Individual Final Report

Cheng Zeng

Overview

Theory and algorithm learning are the most difficult in this project because NLP is the relative new research direction in the machine learning field. Many concepts need to understand. once a step was tackled, following work would be hard to move on. Fortunately, most points in this task are based on the practice. Therefore, it is not hard to understand.

Individual work

The transformer module requires a large amount of energy to know and achieve. At the beginning, the words embedding, and positional encoding aim to transfer the words into vector. The space dimension of vector depends on the longest sentence but when some sentences are extremely long, the curse of dimension also needs to be considered.

Afterwards, input will go to the encoder part. As is known, transformer is the development of the sequence-to-sequence and uses several attention methods. The cores of attention modules are the self-attention, which can well to find the relationship of words in the sentences. Then the multi-head attention is the repeat of the self-attention.

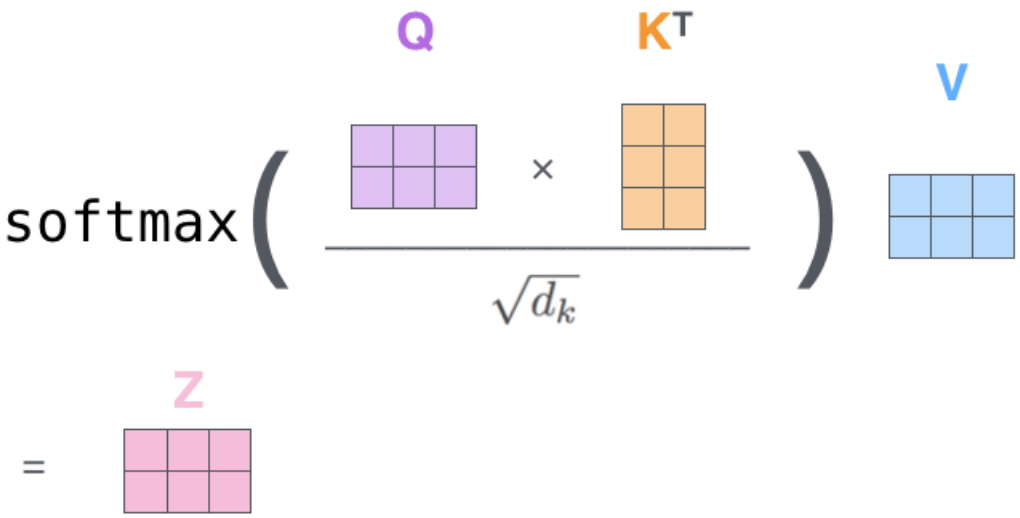

Figure 1 Self-attention equation

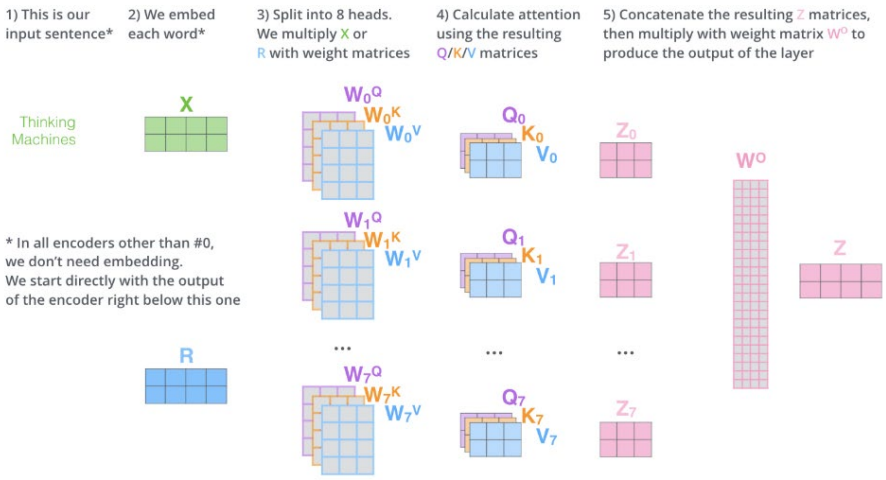

Figure 2 Multi-head attention processing

The BERT model newly released by the Google AI team has shown amazing results in SQuAD1.1, the top-level test of machine reading comprehension: all two metrics surpass humans comprehensively, and it also achieved the best results in 11 different NLP tests. , Including pushing the GLUE benchmark to 80.4% (absolute improvement 7.6%), MultiNLI accuracy reaching 86.7% (absolute improvement rate 5.6%), etc. It is foreseeable that BERT will bring milestone changes to NLP, which is also the most important recent development in the field of NLP.

Training & Results

The data set has over 200,000 records. However, the data is significantly imbalanced. The ratio of positive and negative is 9:1. When train our model, we set the parameters shown in Table 1. And we train split the training set and test set with the ratio 8:2. The result show we achieve 0.96 in accuracy and 0.65 in F1-score.

The result is not much performed well in this task. In the future we can apply many preprocessing methods to improve the performance of our model such as the parameters of models.

Code

$$\frac{100}{100+30} \times 100 = 76.9$$

Reference

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998{6008, 2017.