Data Compression Conference
22-25 March 2022

Less is More: Compression of Deep Neural Networks for Adaptation in Photonic FPGA Circuits

Eftychia Makri, Georgios Agrafiotis, Ilias Kalamaras, Antonios Lalas, Konstantinos Votis and Dimitrios Tzovaras

Information Technologies Institute / Centre for Research and Technology Hellas
6th km Xarilaou – Thermi, Thessaloniki, 57001, Greece

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Information Technologies Institute

Background

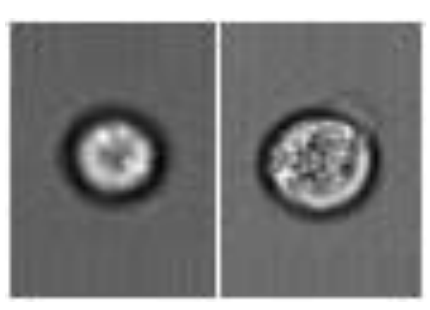
- Photonic circuits pave the way to **ultrafast computing** and **real-time inference** of applications with paramount importance, such as **imaging flow cytometry** (IFC).
- Nevertheless, current photonic FPGA implementations, exhibit **inherent restrictions** that consequently diminish the neural networks (NN) complexity that can be supported: **few inputs** (~50), few number of tunable basic units (**TBUs**) (~200), i.e. multiplication of **small matrices** (e.g. 11x11 for a fully connected layer).

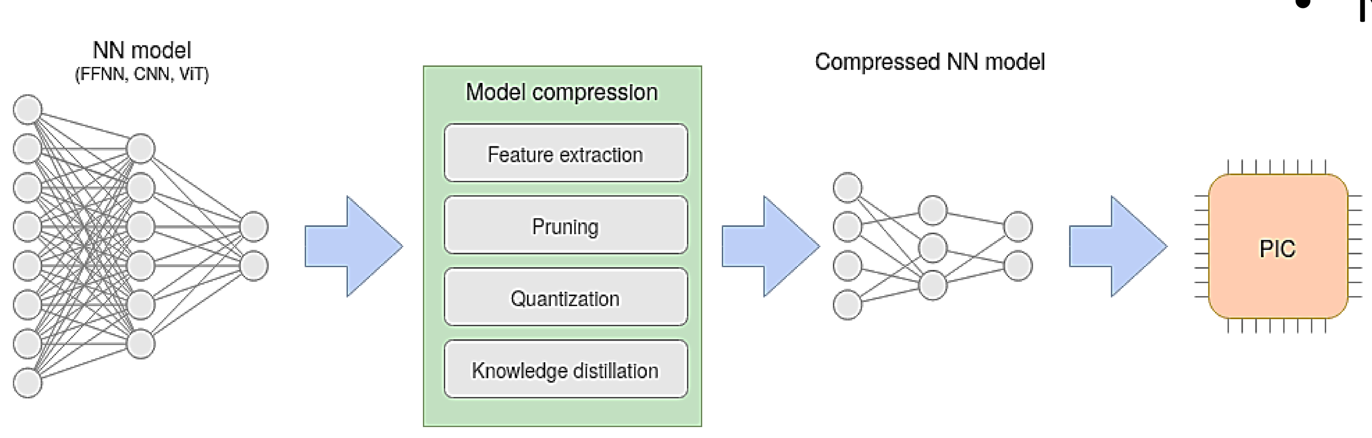
Goal

- Investigation of **compression schemes for NN models' size reduction,** under the prism of photonic FPGA limitations.

Architecture

- Examination of the **cell image classification** problem (small vs. large cells)





- Examined NN models:
  - Feed-forward (**FNN**)
  - Convolutional Neural Networks (**CNN**)
  - Vision transformers (**ViT**)
- Network compression methods:
  - **Feature extraction**
  - **Network pruning**
  - **Network quantization**
  - **Knowledge distillation (KD)**

Feature Extraction

- Convolutional Auto-Encoders (CAE)

Variant of **CNNs**, learning a **compressed representation** of the input (66x66 → 7x7).

✓ **Improves accuracy e.g. ViT: +3%**

✗ **Software preprocess/extra FPGA**

- Downsampling

Usage of **nearest-neighbor** interpolation (66x66 → 7x7)

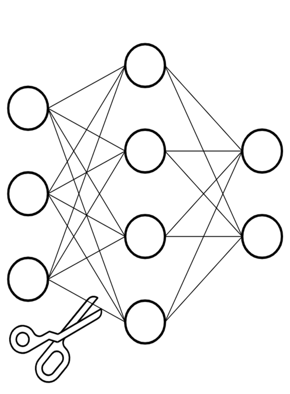
✓ **Completely on photonic FPGA**

✗ **Accuracy metrics decrease e.g. ViT: -9%**

Network pruning

- Neural network pruning, gradually **zeroing out model weights** throughout the training phase, to achieve model sparsity.

Before Pruning

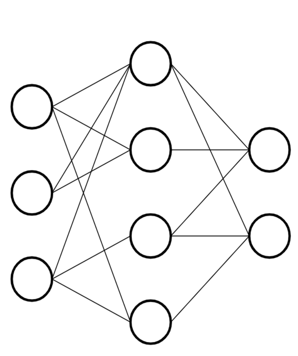


FNN Initial parameters 418 Accuracy: 94.8%

CNN Initial parameters 197 Accuracy: 93.7%

ViT Initial parameters 821 Accuracy: 95.6%

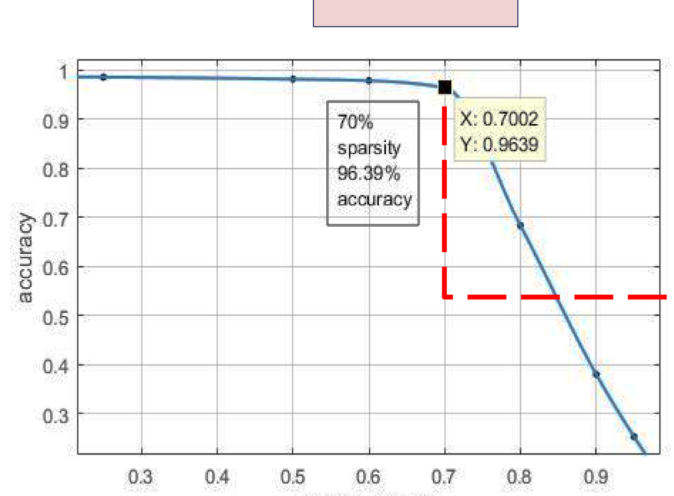
After Pruning



FNN Final parameters 230 Accuracy: 93.6%

CNN Final parameters 168 Accuracy: 91.7%

ViT Final parameters 493 Accuracy: 95.2%



There is a **trade-off** between accuracy and pruning sparsity factor

Network quantization

- **Post-training** model quantization can reduce latency, processing power, and model size with little degradation in accuracy.
- Weights get converted to types with **reduced precision**, such as **16-bit floats** or **8-bit integers**, shrinking models up to 4 times.

Example of **int8** quantization applied to a **FNN** & **ViT**:

FNN	Model Size (bytes)	Accuracy (%)	ViT	Model Size (bytes)	Accuracy (%)
Initial Model	3,735	94.8	Initial Model	13,896	95.6
Pruned k% 0.45	2,054	93.6	Pruned k% 0.40	8,321	95.2
Pruned and Quantized	1,467	**93.5**	Pruned and Quantized	5,943	**95.1**

There is a **trade-off** between accuracy and model size.
There is a little drop in accuracy (~**1.3%** for FNN, **0.4%** for ViT) but **the model is ~x2.5 smaller**, closely to a photonic adaptation.

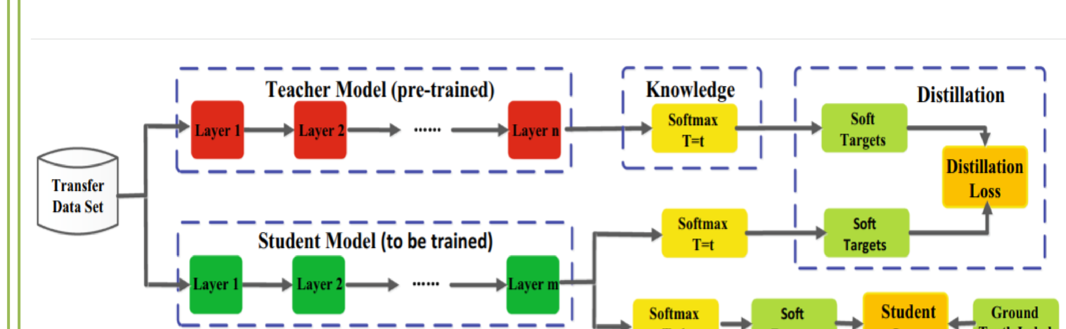
Knowledge Distillation

- KD is the process of training and **improving the performance** of a small deep learning model by utilizing a pre-trained larger one.

CAE & Pruned & Quantized CNN with KD
↓
Accuracy +1%

CAE & Pruned & Quantized FNN with KD
↓
Accuracy +0.4%

CAE & Pruned & Quantized ViT with KD
↓
Accuracy +0.3%



Conclusions

- The models depicted **adequate stability** to the applied compression, revealing NNs size reduction methods, could **significantly decrease the size of a model without sacrificing much the accuracy**, bringing the model closer the photonic FPGA limitations.
- By **combining compression methods**, which is suggested as the **primary technique**, beyond the models size reduction, low energy consumption and noise tolerance are achievable on the photonic FPGA.
  FFNN: -0.8% accuracy, -60% size; CNN: -1.4% accuracy, -39% size; ***ViT: -0.1% accuracy, -57% size***
- **Future work** includes implementation of reduced NN on photonic FPGA simulators, and **classification of time-stretched single pixel cytometry images**, coming from the Single Pixel Time-Encoded Microscopy (**STEM**) Imaging technique of the NEOTERIC project.
- Examination of an **optical patching scheme** and **compression** of it, for optical convolution and recurrence for adaptation on the photonic FPGA.

ACKNOWLEDGEMENT

This work was founded by the European's Union Horizon 2020 Research and Innovation Program through NEuromorphic Reconfigurable Integrated photonic Circuits as artificial image processor (NEoteRIC) under Grant Agreement No. 871330.

NEoteRIC

CONTACT INFORMATION

Dr. Antonios Lalas, Centre for Research and Technologies Hellas, lalas@iti.gr