自然語言處理 HW3

成功大學114資訊系 F74102022唐文蔚

|  |  |
| --- | --- |
| Environment | Ubuntu 20.04.6 LTS |
| CPU | Intel(R) Xeon(R) Gold 6242R CPU @ 3.10GHz |
| GPU | NVIDIA GeForce RTX 3090 |
| Python Version | Python 3.9.20 |

Report：

Q1：Which (pre-trained) model do you use? Why to choose the model?

A1：google-bert/bert-base-uncased

比較一些常用的模型後，我選擇使用google-bert/bert-base-uncased，原因是語意理解任務對大小寫不敏感，可減少詞彙變體影響；Google的預訓練模型品質很高，且相較其他模型，處理簡單、泛化性好，而實驗的結果也證明此模型是較好的選擇，針對nli task fine-tune的模型結果也和基本的模型大致相同，沒有更好的表現。

**google-bert/bert-base-uncased**

Test set - Spearman Corr: 0.828 | Accuracy: 0.877 | F1 Score: 0.869

**google-bert/bert-base-cased**

Test set - Spearman Corr: 0.799 | Accuracy: 0.843 | F1 Score: 0.845

**google-bert/bert-base-multilingual-uncased**

Test set - Spearman Corr: 0.789 | Accuracy: 0.849 | F1 Score: 0.839

**MayaGalvez/bert-base-multilingual-cased-finetuned-nli**

Test set - Spearman Corr: 0.808 | Accuracy: 0.860 | F1 Score: 0.848

**tomaarsen/bert-base-uncased-nli-v1**

Test set - Spearman Corr: 0.825 | Accuracy: 0.873 | F1 Score: 0.867

**Jihyun22/bert-base-finetuned-nli**

Test set - Spearman Corr: 0.791 | Accuracy: 0.839 | F1 Score: 0.837

Q2：Compared with models trained separately on each of the sub-task, does multi-output learning improve the performance?

A2：是，與分別訓練子任務的模型相比，multi-output learning 在 Spearman Correlation (regression)、Accuracy 和 F1 Score 上 (classification)都有小幅改進，代表 multi-output learning應該有提升模型的表現的效果。

**Multi-output learning：**

Test set - Spearman Corr: 0.828 | Accuracy: 0.877 | F1 Score: 0.869

**Trained separately：**(只使用其中一個loss更新來達到分別訓練模型的效果)

Test set - Spearman Corr: 0.827 | Accuracy: 0.843 | F1 Score: 0.838

Q3：Why does your model fail to correctly predict some data points? Please provide an error analysis.

A3：否定詞、動作、空間、同義字等可能影響結果。

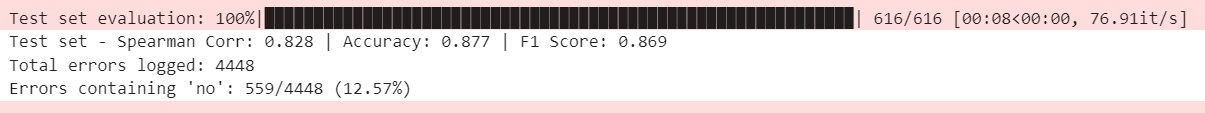
1. 否定詞處理困難

ex.

"there is no man playing an electronic keyboard" vs "a man is playing an electronic keyboard"

"the men are not fist fighting in a ring" vs "the men are fist fighting in a ring"

問題：模型難以準確區分否定句與肯定句的差異，影響判斷。



1. 細微動作差異

ex.

"a young man is getting a motocross bike up a dirt hill" vs "a young man is pushing a motocross bike up a dirt hill"

"a boy is jumping on the side of a stone water fountain" vs "a boy is jumping off the side of a stone water fountain"

問題：對於動作描述的細微差異不夠敏感。

1. 空間位置描述

ex.

"a man is standing next to a bus" vs "a man is standing on a sidewalk"

"a famous singer is dancing on the ceiling" vs "a man is dancing on the floor"

問題：難以捕捉空間位置表達差異。

1. 同義詞和近義詞處理

ex.

"school children" vs "kids"

"orange juice is being drunk" vs "a man is drinking orange juice"

問題：無法處理同義詞和近義詞。

Q4：How do you improve your model performance?

A4：

1. 使用不同Optimizer ( Test Score )

|  |  |  |  |
| --- | --- | --- | --- |
| Optimizer | Spearman Corr | Accuracy | F1 Score |
| Adam | 0.808 | 0.868 | 0.863 |
| AdamW | 0.828 | 0.877 | 0.869 |
| RMSprop | 0.812 | 0.871 | 0.862 |

1. 超參數：嘗試調整epoch、lr等超參數。
2. 使用不同的bert-base模型(如A1所述)
3. 更改模型架構：嘗試為bert-base後面接的regressor和 classifier加上多層次的架構來處理更複雜的任務，但結果顯示最simple的單層架構反而是表現最好的。

<Bonus：Loss的處理> 結果顯示最簡單的兩者相加就可以得到很好的效果。

|  |  |  |  |
| --- | --- | --- | --- |
|  | Spearman Corr | Accuracy | F1 Score |
| (1) | 0.828 | 0.877 | 0.869 |
| (2) | 0.820 | 0.865 | 0.857 |
| (3) | 0.812 | 0.862 | 0.857 |
| (4) | 0.818 | 0.870 | 0.863 |
| (5) | 0.816 | 0.851 | 0.848 |

1. 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述直接相加：將兩者loss直接相加
2. 動態調整 – 1 ：正規化

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

1. 動態調整 – 2 GradNorm



一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

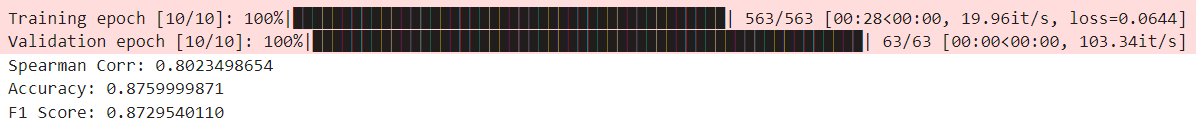
1. 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

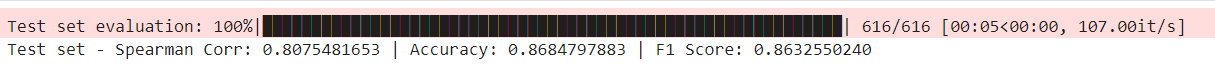
   自動產生的描述交替優化：用兩種loss交替優化，一個epoch只採用其中一種
2. 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

   自動產生的描述Loss Smoothing

<screenshot records>

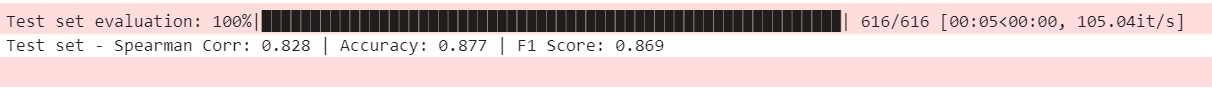
Adam





AdamW





RMSprop



