Martin Jolif
martinjolif@gmail.com

Lab session # 1
ALTEGRAD 2024

10/13/24

1 Question 1

The basic self-attention mechanism seen in the lab as a weighted sum of the annotations is not enough to represent the overall semantics of the sentence. Indeed, the attention vector representation from the basic self-attention only focuses on a specific part of the sentence. However, to be able to understand the overall semantics of a sentence, we need to be able to extract different aspects of the sentence. The authors of the paper [2] proposed a solution to focus on several parts of a sentence: instead of representing the attention by a vector, they represent the attention by a matrix of size $r * n$ (r is the number of different parts to be extracted from the sentence and n the number of tokens of the sentence) in which every row represents the attention vector of a part of the sentence. Moreover, the authors add a penalization term to their work to avoid that all the rows of the matrix will be the same: it discourages the redundancy in the embedding. The association of the use of a matrix instead of a vector and the penalization method makes attention a more complex and sensitive object.

2 Question 2

Recurrent operations such as RNNs (LSTM or GRU) where at that time state of the art approaches to solve sequence modeling like language modeling, for example. However, self-attention has replaced the recurrent operations for different reasons. First, in recurrent neural networks, the result of x_t depends on the result of x_{t-1}, therefore we are not able to parallelize our calculations (step by step calculations). On the other hand, self-attention can process all elements of a sequence simultaneously, allowing for much faster training on modern hardware like GPUs. Moreover, self-attention is able to have access to both past and future elements of a sequence instead of just a representation of the last few elements in the RNN case (problem of long-term dependencies because of vanishing gradient). Therefore, with self-attention, we are more able to make links between words that are far away in a sentence, for example. Finally, the interpretability of the link between the elements of a sequence is much easier through self-attention than through RNN (see figure 1).

Figure 1: Self-attention interpretation from [3]

3 Question 3

I chose to plot the attention coefficient for the last document, like in figure 2 :

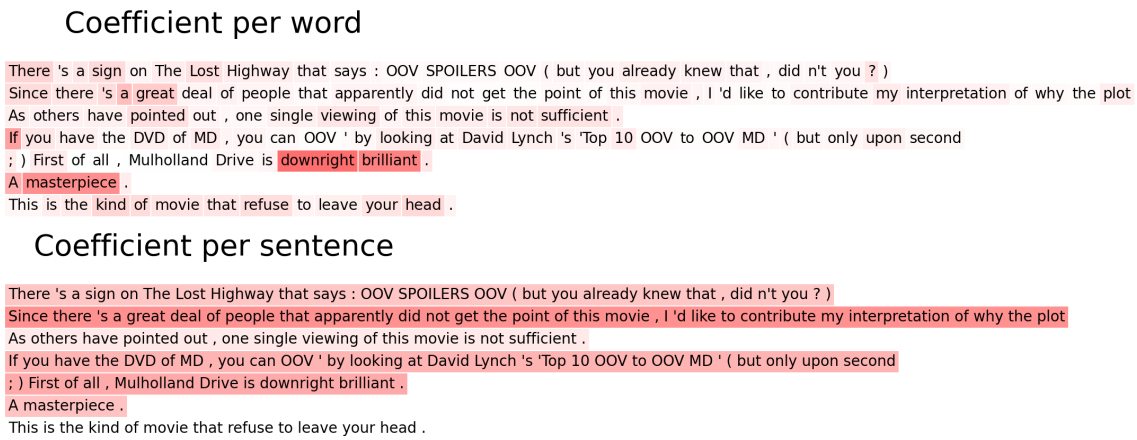

Figure 2: Representation of the self-attention value by word and sentence

The darker the red color behind the word or sentence, the higher his attention value is. The words with higher attention are the words that are the most responsible for the classification of the documents as a positive or negative review. In our case of the last document of the dataset, the output of our model after the sigmoid function is 0.9472. Therefore, the last document is classified as a positive review of a movie. The words ”downright brillant” and ”masterpiece” were the words associated with a high positive sense, which helped our model classify it as positive.

4 Question 4

Unfortunately, the HAN architecture has some limitations. Indeed, at the first step of the model, each sentence of the document is independently encoded with an encoder. However, doing the hypothesis that the sentences are independent is false. Indeed, they are very often links between different sentences in a document. With the HAN architecture, the relationship between two sentences is not taken into account, which can be improved by taking this possible relationship into account. The Context-Aware paper [1] tries to solve this problem and is able to pay high attention to complementary words even if there is some repetition instead of having the same redundant words with high attention for the HAN architecture.

References

[1] Michalis Vazirgiannis Jean-Baptiste Remy, Antoine Jean-Pierre Tixier. *Bidirectional Context-Aware Hierarchical Attention Network for Document Understanding*, 2019.

[2] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *International Conference on Learning Representations (ICLR)*, 2017.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Conference on Neural Information Processing Systems (NIPS)*, 2017.