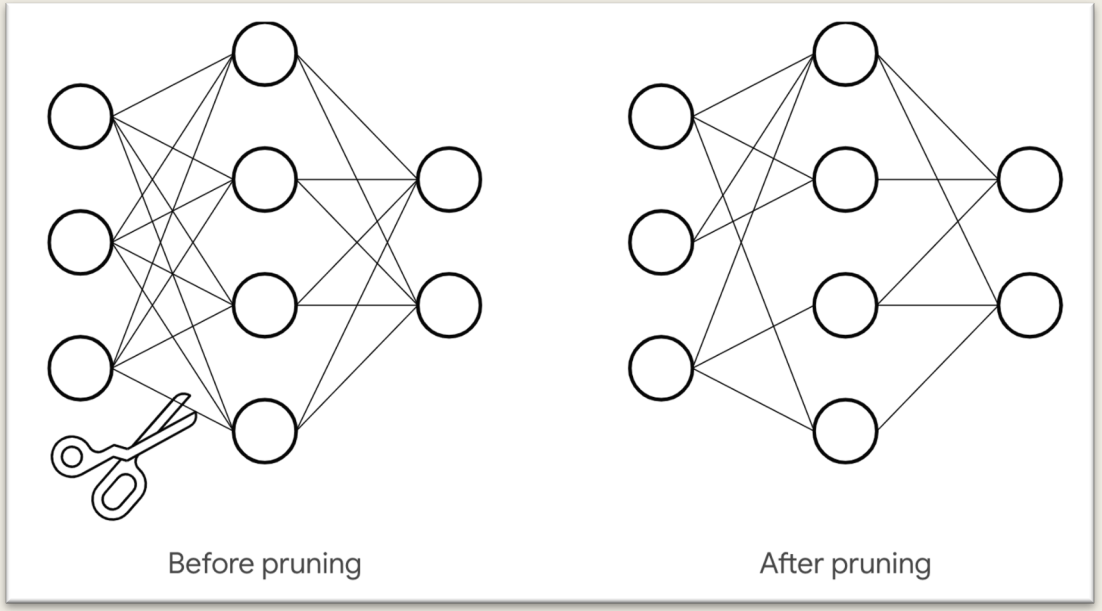
Model Compression with Bayesian Optimisation and PyTorch

Samuil Stoychev

Model Compression

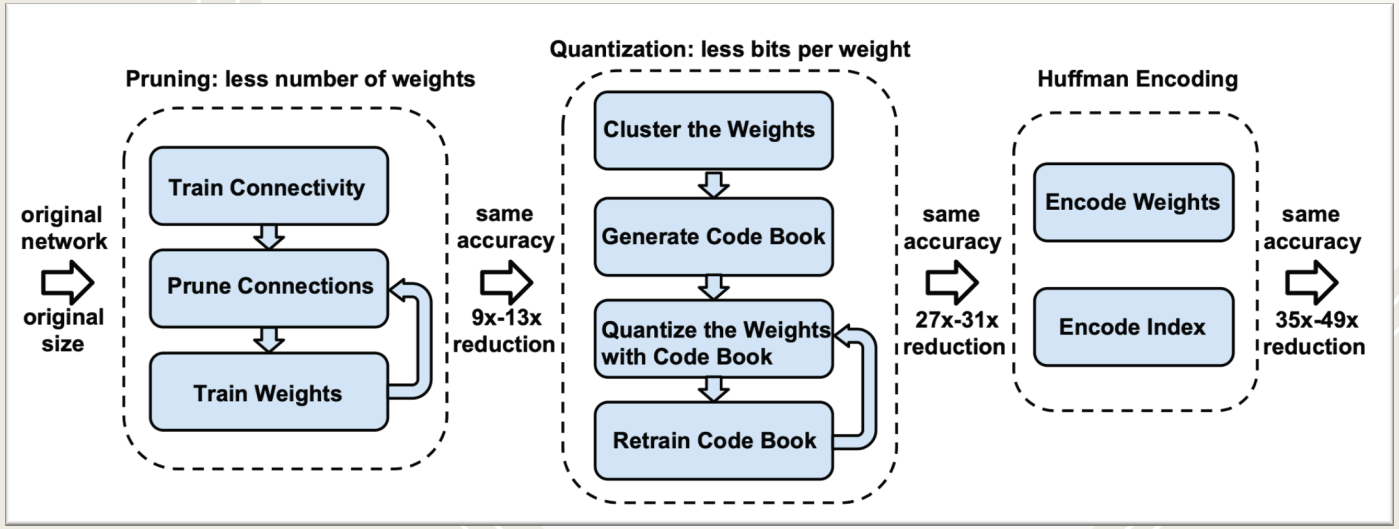
Model compression techniques aim to **reduce the resource consumption** of deep neural networks, while keeping accuracy high. Examples include:

- Pruning
- Quantization
- Huffman Coding



Tuning Model Compression

- In practice, model compression techniques tend to be **combined** into compression pipelines.
- Moreover, different compression techniques need to be configured with **various hyperparameters**.
- This can potentially lead to a **large search space**.



Goal of This Project

- Implement model compression for **PyTorch** models.
- Use **Bayesian optimisation** with **BoTorch** to select the best compression hyperparameters.
- Evaluate the performance of the compressed model against a baseline implementation.
- Evaluate the performance of Bayesian optimisation compared to other optimisation strategies – e.g. exhaustive search, random search.

Work Plan

So far:

- Familiarised myself with PyTorch.
- Checked compression techniques.

Next month:

- Set up baseline PyTorch models and implement the compression techniques.
- Apply Bayesian optimisation with BoTorch.
- Evaluate performance.

Questions?