Neural Network Compression

Reading

Compressing Large-Scale Transformer-Based Models: A Case Study on BERT

Targets

- number of parameters
- static memory
- FLOPS
- runtime memory
- run time

transformer have a relatively high runtime memory requirement because theu are parallel

Approaches

- quantization
- pruning
- knowledge distillation
- matrix decomposition
- dynamic inference acceleration

Quantization

reducing the number of unique values required to represent model weights and activations, which allows: **effect how big each weight is**

- to represent them using fewer bits,
- to reduce the memory footprint, and
- to lower the precision of the numerical calculations

approaches:

- post-training quantization
- quantization-aware training

Pruning

identifying and removing redundant or less important weights and/or components, which

- sometimes even makes the model more robust and better-performing

approaches:

interesting: it can make the model more rubost

- unstructured pruning
- structured pruning

Knowledge distillation

training a smaller model (called the student) using outputs (from various intermediate functional components) of one or more larger pre-trained models (called the teachers)