CPSC 8430 Homework#3

Email: xiaofey@clemson.edu

Name: Xiaofeng Yang

CUID: C73220300

GitHub Link: https://github.com/Xiaoo112/CSPS_8430_Homework3

March 21, 2024

Homework#3 BERT Model on SQUAD dataset

1. BERT Model Architecture

a. Configuration:
   24-layer, 1024 hidden dimension, 16 attention heads, 336M parameters
b. Preprocess Information:

15% of the tokens are masked.

In 80% of the cases, the masked tokens are replaced by `[MASK]`.

In 10% of the cases, the masked tokens are replaced by a random token (different) from the one they replace.

In the 10% remaining cases, the masked tokens are left as is.

c. Pretraining Information:

The model was trained on 4 cloud TPUs in Pod configuration (16 TPU chips total) for one million steps with a batch size of 256. The sequence length was limited to 128 tokens for 90% of the steps and 512 for the remaining 10%. The optimizer used is Adam with a learning rate of 1e-4, $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a weight decay of 0.01, learning rate warmup for 10,000 steps and linear decay of the learning rate after.

2. Preprocess Train and Validation Dataset

a. Hyperparameter:
   The maximum length of input sequence is limited to 512 tokens, and the overlapping stride for each sequence window is set as 64.
b. Padding:
   For the input sequences with token numbers less than the maximum length 512, they are padded with <pad>.
c. If the answer is not within a single window context, then the final label of that certain window is set to (0, 0)
d. Data shape for training and validation dataset as below:

```
Tokenizing and labeling dataset...
Map: 100% 37111/37111 [00:12<00:00, 2885.22 examples/s]
Dataset({
    features: ['input_ids', 'token_type_ids', 'attention_mask', 'start_positions', 'end_positions'],
    num_rows: 37130
})

Processing validation and test datasets for model evaluation...
Map: 100% 5351/5351 [00:02<00:00, 2355.05 examples/s]
Map: 100% 5351/5351 [00:03<00:00, 2282.48 examples/s]
Map: 100% 5351/5351 [00:02<00:00, 2192.30 examples/s]
Dataset({
    features: ['input_ids', 'token_type_ids', 'attention_mask', 'offset_mapping', 'example_id'],
    num_rows: 5376
})
```

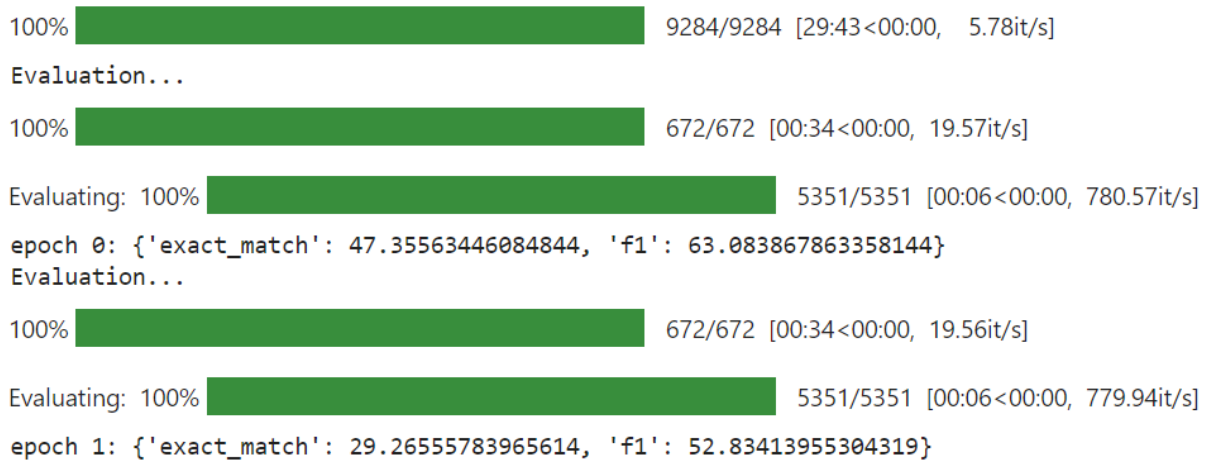
3. Fine-Tuning of the BERT Model

a. Automatic Mixed Precision: enables automatic conversion of certain GPU operations from FP32 to half-precision FP16 to speed up the fine-tuning process while maintaining accuracy from HuggingFace.
b. Applied Linear Learning rate scheduler to let learning rate decrease from 0.00003 to 0 gradually automatically from HuggingFace.
c. Optimizer is AdamW and batch size is 12, and the model was fine-tuned on Squad train split for 2 epochs.

4. Final Results

The final model was tested based on one Spoken-Squad test dataset with no noise, one Spoken-Squad test dataset with noise V1, and one Spoken-Squad test dataset with noise V2.

a. Result of training



b. Result of evaluation