CS 181AI
Lecture 24

Model Compression

Arthi Padmanabhan

Apr 17, 2023

Logistics

- Wednesday: working session
- Next Wednesday (4/26): Project presentations
  - 15 min each group

Today

- Methods of model compression
- Pruning Demo

Compression

- We want to make models smaller so they will lower:
  - Time to run
  - Memory
  - Energy

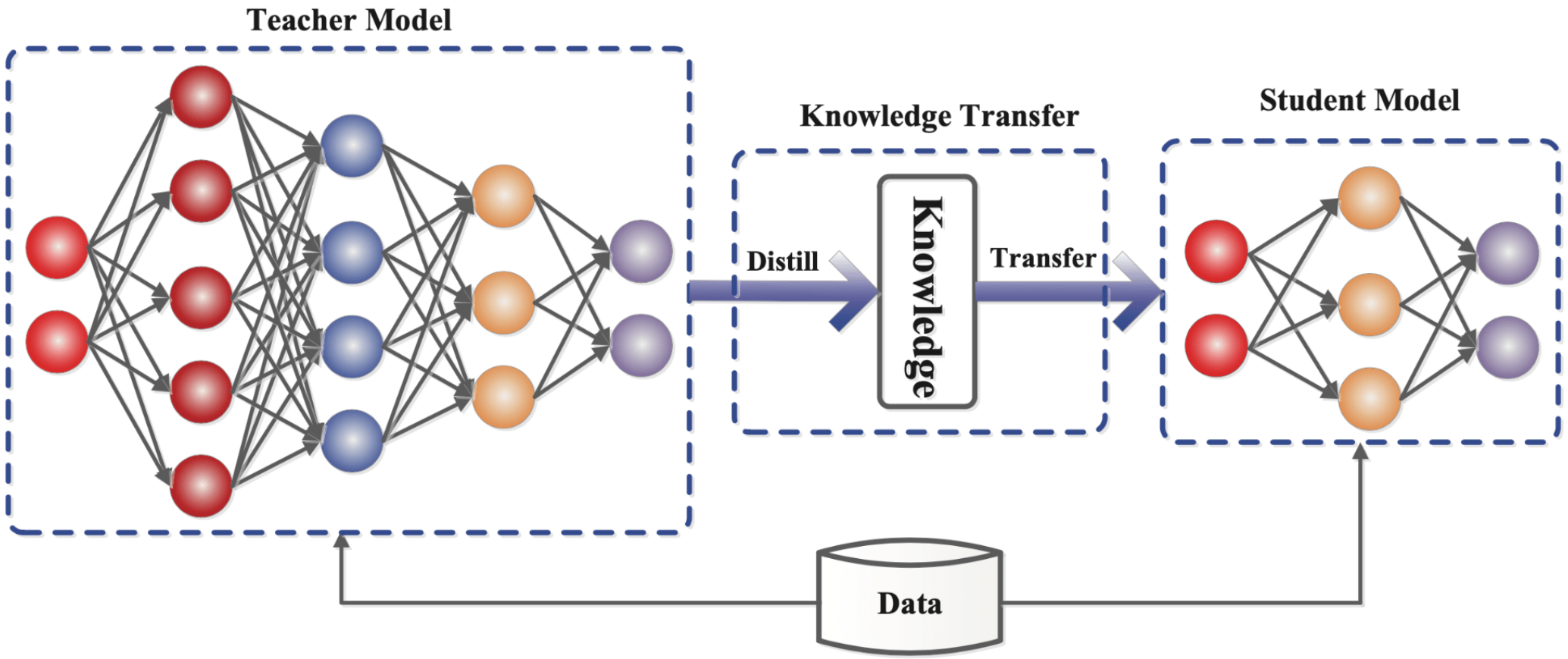
Methods of Model Compression

- Quantization
- Knowledge Distillation
- **Pruning**

Quantization

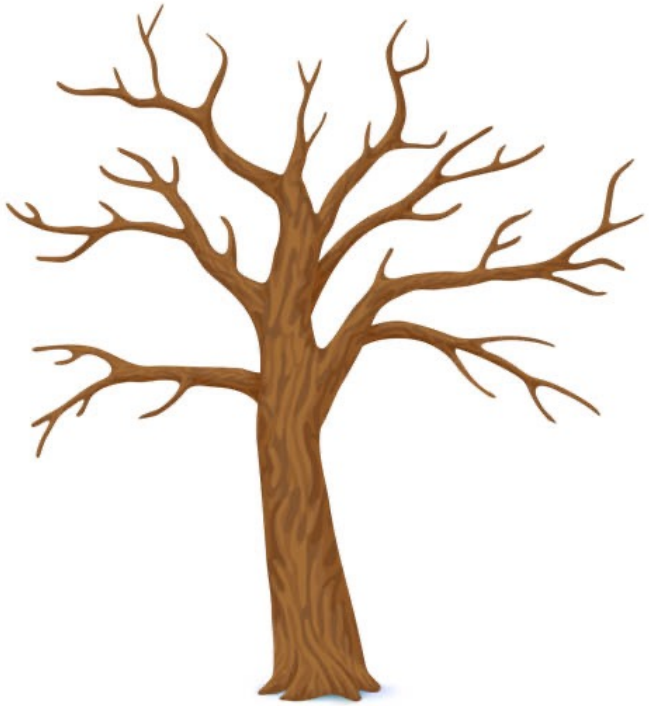
- We can change the preciseness of the numbers used to represent weights

Knowledge Distillation



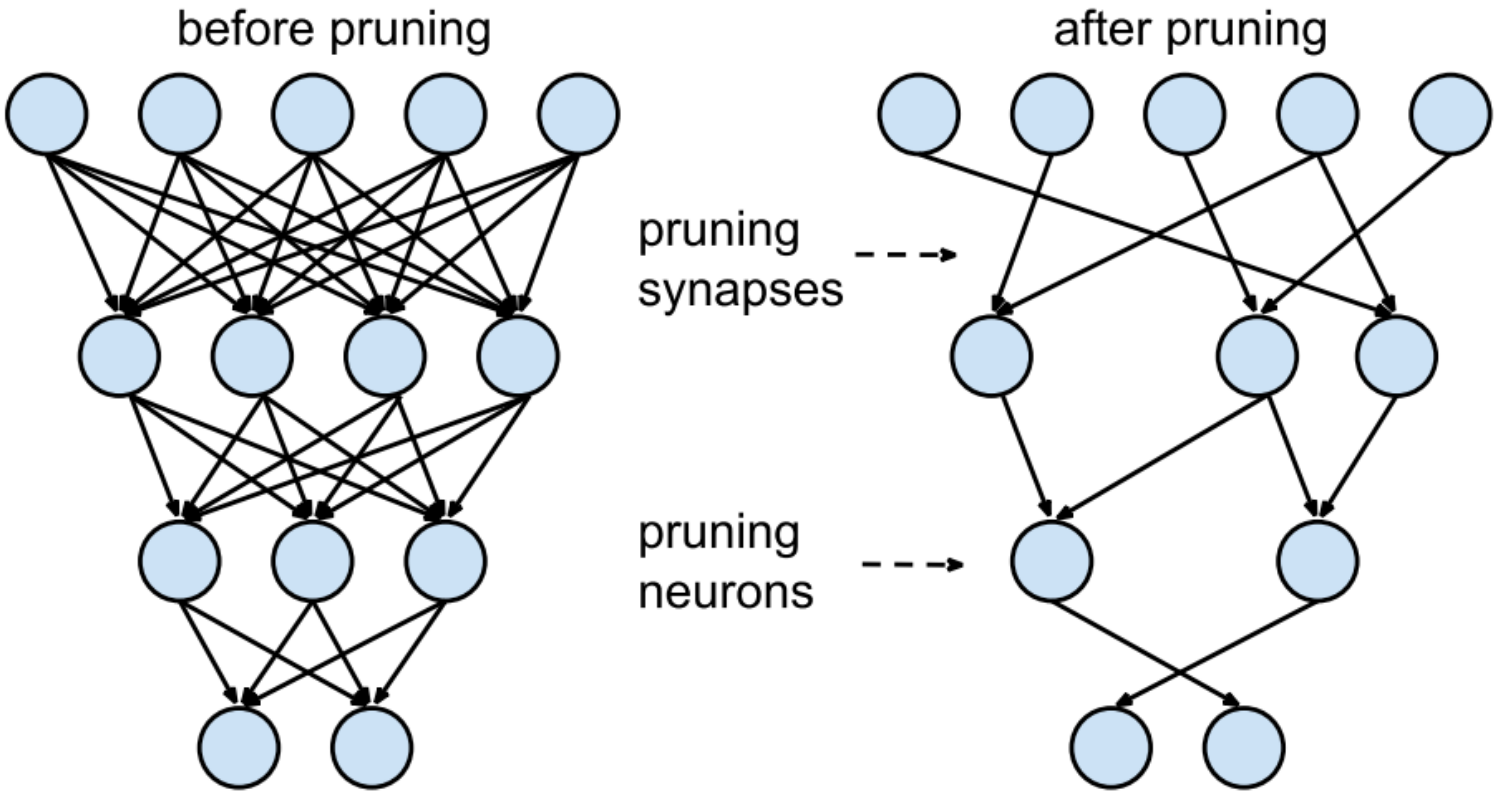
Pruning

- Deep learning involves lots of parameters
- In many cases, many parameters are useless



Reminder: Matrix Multiplication

Pruning Synapses vs. Neurons



Structured vs Unstructured Pruning

- Unstructured – find lowest strength connections and remove
- Structured – remove a larger part of the network, like a neuron or even layer

Retrain to Recover Accuracy

Iterative Pruning

- We might want to prune as an iterative process so we can decide whether accuracy is good enough after retraining

Effectiveness of Pruning

- It’s possible to keep very similar accuracy while removing:
  - 90% of ResNets
  - 75% of MobileNets
  - 60% of transformers (NLP)
- It is becoming more commonly used. It hasn’t because there are many hyperparameters, which makes the process complex

Pruning Demo

- PyTorch does not actually remove weights – it sets us up to understand what would happen if we did