Gabriel Athènes
gabriel.athenes@polytechnique.edu

Lab session # 1
ALTEGRAD 2021

11/16/21

1 Question 1

According to article [1], the first way to improve the self-attention mechanism is to embed sentences not as vectors but as matrixes by using an annotation matrix A. Indeed, a sentence can be seen as different semantic parts seperated by a multiple m of words and adding up together to form the sentence. This can be achieved by extending the second weight vector into a weight matrix, which implies that the annotations are now a matrix A. Now that we have a matrix A, we want the different rows of A to represent different parts of the sentence, otherwise this expansion is less usefull. The article therefore proposes to add a penalization term to the original loss to enforce the matrix A to have a certain shape. The penalization term proposed in the article is the Frobenius norm of $AA^T - I$. Indeed minimizing this norm first ensures that the rows of A are orthogonal, meaning that they carry probabilities for different words and thus are semantically different. Second, it ensures that the norm of the rows of A are close to 1. Since the coefficients of the rows of A can be seen as probabilities on words, this ensures that rows carry weights on a few words only, encouraging the annotation matrix to synthesize the semantic parts of the sentence in the fewer words possible.

2 Question 2

The first reason to replace to replace recurrent operations with self-attention is the total computational complexity per layered which is decreased using Self-attention. According to article [3], self-attention layers are faster than recurrent layers when the sequence length n is smaller than the representation dimensionality d, which is most often the case with sentence representations used by state-of-the-art models in machine translations. The second reason to use self-attention is the path length between long-range dependencies in the network. According to the article [3], learning long-range dependencies is a key challenge in many sequence transduction tasks, meaning that shorter paths between any combination of positions in the input and output sequences make it easier for our model to learn long-range dependencies. A recurrent layer requires $O(n)$ sequential operations to connect all positions, compared to $O(1)$ for a self-attention layer, giving an advantage to self-attention over recurrent operations.

3 Question 3

Let's plot the word coefficients of the 7 sentences of the last review in the given document.

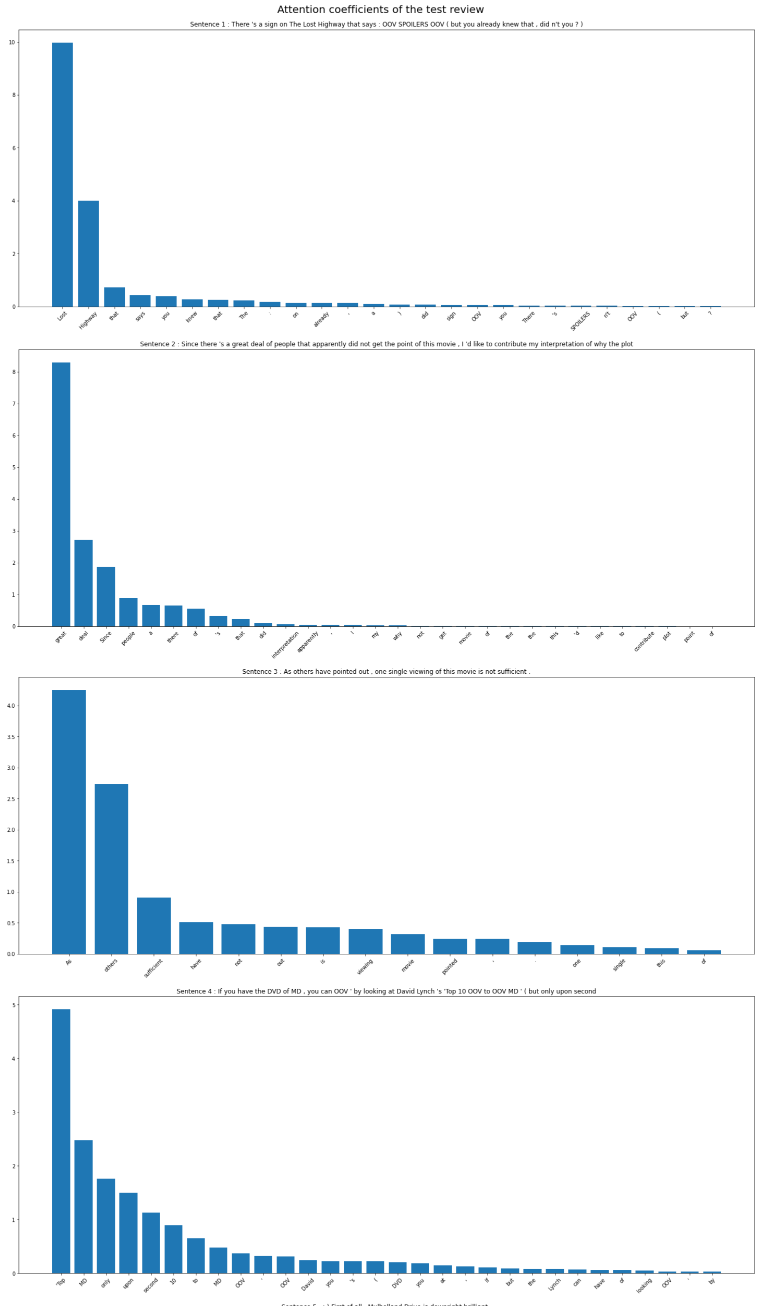

Figure 1: Attentional coefficients of the first four sentences.

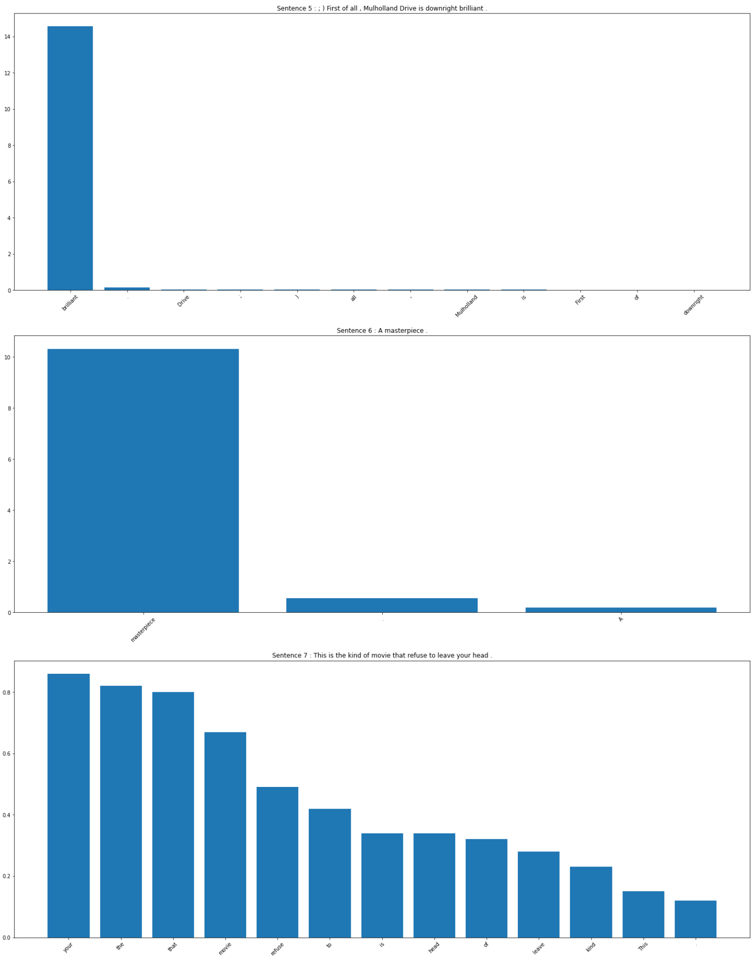

Figure 2: Attentional coefficients of the last 3 sentences.

We can see from these plots that HAN has succeeded in focusing on words that are highly expressive. In sentence 5 and 6 for example, Masterpiece and Brilliant are both heavily coefficiented. In sentence 2 and 4, Top and great are also attributed big coefficients, which is coherent with the fact that it confers the sentence a very positive feeling. What is interesting is that the sentence 7 has very homogeneous coefficients. Looking at the sentence, it corresponds to the fact that the sentence is indeed difficult to interpret. ”This is the kind of movie that refuses to leave your head” is ambiguous and could be interpreted as either negative or positive.

4 Question 4

According to article [2], the main limit of HAN is that sentences are encoded independently in a document, meaning encoded ignoring the other sentences. This implies that HAN spends most of its time encoding the same features of similar sentences, therefore neglecting other aspects of the document. Once several similar sentences with common semantic parts A are passed, HAN is not able to direct its attention to other parts of the sentences that are not A. The context-aware HAN proposed in the article solves this issue. Indeed, as figure 1 and 2 of the article show, CAHAN ignores the common part A of similar sentences once they have been passed several times.

References

[1] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

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

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.