MVA / ALTEGRAD 2022 / Lab session 1: Hierarchical Attention Network

Imane Elbacha

14 November, 2022

1 Self attention:

1.1 Question 1:

How can the basic self-attention mechanism be improved?

We are going to begin by defining the basic self-attention mechanism. The intuition behind this concept is determining weights that will compute how much *attention* the model should giving to elements in the sequential input. The math behind it is the following:

$$u_t = tanh(Wh_t) \tag{1}$$

$$\alpha_t = \frac{exp(u_t^T u)}{\sum_{t'}(exp(u_{t'}^T u)} \tag{2}$$

$$s = \sum \alpha_t h_t \tag{3}$$

With (1) the context vector (2) the alignment coefficients and s the attentional vector.

Using the model presented in this paper we can consider a way to improve the self attention vector. Usually the vector s focuses on specific components of the sentence(like a special set of related words or phrases). However the semantic characteristic of a sentence can be induced from a combination of words and how they were put together. For example: *It's raining cats and dogs* the semantic meaning of the sentenced can only be understood through weighting it the combination of the words. Hence the alignment coefficients can be improved in a way that it characterises the distribution of section of the sentences.

As formulated in the paper we need to perform multiple hops of attention. Say we want r different parts to be extracted from the sentence, with regard to this, we add a vector factor to the alignment coefficients w and the resulting annotations a become annotation matrix vectors. Formally,

$$_t = w\frac{exp(u_t^T u)}{\sum_{t'}(exp(u_{t'}^T u)} \tag{4}$$

1.2 Question 2:

What are the main motivations for replacing recurrent operations with self-attention?

The early suggestions were based on the usage of RNNs in an encoder-decoder architecture for sequence-to-sequence issues like neural machine translation. The ability of these architectures to preserve information from the first elements was lost when new elements were added to the sequence, which is a significant disadvantage when working with extended sequences. Every step of the encoder's hidden state is connected to a specific word in the input sentence, usually the most recent one. Therefore, if the decoder just accesses the decoder's final concealed state, it will miss important information about the sequence's initial elements. The attention mechanism was subsequently created as a solution to this problem.

As opposed to RNNs, which often focus on the last state of the encoder, the decoder will examine all of the encoder's states in each step, giving it access to data on every component of the input sequence. Attention extracts data from the entire sequence, a weighted average of all previous encoder states. This enables the decoder to give each piece of the output more weight or relevance than another element of the input. acquiring the skill of focusing on the appropriate input element at each stage to anticipate the following output element

These are the motivations for replacing recurrent operations with self-attention

2 Hierarchical self-attention

2.1 Question 3:

Paste your attention coefficients for a document of your choice. Interpret your results.
For sentences we get the following scores:

```
11.17 There 's a sign on The Lost Highway that says : OOV SPOILERS OOV ( but you already knew that , did n't you ? )
12.89 Since there 's a great deal of people that apparently did not get the point of this movie , I 'd like to contribute my interpretatic
11.3 As others have pointed out , one single viewing of this movie is not sufficient .
15.09 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
19.67 ; ) First of all , Mulholland Drive is downright brilliant .
16.25 A masterpiece .
13.63 This is the kind of movie that refuse to leave your head .
```

Figure 1: Attention coefficients for sentences

For the words the attention coefficients fall within the same score margin except for words like 'great' and 'masterpiece' that convey a strong sentiment hence get higher attention scores.

2.2 Question 4:

What are some limitations of the HAN architecture?

On six massive sentiment and topic categorization datasets, HAN was extremely successful and set new standards. Its main limitation is that each sentence is encoded separately. That is, HAN entirely disregards the other sentences while constructing the representation of a specific sentence in the document. It is evident that this lack of communication is not ideal, and can't be expected to give the best understanding of the features of the entirety of the document.