**作業二 報告**

**資訊114 H44091196 洪茂菘**

1. **Data**:

* Load the data and split them into train and validation (10 pts). You can use pandas, Dataset and Dataloader.



如上圖所示，我使用Dataset將資料做讀入，並且將其依9:1切分為train和validation。

* Generate your own version of data (10 pts), At least three two-digit addition and subtraction problems, and if you'd like, you can also try multiplication and division

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

我設計了一個隨機產生資料的function，其產生的結果分成expression和result，中間用逗號隔開，例如: ('(17+14)-12=', 19)。

* Tokenize the text (10 pts). You can design your own tokenizer or use any API (if available).

一張含有 螢幕擷取畫面, 文字, 多媒體軟體, 軟體 的圖片

自動產生的描述

一張含有 多媒體軟體, 軟體, 文字, 繪圖軟體 的圖片

自動產生的描述

以上第一張圖是切分token的部分，我是依照助教提供的範例程式碼下去做修改的，可以看到有成功將資料中的算式切成一個個的token；第二張圖則是將資料轉為id的結果。

1. **Generation Models**

* Model design (10 pts). You can use RNN, GRU or LSTM… whatever. (at least one sequential model)

LSTM:

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

RNN(bouns使用):

一張含有 文字, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

* Train(finetune) the model (10 pts).

訓練:

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

* Evaluate your model when you are training and test the model. (10 pts)

節錄部份訓練過程:

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. **Analysis**
2. **Model analysis**

model design

我的模型設計包括以下幾個部分，各層的功能如下：

* Embedding layer：用於將輸入序列中的每個標記映射到指定維度的向量表示，意即將src進行編碼。
* RNN layer：用於處理嵌入後的序列數據，提取序列中的時間訊息。
* Output layer：用於將 RNN 輸出的隱藏狀態映射到輸出標記的 logits。

Loss Reduction:

一張含有 文字, 螢幕擷取畫面, 繪圖, 圖表 的圖片

自動產生的描述

隨著iteration的增加，可以看到training和validation loss兩者皆有顯著地下降，代表訓練正常，該網絡仍在學習，且沒有過擬合(overfitting)的情況發生。

results of validation and testing:

一張含有 文字, 行, 繪圖, 圖表 的圖片

自動產生的描述

上圖是我訓練了50個epoch後所繪製的test accuracy，記錄了每次訓練後的acc率，可以看到其逐漸上升，最終停在約0.8。



一張含有 文字, 功能表, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

關於驗證的部分，上面兩張圖是我做驗證後的結果，其acc率為0.75。第二張圖是將使用模型生成的答案與正確答案做比較的結果，前面的算式是透過訓練產生的，空格後的數值則是正確答案，若前後結果相同我將其標為”正確”，反之則標記成”有誤差”，最後再透過計算出成功算對的次數有幾次來得到validation accuracy，我認為以第一次做這類型的作業來看，結果算是相當不錯了!

1. Dataset analysis

characteristics of the datasets and understanding of it:

本次作業的資料包含數學算式的輸入和輸出序列，其中輸入序列是算式的表達式，輸出序列是算式的結果。測試資料中的算式類型多樣，包含了加法、減法、乘法等不同類型的算式，以下將會對我生成的**兩種資料**下去作分析與比較:

一、20000 筆範圍 0~20 的資料: 二、20000 筆範圍 0~99 的資料:

一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述 一張含有 文字, 螢幕擷取畫面 的圖片

自動產生的描述

Dataset Adjustment and Impact

一張含有 文字, 螢幕擷取畫面, 軟體 的圖片

自動產生的描述

如同助教給予的提示所說，對於加減法的任務，我們須將算是拆解成1+2+3=和6兩部分，其目的是要讓模型去學習，並且在看到等號之後能產生答案及結數字符<eos>；此外，在計算loss時，只需預測等號的後面，故只要計算這部分即可，前面斜線部分可不用採計。

Results:

一、20000 筆範圍 0~20 的資料:

一張含有 螢幕擷取畫面, 文字, 繪圖, 行 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

此時的Loss都正常下降且收斂，代表在訓練得不錯，也無出現過擬合的問題，acc率也穩定上升。

二、20000 筆範圍 0~99 的資料:

一張含有 文字, 螢幕擷取畫面, 繪圖, 行 的圖片

自動產生的描述一張含有 文字, 螢幕擷取畫面, 字型, 繪圖 的圖片

自動產生的描述

可以看到Loss都有下降但無法像第一次完美收斂，且acc率的分數普遍都很低，我推測可能是由於資料數太少或學習率太高所導致。

1. Discussion

Impact of Different Learning Rates

不同的學習率會對模型的訓練過程和結果產生影響。

根據我的實測，學習率較大時，模型可能會快速收斂，但容易導致損失函數波動較大或無法收斂；反之若學習率較小時，模型收斂速度較慢，但可以更穩定地學習和優化模型參數。

而**本次作業我認為學習率在0.001時表現最佳**。

Impact of Different Batch Sizes

依據我多次改動Batch size的結果來看，不同的Batch大小會影響模型的訓練速度和泛化能力；另外較大的Batch大小可以加快訓練速度，但可能導致模型的泛化能力下降；而較小的Batch大小可以提高模型的泛化能力，但訓練速度較慢。本次作業我的Batch size最終採用400。

Model Characteristics

我使用的模型具有適用於處理序列數據的特點，包括嵌入層和循環神經網絡層。這種模型適用於處理具有時間序列結構的數據，如自然語言處理任務中的文本序列，以及這次作業中的數學算式序列。

1. Bonus (10 pts)
2. Compare the performance of multiple models and provide a brief analysis. (10 pts)

一張含有 繪圖, 螢幕擷取畫面, 行, 圖表 的圖片

自動產生的描述

一張含有 文字, 字型, 行, 繪圖 的圖片

自動產生的描述

上圖結果是使用RNN作為模型，可以明顯地觀察到，這裡accuracy rate較低，根據網路上查詢的資料解釋，LSTM訓練效果上更佳主要是因為它能有效解決梯度消失和梯度爆炸的問題，因此，在處理需要記憶較長時間序列訊息的任務時， LSTM通常能達到更好的訓練和預測效果。