921 U2620 HW7

TOTAL POINTS

4 / 4

QUESTION 1

1 After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position? (the rules you applied must be different from the sample code) **2 / 2**

✓ **- 0 pts** *Correct*

- 2 pts Wrong

QUESTION 2

Try another type of pretrained model which can be found in huggingface's Model Hub (e.g. BERT -> BERT-wwm-ext, or BERT -> RoBERTa), and describe 2 pts

2.1 the pretrained model you used **0.5 / 0.5**

✓ **- 0 pts** *Correct*

- 0.5 pts Wrong

2.2 performance of the model you used **0.5 / 0.5**

✓ **- 0 pts** *Correct*

- 0.5 pts Wrong

2.3 the difference between BERT and the pretrained model you used (architecture, pretraining loss, etc.) **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Wrong

1.

使用 topk 個 start_logtis, end_logits 來當作 candidates，然後兩兩組合(e.g. k = 5 個就會有 5*5 種組合)，從裡面挑出最大的 start_prob + end_prob，且其對應的 (start_index, end_index) pair 要滿足 start_index >= end_index and start_index >= paragraph_base and end_index < paragraph_sep and end_idx - start_idx <= 30，其中 paragraph_base 為該 window 中 paragraph 開始的 index; paragraph_sep 為 paragraph [SEP] 的 index。

2.

使用了 BERT-base、RoBERTa

Performace(Kaggle Public Score):

BERT-base: **0.76603**

RoBERTa: **0.82977**

差異:

RoBERTa 使用 dynamic masking，使用 10 組 masking 來做訓練，相比 BERT 的 masking 是靜態的，導致 BERT 的每個 epoch 的 masking 都是一致的，假設今天訓練 40 個 epochs，RoBERTa 中相同 masking 只會看到 4 次，而 BERT 卻看到 40 次。

RoBERTa 比起 BERT 使用更多 dataset 和一些 hyper-parameters 的調整。

1 After your model predicts the probability of answer span start/end position, what rules did you apply to determine the final start/end position? (the rules you applied must be different from the sample code) **2 / 2**

✓ **- 0 pts** *Correct*

- 2 pts Wrong

1.

使用 topk 個 start_logtis, end_logits 來當作 candidates，然後兩兩組合(e.g. k = 5 個就會有 5*5 種組合)，從裡面挑出最大的 start_prob + end_prob，且其對應的 (start_index, end_index) pair 要滿足 start_index >= end_index and start_index >= paragraph_base and end_index < paragraph_sep and end_idx - start_idx <= 30，其中 paragraph_base 為該 window 中 paragraph 開始的 index; paragraph_sep 為 paragraph [SEP] 的 index。

2.

使用了 BERT-base、RoBERTa

Performace(Kaggle Public Score):

BERT-base: **0.76603**

RoBERTa: **0.82977**

差異:

RoBERTa 使用 dynamic masking，使用 10 組 masking 來做訓練，相比 BERT 的 masking 是靜態的，導致 BERT 的每個 epoch 的 masking 都是一致的，假設今天訓練 40 個 epochs，RoBERTa 中相同 masking 只會看到 4 次，而 BERT 卻看到 40 次。

RoBERTa 比起 BERT 使用更多 dataset 和一些 hyper-parameters 的調整。

2.1 the pretrained model you used **0.5 / 0.5**

✓ **- 0 pts** *Correct*

- 0.5 pts Wrong

1.

使用 topk 個 start_logtis, end_logits 來當作 candidates，然後兩兩組合(e.g. k = 5 個就會有 5*5 種組合)，從裡面挑出最大的 start_prob + end_prob，且其對應的 (start_index, end_index) pair 要滿足 start_index >= end_index and start_index >= paragraph_base and end_index < paragraph_sep and end_idx - start_idx <= 30，其中 paragraph_base 為該 window 中 paragraph 開始的 index; paragraph_sep 為 paragraph [SEP] 的 index。

2.

使用了 BERT-base、RoBERTa

Performace(Kaggle Public Score):

BERT-base: **0.76603**

RoBERTa: **0.82977**

差異:

RoBERTa 使用 dynamic masking，使用 10 組 masking 來做訓練，相比 BERT 的 masking 是靜態的，導致 BERT 的每個 epoch 的 masking 都是一致的，假設今天訓練 40 個 epochs，RoBERTa 中相同 masking 只會看到 4 次，而 BERT 卻看到 40 次。

RoBERTa 比起 BERT 使用更多 dataset 和一些 hyper-parameters 的調整。

2.2 performance of the model you used **0.5 / 0.5**

✓ **- 0 pts** *Correct*

- 0.5 pts Wrong

1.

使用 topk 個 start_logtis, end_logits 來當作 candidates，然後兩兩組合(e.g. k = 5 個就會有 5*5 種組合)，從裡面挑出最大的 start_prob + end_prob，且其對應的 (start_index, end_index) pair 要滿足 start_index >= end_index and start_index >= paragraph_base and end_index < paragraph_sep and end_idx - start_idx <= 30，其中 paragraph_base 為該 window 中 paragraph 開始的 index; paragraph_sep 為 paragraph [SEP] 的 index。

2.

使用了 BERT-base、RoBERTa

Performace(Kaggle Public Score):

BERT-base: **0.76603**

RoBERTa: **0.82977**

差異:

RoBERTa 使用 dynamic masking，使用 10 組 masking 來做訓練，相比 BERT 的 masking 是靜態的，導致 BERT 的每個 epoch 的 masking 都是一致的，假設今天訓練 40 個 epochs，RoBERTa 中相同 masking 只會看到 4 次，而 BERT 卻看到 40 次。

RoBERTa 比起 BERT 使用更多 dataset 和一些 hyper-parameters 的調整。

2.3 the difference between BERT and the pretrained model you used (architecture, pretraining loss, etc.) **1 / 1**

✓ **- 0 pts** *Correct*

- 1 pts Wrong