Yunhao CHEN
yunhao.chen@polytechnique.edu

Question 1

We could borrow the idea of Transformers: using multi-head attention to capture different important components in the sentence. In fact, this idea should be originally from CNN, which uses multi-kernel.

Moreover, the paper [1] offers a penalization term to avoid redundancy problems (i.e. the attention mechanism always provides similar summation weights)

Question 2

The recurrent operation precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. One big advantage of Transformer is its parallelization.

Bonus Question

Bonus question: What is the purpose of the parameter my_patience?
Answer: It is a hyperparameter for early stopping (which is a trick to avoid overfitting). If the validation accuracy does not improve for 'my_patience' epochs, the training process will stop immediately.

Bonus task

Bonus task: use tensorboard to visualize the loss and the validation accuray during the training.

Question 3

Fig. 1 represents the sentences in order and their weights in the chosen document. Fig. 2 shows a stem plot of attention weights at sentence level. I found that the 3rd sentence ("That's called pointless foreshadowing.") is the most important. And if we look at the Fig. 3 which shows the attention weights at word level, we could notice that in the 4rd sentence, the most important word is "pointless".
I think HAN works well on this document. It has selected the most discriminatory sentence (the 3rd one), which reflects some negative emotion. At word level, it focuses on the word "pointless".

```
13.33 OOV : Honey , here 's them eggs you ordered .
23.72 Honey , like bee , get it ?
39.62 That 's called pointless foreshadowing .
5.73 Edward Basket : Huh ?
4.38 ( On the road ) Basket : Here 's your doll back , little girl .
4.85 You really should n't be so careless with your OOV .
8.37 Little girl : Yeah , whatever .
```

Figure 1: Sentences and their attention weights

Question 4

In this paper [2], Remy insists that HAN completely ignores the other sentences while producing the representation of a given sentence in the document. The lack of communication between sentences leads to repeat distribution of attention on same salient feature.

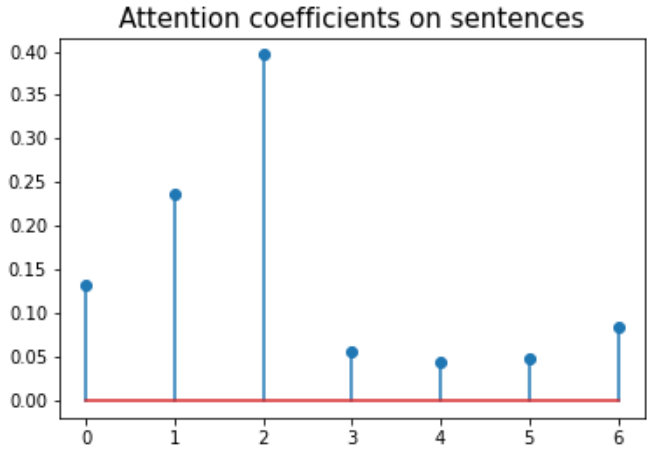

Figure 2: Attention coefficients on sentences

References

[1] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017.

[2] Jean-Baptiste Remy, Antoine Jean-Pierre Tixier, and Michalis Vazirgiannis. Bidirectional context-aware hierarchical attention network for document understanding. *CoRR*, abs/1908.06006, 2019.

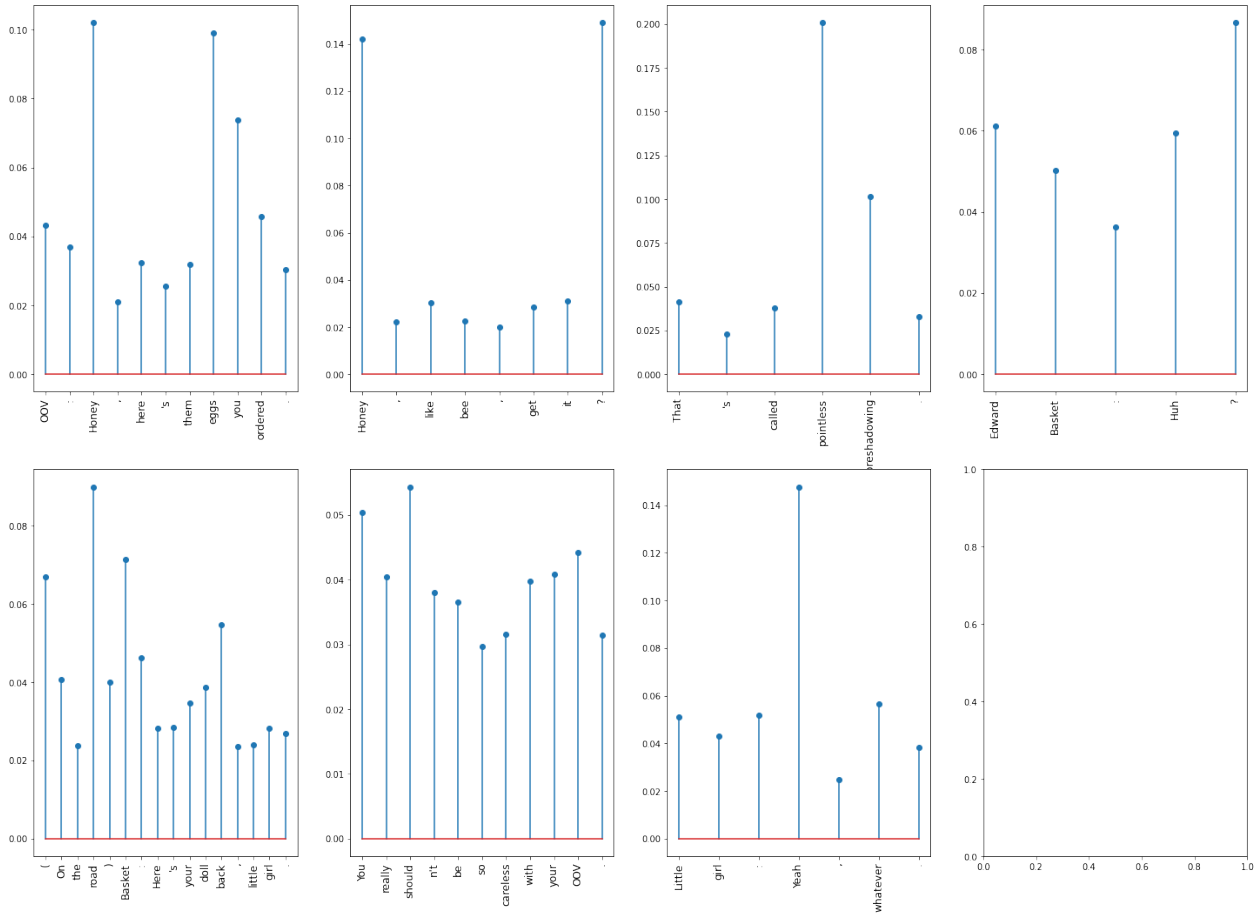

Figure 3: Weights in sentence