Assignment4

April 28, 2020

1 Neural Machine Translation with RNNs

(g) The mask is used for setting padding words' weights(attention distribution score) to $-inf$, i.e. it sets the impact of padding words in source sentence to minimal for decoder. It is useful because the decoder should only pay attention to those real words rather than padding words in source sentence.

(i) Dot product attention vs Multiplicative attention
Dot product attention: Pros: Fewer parameters. Faster and space-efficient. Cons: If attention output vector is not in the same space as decoder's hidden state, the learned result may be not that optimal.

Multiplicative attention vs additive attention.
Additive attention: Pros: Captures non-linear interaction between decoder's attention output and encoder's hidden state. Cons: Computation is slower.

2 Analyzing NMT Systems

(a)
i. There is an unnecessary repeated word "favorite". Possible reason is that the middle word "favorite" is optimal for t timestamp but not for global optimal. Possible way to improve it is using beam search when decoding.
ii. The order of the target sentence is a bit weird, "more reading" should be put before author as adjective modifiers for author. This target sentence's word order is influenced by source sentence's word order. We may increase the decoder's hidden state dimension for learning a better language model.
iii. The word Bolingbroke is wrongly translated. Potential reason: The word Bolingbroke is not in the English corpus. Solution: Add more words in English corpus.
iv. The word "manzala" is translated as "apple", which is incorrect. The reason of this error is that, this word has multiple meanings in Spanish, and the model doesn't capture its context meaning. In this sentence the context shows that the meaning should be "block" rather than "apple". We may improve it by increasing hidden state dimension.
v. The model translate "profesores" to "she", but from the source sentence there is no gender information about the "profesores". This error may be caused by model bias. Possible solution, pretrain the word embedding to reduce the bias causing from the corpus distribution.
vi. The unit conversion is wrongly. The potential reason is that numbers are not interpreted correctly. Maybe we can apply character-level RNN(not quite sure about the solution...)?

(b) The predicted text: ”I was making it secret.” Actual English text: ”She did it in secret.” Spanish text: ”Lo haca en secreto”. The subject is wrong. I guess it is because ”Lo” in Spanish is a general pronoun and the model doesn’t capture the real subject. Maybe we can increase the hidden state dimension for learning more information from context.
The predicted text: ”I was for the first time to Antarctica about 10 years ago , and I saw my first bike.”, actual English text: ”I first went to Antarctica almost 10 years ago, where I saw my first icebergs.” Spanish: ”Fui por primera vez a la Antrtida hace casi 10 aos, y all vi mis primeros tmpanos.” The predicted text for the last word is wrong, its meaning is ”iceberg”, which is shown in the later sentences. To resolve such issue, the model may need to refer to multiple sentences when learning a single sentence.

(c)
i. For c_1: $p_1 = \frac{min(max(3,1),5)}{5} = 0.6$, $p_2 = \frac{min(max(2,0),)}{4} = 0.5$
For c_2: $p_1 = \frac{min(max(2,3),5)}{5} = 0.6$, $p_2 = \frac{min(max(1,1),4)}{4} = 0.55$
Both sentence lengths are 5. So $BP = 1$ for all of them since the reference sentence are with the same length. So c_1: $BLEU = exp(0.5 * (log(0.8) + log(0.5))) = 0.63$
c_2: $BLEU = exp(0.5 * (log(0.8) + log(0.25))) = 0.38$
So c_1 has higher score. I don’t agree with the score, since c_2 is actually a better translation.
ii. For c_1: p_1, p_2 are unchanged. For c_2: $p_1 = \frac{min(2,5)}{5} = 0.4$, $p_2 = \frac{min(1,4)}{4} = 0.28$
Both sentence lengths are 5. So $BP = 1$ for all of them since the reference sentence are with the same length. So c_1: $BLEU = exp(0.5 * (log(0.8) + log(0.5))) = 0.63$
c_2: $BLEU = exp(0.5 * (log(0.8) + log(0.25))) = 0.44$
So c_1 has higher score. I don’t agree with the score, since c_2 is actually a better translation.
iii. As we see in ii., the results are becoming less accurate after losing the second reference sentence. Actually predicted sentence 2 is better, and it is closer to the reference sentence 2. But after we lost reference sentence 2, the comparing between c_1 and c_2 become more unfair.
iv. pros: 1) faster computing compared with human eval. 2) More objective
cons: 1) The comparison may not reflect the real performance of the model results. c_1, c_2 above is a good example. 2) The evaluation metrics is heavily depends on the evaluating corpus, which is less stable compared with human evaluation method.