Assignment II NLP: Machine Reading Comprehension

SUN Yibo 3036047195

**1 Hyperparameters Tunning**

**1.1 Hyperparameters Setting**

For the first model, I didn’t change a lot of parameters except the initial learning rate.

* **MODEL 1 (BERT)**
* learning\_rate 0.00003
* Others default

**1.2 Improvement**

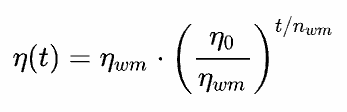
In the following training, I changed a lot of parameters. On the one hand, this is to improve the accuracy of the results, on the other hand, it can effectively prevent overfitting.

* **MODEL 2 (BERT)**
* learning\_rate 0.00003
* adam\_epsilon 0.000001
* weigh\_decay 0.001
* warmup\_step 4500
* per\_gpu\_train\_batch\_size 12

In details, batch size was improved from 8 to 12. Larger batch size can reduce the impact of individual cases on training. However, batch size cannot be set to 16 due to the limitations of GPU memory.

Weight decay is a kind of penalty to the size of weights, to prevent errors and noise amplification and the prediction will be inaccurate.

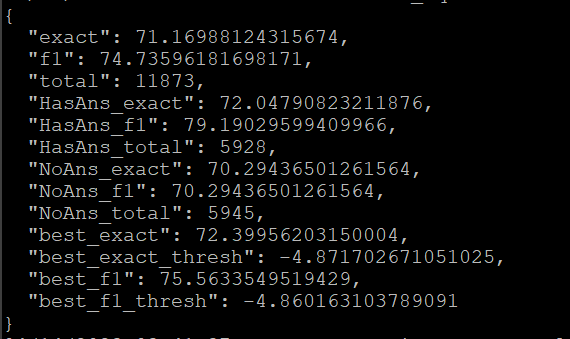
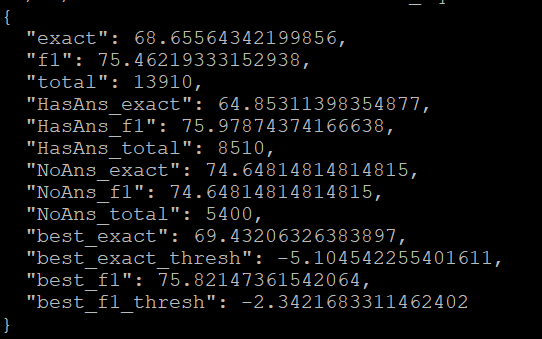
Warmup is a strategy for adjusting learning rates. It can make model parameters converge faster and achieve higher precision.



Before the set warmup step, the learning rate increases to *learning\_rate*. Then as the training progresses, the learning rate gradually declines

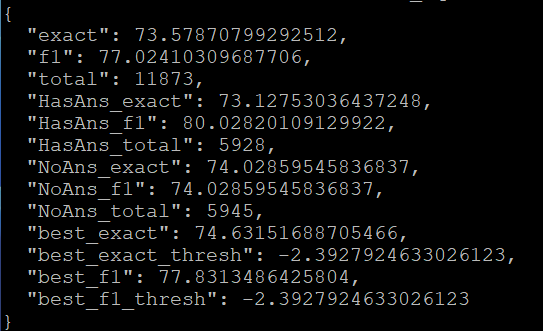
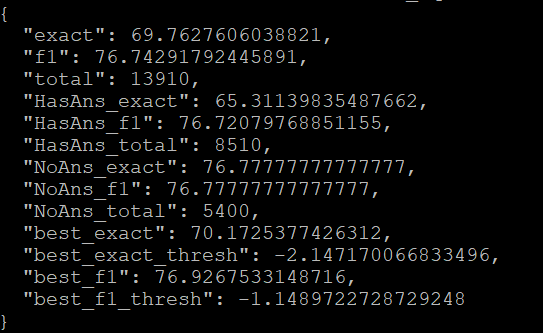
**2 Experiment Results**

For **MODEL 1**, the results on validation dataset and test dataset are following:



The performance on test data is 0.5\*EM+0.5\*F1=72.95.

For **MODEL 2**, the results on validation dataset and test dataset are following:



The performance on test data is 0.5\*EM+0.5\*F1=75.30.

It can be seen that after parameter tunning, the performance of the trained model has been significantly improved.

**3 Improvement Trying**

**3.1 RoBERTa**

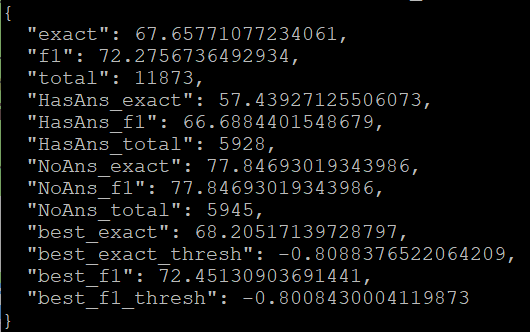
Here, I was trying to use another pre-trained model RoBERTa. Compared with BERT, RoBERTa improved in model size, batch size and data volume.

For inplement, we can use *transformers* to import.

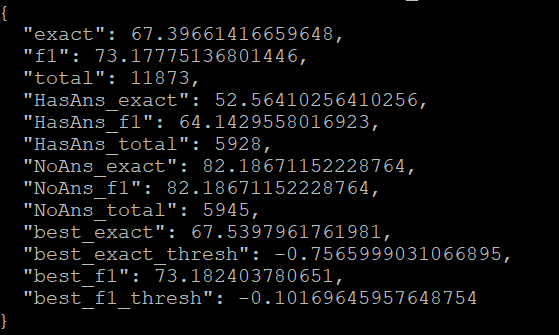
**3.2 Results**

However, the results were not better than in MODEL 1/2. Performances are showed as follows:

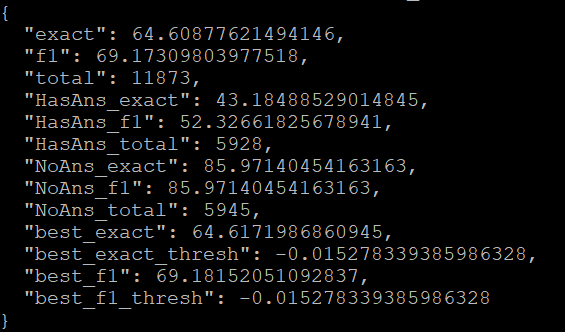
* **MODEL 3 (RoBERTa)**
* The same parameters as MODEL 2



* **MODEL 4 (RoBERTa)**
* warmup\_step 0



* **MODEL 5 (RoBERTa)**
* learning\_rate 0.00005
* adam\_epsilon 0.00000001
* warmup\_step 0
* weigh\_decay 0



The model based on pre-training RoBERTa still has a lot of space for adjustment and improvement.