LACOMBE Yoach
yoach.lacombe@gmail.com

1 Question 1

The self-attention mechanism can be improved in a number of ways, one of them being providing multiple attentional vectors. This is done in order to have multiple focus on different parts of the same sentence (each part being semantically different from the others).
[5] introduced this idea and added a regularization term on the matrix of attentional vectors in order to have each attentional vector focusing on different parts of the sentence (forcing each attentional vector to be orthogonal to the order by the penalization term $P = ||AA^T - I||_F^2$ where A is the matrix of attention vectors).

2 Question 2

RNN requires O(n) sequential operations, n being the length of the sentence, because we have to go 0(n) times throughout the RNN layer to get the context vector, whereas the self-attention layer has O(d) operations, d being the number of attentional vectors.
Moreover, self-attention layers can be more interpretable layers and are more prone to seize long-ranges dependencies and complex semantical structures.
Finally, the attentional vectors computation can be parallelized as compared to the sequential computation of RNN propagation.

3 Question 3

See Fig. 1. I've plotted the word attentional vectors of the first 5 sentences (out of 7) of the last review. The 6th sentence is "A masterpiece.".
The first sentence is deemed the most important by the hierarchical self-attention. However, it is clear that it should be (in terms of sentiment analysis) the fifth or the sixth sentences, since they use the terms "brillant" or "masterpiece". So the model is not able to detect important sentences.
Moreover, the model often focuses on generics words (or even on commas and strophes) of a given sentence, especially when no sentiment can be given from this sentence (ex: 1st sentence just warns about spoilers but the model still focuses on certain words).
Finally, the model totally overlook the sense of certain sentences and words (ex: 2nd sentence, it focuses on 'great' which is not giving the sentiment of the sentence).

4 Question 4

From [6]'s observations, the main limitation of the HAN architecture is that it considers each sentence out of context while computing the attentional vector of a given sentence. It is suboptimal in certain situations such as:

- It spends its attentional budget on the most noticeable words even when these words appeared many times on different sentences (it could have focused on other words after the first appearance).
- If the same sentence is repeated throughout the text (redundant information), it stills give the same vector to the second level of the architecture.

References

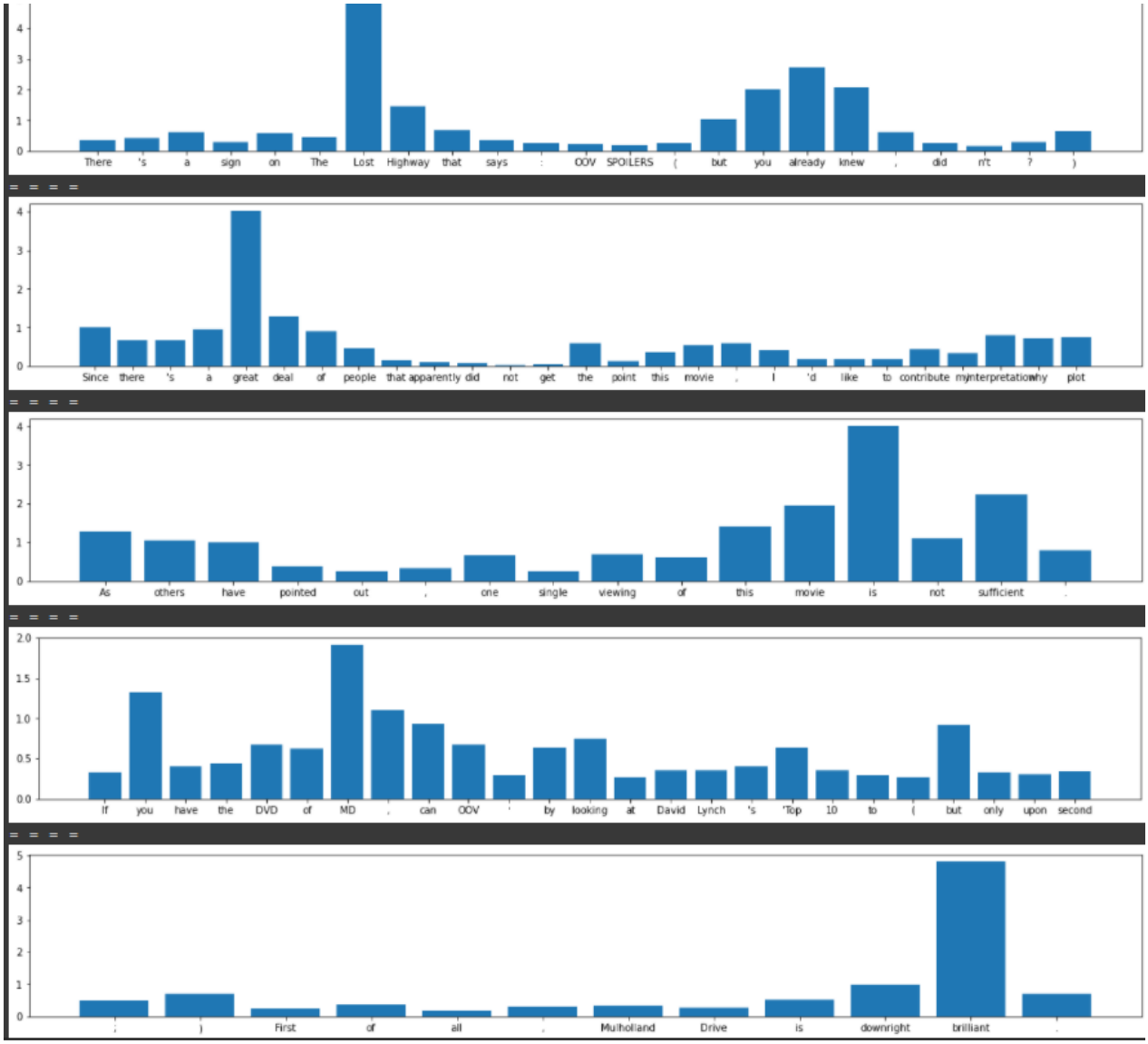

Figure 1: Histogram of the attentional vectors of the 5 first sentences of the last review.