Report

Q1: Data processing

Tokenizer

1. **Describe in detail about the tokenization algorithm you use. You need to explain what it does in your own ways.**

   The pre-trained model I used is `hfl/chinese-roberta-wwm-ext-large`. Roberta's tokenizer is derived from GPT-2 tokenzier, it's a BPE(Byte-Pair-Encoding). BPE first splits text into words and calculate the frequency, then splits words into chars. After that, add the most-frequent chars sequence into vocabs. Continue doing above operation until vocal size reaches `vocab_size` hypermeter.

Answer Span

1. **How did you convert the answer span start/end position on characters to position on tokens after BERT tokenization?**

   While doing tokenization, we can enable two flags: `return_overflowing_token` and `return_offsets_mapping`. Then tokenizer will return `overflow_to_sample_mapping` and `offset_mapping` these two data for us to track text and input_ids. Then we can simplying use while loops to find the positions of answer span.

2. **After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position?**

   Select the start/end positions below `n_best_size` parameter. Then map both start/end positions combinations back into text. Select the prediction with most score: probability score of start position + probability score of end position.

Q2: Modeling with BERTs and their variants

Describe

1. **Your model**
2. **The performance of your model.**
3. **The loss function you used.**
4. **The optimization algorithm (e.g. Adam), learning rate and batch size.**

Try another type of pre-trained LMs and describe

1. **Your model**
2. **The performance of your model.**
3. **The difference between pre-trained LMs (architecture, pretraining loss, etc.)**

Q3: Curves

1.

References

1. https://huggingface.co/docs/transformers/tokenizer_summary#bytepair-encoding-bpe
2.