Mercier Marine
marine.mer98@gmail.com

Lab session # 1
ALTEGRAD 2021

11/15/21

1 Question 1

The basic attention mechanism may provide similar summation weights for all the r hops, then the attentional vector would suffer from redundancy. Therefore we need to encourage diversity of summation weights for each hop, thanks to a penalization.

The Kullback Leibler divergence between any 2 of the summation weight vectors can evaluate diversity, we can't encourage each row to focus on a single semantic aspect and this metric might be unstable according to [3].

[3] introduces the following penalization : $P = ||\alpha\alpha^T - I||_F^2$ with α the alignement coefficients, I the idendity matrix and $||.||_F$ the Frobenius norm. Then we add P, multiplied by a coeffecient, to the loss and we minimize it.

Indeed if we take α_i and α_j, two summation weights, then $1 \geq \alpha_i\alpha_j \geq 0$. If $\alpha_i\alpha_j = 1$, the distributions are identical and concentrated on a single word, therefore we substract the Identity matrix. Therefore the minimization of P will generate diversity while focusing the distribution on a single word.

2 Question 2

Recurrent neural network returns a sequence of hidden reprsentation $(h_1, ..., h_T)$. h_t is computed at time t thanks to the previous hidden representation h_{t-1} and input x_t.

As underlined in [2], the sequential nature of recurrent operations precludes parallelization. It creates problems with longer sequences in terms of computational time and memory. Therefore the use of batches is also limited.

This sequential nature also limits the ability of the model to handle word inversions, which can be a problem in translation for example.

3 Question 3

The first few sentences have more weight overall than the last few sentences. The sum of the weights of the first sentence is equal to 19, while that of the last sentence is equal to 1.69. However, it is not the one that contains the most meaning.

Many meaningless words have large weights like "to", "is", "that", "'s" or punctuation. It shows that the model does not focus on the most important words. As a result, some subtleties are not understood or some important passages are completely omitted.

4 Question 4

The main drawback of the HAN architecture is the isolation of sentences. Sentences are taken out of any contexts. This lack of context is obviously a weakness. [1] shows that it can be a problem in the following situations :

- If the same sentence is repeated, HAN will focus on the strongest words every time. All the attentional budget is allocated to these salient features and some parts of the sentence are not covered. Therefore some nuances in the sentence are missed.

- Furthemore, if the same sentence is repeated in the document, the output of HAN will always be the same for every sentence. Some other algorithm might be able to extract complementary information rather than redundant information from each apparition of the sentence.

References

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

[2] 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.

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