HW4 NLP

Baktash Ansari

May 2023

1 NMT

g) Pad token's positions contains 1s in enc_mask so when we want to calculate e_t they convert to $-\infty$ and then in the calculation of α_t we have : $exp(-\infty) = 0$ so it doesn't have any effect in attention output a_t. Using 1s in enc_mask is necessary because $< pad >$ tokens don't have any meaningful effect on sentences and must be remove from attention output.

h) The BLUE score is 19.88962527.

I) i : One advantages of dot product is efficiency in computing. Dot product attention can be computed only using simple matrix multiplication which is faster than element-wise multiplication. So it is faster and requires less computation. One disadvantages of dot product is that it doesn't have any weight parameter and learning rate like multiplicative attention which leads to less accuracy. **ii :** Additive and multiplicative attention are similar in complexity, although multiplicative attention is faster and more space-efficient in practice as it can be implemented more efficiently using matrix multiplication. Both variants perform similar for small dimensionality d_h of the decoder states, but additive attention performs better for larger dimensions.

2 Analyzing NMT Systems

a) The 1D Convolutional layer can help capture the local dependencies between adjacent characters or morphemes. This can be useful for capturing important linguistic features, such as prefixes or suffixes, that are often used to form new words or modify the meaning of existing words. By learning these local dependencies, the model can better understand the structure and meaning of the input sequence, leading to better translation performance. Based on hint adding a 1D Convolutional layer after the embedding layer in a neural machine translation (NMT) system could help capture the local context of the input sequence. Since each Mandarin Chinese character can represent either an entire word or a morpheme in a word, the 1D Convolutional layer could help capture the relevant morphemes and words that are important for translation.

For example if we focus on example of hint, by using a 1D Convolutional layer, the model can capture the relationship between the two characters and learn to recognize them as a single unit.

b) **i.** Error: The NMT translation incorrectly uses singular "culprit" instead of plural "culprits". Possible Reason: The model may have relied too heavily on the word (culprits) in the source sentence and failed to correctly identify it as a plural noun. Possible Solution: Increase the model's awareness of grammatical number by incorporating information about noun plurality into the input representation or by explicitly modeling plurality in the decoder.

ii. Error: The NMT translation repeats the word "resources" instead of translating the second part of the sentence correctly. Possible Reason: The model may have failed to identify the second part of the sentence as a separate clause and instead treated it as part of the first clause, leading to the repetition of the word "resources". Possible Solution: Improve the model's ability to identify and parse sentence structure by incorporating syntactic information, such as part-of-speech tags or dependency parsing, into the input representation or by using a syntax-aware attention mechanism.

iii. Error: The NMT translation incorrectly uses "administration" instead of "authorities". Possible Reason: The model may have incorrectly interpreted the Chinese word (dāngjú) as referring specifically to a government administration rather than a more general term for authorities. Possible Solution: Increase the model's understanding of the broader context in which words are used by incorporating contextual information, such as topic modeling or document-level attention, into the model architecture.

iv. Error: The NMT translation incorrectly translates the Chinese idiom as "it's not wrong" instead of "act not, err not". Possible Reason: The model may have relied too heavily on the literal meaning of the Chinese characters and failed to correctly interpret the idiomatic meaning of the phrase. Possible Solution: Incorporate a pre-processing step to explicitly identify and translate idiomatic expressions or incorporate knowledge of idiomatic expressions into the model's attention mechanism.

c) C_1**:**

We have : $P_1 = \frac{1+1+1+1}{9} = \frac{4}{9}$ and $P_2 = \frac{1+1+1}{8} = \frac{3}{8}$. $len(r_1) = 11$ and $len(r_2) = 6$ and $len(C_1) = 9$ so we have $BP = exp(1 - \frac{11}{9})$ and we have $BLEU = exp(\frac{-2}{9}) * exp(0.5 * log(\frac{4}{9}) + 0.5 * log(\frac{3}{8}) = 0.542$

C_2 **:**

we have : $P_1 = \frac{1+1+1+1+1+1}{6} = 1$ and $P_2 = \frac{1+1+1}{5} = \frac{3}{5}$. $len(r_1) = 11$ and $len(r_2) = 6$ and $len(C_2) = 6$ so we have $BP = 1$ and we have $BLEU = 1 * exp(0.5 * log(1) + 0.5 * log(\frac{6}{10}) = 0.895$. Based on BLEU score the second one is better but I disagree because the second translations is not complete and doesn't have any reasonable meaning.

ii. C_1 **:**

we have : $P_1 = \frac{1+1+1+1}{9} = \frac{4}{9}$ and $P_2 = \frac{1+1+1}{8} = \frac{3}{8}$. $len(r_2) = 6$ and $len(C_1) = 9$ so we have $BP = 1$ and we have $BLEU = 1 * exp(0.5 * log(\frac{4}{9}) + 0.5 * log(\frac{3}{8}) = 0.677$

C_2 **:**

we have : $P_1 = \frac{1+1+1}{6} = \frac{1}{2}$ and $P_2 = \frac{1}{5}$. $len(r_2) = 6$ and $len(C_2) = 6$ so we have $BP = 1$ and we have $BLEU = 1 * exp(0.5 * log(\frac{1}{2}) + 0.5 * log(\frac{1}{5}) = 0.606$ At this time the first one receive higher score and I agree with it.

iii. Evaluating NMT systems with respect to only a single reference translation can be problematic because a single reference may not capture the full range of acceptable translations for the given source sentence. There may be different valid ways to translate the source sentence, each with its own subtle nuances, idiomatic expressions, or variations in word choice or sentence structure. By using only one reference, the evaluation may not accurately capture the quality of the NMT system's output and may penalize translations that are still valid but do not match the reference exactly. The BLEU score metric is designed to handle multiple reference translations by comparing the candidate translation against all reference translations and computing the modified n-gram precision scores for each reference. BLEU then computes the geometric mean of these n-gram precisions, weighted by a set of weights that sum to 1. The BLEU score thus rewards translations that are closer to any of the references, rather than penalizing translations that do not exactly match any single reference. As a result, BLEU is more robust to variations in reference translations and can better capture the overall quality of the candidate translation.

iv : For advantages we can say that it is faster and less expensive than human evaluation and can be used to evaluate a large number of translations in a short amount of time. For disadvantages BLEU evaluates machine translations based on n-gram precision, which may not capture all aspects of translation quality. BLEU is based on n-gram matching and does not take into account the meaning or context of the translated sentences. As a result, BLEU scores may be inflated for translations that use similar words or phrases as the reference translations but do not capture their meaning accurately. BLEU also does not account for paraphrases or variations in word order or sentence structure that may still be valid translations.