NTU ADL Homework 1, 2023 Fall.

Data processing

Tokenizer:

- I choose the “Word Piece” tokenization algorithm as my primary method. To explain this tokenization algorithm, you can draw a simple parallel to a well-known method known as “divide and conquer.” While they are not identical, there are some commonalities between them.

- The “Word Piece” method doesn’t solely focus on entire words; instead, it places emphasis on understanding the constituent parts of a word and their meanings. Let’s take the word “tokenization” as an example: “token” may help the model grasp the core meaning of the word, “niz(e)” might guide the model in transforming the word into a verb, and finally, “ation” performs a similar role by turning the verb into a noun.

Answer Span:

- By using an off mapping techenique, mapping characters position from token to original context.

- After computing all the start/end sets’ probability, check each of them, remove the set that violate rules - end_pos before start_pos, answer out of the span, etc.

Modeling with BERTs and their variants

First Model: Bert-Chinese-Base

- Exact Match: 78.598.

- Cross-entropy loss.

- AdamW, LR = 3e-5, Effective Batch Size = 32 = 1(batch_size_per_device) * 32(gradient_accmulation_steps)

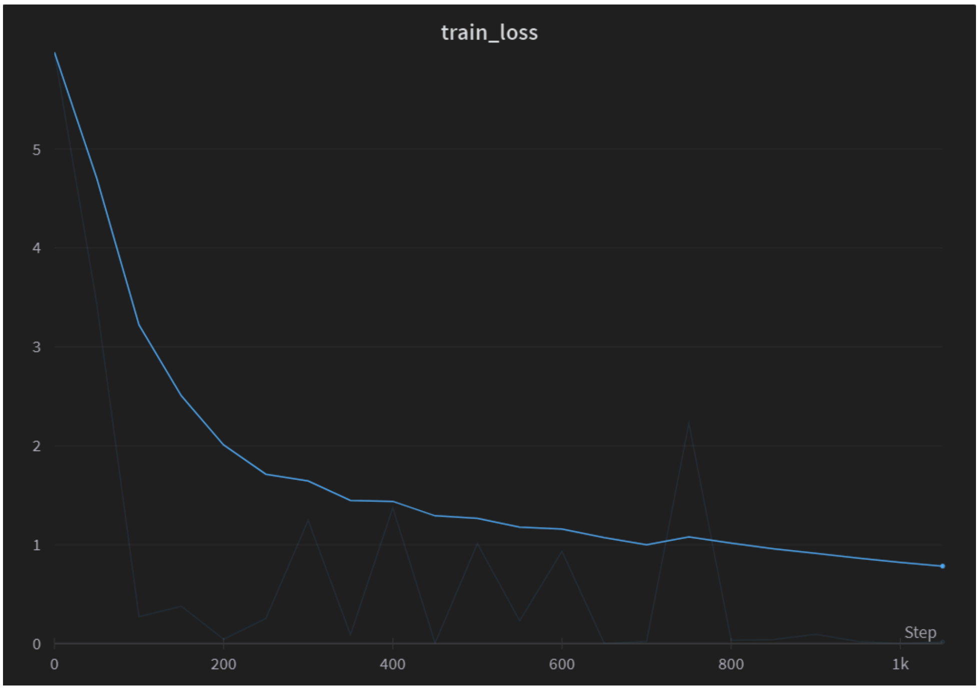
Another Model: Chinese-Roberta-wwm-ext-large

- **Exact Match: 82.186**

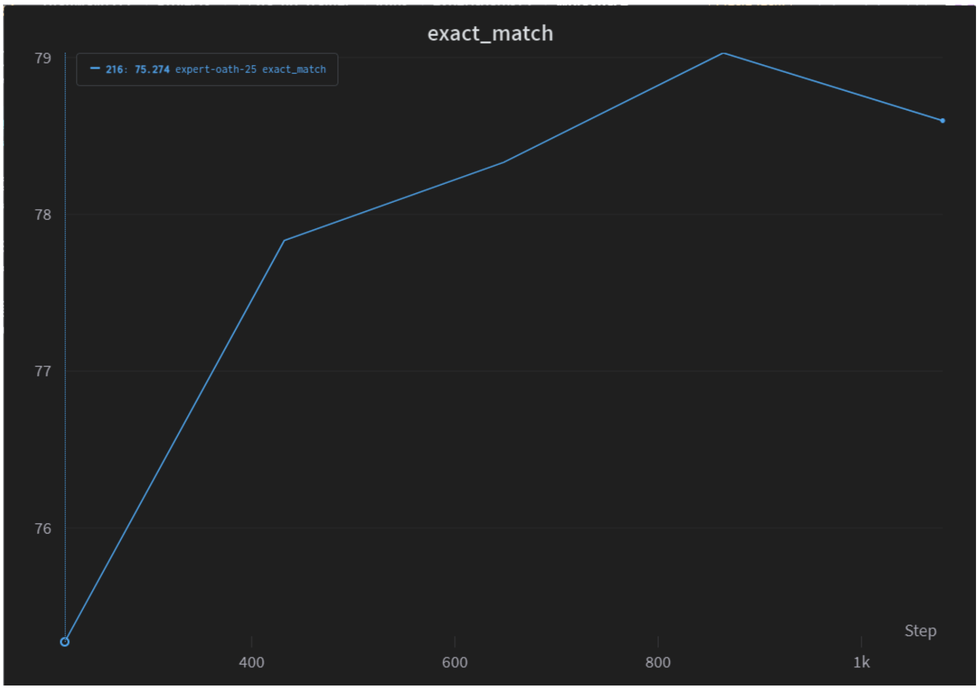
- **RoBERTa model adopts the original BERT model architecture, trains for longer durations with larger batches and longer sequences, uses more data, and removes the next sentence prediction while employing dynamic masking, and using different tokenize strategy called “BPE”.**

Curves

Loss:



Exact Match:



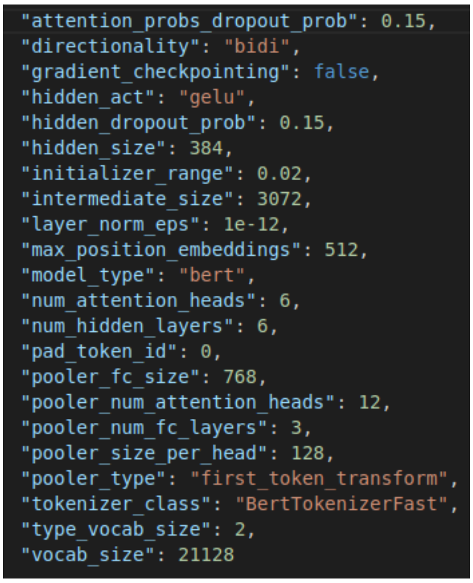
Pre-trained vs Not Pre-trained

Configuration & Parameters:

- **Max Sequence Length = 512**

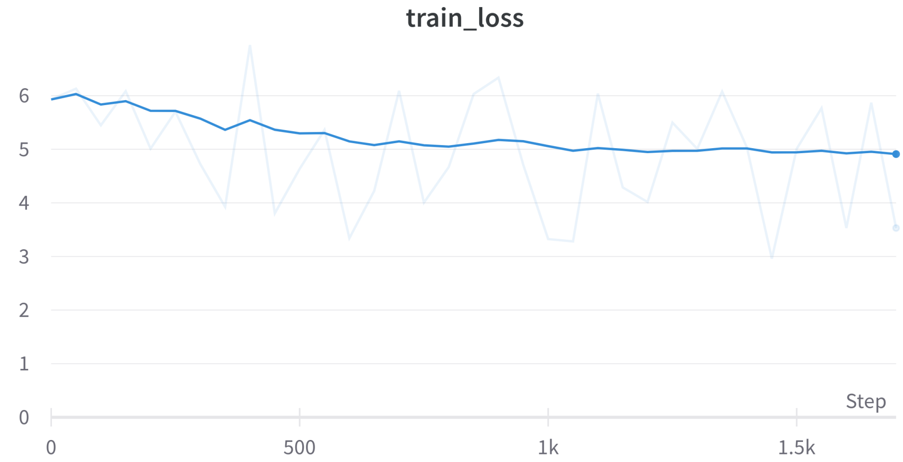
- **Effective Batch Size = 32 = 1(batch_size_per_device) * 32(gradient_accmulation_steps)**

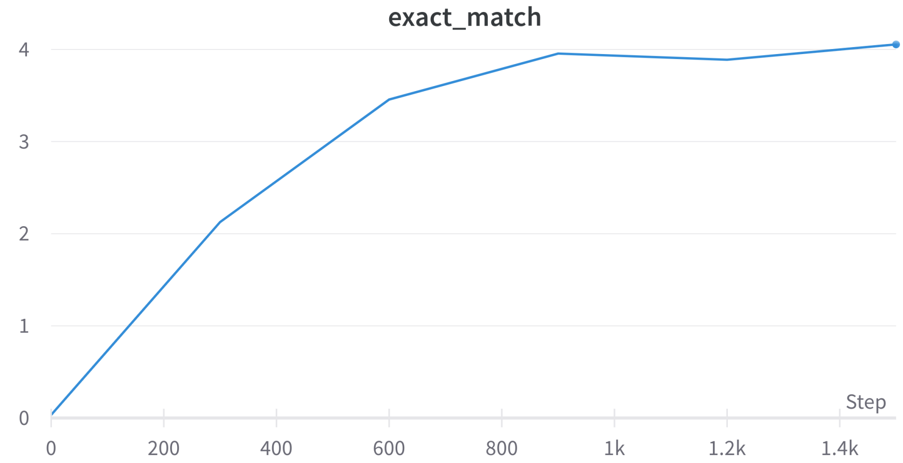
- **Epochs = 2**



```
"attention_probs_dropout_prob": 0.15,
"directionality": "bidi",
"gradient_checkpointing": false,
"hidden_act": "gelu",
"hidden_dropout_prob": 0.15,
"hidden_size": 384,
"initializer_range": 0.02,
"intermediate_size": 3072,
"layer_norm_eps": 1e-12,
"max_position_embeddings": 512,
"model_type": "bert",
"num_attention_heads": 6,
"num_hidden_layers": 6,
"pad_token_id": 0,
"pooler_fc_size": 768,
"pooler_num_attention_heads": 12,
"pooler_num_fc_layers": 3,
"pooler_size_per_head": 128,
"pooler_type": "first_token_transform",
"tokenizer_class": "BertTokenizerFast",
"type_vocab_size": 2,
"vocab_size": 21128
```

Performance


train_loss
Step
1k
1.5k


exact_match
Step
1k
1.2k
1.4k