**GAI Project 2.b Text summarization**

**F74101254 資訊系 張暐俊**

* **Model**分別使用了兩種模型：T5、GPT2，前者是助教在範例code中使用，後者為建議我們在訓練比較時使用。為了比較兩者的測驗結果，因此兩者都有使用。  
  在選擇模型上，皆是選擇已經pretrained的model去做finetune。
  1. **T5  
     一張含有 文字, 螢幕擷取畫面, 字型 的圖片

     自動產生的描述**一張含有 文字, 行, 繪圖, 圖表 的圖片

     自動產生的描述
  2. **GPT2  
     一張含有 文字, 螢幕擷取畫面, 字型 的圖片

     自動產生的描述**一張含有 文字, 行, 繪圖, 圖表 的圖片

     自動產生的描述

首先，以loss rate來看的話，GPT2的表現相對較優。另外，在訓練時長上，T5模型的訓練時長相較於GPT2來說得較長，但這可能與選擇的pretrain model有關係。

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  
上圖為兩者模型的rouge分數比較圖。

而根據T5和GPT2兩個模型，前者是Seq2Seq模型，後者更擅長處理text generation的作業。而在我查到的資料中，也說明以text summaration的部分來說的話，T5模型似乎更拿手，從上圖的評測結果或許可見一斑。

* **Dataset**

**一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述**使用load\_dataset將資料集從hugging face抓下來之後再對各種類別的資料集做切割。要是不做切割的話，光是train的dataset就有2,400,000筆左右的資料，訓練時長上至少要10小時以上起跳。

* 1. **T5  
     一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

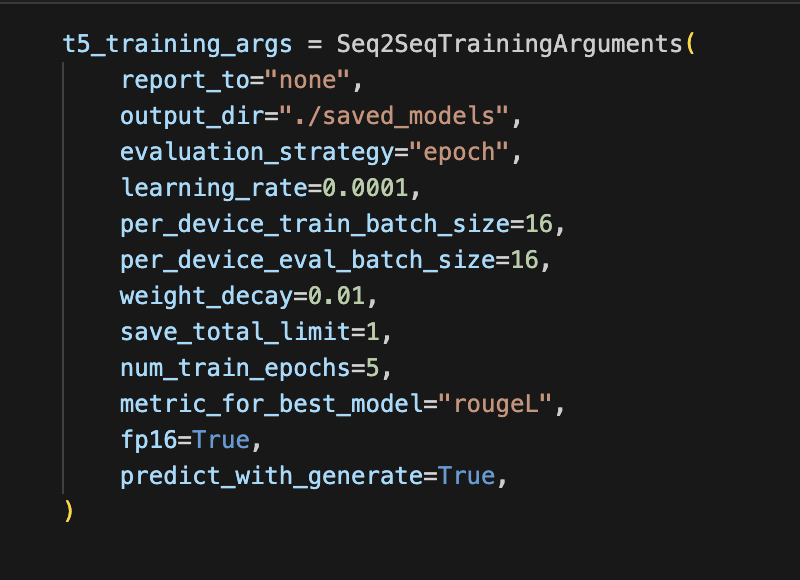
     自動產生的描述**將「总结：」後方加上我們的text欄位後，標記為input，並將我們的summary欄位標記為labels，讓模型去做比較與計算。  
     以train dataset的第一筆資料為例：  
     **一張含有 文字 的圖片

     自動產生的描述**
  2. **GPT2  
     一張含有 文字, 軟體, 多媒體軟體, 螢幕擷取畫面 的圖片

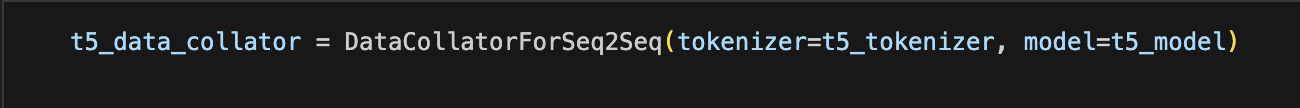
     自動產生的描述**首先在text欄位的後方加上「总结：」，並將他跟summary欄位一併輸入進tokenizer裡做padding。不先結合在一起再tokenized的原因是希望兩者之間不是默認唯一整個句子，而是兩個句子。將此新生成之結果標記為input。一樣，將summary欄位tokenized之後，設定為label，讓模型去做比較。一張含有 文字, 收據, 代數 的圖片

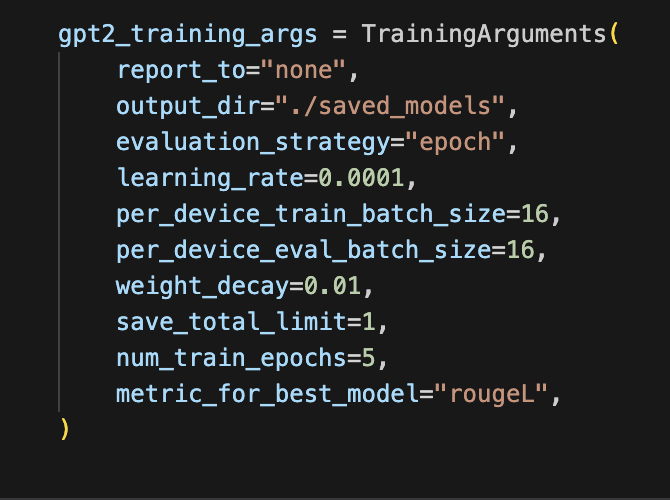
     自動產生的描述
* **Train**

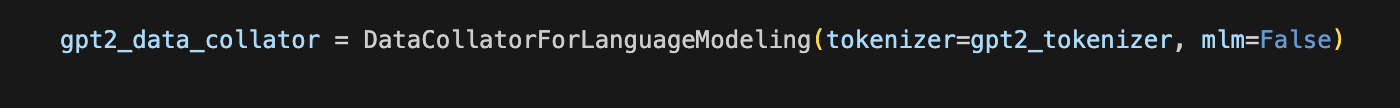
訓練方式是使用trainer以及fine tune的方式，而非從助教的範例code去做改動。

* 1. **T5**下圖為trainer的各個參數設定  
     **** **一張含有 文字, 螢幕擷取畫面, 字型 的圖片

     自動產生的描述**

要提到的是，使用的是Seq2Seq的training argument跟trainer，而非一10般情況下使用的trainer。  
同樣的，data collector也要配合T5模型的特性，去使用Seq2Seq的。  


* 1. **GPT2**下圖為trainer的各個參數設定 ** 一張含有 文字, 螢幕擷取畫面, 字型 的圖片

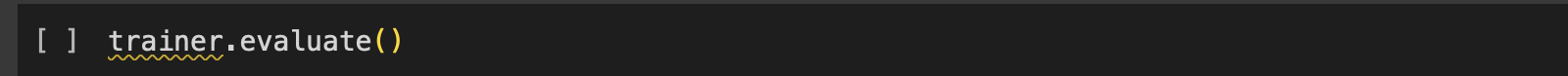
     自動產生的描述**GPT2則是使用一般的training argument跟trainer即可。  
     而在data collector的function上，我們選用的是生成語言使用的。  
     

最後，則是用trainer去呼叫train() 這個function，就可以開始訓練了。  
一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* **Evaluation**在計分方式上選用的是範例code中使用的「rouge」。  
  一張含有 字型, 螢幕擷取畫面, 文字, 圖形 的圖片

  自動產生的描述

並一樣使用trainer的function，也就是呼叫evaluate()。

* 1. **T5**下圖為T5使用的compute metrics  
     **一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

     自動產生的描述**  
     evaluate function評測出來的結果  
     一張含有 文字, 螢幕擷取畫面, 字型 的圖片

     自動產生的描述
  2. **GPT2**下圖為GPT2使用的compute metrics **一張含有 文字, 螢幕擷取畫面, 軟體, 字型 的圖片

     自動產生的描述**  
     evaluate function評測出來的結果  
     **一張含有 文字, 螢幕擷取畫面, 字型, 黑色 的圖片

     自動產生的描述**
* **Bonus** 
  1. **比較不同epochs數量下 loss rate的不同**
     1. **T5**

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, 行, 繪圖, 圖表 的圖片

自動產生的描述 一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

* + 1. **GPT2**

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, 行, 繪圖, 圖表 的圖片

自動產生的描述一張含有 文字, 行, 繪圖, 數字 的圖片

自動產生的描述

* 1. **比較不同epochs數量下 rouge的評分結果**
     1. **T5**

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

* + 1. **GPT2**

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

總結來說，無論是T5還是GPT2哪一個模型，並不是epochs數量愈大，效果就越好。