Quantization Fundamentals

These slides are based on the content presented in Deep Learning AI's Quantization Fundamentals Course:

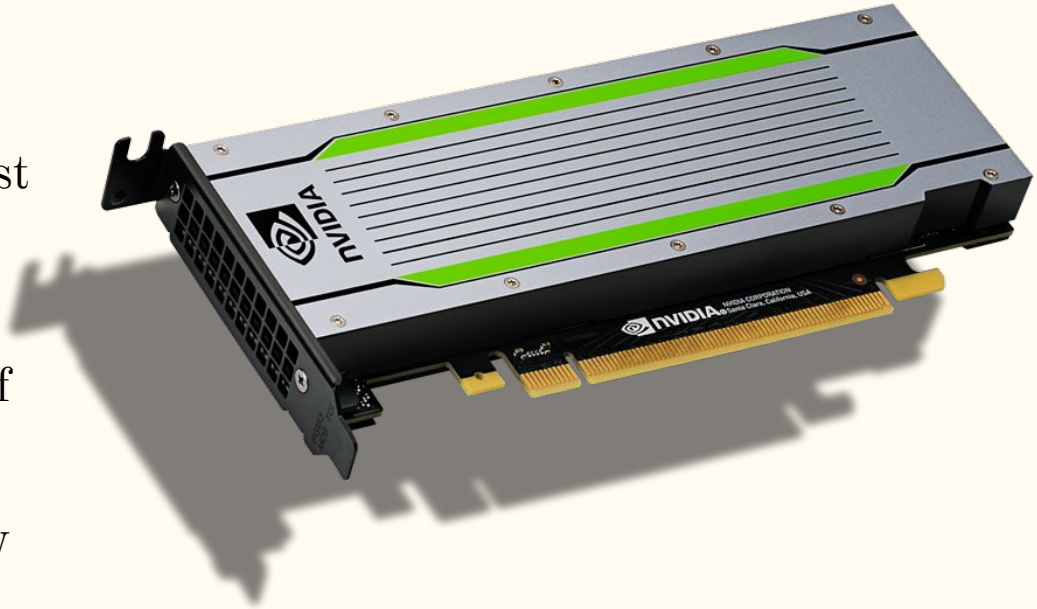
@melanimaheswar1

A gap exists between the largest models and the largest hardware.

A 7B model would need ~280 GB just to make the model fit on hardware.

Consumer-type hardware, such as NVIDIA T4 GPUs have just 16GB of RAM.

Thus, running these models efficiently is of great importance.



Quantization involves representing model weights in a lower precision.

Let's start by considering the small matrix on which stores some parameters of a small model.

13.5	14.3	8.5
-4.7	-3.2	-6.4
-0.4	1.3	3.73

FP32: 4 bytes to store each value, the default storing data type for most models

The matrix is stored in float32 and therefore has to allocate **4 bytes per parameters.**

4 times **8-bit precision**, meaning the total memory footprint of the matrix would be **36 bytes.**

Quantization involves representing model weights in a lower precision.

If we quantize the weight matrix in 8-bit precision (int8), we allocate 1 byte per parameter.

13.5	14.3	8.5
-4.7	-3.2	-6.4
-0.4	1.3	3.73

FP32: 4 bytes to store each value

FP32 to INT8

13	14	8
-5	-3	-6
-0	1	4

INT8: 1 byte to store each value

Quantization comes with a price!!

13	14	8
-5	-3	-6
-0	1	4

INT8: 1 byte to store each value

Thus, we'll in total need **only 9 bytes** to store the entire weight matrix.
However, this comes with a price, the quantization error.

0.5	0.3	0.5
-0.3	-0.2	-0.4
-0.4	0.3	0.23

The goal of SOTA quantization methods is to keep this error to a minimum and avoid any performance degradation.

Model compression techniques aside from quantization:

- *Pruning*
- *Knowledge distillation*

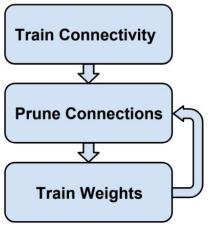


Figure 2: Three-Step Training Pipeline.

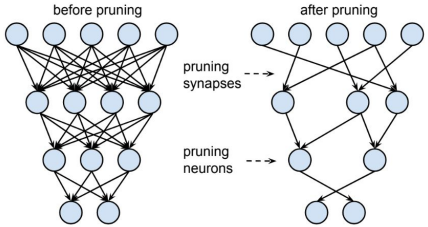


Figure 3: Synapses and neurons before and after pruning.

Pruning involves removing connections that do not improve the model

Layers are removed, based on metrics (ex: magnitudes of the weights)

Learning both Weights and Connections for Efficient Neural Networks

Song Han
Stanford University
songhan@stanford.edu

Jeff Pool
NVIDIA
jpool@nvidia.com

John Tran
NVIDIA
johntran@nvidia.com

William J. Dally
Stanford University
NVIDIA
dally@stanford.edu

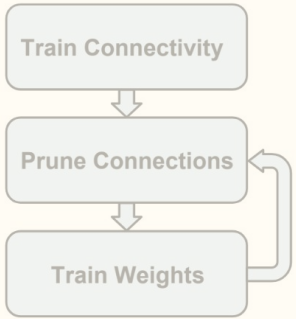


Figure 2: Three-Step Training Pipeline.

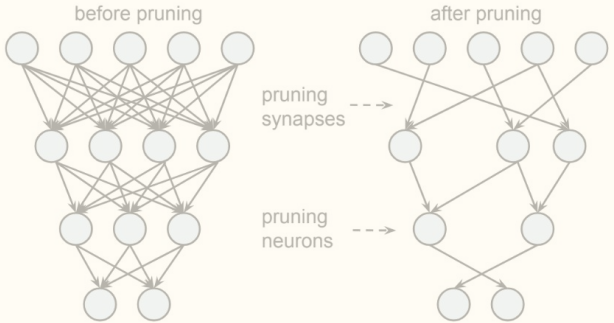
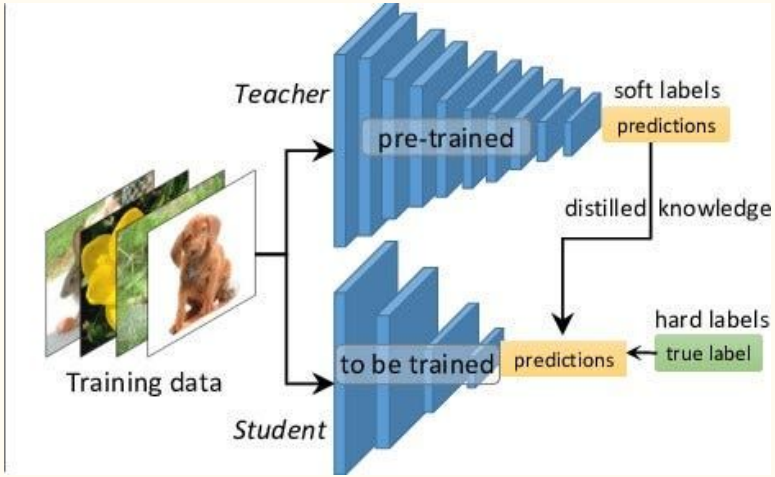


Figure 3: Synapses and neurons before and after pruning.

Knowledge Distillation



With **knowledge distillation**, a student model (the target-compressed model) is **trained with the use of the output from the teacher model in addition to the main loss term.**