cs224n

(arthurg)

May 26, 2020

Question 1.

The masks is 1 where the padding occurs, 0 elsewhere. In the attention function, we give an attention score of -infinity where the padding occured. Then, after we apply softwmax function, whever the padding occurs will always have a very low probability close to 0.

This is necessary because we don't want to give attention to padding tokens during the decoding process (as their corresponding encoder states contain no real information).

Question 2.

My model's BLEU score is 35.69179114233418

Question 3.

The NMT translated mistook "one of my favourites" for "favourite of my favourites."

In the third decoding step (when trying to decode "one"), the attention was improperly focused on "favoritos" instead of "otro". To fix, we need better attention scores. This could be done by stacking more layers in our attention mechanism (currently we have $W_{attProj} \odot h_i^{enc}$ which is technically only 1 linear layer). In addition, we might also be able to fix it by getting more training data.

Question 4.

The NMT outputted "the author for children, more reading in the U.S." instead of "Americas most widely read childrens author, in fact."

The issue is that the words are mostly accurate, however the order and the grammar seems slightly off. To fix this issue, i would incorporate a pre-trained language model into the model, similar to what is done in SMT.

Question 5.

The NMT system outputted "Richard <unk>" instead of "Richard Bolingbroke"

The current NMT system is not programmed to handle OOV words. When ever the decoder is expecting to output an OOV word, we could instead output the corresponding input word with the highest attention score. Many OOV words are things that like proper nouns (eg: names), which don't change between different languages.

Question 6.

The NMT outputted "back to the apple" instead of "go around the block"

The issue is that in spanish "manzana" can either translate to "apple" or "block". We can also fix this with a language model as suggested 2 questions ago. We can also try training the NMT on less formal text (eg: casual conversations) to try to pick up figures of speech such as "around the block"

Question 7.

The NMT outputted "womens room" instead of "teachers lounge."

There could be multiple issues that could be causing this translation. One explanation is that there is bias in our word embeddings, where "women" and "teacher" were associated closely (as discussed in A1). Another explanation is that the attention mechanism is not working properly, and the attention score for "Elle" and "al bano" were too high when trying to decode "teachers".

The fix for the first issue is to get better pre-trained embeddings. The fix for the second issue is to get better attention mechanism as previously discussed.

Question 8.

The NMT outputted "250 thousand acres" instead of "100,000 acres"

The main issue is that the NMT probably has not seen enough example of translating roman numbers in the training data. Depending on how the numerical "words" are structured, we could be translating each arabic numeral as a word. This means that when we use our attention mechanism has to be know how to handle every single power of 10 (eg: diff attention mechanism needs to know about 1-9, 10-99, 100-999, 1000-9999).

To handle this better, we could probably write a rule keep roman numbers as OOV and translate them by copying the input directly to the output if the attention is high on the number (similar to what is proposed for proper nouns).