Homework1 Report

R07922058 張友誠

Q1. Data preprocess

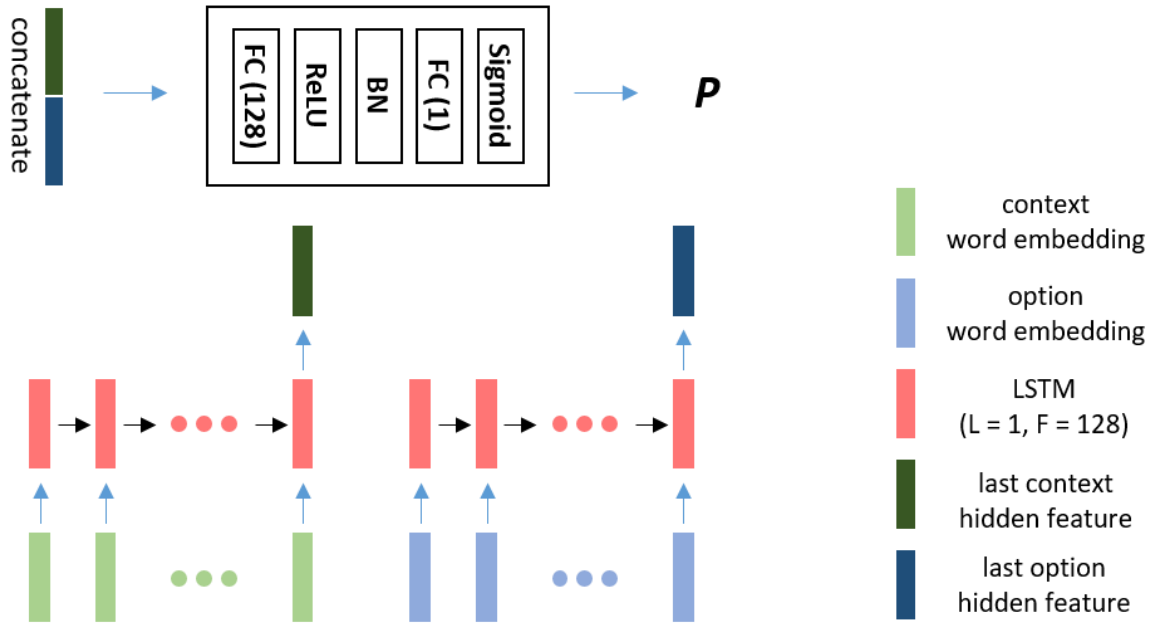
a. Use **nltk package** to tokenize the sentence

b. Positive sample : Negative sample = 1 : 9

c. Utterance length's maximum in the training process is **350**, and option's is **50**. In the validation and testing process, they are 300 and 50 respectively.

d. FastText crawl-300d-2m.vec

Q2. RNN w/o attention model



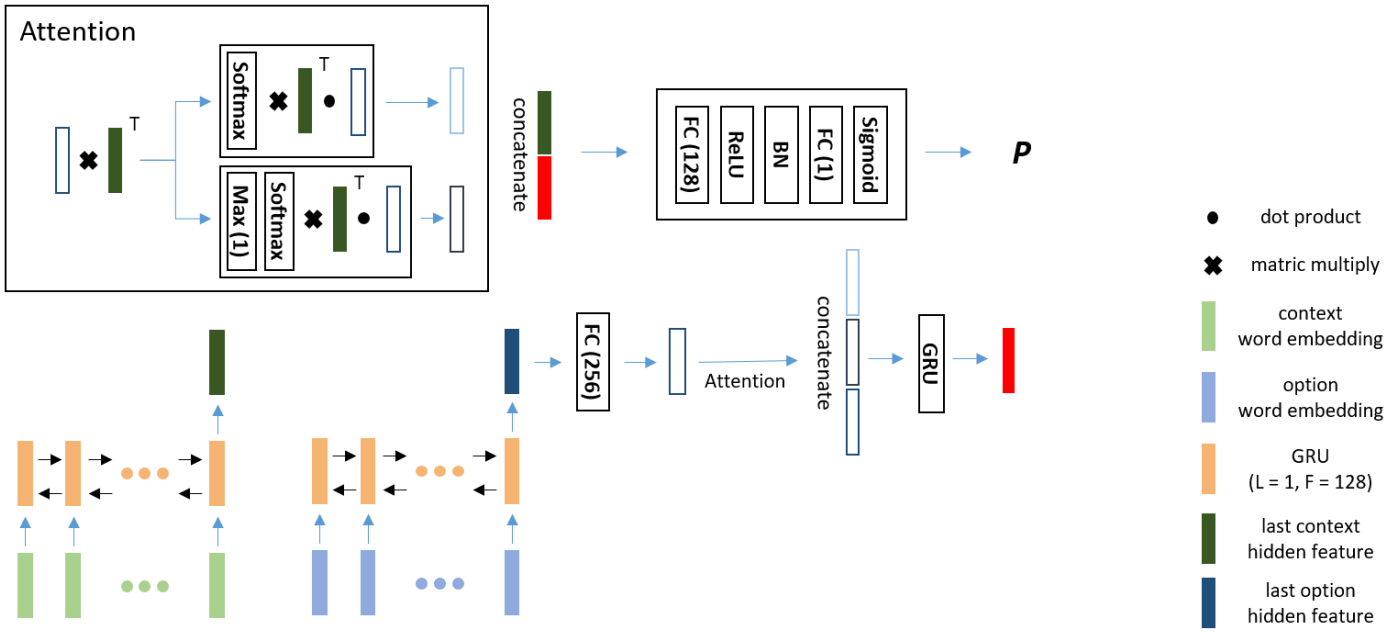
Public score: 9.66666

Validation recall@10: 0.6186

Loss function: Binary Cross Entropy

Optimizer: Adam (lr = 1e-3, bs = 128)

Q3. RNN w/ attention model



Public score: 9.41333

Validation recall@10: 0.7246

Loss function: Binary Cross Entropy

Optimizer: Adam (lr = 1e-3, bs = 128)

Ref. BiDAF (https://arxiv.org/pdf/1611.01603.pdf)

Q4. Best model

The best model is similar to the attention model described previously. However, the difference is that it applies the **LSTM cell** in the model instead of the GRU cell. And it not only uses the last hidden features but also **the mean of each hidden feature** to calculate the similarity between the contexts and the options. With more information from the utterances and options, the MLP used to calculate the similarity can perform better. That's why this model outperforms the RNN w/ attention model.

Public score: 9.3800

Validation recall@10: 0.7328

Loss function: Binary Cross Entropy

Optimizer: Adam (lr = 1e-3, bs = 128)

Q5. Compare LSTM and GRU

The following experiments are based on the model **without** attention layers.

a. Validation recall@10

GRU: 0.6380

LSTM: 0.6186

b. Public score

GRU: 9.5666

LSTM: 9.6666

c. Required GPU memory

GRU: 1024 MB

LSTM: 1713 MB

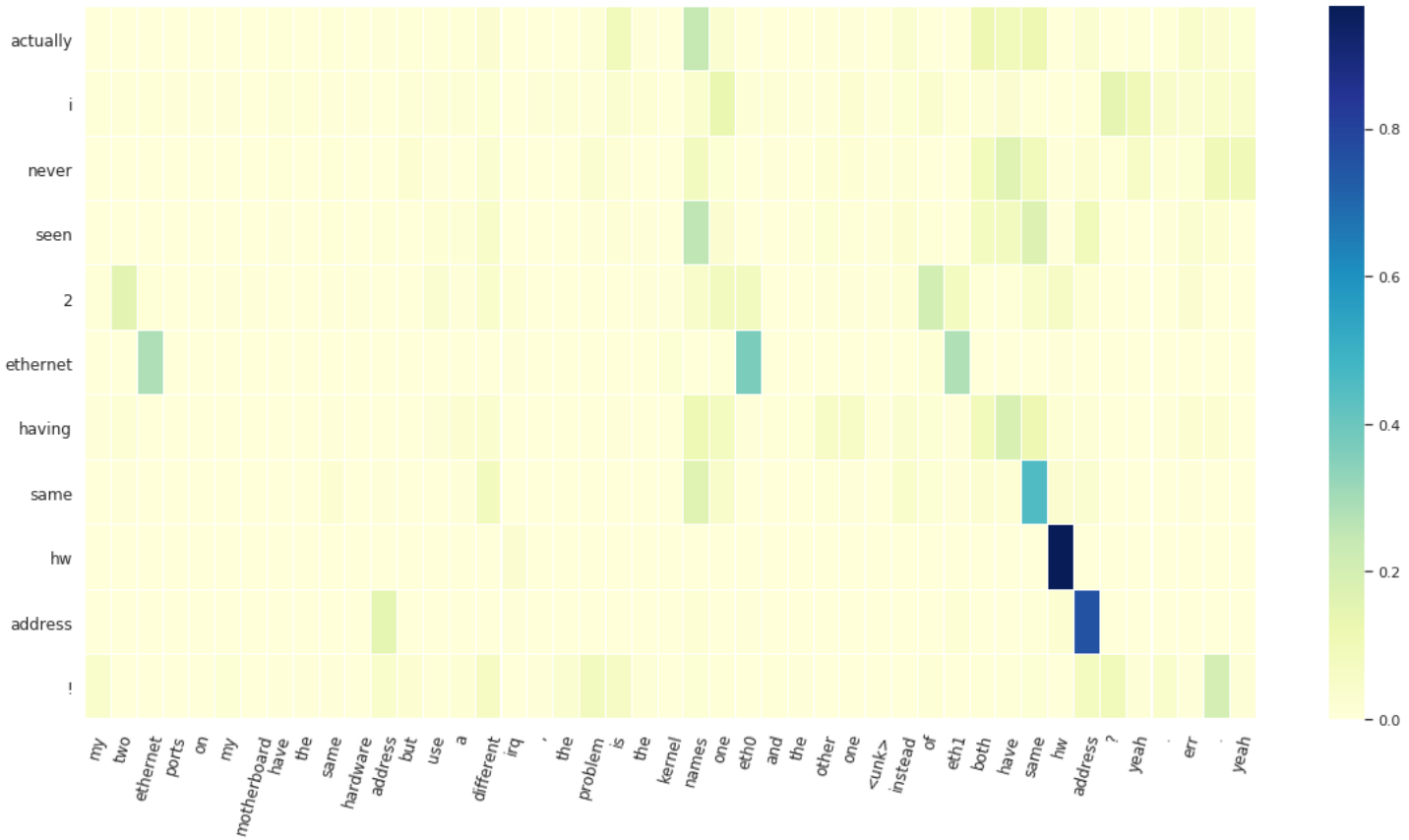
d. Training / testing speed

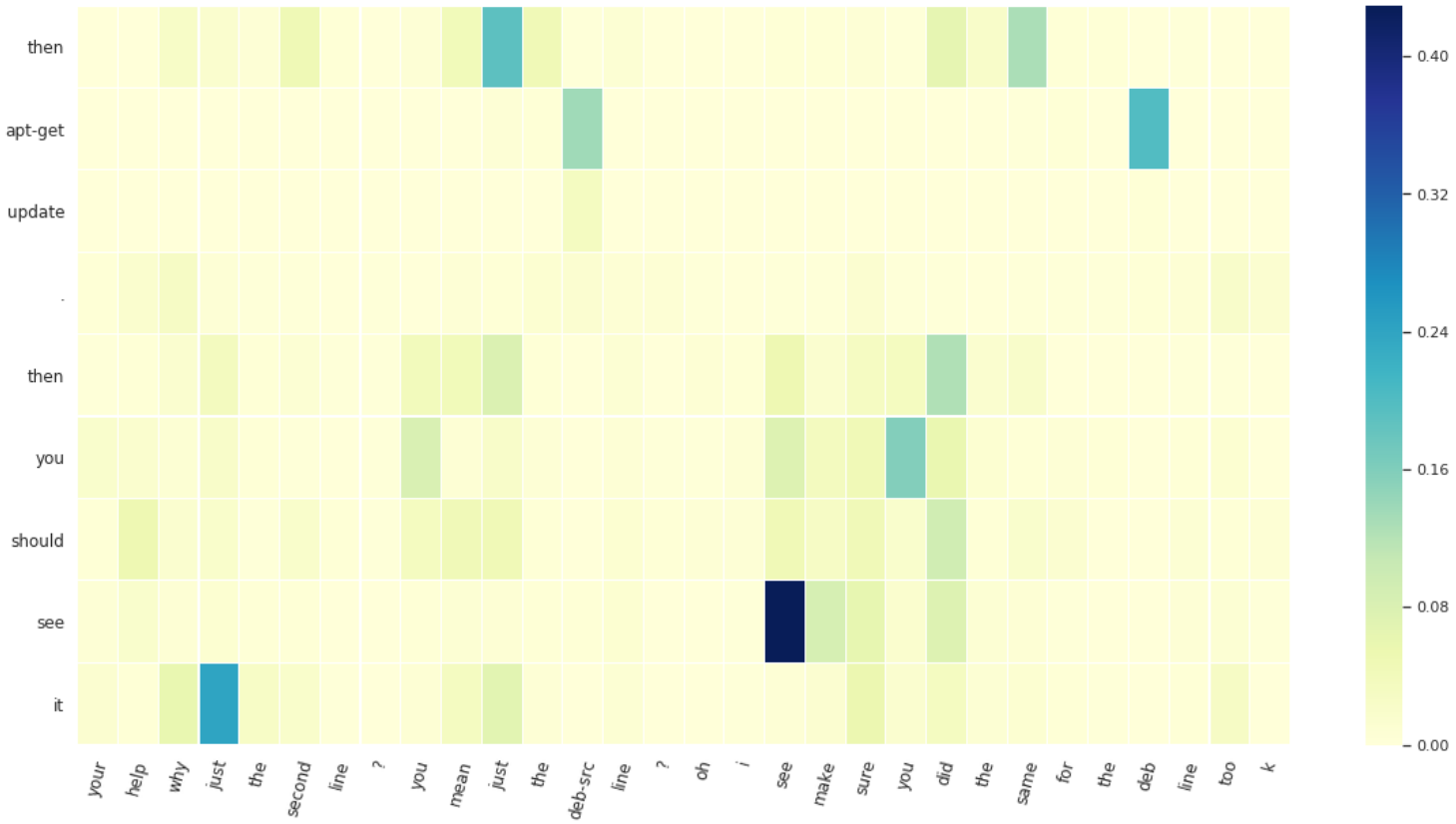
GRU: train 4.22 it/s; test 2.44 s/it

LSTM: train 3.91 it/s; test 3.76 s/it

Q6. Visualization of the attention map

(package: seaborn.heatmap)





Finding:

a. **2 : two**

numbers of the digit form and in English have the strong attention score

b. **apt-get : deb-src, deb**

linux install command have strong score with the package file format

c. **Overall**

- The words which are **the same or similar semantic meaning** are always have high score
- Verb and corresponding noun would get the high score

Q7. Compare different setting

a. **Different negative samples**

(Positive: Negative) The performance of 1:4 is worse than 1:9.

And the setting of the 1:4 will cause overfitting seriously and early.

b. **Different number of utterances in dialog**

The length 350 (or 400) is long enough to cover almost contexts in the data.

So if the length is too long (> 400), it would lead to a poor outcome.