Megi Dervishi
megi.dervishi@ens.fr

Lab session # 1
ALTEGRAD 2021

11/16/21

Question 1

The basic attention mechanism can be improved by using a matrix representation and a penalization term which gives the model a greater capacity to disentangle the latent information from the input sentence [3]. The reason why using a matrix representation is better because the vector representation usually focuses only on a specific part/parts of the sentence that are semantically related to one-another. However since there can be multiple occasions of such in a sentence, to represent better the overall semantics of a sentence one needs to perform multiple s vectors for different parts of the sentence. And hence we can represent it as a matrix $S = AH$ where A is the matrix of all α coefficients. But this matrix can suffer from redundancy problems and to avoid such cases the paper introduce a penalization term P which is $P = \|AA^T - I\|_F^2$. Another improvement that the paper [3] introduces is adding an additional weighted layer to the attention mechanism:

$$u'_t = W_2 u_t = W_2(\tanh(W h_t))$$
$$\alpha_t = \frac{\exp u_t'^T u'}{\sum_{t'=1}^{T} \exp u_{t'}'^T u'}$$

Question 2

The main disadvantage that RRN have over Transformers(self-attention)[2] is their sequential nature of processing data, which becomes critical when processing large sequences of data as memory constraints limit batching across different examples. On the other hand using self-attention mechanism allows parallelization of data hence more powerful GPUs can be used. Another challenge appearing in RRNs is the vanishing/exploiding gradient problem which makes it more difficult to learn dependencies in very long sequences of words.

Question 3

The following document has 7 sentences whose attention coefficients for each word are shown in Figure 1. The sentences with labels from (a) to (g) of the document are:

(a) There ’s a sign on The Lost Highway that says : OOV SPOILERS OOV (but you already knew that , did n’t you ?)

(b) Since there ’s a great deal of people that apparently did not get the point of this movie , I ’d like to contribute my interpretation of why the plot

(c) As others have pointed out , one single viewing of this movie is not sufficient .

(d) If you have the DVD of MD , you can OOV ’ by looking at David Lynch ’s ’Top 10 OOV to OOV MD ’ (but only upon second

(e) ;) First of all , Mulholland Drive is downright brilliant .

(f) A masterpiece .

(g) This is the kind of movie that refuse to leave your head .

Generally speaking from the plots we can conclude that in a sentence not all words contribute the same importance when creating context. For example as humans when reading sentence (c) the word ”movie” would be more important than the word ”of”. We can also observe such conclusion in the above plots as the word ”movie” has a higher attention coefficient than that of the word ”of”. Also, we can notice that in each sentence apart from sentence (g) there is one/two words which have a much higher attention coefficient than the other ones. They are usually either adjectives such as ”Lost, great, sufficient, brilliant” or key words such as ”Highway, Top, masterpiece”. This means that the model is able to capture the most important words in a sentence which provide him with a good enough context. Among all the words in the document the ones who have the highest attention coefficients are ”Top, brilliant, masterpiece, MD, 10, great, deal”. The highest coefficient of sentence (d) is much greater in value than the highest coefficient in sentence (g). That is because sentence (d) is ranked with a greater importance score from the model compared to sentence (g). Therefore the model will pay more attention to build context to the words coming from sentence (d) compared to that of sentence (g).

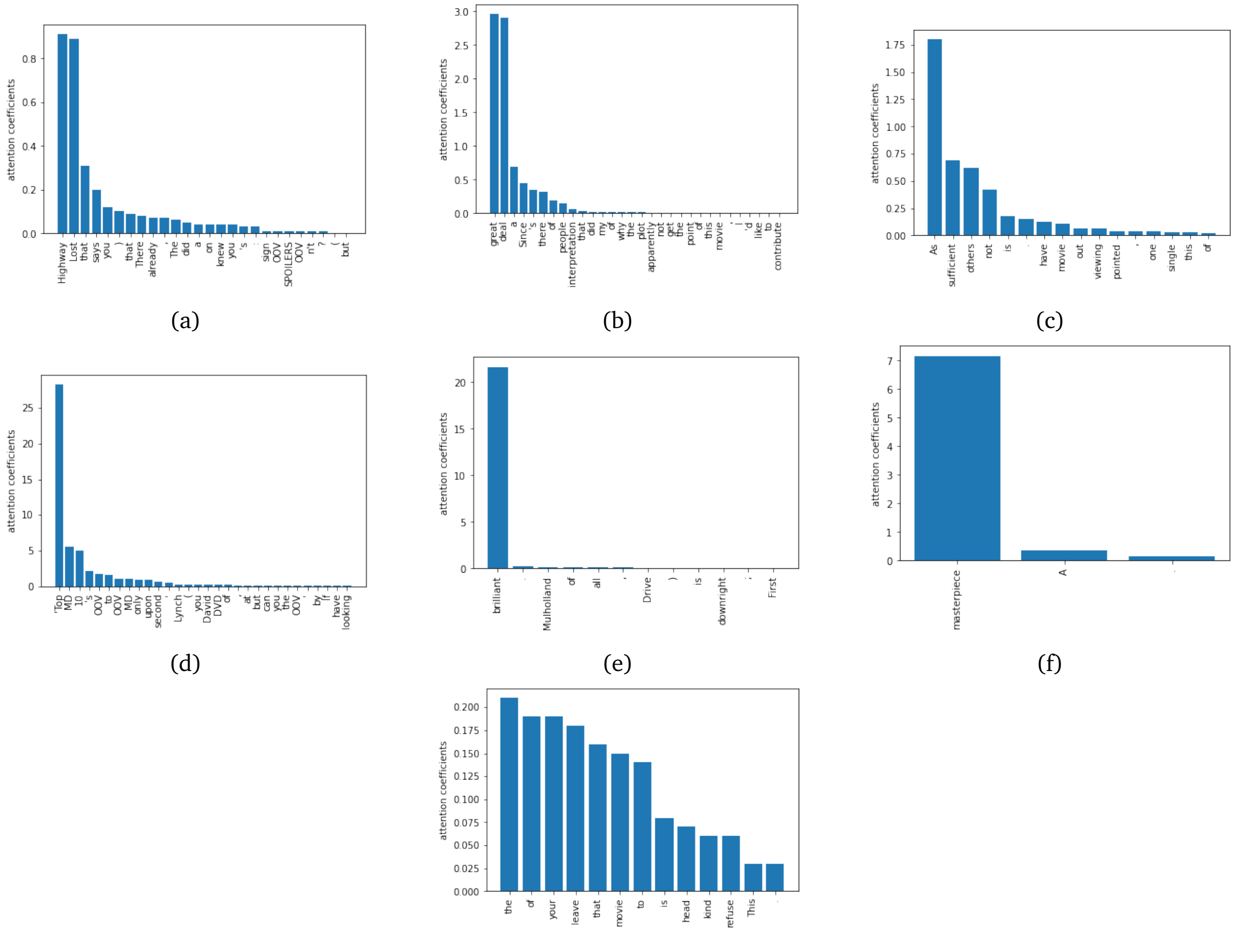

Figure 1: The attention coefficients for each word and for each sentence in the document.

Question 4

HAN encodes each sentence independently, without considering any kind of contextual information. This means that when encoding a sentence the architecture ignores the other sentences[1]. This makes the model suboptimal in capturing context well. At level 2 HAN only assigns importance scores to sentences, but given that sentences were independently encoded and when doing such rankings the document encoder does not modify the sentence vectors, certain parts of context may be lost. Hence the encoder may not address issues like high redundancy. For example let us say we have the following two sentences: ”The course is brilliant. It is brilliant because of its lectures.” As sentences are independantly encoded in both cases the model would choose to pay more attention to the word ”brilliant”. While it is true in the first sentence, in the second sentence the word ”lectures” should be more important than that of ”brilliant”, due to the context coming from the first sentence. This is something that HAN cannot percieve.

References

[1] Remy Jean-Baptiste, Jean-Pierre Tixier Antoine, and Vazirgiannis Michalis. Bidirectionalcontext-awarehierar- chical attention network for document understanding. *arXiv preprint arXiv:1908.06006*, 2019.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

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