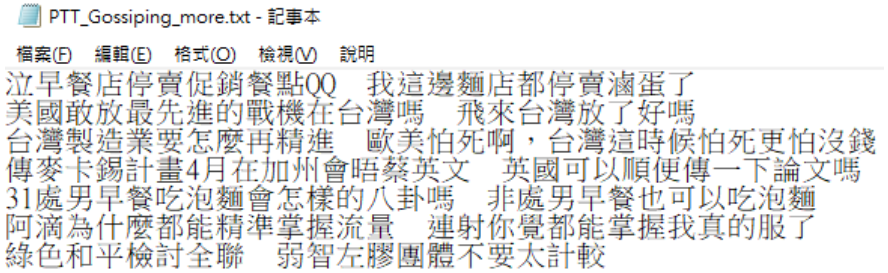
Dataset

- Collect from PTT Gossiping.
- Use CKIP to do word segmentation.
  - title & content & comment.
- The more times a word appears in a comment,

  the more it is mentioned in the content and title,

  this comment will be selected.
  - If not, randomly pick one.
- Total: 14,7882
  - Max sequence length = 41
  - Vocab size = 5289


PTT_Gossiping_more.txt - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

泣早餐店停賣促銷餐點QQ 我這邊麵店都停賣滷蛋了
美國敢放最先進的戰機在台灣嗎 飛來台灣放了好嗎
台灣製造業要怎麼再精進 歐美怕死啊，台灣這時候怕死更怕沒錢
傳麥卡錫計畫4月在加州會晤蔡英文 英國可以順便傳一下論文嗎
31處男早餐吃泡麵會怎樣的八卦嗎 非處男早餐也可以吃泡麵
阿滴為什麼都能精準掌握流量 連射你覺都能掌握我真的服了
綠色和平檢討全聯 弱智左膠團體不要太計較

Q


作者 Gshan (ccshpengyuyan)
標題 [問卦] 3C回收商賺什麼
時間 Wed Jun 14 10:30:32 2023

是這樣的啦 最近想買新macbook 好讓我能繼續坐在星巴克爽

想說試試看舊機換新機 加減省 就可以多買幾杯星冰樂了耶

結果查了一下民間回收商報價 一查不得了

macbook pro 2017 128G還有12000的報價是怎樣啊

是報價報爽的 還是這個真的有賺頭啊

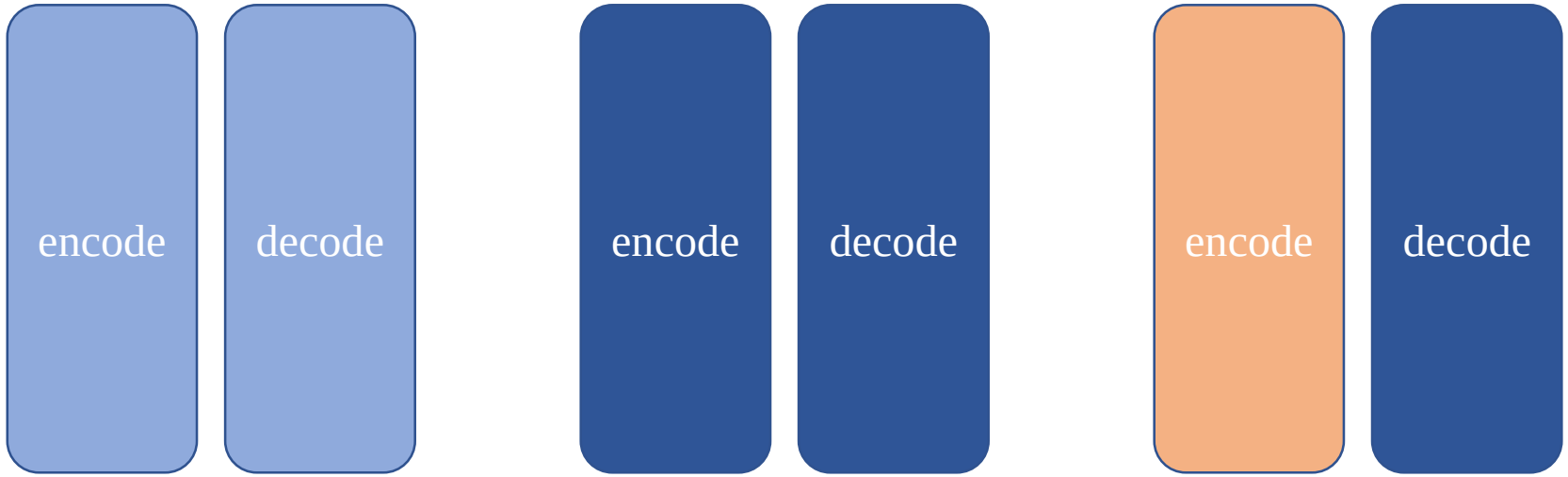
回收商都賺啥 有掛嗎

--
Sent from nPTT on my iPhone X

--
※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 42.72.136.84 (臺灣)
※ 文章網址: https://www.ptt.cc/bbs/Gossiping/M.1686709834
→ bill403777: 都在做慈善的 沒賺你錢啦
推 brianuser: 拿去賣20000啊
→ vowpool: 做逆向工程吧 然後剩餘的拿去殺肉
→ yesonline: 報價僅供參考. 看實機時會東扣西扣....

A

Model Architecture(PyTorch)

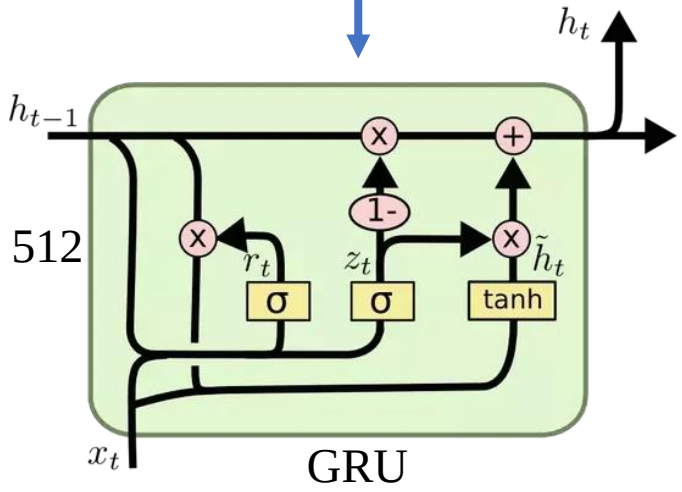


GRU-GRU LSTM-LSTM Bert-RNN

Train RNN based

Q:你好嗎? → index[10,12,32,..,0]

Embedding(size: 200) shape: (5289,200)



GRU

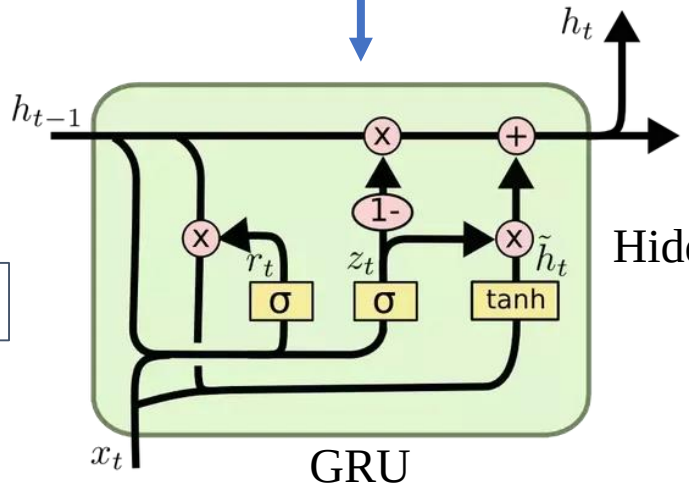
Hidden size = 512

Hidden (Batch, 41, 512)

Last hidden(Batch, 1, 512)

A:不好→[<sos>,25,12,<end>, … ,0]

Embedding(size: 200) shape: (5289,200)



GRU

Hidden size = 512

Output (Batch, 41, 512)

Liner (5289)

Train Bert Encode、RNN Decode v1

Q:你好嗎?

→

index[20,24,64,..,0]

Bert tokenizer

Untrainable
Bert Model

Last hidden(Batch, 1, 768)

Cell State

Hidden State

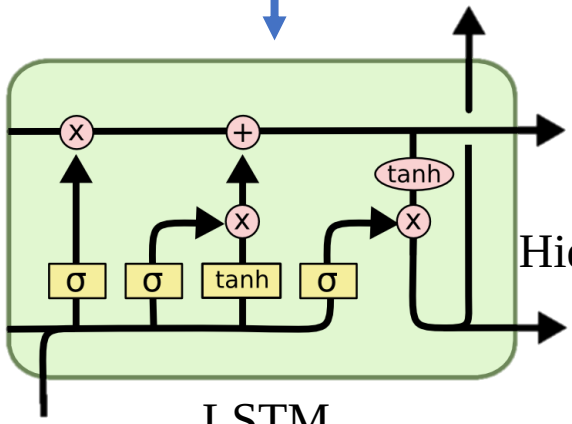
A:不好

→

index[<sos>,3,24,<end>, … ,0]

Bert tokenizer

Embedding(21128,100)



Hidden size = 768

LSTM

Output (Batch, 41, 768)

Liner (21128)

May not realize the Question.

Train Bert Encode、RNN Decode v2

Q:你好嗎?

→

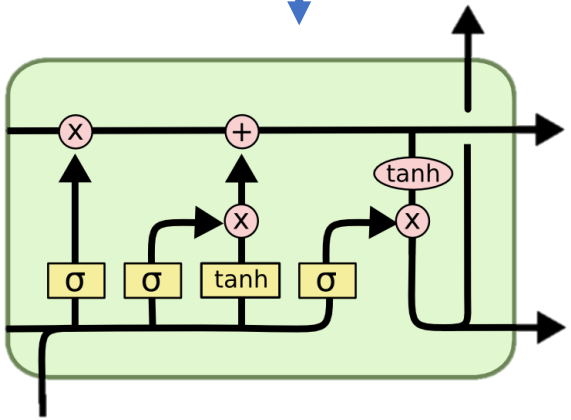
index[20,24,64,..,0]

Bert tokenizer

Untrainable
Bert Model

Hidden (1, 41, 768)

Hidden size = 512



Cell State

Hidden State

Last hidden(Batch, 1, 512)
Last cell(Batch, 1, 512)

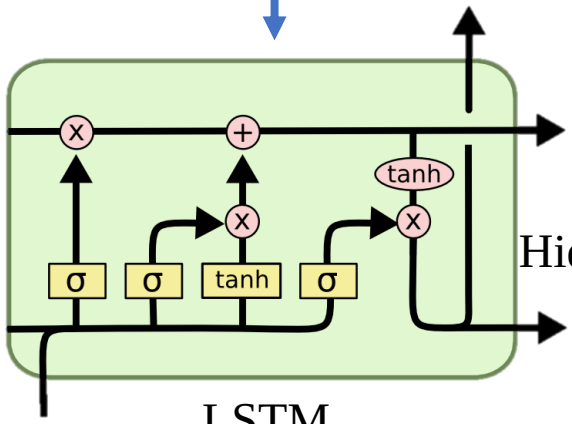
A:不好

→

index[<sos>,3,24,<end>, … ,0]

Bert tokenizer

Embedding(21128,100)



Hidden size = 512

LSTM

Output (Batch, 41, 512)

Liner (21128)

Train Bert Encode、RNN Decode v3

Q:你好嗎?

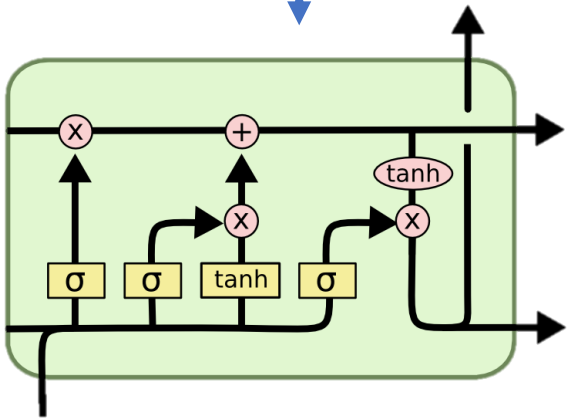
→

index[20,24,64,..,0]

Bert tokenizer

Untrainable
Bert Model

Hidden (1, 41, 768)



Hidden size = 512

Cell State

Hidden State

Last hidden(Batch, 1, 512)
Last cell(Batch, 1, 512)

A:不好

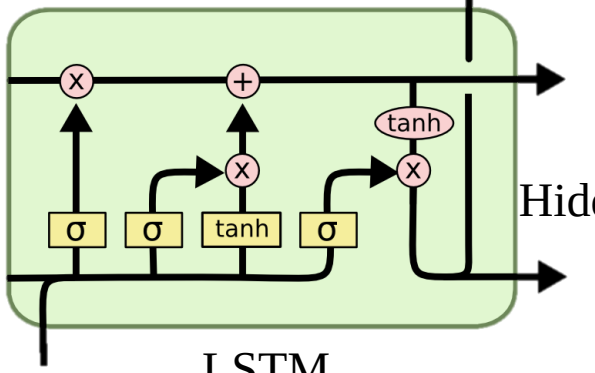
→

index[<sos>,3,24,<end>, … ,0]

Bert tokenizer

Untrainable
Bert Model

Hidden (1, 41, 768)



Hidden size = 512

LSTM

Output (Batch, 41, 512)

Liner (21128)

Accuracy self define

- PyTorch accuracy calculation needs to write by self.
- Because the model always predicts zero very correctly.
  - Including zeros for correct accuracy may lead to very high accuracy.

  - EX: Total sequence length = 41
    - Model Predict : [10,11,12,13,14,15,16,17,18,19,20, …, 0] (11 elements not zero, 30 zeros)
    - Target Answer : [10,11,12,17,17,18,19,17,18,19,20, …,0]

    - Original accuracy = $\frac{37}{41}$ = 0.902
    - Self define accuracy = $\frac{37-30}{41-30} = \frac{7}{11}$ = 0.636 (this may be easier know predict result)

Model details

	GRU-GRU	LSTM-LSTM	Bert-LSTM v1	Bert-GRU v2	Bert-LSTM v2	Bert-LSTM v3
LOSS	nn.CrossEntropyLoss					
Optimizer	Adam					
Embedding	200	200	100	100	300	x
RNN hidden	512	512	768	512	512	512
Output Liner	5289	5289	21128	21128	21128	21128

- 為什麼韓國那麼多瑜伽老師 =>https://i.imgur.com/WYEOIze.gif今年訕

```
  0%|          | 0/14 [00:00<?, ?it/s]
Q: 為什麼人工智慧在這今年炸開了
A: https://i.imgur.com/WYhNEjl.jpg
```

```
  0%|          | 0/17 [00:00<?, ?it/s]
Q: 王建民現在出來投還是屌打一堆投手嗎
A: 問他
```

```
  0%|          | 0/11 [00:00<?, ?it/s]
Q: 肥宅BMI值是多少才算
A: 肥宅滾
```

```
  0%|          | 0/5 [00:00<?, ?it/s]
Q: 高雄發大財
A: 高雄起飛
```

```
  0%|          | 0/13 [00:00<?, ?it/s]
Q: 為什麼韓國那麼多瑜伽老師
A: https://i.imgur.com/yfasGkf.jpg
```