MiniLM

04 January 2024 18:45

MiniLM - Deep self-attention distillation

BERT learns context based text representation by predicting words given their context using large scale text corpora

There are millions of parameters which brings challenges to fine tuning and online serving in real life applications and causes latency issues and time/memory constraints

- Deep self-attention distillation for compressing large pre-trained transformers
  - Deeply mimic the self-attention modules which are fundamentally important
  - Traditional distillation requires matching layers between the teacher and student models leading to limitations in student design and could hinder performance
  - MiniLM's approach: It focuses on the last layer of the transformer, which holds the refined representations of the words after layers of processing by the previous layers
  - Advantages:
    - No more mapping the layers in the teacher model to that of the student model.
    - Focusing on a single layer simplifies the distillation process
    - Flexibility in the student design(allows for a smaller model)
  - Introduces scaled dot-product between values in the self-attention module as the new deep self-attention knowledge
    - The dot product reveals how closely the words are related to each other
    - Deep self-attention knowledge is an intricate form of the information distilled from the teacher network

1. train the student by deeply mimicking the self-attention module, which is the vital component in the Transformer, of the teacher's last layer
2. Transferring relation between values to achieve deeper mimicry
3. Also perform attention distribution transfer in the self-attention module
4. Makes use of teaching assistant helps the distillation of large pre-trained Transformer models when the size gap between the teacher model and student model is large.

Model	No. of layers of transformer	Hidden size
Teacher model	L	d
Teaching assistant	L	d'
Student model	M	d'

Why use Distillation technique used in MiniLM rather than the one used in DistilBERT:
- DistilBERT primarily focuses on final predictions which can lead to a shallow understanding of the language. MiniLM's deep self-attention distillation captures complex relationships
- Smaller and more efficient student model
- MiniLM has shown to have better generalization

- Self-Attention module has the following
  - Queries - act like pointers searching for information
  - Keys - They highlight the relevant feature
  - Values- hold the actual meaning of the word
- Attention distribution is the dot product between queries and keys