

CPSC 8430 Homework#3

Email: xiaofey@clemson.edu

Name: Xiaofeng Yang

CUID: C73220300

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

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

- 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.
- Applied Linear Learning rate scheduler to let learning rate decrease from 0.00003 to 0 gradually automatically from HuggingFace.
- 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

100%  9284/9284 [29:43<00:00, 5.78it/s]

Evaluation...

100%  672/672 [00:34<00:00, 19.57it/s]

Evaluating: 100%  5351/5351 [00:06<00:00, 780.57it/s]

epoch 0: {'exact_match': 47.35563446084844, 'f1': 63.083867863358144}

Evaluation...

100%  672/672 [00:34<00:00, 19.56it/s]

Evaluating: 100%  5351/5351 [00:06<00:00, 779.94it/s]

epoch 1: {'exact_match': 29.26555783965614, 'f1': 52.83413955304319}

b. Result of evaluation

Performing evaluation on the Test Set

100%  672/672 [00:34<00:00, 19.66it/s]

Evaluating: 100%  5351/5351 [00:06<00:00, 779.49it/s]

Performing evaluation on the Test V1 Set with V1 noise

100%  672/672 [00:34<00:00, 19.61it/s]

Evaluating: 100%  5351/5351 [00:06<00:00, 780.26it/s]

Performing evaluation on the Test V2 Set with V2 noise

100%  672/672 [00:34<00:00, 19.64it/s]

Evaluating: 100%  5351/5351 [00:06<00:00, 781.40it/s]

Test Set (No Noise - 22.73% WER) - Exact Match: 29.26555783965614, F1 Score: 52.83413955304319

Test V1 Set (V1 Noise - 44.22% WER) - Exact Match: 20.08970285927864, F1 Score: 39.25835101485107

Test V2 Set (V2 Noise - 54.82% WER) - Exact Match: 14.44589796299757, F1 Score: 29.184976143129315