Lab session 1: Hierarchical Attention Network

Maxence Lasbordes

Abstract

This lab session introduces Hierarchical Attention Networks (HAN) for document classification.

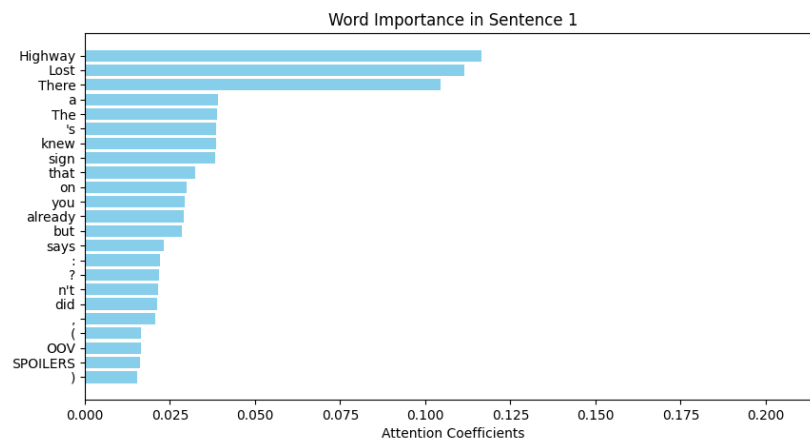## 1   Questions

[Question 1]
The paper by Zhouhan Lin et al. (2017) on structured self-attentive sentence embeddings suggest to use a 2-D matrix instead of a vector to represent the embedding, which makes the embedding more interpretable by allowing each row to attend different part of the sentence.

[Question 2]
Some of the main motivations for using transformers include their ability to do parallel computation (since RNNs are sequential and slower) and in capturing long-range dependencies, as each token in a sentence can attend to every other token. Additionally, transformers do not suffer from the exploding and vanishing gradient issues that RNNs face.

[Question 3]
This diagram illustrates the impact of each word on the sentence. For instance, the word "Highway" provides significantly more context and meaning than the word "did".

[Question 4]
In HAN, the document encoder only processes the document in a forward direction, limiting its ability to capture long-range dependencies. Also, there is a lack of context awareness as each sentence is encoded in isolation of surrounding sentences.

[Bonus Question]
The $my\_patience$ parameter is used for early stopping, this technique is used to prevent over-fitting by stopping the training process after a certain number of epochs if the validation accuracy does not improve.