Assignment3_Part3_Transformer

Analysis on hyper-parameter

2020-21717 Kim Jongwan

N0.	emsize	nhid	nlayers	dropout	batchsize	test loss	test ppl
1	512	256	4	0.1	512	5.19	178.69
2	512	256	4	0.2	512	4.91	135.02
3	512	256	4	0.3	512	4.89	132.50
4	256	256	4	0.1	512	4.89	132.98
5	256	256	4	0.2	512	4.87	130.28
6	256	256	4	0.3	512	4.93	138.20
7	256	256	5	0.1	512	4.87	130.61
8	256	256	5	0.2	512	4.89	132.66
9	256	256	5	0.3	512	4.95	141.26
10	256	256	6	0.1	512	4.87	130.15
11	256	256	6	0.2	512	4.93	138.19
12	256	256	6	0.3	512	5.03	152.64
13	256	512	4	0.1	512	4.88	131.05
14	256	512	4	0.2	512	4.88	131.90
15	256	512	4	0.3	512	4.93	134.81
16	256	512	5	0.1	512	4.88	131.02
17	256	512	5	0.2	512	4.86	129.34
18	256	512	5	0.3	512	4.91	136.28
19	256	512	6	0.1	512	4.86	129.57
20	256	1024	4	0.1	512	4.87	138.85
21	256	1024	4	0.2	512	4.84	127.02
22	256	1024	5	0.1	512	4.87	130.38
23	256	1024	5	0.2	512	4.86	128.79
24	256	2048	4	0.1	512	4.88	132.88
25	256	2048	4	0.2	512	4.84	126.69
26	256	2048	5	0.1	512	4.87	130.18
27	256	2048	5	0.2	512	4.91	135.41

The baseline of the parameter was conducted by referring to the paper "Attention is All you Need[1]".

Perspective of embedding dimesion

Comparing the 1,2.3 experiment and the 4,5.6 experiment, the test loss increases as the embedding dimension

[1] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.

increases under the same conditions. So, in subsequent experiments, it is fixed at 256.

Perspective of the dimension of the feedforward network model in Transformer Encoder

Overall results show generally good results as the dimension increases.

Perspective of the number of Transformer Encoder Layer

According to the change in the number of layers in the same dropout value from experiment 4 to experiment 12, there are good results in many cases with a value of 4 or 5. After several experiments, I set the values of 4 or 5. As a result, the best model is the case with the value of 4.

Perspective of the dropout

The drop out value changes by 0.1 unit. In many cases, values of 0.1 and 0.2 give good results. Even in experiment 25, which showed the best performance, the drop out value is 0.2.

Others

Prior to running the above experiment, I tried various values of hyper-parameter.

Although out of the scope of this task, I used Bayesian Optimization hyper-parameter method for searching the approximate value. I set the range of values by referring to the paper. Due to the constraints on the experimental environment, the batch size is 512. Although the batch size was changed to 1024, the performance decreased. I changed the number of heads as well, but I thought 8 was appropriate. So I experimented without making any changes. The evaluation batch size is also fixed to 64.