# Benchmarking New Hardware for Machine Learning for Particle Physics

**Stefano Vergani**

Cavendish Laboratory, Department of High Energy Physics, University of Cambridge

## 1. LArTPC Detectors

- **Liquid argon time projection chamber (LArTPC) detectors measure ionisation tracks produced by charged particles inside a cryostat filled with liquid argon.**
- **The ionisation electrons drift in an electric field towards wire planes, where their charge is collected and measured.**
- **For this work, a fictitious LArTPC detector has been simulated with GEANT4.**

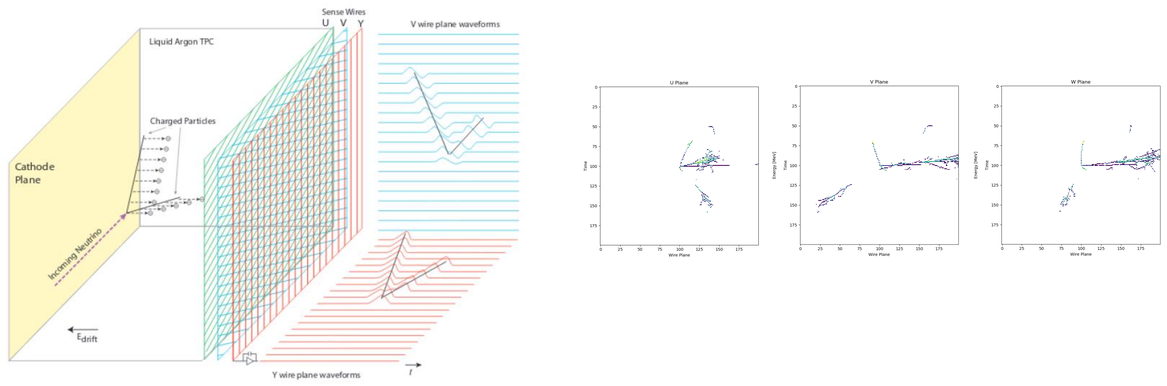


*Figure 1: On the left, sketch of a LArTPC taken from [1]. On the right, ionisation $e^-$ collected in the three different wire planes, from a fictitious LArTPC simulation.*

## 2. Processing Units

**A Processing Unit (PU) is a circuit which performs operations. There are several of them:**

- **Central Processing Unit (CPU): processes the basic instructions that drive a computer**
- **Graphics Processing Unit (GPU): used in graphic rendering software and in video games. In recent years, it has been used also to perform matrix calculations for Deep Learning (DL)**
- **Neural Processing Unit (NPU): a special groups of PUs designed specifically to perform matrix calculations for DL**

## 3. Tensor Processing Unit


**Tensor Processing Units (TPUs) are a subset of NPUs designed by Google to do tensor calculations. Edge TPU has portable sizes (65 mm x 30 mm). They have a very low power consumption and come at a cheap price (~80 USD).**

*Figure 2: Picture of an Edge TPU compared to the size of a clip. Taken from [2].*

## 4. Specs of Different PUs

**For this study, three different types of PUs will be used. They have different specifics, prices, and energy consumption and all of them must be taken into account when choosing the right one.**

| Processor Type | CPU | GPU | EDGE TPU |
|---|---|---|---|
| Model | Intel ® Core ™ i7-8550U @ 1.8 GHz | NVIDIA Tesla P100 16 GB | Coral Edge TPU |
| TDP* (in w) | 15 | 250 | 2 |
| Commercial Price | 99 USD | 7500 USD | 80 USD |

*Table 1: Specifics of all the PUs used for this study.*
* Thermal Design Power (TDP) represents the average power, in watts, the processor dissipates when operating at base frequency with all cores active under Intel-designed, high-complexity workload.

## 5. Neural Networks

**Only a certain subset of possible Neural Networks (NNs) can run on an Edge TPU. Among them, Residual Network 50 (ResNet-50) V2 and DenseNet-169 have been identified as the most promising. They have been trained with 10k images equally divided between neutrino Neutral Current (NC), $\nu_\mu$ Charged Current (numu CC), and $\nu_e$ Charged Current (nue CC) events. Validation has been done with another set of 2k images equally divided between NC, numu CC, and nue CC events. Training and validation have been done using the popular TensorFlow software.**

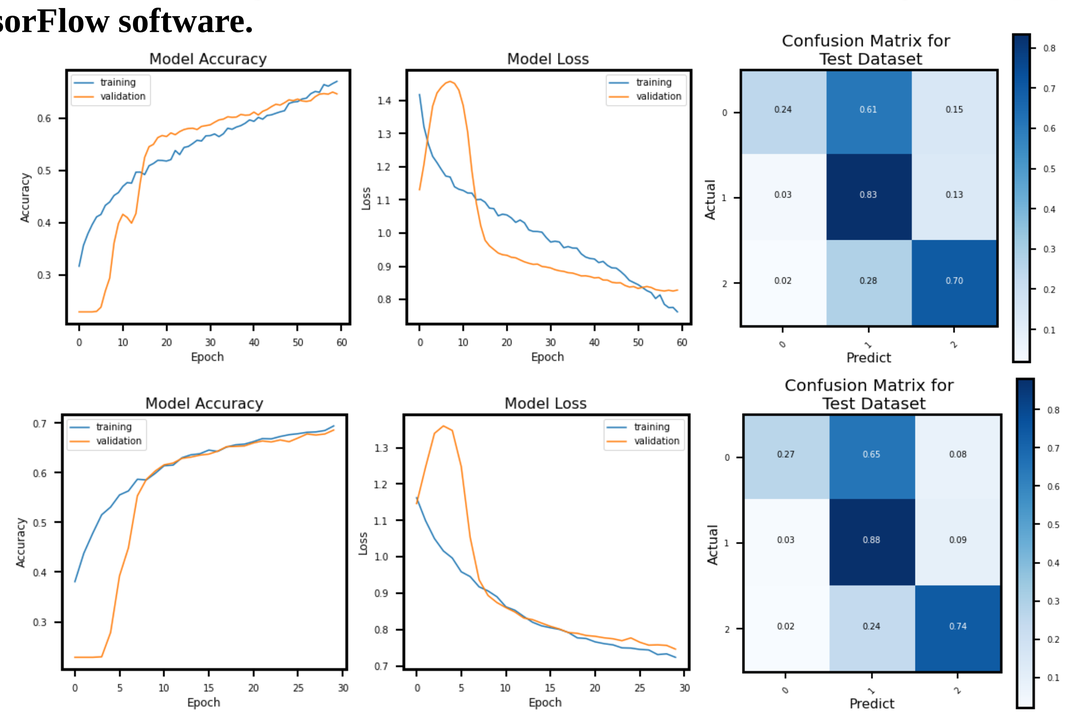


*Figure 3: Top from left to right: training/validation accuracy, loss, and confusion matrix for ResNet-50 V2. Bottom from left to right: the same for DenseNet-169.*

## 6. TensorFlow Lite

**The TensorFlow (TF) trained models have been subsequently converted to all the available TF Lite optimizations to test their speed and accuracy on CPU, GPU, and Edge TPU. The idea behind TF Lite is to make a lighter and faster model using quantization, that is converting weights and/or activations into 8 bits or float 16. This could lead to some accuracy loss. Edge TPU works only with quantized models. Table 2 presents a comparison between different versions of TensorFlow Lite.**

| TF Lite Optimization | Supported Hardware | Size Reduction with respect to TF | Input/Output Tensor Type | Is the model quantized? |
|---|---|---|---|---|
| Not Optimized | CPU, GPU | ~70% | Float 32 | No |
| Post-Training Dynamic Range Quantization | CPU, GPU | ~90% | Float 32 | Only weights to 8 bits |
| Post-Training Float16 Quantization | CPU, optimized for GPU | ~80% | Float 32 | Only weights to float 16 |
| Post-Training Integer Quantization | CPU, GPU, Edge TPU (model must be specifically compiled) | ~90% | Float 32 / Int 8/ Uint 8 (Edge TPU only with Uint 8) | Weight and activations to 8 bits |

*Table 2: Comparison between features of different TF Lite Optimizations.*

## 7. Results and Future Work

- **TensorFlow Lite appears to be particularly useful for users with limited RAM on the CPU/GPU.**
- **Edge TPU could be an interesting solution for users who cannot afford a GPU. Research centers could equip each workstation with one. Interesting for future R&D in edge computing for particle physics.**
- **Future plans comprehend testing more NNs and a publication.**

| Model | Size (MB) | Accuracy (same for all hardware) | Speed on CPU | Speed on GPU | Speed on Edge TPU |
|---|---|---|---|---|---|
| TensorFlow | 274 | 64.15% | 40.88 ms | 4.91 ms | N.A. |
| TF Lite Not Optimized | 90 | 64.15% | 89.36 ms | 94.4 ms | N.A. |
| Post-Training Dynamic Range Quantization | 23 | 54.35% | 548.54 ms | 314.76 ms | N.A. |
| Post-Training Float 16 Quantization | 45 | 64.15% | 114.16 ms | 64.05 ms | N.A. |
| Post-Training Integer Quantization | 24 | N.A.* | 6.1 s | 9.3 s | 42.41 ms |

*Table 3: Results of the benchmark with ResNet-50 V2.*

| Model | Size (MB) | Accuracy (same for all hardware) | Speed on CPU | Speed on GPU | Speed on Edge TPU |
|---|---|---|---|---|---|
| TensorFlow | 157 | 68.6% | 48.04 ms | 1.83 ms | N.A. |
| TF Lite Not Optimized | 48 | 68.6% | 81.02 ms | 112.14 ms | N.A. |
| Post-Training Dynamic Range Quantization | 13 | 63.65% | 530.95 ms | 296.88 ms | N.A. |
| Post-Training Float 16 Quantization | 25 | 68.6% | 81.92 ms | 114.21 ms | N.A. |
| Post-Training Integer Quantization | 13 | N.A.* | 7.6 s | 7.2 s | 23.67 ms |

*Table 4: Results of the benchmark with DenseNet-169.*

**References:**

- [1] MicroBooNE Collaboration, JINST 12, P02017 (2017), ArXiv: 1612.05824v2
- [2] https://coral.withgoogle.com/products/accelerator

**Notes:**

*There is currently a known issue in the quantization of TF Lite models meaning that the accuracy of inference on the Edge TPU is not representative, and is hence not given here at this stage. The accuracy is expected to be identical between the CPU, GPU (as demonstrated) and Edge TPU, and this will be verified in the near future

Science & Technology Facilities Council