Romy BEAUTÉ
romybeaute.univ@gmail.com

Lab session # 1
ALTEGRAD 2021

11/16/21

1 Question 1 : How can the basic self-attention mechanism be improved?

Self-attention mechanisms have enable strong improvements in NLP tasks, by considering subsets of input’s hidden representations at all time steps in order to limit the loss of information. However the ”memory” architecture of self-attention mechanisms are quite primitive, and doesn’t provide any information about ”word content” and therefore cannot establish a meaning of the word in a sentence. Such limitation can be quite problematic for example when dealing with polysemous words that have different meaning for a given context. Attention models could therefore be completed with models allowing extra source of information in order to provide support for the extraction of sentence embedding [1]. More complex models should consider not only the influence of the past hidden layer to compute current input’s representation, but also the influence of further (past) and future words in the sentence to mutually influence each other’s representations. Bidirectional architecture are interesting to this extent. Another way to improve self-attention mechanism has been proposed by Zhouhan Lin et al. (2017) [2] and consists in moving towards an unsupervised learning by using a sequential decoder on top of the sentence embedding.

2 Question 2 : What are the main motivations for replacing recurrent operations with self-attention?

The main problem of recurrent operations networks is that they are sequential, which restrict parallelization. By freeing themselves from recurrence and convolutions, self-attention architecture present the advantages of being more parallelizable and of being more efficient on the training steps (faster) while showing better results over recurrent operations.

3 Question 3 : Paste (or even better, plot) your attention coefficients for a document of your choice. Interpret your results.

Here I chose to interpret the coefficient of the document we worked on though this Lab session, i.e the movie review from the Internet Movie Database (IMDB) dataset. First, we printed the accuracy of the trained model to evaluate its accuracy on the review text : accuracy = 97.1 which is very good. We therefore expect that the coefficients associated with words are consistent to what we could expect from them, leading to a good classification of the document.

Here are the coefficient we obtained : (’There’, 0.0) (”’s”, 0.0) (’a’, 0.0) (’sign’, 0.0) (’on’, 0.01) (’The’, 0.02) (’Lost’, 26.36) (’Highway’, 0.91) (’that’, 0.24) (’says’, 0.15) (’:’, 0.07) (’OOV’, 0.02) (’SPOILERS’, 0.03) (’OOV’, 0.01) (’(’, 0.01) (’but’, 0.01) (’you’, 0.17) (’already’, 0.04) (’knew’, 0.05) (’that’, 0.07) (’,’, 0.03) (’did’, 0.02) (”n’t”, 0.01) (’you’, 0.03) (’?’, 0.0) (’)’, 0.07) = = = = (’Since’, 0.13) (’there’, 0.1) (”’s”, 0.14) (’a’, 0.44) (’great’, 22.36) (’deal’, 5.57) (’of’, 0.33) (’people’, 0.2) (’that’, 0.09) (’apparently’, 0.03) (’did’, 0.05) (’not’, 0.01) (’get’, 0.02) (’the’, 0.01) (’point’, 0.0) (’of’, 0.0) (’this’, 0.0) (’movie’, 0.01) (’,’, 0.01) (’I’, 0.01) (”’d”, 0.0) (’like’, 0.0) (’to’, 0.0) (’contribute’, 0.0) (’my’, 0.04) (’interpretation’, 0.12) (’of’, 0.0) (’why’, 0.0) (’the’, 0.0) (’plot’, 0.0) = = = = (’As’, 1.75) (’others’, 1.18) (’have’, 0.56) (’pointed’, 0.22) (’out’, 0.58) (’,’, 0.48) (’one’, 0.37) (’single’, 0.16) (’viewing’, 1.12) (’of’, 0.13) (’this’, 0.11) (’movie’, 0.16) (’is’, 0.27) (’not’, 0.13) (’sufficient’, 0.11) (’.’, 0.25) = = = = (’If’, 0.27) (’you’, 0.87) (’have’, 0.22) (’the’, 0.93) (’DVD’, 2.01) (’of’, 0.59) (’MD’, 3.09) (’,’, 0.51) (’you’, 0.77) (’can’, 0.21) (’OOV’, 0.07) (”””, 0.05) (’by’, 0.07) (’looking’, 0.11) (’at’, 0.27) (’David’, 0.88) (’Lynch’, 0.3) (”’s”, 0.2) (”’Top”, 2.27) (’10’, 0.12) (’OOV’, 0.07) (’to’, 0.08) (’OOV’, 0.09) (’MD’, 0.44) (”””, 0.05) (’(’, 0.04) (’but’, 0.04) (’only’, 0.16) (’upon’, 0.09) (’second’, 0.32) = = = = (’;’, 0.02) (’)’, 0.03) (’First’, 0.01) (’of’, 0.05) (’all’, 0.06) (’,’, 0.05) (’Mulholland’, 0.1) (’Drive’, 0.01) (’is’, 0.02) (’downright’, 0.01) (’brilliant’, 10.38) (’.’, 0.07) = = = = (’A’, 0.17) (’masterpiece’, 4.4) (’.’, 0.05) = = = = (’This’, 0.15) (’is’, 0.2) (’the’, 0.26) (’kind’, 0.05) (’of’, 0.11) (’movie’, 0.08) (’that’, 0.11) (’refuse’, 0.07) (’to’, 0.06) (’leave’, 0.04) (’your’, 0.28) (’head’, 0.07) (’.’, 0.07) = = = = (’Lost’, 26.36) (’Highway’, 0.91) (’that’, 0.24) (’you’, 0.17) (’says’, 0.15) (’:’, 0.07) (’that’, 0.07) (’)’, 0.07) (’knew’, 0.05) (’already’, 0.04) (’SPOILERS’, 0.03) (’,’, 0.03) (’you’, 0.03) (’The’, 0.02) (’OOV’, 0.02) (’did’, 0.02) (’on’, 0.01) (’OOV’, 0.01) (’(’, 0.01) (’but’, 0.01) (”n’t”, 0.01) (’There’, 0.0) (”’s”, 0.0) (’a’, 0.0) (’sign’, 0.0) (’?’, 0.0) = = = = (’great’, 22.36) (’deal’, 5.57) (’a’, 0.44) (’of’, 0.33) (’people’, 0.2) (”’s”, 0.14) (’Since’, 0.13) (’interpretation’, 0.12) (’there’, 0.1) (’that’, 0.09) (’did’, 0.05) (’my’, 0.04) (’apparently’, 0.03) (’get’, 0.02) (’not’, 0.01) (’the’, 0.01) (’movie’, 0.01) (’,’, 0.01) (’I’, 0.01) (’point’, 0.0) (’of’, 0.0) (’this’, 0.0) (”’d”, 0.0) (’like’, 0.0) (’to’, 0.0) (’contribute’, 0.0) (’of’, 0.0) (’why’, 0.0) (’the’, 0.0) (’plot’, 0.0) = = = = (’As’, 1.75) (’others’, 1.18) (’viewing’, 1.12) (’out’, 0.58) (’have’, 0.56) (’,’, 0.48) (’one’, 0.37) (’is’, 0.27) (’.’, 0.25) (’pointed’, 0.22) (’single’, 0.16) (’movie’, 0.16) (’of’, 0.13) (’not’, 0.13) (’this’, 0.11) (’sufficient’, 0.11) = = = = (’MD’, 3.09) (”’Top”, 2.27) (’DVD’, 2.01) (’the’, 0.93) (’David’, 0.88) (’you’, 0.87) (’you’, 0.77) (’of’, 0.59) (’,’, 0.51) (’MD’, 0.44) (’second’, 0.32) (’Lynch’, 0.3) (’If’, 0.27) (’at’, 0.27) (’have’, 0.22) (’can’, 0.21) (”’s”, 0.2) (’only’, 0.16) (’10’, 0.12) (’looking’, 0.11) (’OOV’, 0.09) (’upon’, 0.09) (’to’, 0.08) (’OOV’, 0.07) (’by’, 0.07) (’OOV’, 0.07) (”””, 0.05) (”””, 0.05) (’(’, 0.04) (’but’, 0.04) = = = = (’brilliant’, 10.38) (’Mulholland’, 0.1) (’.’, 0.07) (’all’, 0.06) (’of’, 0.05) (’,’, 0.05) (’)’, 0.03) (’;’, 0.02) (’is’, 0.02) (’First’, 0.01) (’Drive’, 0.01) (’downright’, 0.01) = = = = (’masterpiece’, 4.4) (’A’, 0.17) (’.’, 0.05) = = = = (’your’, 0.28) (’the’, 0.26) (’is’, 0.2) (’This’, 0.15) (’of’, 0.11) (’that’, 0.11) (’movie’, 0.08) (’refuse’, 0.07) (’head’, 0.07) (’.’, 0.07) (’to’, 0.06) (’kind’, 0.05) (’leave’, 0.04) = = = =

We note that all the preposition and punctuation are associated with a very weak attention coefficients. However, adjectives present the highest coefficients. This is even more true with nouns that present an emotional valence such as (’Lost’, 26.36) or (’brilliant’, 10.38).

4 Question 4 : What are some limitations of the HAN architecture?

In the Hierarchical Attention Network (HAN) architecture, learning of input’s representation is supported by a different encoder at each level. At each level, the input of the encoder is defined as the output of the preceding (lower) level. This lead to the consequence that at level 1, each string in the document is encoded with the same string encoder individually, resulting in a sequence of string vectors. As pointed out in a previous article [3], this fact lead to an important limitation of the HAN architecture : each word is as level 1, each sentence is processed independently from the other sentences. This lead to a contextual and coherence problem : the meaning of the sentences are totally de-correlated through the text, and isolated between each others. Finally, this architecture is sub-optimal when dealing with redundancy in the text. Improvements should therefore focus on enhancing the learning of contextual capabilities. This could be enhanced by providing the model with explicit knowledge about coverage, diversity, and redundancy. On the experimental side, one should test the performance of the model when dealing with ambiguity such as in texts that require a deep understanding, and particularly when the meaning of polysemous words have to be clarified. Bidirectional encoders, taking both past and future words in the sentence, should provide good tools for a finer-grained understanding of the text meaning.

References

[1] Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. Attentive pooling networks, 2016.

[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. *CoRR*, abs/1908.06006, 2019.