Applied Data Science 2

DR ASHLEY SPINDLER

SEMESTER B, 2022

Lecture 9 - Outcomes

- Understand the principles of GPU Acceleration
- Explore computational bottlenecks, and ways to mitigate them

GPUs

What are GPUs?

- GPU – Graphics Processing Unit
- Most often related to videos games, photography and video production
- Often called the “soul” of a PC
- Now form the backbone of ML development

CPU vs GPU

CPU – Central Processing Unit

- Composed of a handful of “cores”
- Low Latency
- Perform serial calculations
- Can only perform a few tasks at once

GPU Graphics Processing Unit

- Large number of cores
- High memory throughput
- Performs parallel calculations
- Can perform 1000s of tasks at once

CPU vs GPU

CPU – Intel i9 series

- ~800 GFLOPS calculation speed
- 18 cores

GPU – NVIDIA Tesla A100

- 10-300 TFLOPS calculation speed
- Equivalent of 6912 cores

GPUs are specialists

- GPUs are highly specialised in performing very specific, often very simple calculations incredibly fast

- E.g. multiplying together matrices, performing convolutions—the GPU can perform thousands of these a second

GPUs are ideal for ML

- A neural network is a graph of often very simple calculations—a Dense layer is just a matrix multiplication and addition—which is performed many, many times
- Even simple models can achieve massive speed ups compared to CPUs—e.g. the MLP from tutorial 8 runs 3x faster on a Tesla K80, compared to CPU

Multiple GPUs can be used together

- Because GPUs perform operations in parallel by design, you can link up multiple GPUs together and increase calculation speed

- Not always beneficial, as you can hit slow downs from data transfer between GPUs

GPU Summary

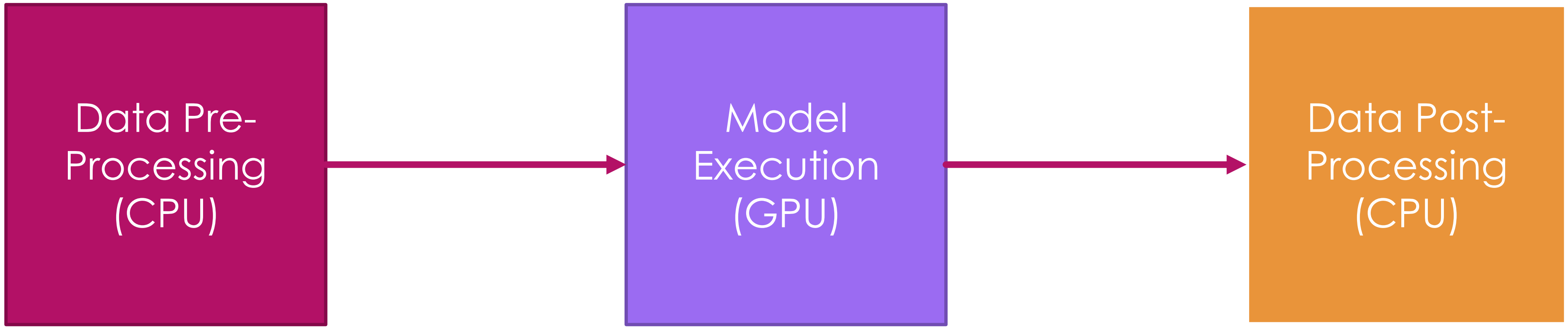
- GPUs utilise specialist hardware and algorithms to perform many thousands of simple calculations at the same time
- This makes them ideal for neural networks, which require many applications of matrix multiplications, additions, and other similar operations
- TensorFlow will automatically use a GPU if it is available, but we can also use multiple GPUs at once

Bottlenecks

In a perfect world…

- We would have very high utilisation of the CPU and GPU resources
- Our input pipelines would efficiently load, transform and move our data from the storage onto the CPU and GPU
- There would be no lag between GPU operations finishing on one batch of data, and the next batch arriving

Computation Pipelines

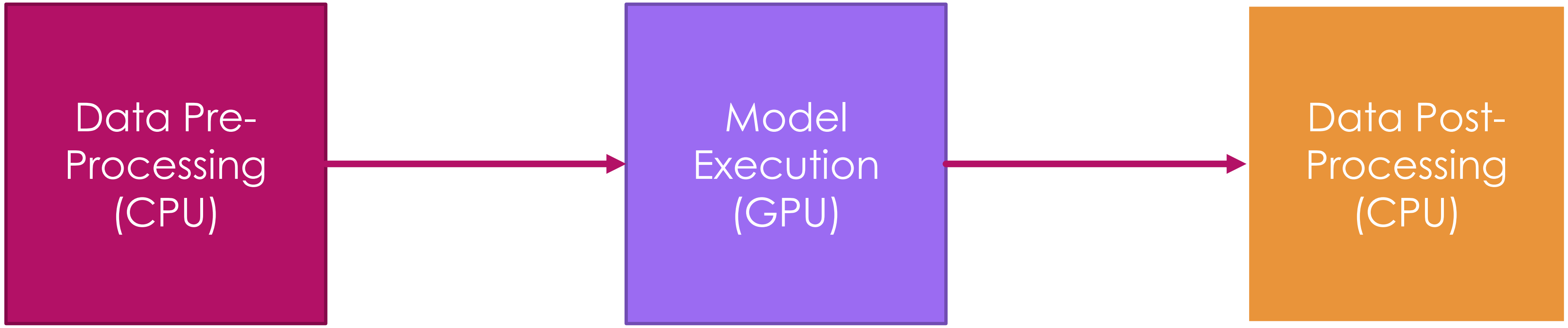


Data Pre-Processing (CPU)

Model Execution (GPU)

Data Post-Processing (CPU)

Computation Pipelines



Data Pre-Processing (CPU)

Model Execution (GPU)

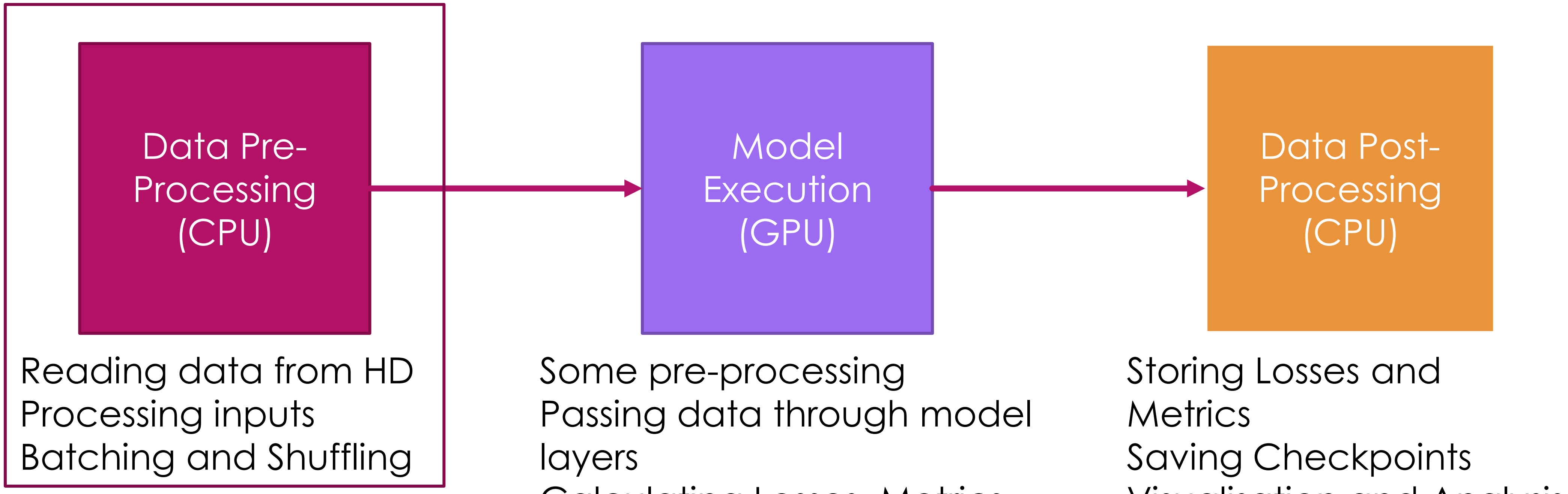
Data Post-Processing (CPU)

Reading data from HD
Processing inputs
Batching and Shuffling

Some pre-processing
Passing data through model layers
Calculating Losses, Metrics and Gradients
Updating Model Weights

Storing Losses and Metrics
Saving Checkpoints
Visualisation and Analysis

Computation Pipelines



Data Pre-Processing (CPU)

Model Execution (GPU)

Data Post-Processing (CPU)

Reading data from HD
Processing inputs
Batching and Shuffling

Some pre-processing
Passing data through model layers
Calculating Losses, Metrics and Gradients
Updating Model Weights

Storing Losses and Metrics
Saving Checkpoints
Visualisation and Analysis

Most Common Bottleneck is in CPU Pre-Processing

Processing Bottlenecks

- Idle time on GPU while CPU is fetching data, likewise CPU idle when GPU is running
- Serial data extraction where only a single file is loaded at a time
- Data transformations performed sequentially on data samples
- Repeating calculations each training step
- Transforming an entire dataset instead of individual batches

Processing Bottlenecks

- Prefetch the next batch of data while model is executing
- Extract data in parallel instead of in sequentially
- Parallelise data transformations
- Cache the results of time consuming transformations
- Batch datasets before performing transformations

TensorBoard Profiler Demo

Further Reading

- TensorFlow Data Performance Guide: https://www.tensorflow.org/guide/data_performance
- TensorBoard Profiler: https://www.tensorflow.org/guide/profiler
- TensorFlow GPU Performance: https://www.tensorflow.org/guide/gpu_performance_analysis

Questions???