MARBOUH Oumaima
oumaima.marbouh@polytechnique.edu

Lab session # 1
ALTEGRAD 2021

12/03/21

1 Question 1

After printing the coefficients of the several words in each sentence of a document in the notebook. We noticed that there is some words that have two or more coefficients in the same sentence, which can be a source of confusion for the model. For that, one of the methods to improve the attention mechanism as mentioned in article [1] is to use a mean pooling layer over the bidirectional GRU states as the attention source.
Another problem that can occur is ”redundant information” mentioned by the authors of article [1], and that is when the attention cells use the same word weights to compute the linear combination of the words (level of importance of each word to build the new information). This problem was solved by using a penalization term to encourage the diversity of information. This gives the model a greater capacity to remove the latent information from the input sentence.
A third method is also proposed a very successful intra-sentence level attention mechanism proposed by [1]. In this method, the bidir-gru produces an attention vector for each of its hidden states during the recurrent iteration, which is sort of an ”online updating” attention.

2 Question 2

Based on article [3] the idea behind the attention mechanism was to permit the decoder to utilize the most relevant parts of the input sequence in a flexible manner, by a weighted combination of all of the encoded input vectors, with the most relevant vectors being attributed the highest weights.
The self attention mechanism helps in drawing global dependencies between inputs and outputs. For instance using self attention mechanism in the encoder helps each position in the encoder to attend all positions in the previous layer of the encoder. Similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder up to and including that position.
Finally, ”encoder-decoder attention” allows every position in the decoder to attend over all positions in the input sequence.

3 Question 3

After obtaining the coefficients that the model gave to each word in each sentence in the last document in the test set (see Figure 1). We did a visualization of these words versus their coefficients, What wee can see is that some words are given more importance than other since they have higher coefficients. For instance, in the second sentence the words ”deal” and ”great” were given higher coefficients compared to others so they will contribute more to the final prediction, while other words such as ”like” were given coefficients that are equal to zero or near zero which means their embiddings will not contribute to the final prediction. In addition to that, some similar words such as ’OOV’ have different coefficients depending on the sentence and the importance of the sentence in the document. The model therefore decides to pays attention to some words and some sentences more than others depending on the context.

4 Question 4

HAN has a major limitation which is not taking the context into account according to article [2], which precise that context refer to the cross-sentence, external, or document-level context. Indeed, At level 1 each sentence of the document is encoded in complete isolation. Which means that HAN ignores other sentences while doing the representation of a given sentence in the document.
In the example given in article [2] in Fig 1 we notice that HAN encodes each sentence independently, which has lead to giving several times the same salient feature the highest coefficients although it has bean already covered, while neglecting other aspects of the document.
In the case of redundant information, HAN instead of extracting complementary information that will help in better coverage, in a richer document representation and therefore in a more accurate prediction, it produces the same embedding for each instantiation of the repeated sentence which results in the same alignment decisions.

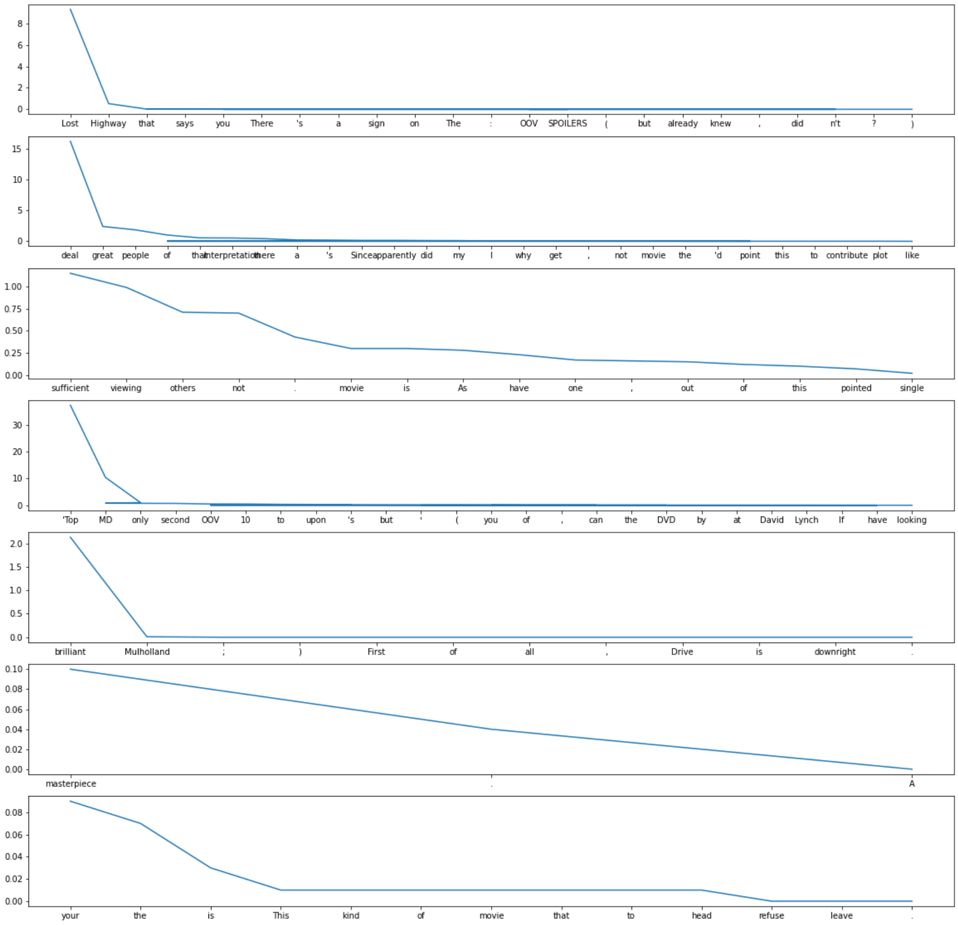

Figure 1: Words Vs their Coefficients in the sentences of a chosen document.

References

[1] Zhouhan Lin, Minwei Feng, Cicero Dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Y. Bengio. A structured self-attentive sentence embedding. 03 2017.

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

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.