Report of Natural Language Processing, homework 3

109550027 紀竺均

1. Describe all the methods you have implemented. (60%)

I. Transformer version == 4.15.0

II. Model

```
model = BertForSequenceClassification.from_pretrained('bert-base-
chinese', return_dict=True, num_labels=4)
model.to(device)
model.classifier = nn.Linear(model.config.hidden_size,
1).to(device)
```

I use transformers.BertForSequenceClassification() with pretrain Bert model "bert-base-chinese" (Chinese Simplified and Traditional, 12-layer, 768-hidden, 12-heads, 110M parameters) to initialize.

III. Optimizer = AdamW(): Use this optimizer to do gradient descent bias and also apply weight decay. I apply weight decay to all parameters other than bias.

IV. Data processing: [Detailed in part 2.]

V. Training process:

```
outputs = model(input_ids=input_ids, token_type_ids=segment_ids,
attention_mask=input_mask, labels=label_ids)
loss = outputs.loss
loss.backward()
optimizer.step()
model.zero_grad()
```

VI. Evaluation process (for training / validation data)

```
outputs = model(input_ids=input_ids, token_type_ids=segment_ids,
attention_mask=input_mask, labels=label_ids)
tmp_eval_loss = outputs.loss
logits = outputs.logits # 1D prediction
```

# logits is an array with shape (BATCH*4 options, 1)

# labels is an array with shape (BATCH, 1)

```
accuracy(logits.reshape((-1,n_class)), label_ids.reshape(-1))
def accuracy(out, labels):
    outputs = np.argmax(out, axis=1) return np.sum(outputs==labels)
```

2. Did you preprocess your data from the dataset? Why? And how?
(Did you encounter the problem that the input length is longer than the maximum sequence length of the model you use? How did you solve this problem?) (30%)

Yes, I did. Because I need to convert the data into features that can be fed into my model. Below shows the preprocess procedures:

i. tokenize: Initialize a BertTokenizer class with a vocab file.

```
tokenizer = BertTokenizer(
        vocab_file=vocab_file, do_lower_case=False)
```

Then tokenize the three sequence {article, choices, question}

ii. If the input length is greater than max_seq_length (set it to 128) >> enter iii.

iii. truncate function: Pop out one word of the longest sequence in {article, choices, question} every time until total length is smaller than max_seq_length.

iv. Add “[CLS]” at the begin. Add “[SEP]” between different choices and sequences.

v. Convert tokens to ids:

```
input_ids = tokenizer.convert_tokens_to_ids(tokens)
```

3. What difficulties did you encounter in this assignment? How did you solve it? (10%)
I think the most difficult part is that I don’t know which model to choose in Transformers. At first, I use BertForQuestionAnswering() to be my model. But this model is used to find the answer in the article, and didn’t fit our purpose—solving multiple choice question—perfectly. Finally, I’ve read some reference that use BertForSequenceClassification() and I think it fits our task well.
Reference: https://github.com/nlpdata/c3