FLIP01 Final Assessment

Cong Ma

[1] QingDao Technological University, China

MODEL:BERT

- A robust method to optimize the best pre training.

The essence of best is to run self supervised learning method on the basis of massive corpus to provide a good feature representation for word learning. The so-called self supervised learning refers to the supervised learning on the data without manual annotation. In future specific NLP tasks, we can directly use the feature representation of Bert as the word embedding feature of the task. So what Bert provides is a model for other task transfer learning, which can be used as feature extractor after task fine-tuning or fixed.

BERT:MLM Shortcoming

- MLM means that during training, some words are masked from the input, and then the words are predicted by the context of the input,
- In Bert's experiment, 15% of wordpiece tokens were randomly masked out. When training the model, a sentence will be fed to the model many times for parameter learning. However, Google does not mask these words every time. Instead, it will directly replace them with [mask] 80% of the time, replace them with any other words 10% of the time, and retain the original token 10% of the time.
- For example:
  80% : "my dog is hairy" -> my dog is [mask]
  10% : "my dog is hairy" -> my dog is apple
  10% : "my dog is hairy" -> my dog is hairy

BERT:NSP

The task of next sense prediction (NSP) is to determine whether sentence B is the following part of sentence a. If yes, output 'isnext', otherwise output 'notnext'. The training data is generated by randomly extracting two consecutive sentences from the parallel corpus, 50% of which keep the extracted two sentences, which conform to the isnext relationship, and the other 50% of the second sentences are randomly extracted from the expectation, whose relationship is not next.

roBERTa:An improved BERT algorithm

- Roberta makes the following optimization for Bert.
  - Use bigger batch and more data to make the model train longer.
  - Removed the NSP (next sense prediction) task.
  - Train on a longer sequence.
  - Mask mechanism for dynamically modifying training data.

Static masking vs dynamic masking

Static making: during data preprocessing, the mask matrix has been generated. Each sample will be randomly masked only once, and each epoch is the same.

Modified version of static making: during preprocessing, copy 10 copies of data, each copy uses a different mask, that is to say, there are 10 different mask methods in the same sentence, and then train n / 10 epochs for each data.

Dynamic masking: every time a sequence is input to the model, a new maks method is generated. That is, the mask is not generated during preprocessing, but dynamically generated when the input is provided to the model.

A larger BPE vocabulary

The vocabulary size of the original Bert is 30K, which is increased to 50K in this paper.

Summary

*Change the mask strategy dynamically, copy 10 copies of data, and then do random mask uniformly.

*The peak value of learning rate and the number of warm up update steps are adjusted.

*Training on longer sequences: do not truncate the sequences, use full length sequences.