Denis Duval
denis.duval13@gmail.com

Lab session # 1
ALTEGRAD 2022

11/14/22

# 1 Question 1

Using the equations given in the handout, the basic self-attention mechanism is described by:

$$u_t = \tanh(Wh_t), a_t = \frac{\exp(u_t^T u)}{\sum_{t'} \exp(u_{t'}^T u)}, s = \sum_t a_t h_t \tag{1}$$

with the trainable parameters being the context vector $u$ and the weight matrix $W$. As said in [2], this attention vector will only focus on a specific component of the sentence, like a special set of related words or phrases. In order to capture all the aspects or components of the sentence(that is either a sequence of words or a sequence of phrases) we could use this basic self-attention multiple times on the same sequence of annotations. Formally, our sentence embedding would now be a matrix of shape $r * u$ where $r$ is the number of times we apply self-attention and $u$ is the dimension of the hidden states $h_t$, containing all the self-attention vectors of the hops.

Intuitively this is only likely to improve basic self-attention, meaning that it would capture more aspects and thus be more attentive on the overall semantic of the sentence, if the several forwards of self-attention are not attentive to the same components. We want to constraint the parameters so that each forward of self-attention will stare at different aspect or component of the sentence. A natural solution for this is adding a penalization term [2]:

$$||AA^T - I||_2^2$$

where A is the matrix of all the annotation weights. Its form can be simply explained : non-diagonal elements of $AA^T$ will be as the more costly as the covariance between two annotation weight vectors is high; diagonal elements of $AA^T$ will be as the more costly as they deviate from $1$ : remembering that the sum of the coefficients of an annotation weight vector must be $1$ due to softmax, this means it encourages the vector to have the least non-zero elements hence capturing the least number of aspects of the sentence.

# 2 Question 2

There are three main motivations to replace recurrent operations such as RNNs or CNNs with the attention mechanism [4]. These motivations are essentially a matter of computational cost.

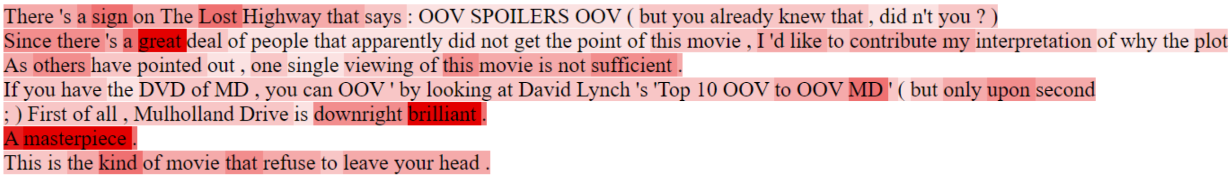
The first one is the computational complexity per layer. If $n$ is the length of the input sequence and $d$ is the dimension of the embedding space then the complexity per layer of a simple self-attention mechanism layer is a $O(n^2 \cdot d)$ whereas the complexity per layer of a RNN is a $O(n \cdot d^2)$. In NLP especially the dimension $d$ of the embedding space is often higher than $n$, so self-attention is from this perspective, more efficient.

Another main reason is the number of operations that can be parallelized. Because recurrent operations need all the previous hidden states to output the next hidden state they can only be parallelized at the level of the whole sentence while a self-attention layer consists in a constant number of steps (that need to be executed one after the other) all of which can be parallelized.

The third reason is the path length between long-range dependencies in the network [4]. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse in the network, due to gradient vanishing. The shorter these paths between any position in the input and any position in the output, the easier it is to learn long-range dependencies [1]. We can upper bound this length with the maximum number of operations an output state is far from an input state it depends on. Again this is a $O(1)$ for self-attention and a $O(n)$ for a RNN.

# 3 Question 3

See 1. We can see that overall the encoder gives a lot of importance to the adjectives and can decide to what extent they bring a positive (or negative) sentiment. Although, we're able to see some limitations as depicted in question 4. In fact the model gives importance to words like 'kind' (last sentence) or 'Lost' (first sentence) as they generally convey great information though in the context of the example they shouldn't have that much importance.



There 's a sign on The Lost Highway that says : OOV SPOILERS OOV ( but you already knew that , did n't you ? )
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
As others have pointed out , one single viewing of this movie is not sufficient .
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
; ) First of all , Mulholland Drive is downright brilliant .
A masterpiece .
This is the kind of movie that refuse to leave your head .

Figure 1: Word importance, encoder level

# 4 Question 4

One limitation of the basic HAN architecture is that the sentence encoder encodes each sentence (of a document) independently [3], thus if a strong negative or positive aspect is repeated in several sentences of a same document then it will hide other important aspects for the overall semantic. It is also proposed to treat in a bidirectional manner the documents at the level of the document encoder. The idea behind is that the context of a sentence should not only be formed by the previous sentences but also by the successive ones.

# References

[1] John F. Kolen and Stefan C. Kremer. *Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies*, pages 237–243. 2001.

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

[3] Jean-Baptiste Remy, Antoine Jean-Pierre Tixier, and Michalis Vazirgiannis. Bidirectional context-aware hierarchical attention network for document understanding, 2019.

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