DistilBERT

27 December 2023 06:30

DistilBERT - https://arxiv.org/pdf/1910.01108.pdf

- Operating these large models on the edge and on power constrained devices is a challenge
- DistilBERT is a faster and smaller language representation model which can be fine-tuned for a wide variety of tasks
- Reduce the size of the BERT model by 40% , while retaining 97% of its language understanding capabilities
- and it is 60% faster
- Language models lead to significant improvement but they also have millions of parameters to train the model.
- With the use of knowledge distillation, the model is able to reach the similar performance on many downstream tasks. Resulting in models that are lighter and faster

Knowledge Distillation

- A smaller model is trained to reproduce the behaviour of a larger model

Training Loss

- The model is trained to use a special loss function called distillation loss, which measures how closely its predictions match the soft probabilities generated by the larger model (BERT)

$$L_{ce} = \sum t_i + \log(s_i)$$

- Softmax-Temperature:
  - Adjusting the output probabilities of both student and teacher models using a temperature parameter T
  - Higher T values make the probabilities more smoother, encouraging student to learn more
- Combined Loss
  - Distillation Loss
  - MLM loss
  - cosine embedding loss

DistilBERT - https://huggingface.co/docs/transformers/model_doc/distilbert

- DistilBERT has the same architecture as BERT, but the token-type embedding and the pooler layers are removed
- Impact of increasing the Hidden Size increases the computational cost but not so drastically when compared to change in the number of layers
- The number of layers have been reduced by a factor 2
- Most of the operations used in the transformer architecture are highly optimized

Initialization

- Initialization is very important for the sub-network (Distilled Model) to converge
- We initialize the student from the teacher by taking one layer out of two