Matias Etcheverry
matias.etcheverry9@gmail.com

Lab session # 1
ALTEGRAD 2022

10/31/22

1 Question 1

The self-attention can be improved as for now, we don't force the weights to be highly differents. Therefore, the current model could lead to an optimized situation where all the weights are equal, which is obviously not an "attention" mechanism.
The paper [1] proposed a new regularization to force weights to have diversity and thus highlights interesting parts of a sentence. Let A be the matrix of normalized weights, and the $||.||_F$ Frobenius norm, they introduce a new penalization as:

$$P = ||AA^T - I||_F^2 \tag{1}$$

Minimizing the penalization allows to have a matrix A as heterogeneous as possible. The model focuses on as few number of words as possible, forcing each sentence to be focused on a single aspect, which punishes redundancy between different summation vector.

2 Question 2

As stated by [3], the problem about recurrent networks is that the result of input x_t is necessary for computing the result of input x_{t+1}. Thus, we can't parallelize much of the recurrent model's training.
Replacing recurrent operations with self-attention allows to draw global dependencies between input and output, and thus reaching state-of-the-art results in a minimum training time.

3 Question 3

Figure 1 shows the score obtained on a test document. The more red the word/ sentence is, the higher coefficient it has. This document was correctly assigned the value 0, which is a bad review.

We can see that the model set highly diversified coefficients on the document. For instance, **WAY** which often interprets exaggeration receives a high score. We also notice that the model learns from proper noun and localization. Thus, **Hollywood** and **New Jersey** may present high coefficients. This is related to the fact Hollywood or New Jersey is often depicted in movies.

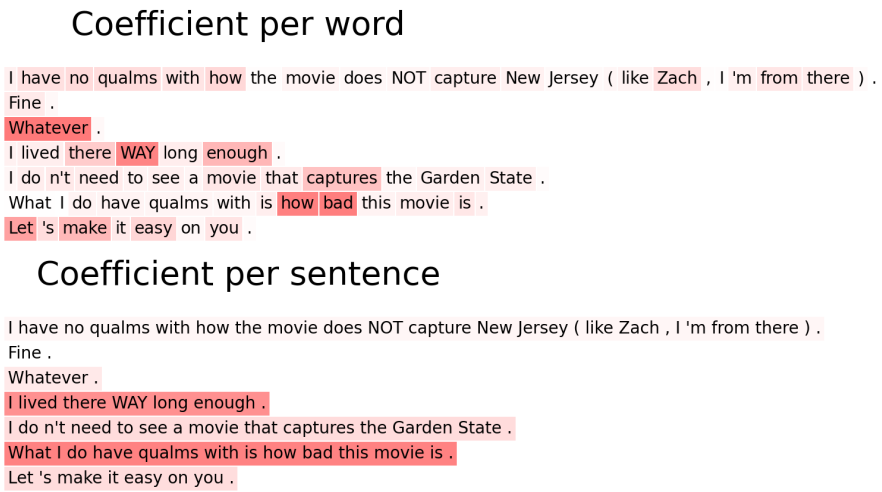

Figure 1: Coefficients obtained per word and per sentence.

4 Question 4

The main problem with HAN architecture is that the sentences are encoded in isolation. Their encoding don't depends on other sentences at all. Thus, it may be difficult to understand logical process in a paragraph. Another problem due to this isolation is that the sentence encoding doesn't prevent redundancy. If a paragraph contains a lot of bad sentences, then the coefficients of those sentences will "add" up. In real life, the redundancy shouldn't be taken into account.
In this context, [2] proposed an updated version of HAN, where a context vector is injected into the self-attention mechanism, to guide the model during the computation of the word alignment coefficients.

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.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.