CS224n Assignment4 Solution

shawn

August 2020

1 Neural Machine Translation with RNNs

(g)

In step() function, the masks are used to fill the value of e_t to be $-inf$ where masks value is 1. As softmax function is $softmax(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)}$, for $x_i = -inf$, the corresponding probability $\alpha_{t,i} = 0$, then it would not contribute to the attention output $a_t = \sum\limits_{i=1}^{m} \alpha_{t,i} h_i^{enc}$.

Masks with value 1 indicates the corresponding word of the input source sentence is pad token. And the hidden state of pad token should not affect the prediction of decoder, so it is necessary for computing the attention.

i

The test BLEU score is 35.76.

j

Dot product attention has one advantage that the computation is simple and efficient. One disadvantage is that it requires the feature size of s_t and h_i to be the same.

Multiplicative attention has one advantage that feature size of s_t and h_i can be arbitrary and different. One disadvantage is that the attention scores are simple linear function of the s_t and h_i.

Additive attention has one advantage that the attention score is learned as non-linear function of s_t and h_i. One disadvantage is that the computational complexity is higher.

2 Analyzing NMT Systems

(c)

(i) for c_1, the $p_1 = \frac{3}{5}$, $p_2 = \frac{2}{4}$, $len(c) = 5$, $len(r) = 5$, $BP = 1$, so the BLEU score for c_1 is $BLEU = exp(0.5 * log(p_1) + 0.5 * log(p_2)) = 0.5477$; for c_2, the $p_1 = \frac{4}{5}$, $p_2 = \frac{2}{4}$, $len(c) = 5$, $len(r) = 5$, $BP = 1$, so the BLEU score for c_2 is $BLEU = exp(0.5 * log(p_1) + 0.5 * log(p_2)) = 0.6324$.

According to the BLEU score, c_2 is the better translation, and I agree with this.

(ii) With respect to r_1 only, the BLEU score for c_1 is still 0.5477. the BLEU score for c_2 is 0.3162. Now c_1 receives the higher score, but it is obvious that c_2 is the better translation.

(iii) A good translation would be one that clearly express the same meaning of the original sentence, but it may not same as the reference translation word by word because there are variations of phrase with similar meaning in the language. When there is only a single reference translation, a bad NMT generated translation could get higher BLEU score, as what we have seen in question (ii). So multiple reference translations will make the BLEU score to be a more comprehensive metric.

(iv) Advantages of BLEU: it is easy and cheap to compute compared with employing people for performance evaluation; Disadvantages: BLEU score can not make a comprehensive evaluation with a single reference as we have discussed. Engineers probably tune the model to get higher BLEU score while it dose not reflect a truly better translation, thus the final NMT model would be bad.