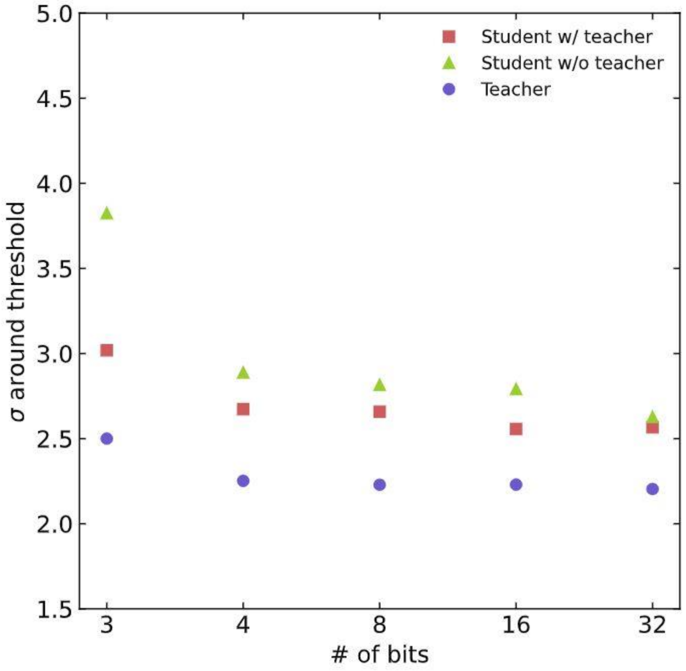


The more aggresive the quantization is, the more helpful the Teacher hints are.

Quantities relative to physics performance are used as test bench for the comparison between Student models with increasing n_{bits}.

Resolution around threshold

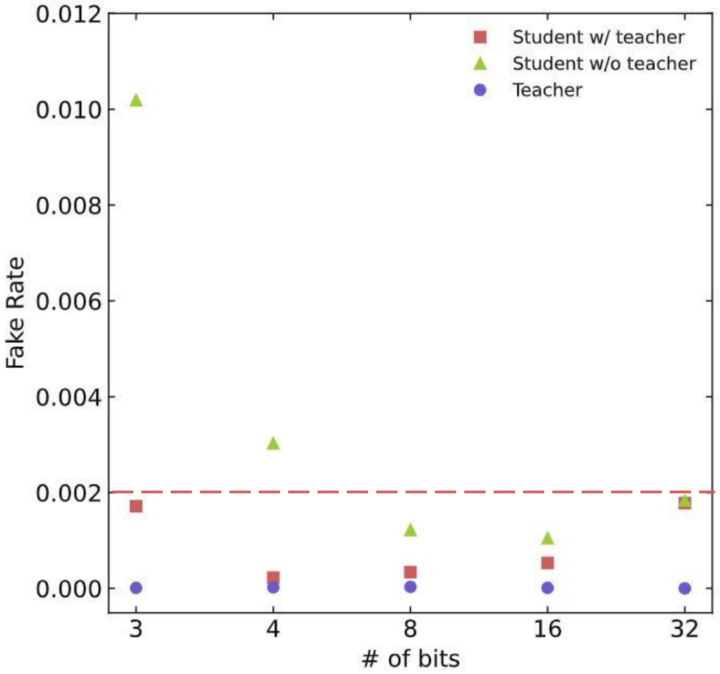
$7{,}5\ \text{GeV} < \sigma_{p_T} < 12{,}5\ \text{GeV}$



With aggressive quantization the resolution gets worse; the hints of the Teacher helps mitigate the resolution degradation.

Fake Rate

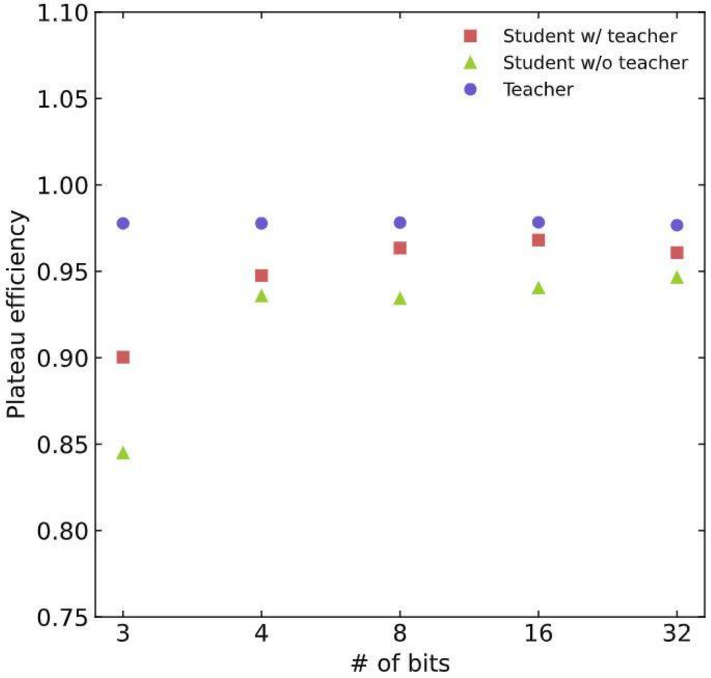
Trigger rate of background only events



The Teacher helps all the Students in reaching a Fake Rate lower than 2‰, as imposed by the experimental framework.

Plateau efficiency

Mean efficiency when $p_T > 17$ GeV



The improvements from the Knowledge Distillation are clearly visible.

Implementation on FPGA

The implementation of the different architectures is performed through HLS4ML library and Vivado HLS tool which translate a Tensorflow model into VHDL code.

Model (9×16)	BRAM	DSPs	FF	LUT	Latency (cycles)
Teacher	1123	31.7 k	2.4 M	265.6 k	640
Student 32 bit	171	3.8 k	247 k	31 k	222
QStudent 4 bit	11	6	14.3 k	29.5 k	183
QStudent 3 bit	11	0	11.6 k	23.3 k	182

Assuming a typical 2 ns clock cycle, the latency requirement is reached only by Quantization with $n_{bits} \leq 4$.

Model (9×16)	BRAM	DSPs	FF	LUT
Teacher (%)	20	258	69	15
Student 32 bit (%)	3	31	7	1
QStudent 4 bit (%)	$\sim$0	$\sim$0	$\sim$0	1
QStudent 3 bit (%)	$\sim$0	$\sim$0	$\sim$0	1

The Student models have a very low percentage of occupancy on the FPGA in question; the Teacher model, instead, cannot be implemented.

Model (9×16)	BRAM	DSPs	FF	LUT	Latency (cycles)
Student 32 bit	6.6	8.3	9.7	8.5	2.9
QStudent 4 bit	102	5.28 k	168	9	3.5
QStudent 3 bit	102	nd	218	11.4	3.5

Compression factors of the Students with respect to the Teacher model.

Conclusions

- **Fragmentation of the Input** → The input size is reduced considerably without significant efficiency losses.

  Latency requirement ✓

- **Quantization of the model** → Aggressive quantisation helps in greatly reducing the resources occupancy.

  Resources occupancy ✓

- **Knowledge Distillation** → The Teacher supports the Student models during training; the losses in the efficiency curve due to degradation are recovered.

  Fake Rate < 2‰ ✓

In conclusion, only with a mix of these different techniques all the requirements can be met.