Machine Learning HW8

ML TAs
ntu-ml-2020spring-ta@googlegroups.com

Slide update

- 5/7 : Policy (p19)
  FAQ (p22, p23, p24)
- 5/8 : 錯誤修正 (p19) <data directory> 路徑修正
- 5/10 : 新增 FAQ (p25)

Outline

- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

Outline

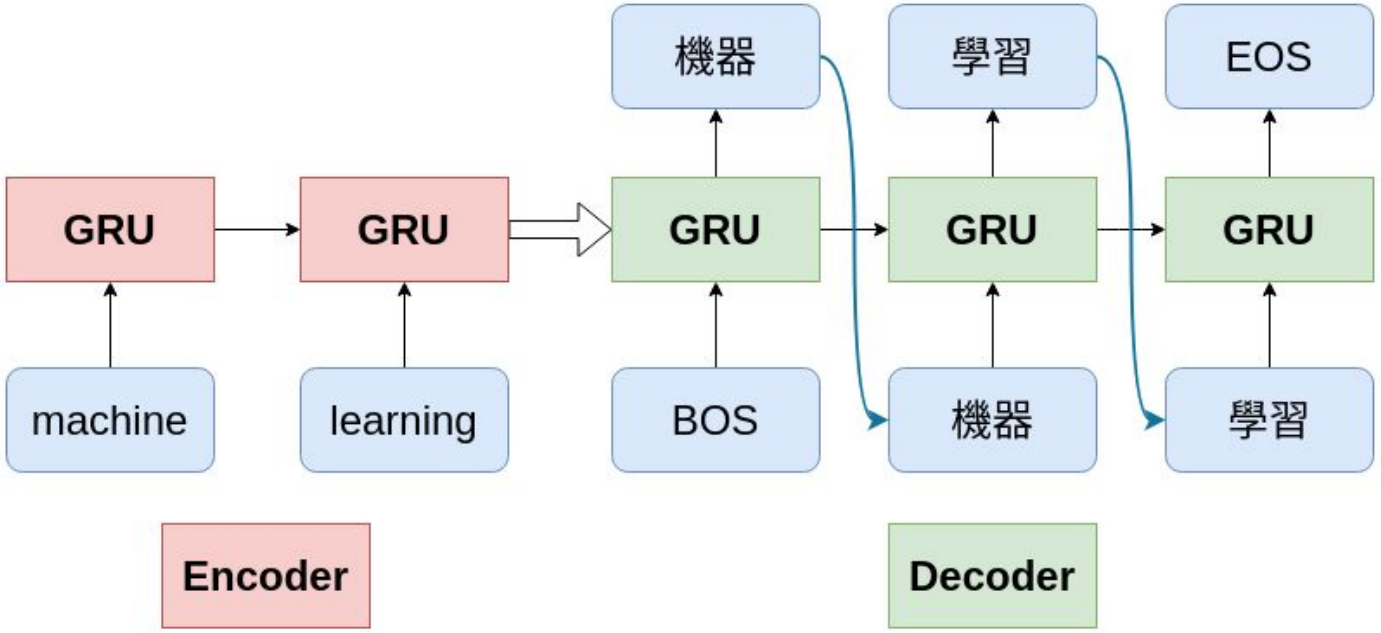
- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

Introduction

- 英文翻譯中文
  a. 輸入:
     一句英文（e.g., Tom is a student .）
  b. 輸出:
     中文翻譯(e.g., 湯姆 是 個 學生 。)

Sequence-to-Sequence Model

- **兩個 recurrent neural networks (RNNs)**
  - 第一個 RNN 為 **Encoder**
    - 將一句英文句子以一個向量表示
  - 第二個 RNN 為 **Decoder**
    - 根據 Encoder 的資訊遞迴輸出中文翻譯



Data Preprocess (1/2)

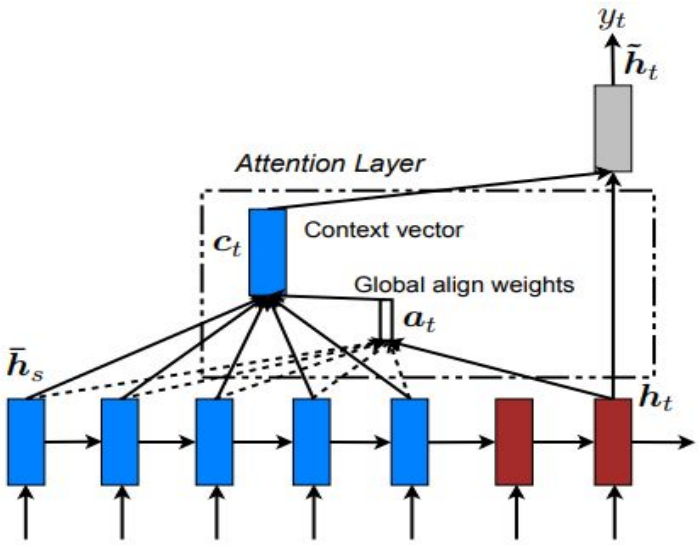
- **英文:**
  - 用 subword-nmt 套件將 word 轉為 subword
  - 建立字典:取出標籤中出現頻率高於定值的 subword
- **中文:**
  - 用 jieba 將中文句子斷詞
  - 建立字典:取出標籤中出現頻率高於定值的詞

Data Preprocess (2/2)

- 特殊字元 : <PAD>, <BOS>, <EOS>, <UNK>
  - <PAD> : 無意義，將句子拓展到相同長度
  - <BOS> : Begin of sentence, 開始字元
  - <EOS> : End of sentence, 結尾字元
  - <UNK> : 沒有出現在字典裡的詞
- 將字典裡每個 subword (詞) 用一個整數表示，分為英文和中文的字典，方便之後轉為 one-hot vector
  reference

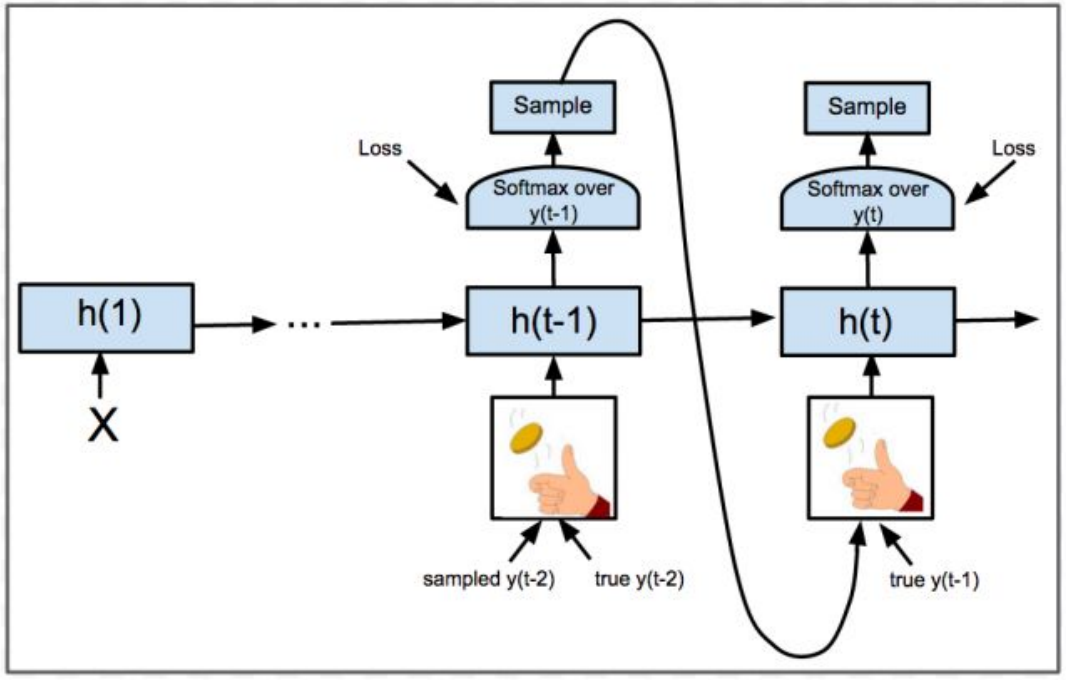
TODO - Attention (1/3)

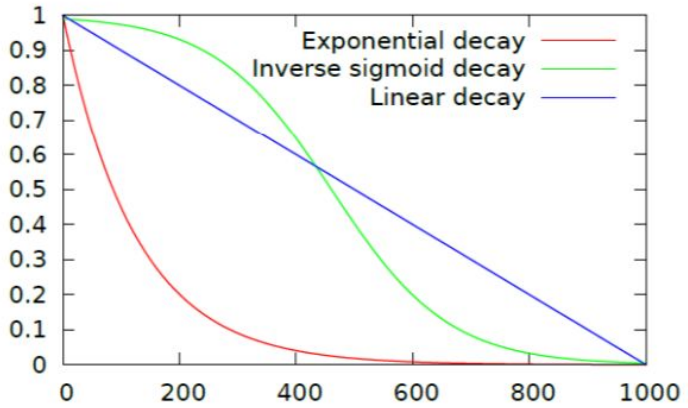
- 取出 Decoder 的隱藏向量與 Encoder 的隱藏向量做運算得到 attention weight
- 根據 attention weight[OBJ][OBJ] 對 Encoder 的隱藏向量做 weighted sum 得到 attention vector
- 將 attention vector 傳入 Decoder (相加或接在一起)



TODO - Schedule Sampling(2/3)

- 解決訓練和測試的不一致
- Decoder 的輸入有一定的機率使用模型本身預測的輸出

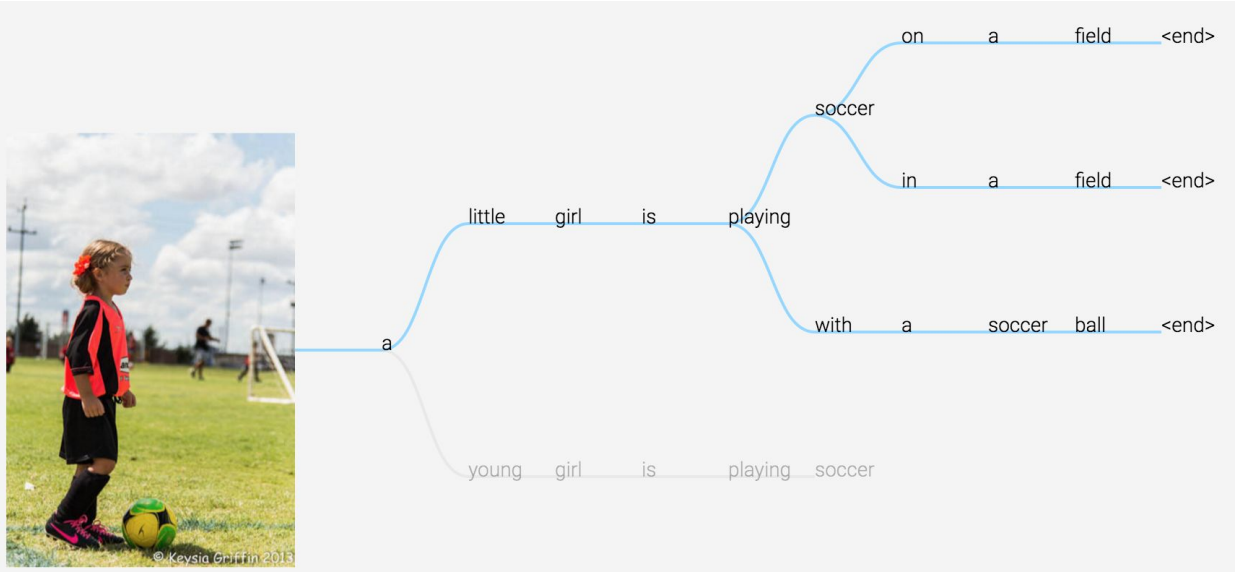




https://arxiv.org/abs/1506.03099

TODO - Beam search (3/3)

- 不在每次取機率最大的字當答案，因為可能產生區域最佳解而非全域最佳解
- 窮舉所有可能現實中不太可行，所以每個 Decoder step 固定取當前生成句子機率前 K 大的句子



Demo：http://dbs.cloudcv.org/captioning

Evaluation Metrics

- BLEU@1
  - Precision = 正確字數 / c

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c 是要計算的句子長度, r 是目標句子的長度

  - BLEU@1 = BP * Precision
  - e.g.:
    正解: ['我', '不', '知', '道', '我', '有', '沒', '有', '時', '間', '。']
    預測: ['我', '不', '知', '道', '我', '是', '否', '時', '間', '。']

BLEU@1: $e^{1-\frac{11}{10}} * \frac{8}{10} = 0.723869$

Outline

- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

Data & Format(1/2)

- Data (出自 manythings 的 cmn-eng):
  - 訓練資料:18000句
  - 檢驗資料:　500句
  - 測試資料: 2636句
- Format:
  - 不同語言的句子用 TAB ('\t') 分開
  - 字跟字之間用空白分開

```
his house is across from mine .	他 的 房子 在 我 的 對面 。
her only pleasure is listening to music .	她 唯一 的 樂趣 就是 聽 音樂 。
they go to church every sunday .	他們 每個 星期天 上 教堂 。
do you want to die here ?	你 想 死 在 這裡 嗎 ？
he did n't study at all .	他 根本 就 沒有 學習 。
my wallet and passport are missing .	我 的 錢包 和 護照 不見 了 。
it 's very hot in this room .	這間 房裡 很熱 。
i 've just been to my uncle 's house .	我 剛剛 去 了 我 叔叔 家 。
tom has loved mary for a long time .	湯姆 愛 瑪麗 很久 了 。
when are you coming home ?	你 什麼 時候 回家 ？
sit wherever you like .	你 愛坐 哪裡 就 坐 哪裡 。
will six o'clock suit you ?	六點鐘 你 方便 嗎 ？
he threw a stone at the dog .	他 朝着 狗 扔 了 塊 石頭 。
i had to ab@@ sta@@ in from smoking while i was in the hospital .	在 醫院 的 時候 ， 我 不得不 戒菸 。
```

Data & Format (2/2)

- 詞庫：
  - int2word_*.json: 將整數轉為文字


```
{"0": "<PAD>", "1": "<BOS>", "2": "<EOS>", "3": "<UNK>", "4": ".", "5": "i", "6": "the", "7": "to", "8": "you", "9": "a", "10": "?",
"11": "is", "12": "he", "13": "n't", "14": "tom", "15": "do", "16": "in", "17": "it", "18": "'s", "19": "of", "20": "my", "21": "she",
"22": "have", "23": "me", "24": "this", "25": "that", "26": ",", "27": "was", "28": "for", "29": "we", "30": "are", "31": "what", "32":
```


  - word2int_*.json: 將文字轉為整數


```
{"<PAD>": 0, "<BOS>": 1, "<EOS>": 2, "<UNK>": 3, ".": 4, "i": 5, "the": 6, "to": 7, "you": 8, "a": 9, "?": 10, "is": 11, "he": 12,
"n't": 13, "tom": 14, "do": 15, "in": 16, "it": 17, "'s": 18, "of": 19, "my": 20, "she": 21, "have": 22, "me": 23, "this": 24, "that":
25, ",": 26, "was": 27, "for": 28, "we": 29, "are": 30, "what": 31, "your": 32, "on": 33, "his": 34, "at": 35, "like": 36, "did": 37,
```


  - *：分為英文(en)和中文(cn)

Outline

- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

Submission Format - GitHub

- GitHub 上的 hw8-<account> 必須包含（注意格式）:
  - report.pdf
  - hw8_train.sh
  - hw8_test.sh
  - other python code

  - 請不要上傳 dataset 和 output files
  - model file 請上傳至雲端 (Dropbox, ...)，在 script 中寫好下載的指令，並寫好載入路徑

Outline

- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

Policy

1. 資料路徑，助教在跑的時候會另外指定，請保留可更改的彈性，不要寫死。
2. Script usage:
   bash hw8_train.sh <data directory>
   bash hw8_test.sh <data directory> <output path>
   <data directory> 為所有需要資料的資料夾目錄，也就是cmn-eng 這個資料夾，training 跟 test 的檔名皆跟公佈的一樣，只是 cmn-eng 資料夾的位置需要動態
   <output path> 皆為檔案路徑，並不需要再加其他路徑，
   e.g. <output path> 為 './b0xxxxxxx/output.txt'
3. 除非有狀況，不然原則上助教只會跑 testing，不會跑 training，因次請用讀取 model 參數的方式進行預測。
4. Testing 時間限制是在二十分鐘之內跑完。

Report

1. **Teacher Forcing:**
   a. 請嘗試移除 Teacher Forcing，並分析結果。
2. **Attention Mechanism:**
   a. 請詳細說明實做 attention mechanism 的計算方式，並分析結果。
3. **Beam Search:**
   a. 請詳細說明實做 beam search 的方法及參數設定，並分析結果。
4. **Schedule Sampling:**
   a. 請至少實做 3 種 schedule sampling 的函數，並分析結果。

Outline

- Task Description
- Data Format
- Submission Format (Code, Report)
- Policy
- FAQ

FAQ

- Q ：需要寫算 BLEU@1 的部份嗎？
- A ：不用

- Q ：下載 model 的時間算在20分鐘內嗎？
- A ：原則上，助教只會跑 attention + bean search 版本 (最完整的版本)
  只要確定該 model 能在 20 分鐘內跑完即可

FAQ

- Q ：test 是要把 prediction 直接輸出到 stdout 嗎?
  還是要寫到某個特定對檔案？
- A ：~~stdout, 不要有其他的輸出, 一行一句~~
  為避免有人會不小心在 stdout 輸出其他資訊
  因此 prediction 改為以 bash 指定輸出檔案
  仍然為一行一句

- Q ：Attention 一定要接在 decoder 的 input 後面嗎？
- A ：不一定，接在 decoder 的 input 後面是較為常見的作法，同學們可以嘗試其他接法，並可以在 report 裡討論

FAQ

- Q ：關於 hw8_train.sh 要訓練的模型是需要各自部分 (attention or beam search or schedule sampling) 的訓練 code 還是只需要 hw8_test.sh 所使用模型的訓練 code
- A ：只需要 hw8_test.sh 所使用模型的訓練 code

- Q ：請問更改的輸出範例中 "bash hw8_test.sh <data directory> <output path>" 是指寫一個results.csv到<output path>的資料夾內嗎?
- A ：因為輸出為一行一句，所以不用存成 .csv 格式 <output path> 本身就是檔案名字，所以直接寫進去即可，e.g. 助教可能會輸入 ’./b0xxxxxx/output.txt’

FAQ

- Q ：如果在助教這跑超過 20 分鐘會如何?
- A ：會斟酌扣分，但在助教這跑的 test data 大約是 200 筆，所以應該不太擔心

FAQ

- 若有其他相關問題，請在 FB 社團貼文或寄信至助教信箱，**請勿直接私訊助教**。
- 助教信箱：ntu-ml-2020spring-ta@googlegroups.com

Link

- Colab: http://bit.ly/2uIq107
- Data: http://bit.ly/2wWLkfa
- Report template: http://bit.ly/32HaREZ
- 遲交表單: https://bit.ly/39d2x2m