NLP HW3

NCKU F64091130 楊采語

|  |  |  |
| --- | --- | --- |
| 訓練模型 | Running environment | Python version |
| Colab  T4 GPU | Colab  Python3 |

大部分code皆有與Claude討論過再修改

1. Which (pre-trained) model do you use? Why to choose the model?

我使用的pre-trained model是bert-base-uncased，因為我認為本次的任務不會出現太多專有名詞，而且本次任務也主要是語意理解任務，所以用bert-base-uncased就能取得不錯的效果。

1. Compared with models trained separately on each of the sub-task, does multi-output learning improve the performance

|  |  |  |
| --- | --- | --- |
|  | Spearman correlation | |
| Task | Multi-Output | Similarity-Only |
| epoch 1 | 0.8340 | 0.7998 |
| epoch 2 | 0.8206 | 0.8178 |
| epoch 3 | 0.8250 | 0.8333 |

在相關性任務中會發現，如果是只單獨訓練一個task，Spearman correlation會較大幅度的上升，而在Multi-Output learning的時候，一開始就會取得不錯的Spearman correlation分數，但後續訓練到第2、第3個epoch時，成果都差不多，沒有顯著的提升。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | | F1 Score | |
| Task | Multi | Entailment | Multi | Entailment |
| epoch 1 | 0.8700 | 0.8220 | 0.8689 | 0.8276 |
| epoch 2 | 0.7940 | 0.8380 | 0.7996 | 0.8381 |
| epoch 3 | 0.8600 | 0.8400 | 0.8620 | 0.8337 |

而在蘊含任務中會發現，如果是只單獨訓練一個task，Accuracy與F1 Score會較穩定的上升，而在Multi-Output learning的時候，一開始就會取得不錯的Accuracy與F1 Score，但在訓練epoch 2時，會突然出現一個不穩定的大幅下降。

綜合上述實驗的觀察，Multi-Output learning可能會因為兩個任務的類似特徵，一開始就取得較好的訓練成果，但可能中途會遇到兩個任務的互相干擾，導致訓練成果會出現較大的波動；此外，可能模型為了平衡兩個任務，在任一個任務都會無法取得最高的分數，但同時也能避免模型出現過度擬合資料的狀況。

1. Why does your model fail to correctly predict some data points?

|  |  |
| --- | --- |
| Similarity Errors | |
| 一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片  自動產生的描述 | Error distribution:  Count 27.000000  Mean 1.220200  Std 0.190317  Min 1.001996  25% 1.067471  50% 1.155431  75% 1.342036  Max 1.649023 |
|  | |
|  | |

由上圖可以發現，在真值3-4之間時預測偏低，也最不穩定(前五大的誤差就有4個的真值是在3-4之間)。而且在句子長度越長的時候，相似度任務的錯誤率會逐漸提高，而在句子的單詞平均長度增加到最高範圍時，也會使得錯誤率變高。

|  |
| --- |
| Entailment Task Error Analysis |
|  |
|  |

由上圖可以看到，當真值為0(NEUTRAL)時，常被預測成1(ENTAILMENT)或2(CONTRADICTION)；而真值為1時全被預測成0；而真值為2時主要被預測成0。

看起來我的模型有學會分辨ENTAILMENT與CONTRADICTION的關係，但在選項可能是NEUTRAL時，模型會比較難判斷。

1. How do you improve your model performance?

針對模型可能會出現不穩定的大幅下降的問題，我增加scheduler讓模型不要太大幅度的調整，另外嘗試選擇只更新某些層，不要更新全部的模型參數，避免搞爛整個模型。

另外根據預測的結果，句子的長度及單字的複雜度也會導致模型的相似度任務的錯誤率提高，因此嘗試加入針對長句子的處理；另外針對蘊含關係任務增加權重，讓ENTAILMENT與NEUTRAL的權重加重一些看模型是否能平衡類別的準確度。

經過測試後發現，其實本次任務不太會造成模型被搞爛的情況，所以後來不採用此策略(因為在同訓練相同次數的情況下，誤差明顯會變大)。

|  |  |  |  |
| --- | --- | --- | --- |
| 未調整前 | | | |
|  | Spearman | Accuracy | F1 Score |
| Epoch1 | 0.8340 | 0.8700 | 0.8689 |
| Epoch2 | 0.8206 | 0.7940 | 0.7996 |
| Epoch3 | 0.8250 | 0.8600 | 0.8620 |
| 調整後 | | | |
|  | Spearman | Accuracy | F1 Score |
| Epoch1 | 0.7712 | 0.8440 | 0.8416 |
| Epoch2 | 0.8278 | 0.8600 | 0.8625 |
| Epoch3 | 0.8321 | 0.8660 | 0.8687 |

由上面成果可以發現，其實調整前後的評估分數沒有大幅的差距，但評估分數有穩定的向上提升，而在調整更新權重後，Entailment Task Error可以看到1 (ENTAILMENT)的部分下降了一半，但相對的真值為0(NEUTRAL)時，就會更容易被判斷成1 (ENTAILMENT)。

|  |  |
| --- | --- |
| 一張含有 文字, 圖表, 繪圖, 行 的圖片  自動產生的描述 |  |

1. 針對高相似度進行特別處理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 一張含有 文字, 圖表, 繪圖, 行 的圖片  自動產生的描述 | | | 在右圖可以看到，在相似度在3~5之間的值，很容易出現高估的情形。  所以可能可以特邊針對高相似度的預測，讓模型增加懲罰制度，讓模型不要預測得太高。 | |
| 調整後 | | | | | |
|  | | Spearman | Accuracy | | F1 Score |
| Epoch1 | | 0.7874 | 0.8360 | | 0.8363 |
| Epoch2 | | 0.8202 | 0.8480 | | 0.8482 |
| Epoch3 | | 0.8446 | 0.8640 | | 0.8638 |

|  |  |
| --- | --- |
| 一張含有 文字, 行, 繪圖, 圖表 的圖片  自動產生的描述 | 一張含有 文字, 螢幕擷取畫面, 圖表, 字型 的圖片  自動產生的描述 |

模型調整後，發現在高相似度的地方還是會有過高的預測值，甚至比沒調整前還要慘，在較低相似度也出現了高估的狀況，導致整個誤差更大了，所以可能還需要調整一下懲罰制度讓模型不要有過高的預測值。

再調高懲罰制度後，整體有變的分散的比較平均，但造成在低相似度的地方會出現過低的預測值，且整體的誤差也變得更大了，或許讓模型自己學習的成效比人為的干預來的更好。

一張含有 文字, 螢幕擷取畫面, 圖表, 繪圖 的圖片

自動產生的描述