CS 224n: Assignment 4 Written Part

May 1, 2022

Question 1: Neural Machine Translation with RNNs (45 points)

(g) The generate_sent_masks() function in nmt_model.py produces a tensor called enc masks. It has shape (batch size, max source sentence length) and contains 1s in positions corresponding to 'pad' tokens in the input, and 0s for non-pad tokens. Look at how the masks are used during the attention computation in the step() function.
First explain (in around three sentences) what effect the masks have on the entire attention computation. Then explain (in one or two sentences) why it is necessary to use the masks in this way.

Solution:
We generate masks for the paddings as 1. In the step function, we fill the attention scores e_t that equal to 1 with negative infinity. When we pass e_t to the softmax function, because $exp(-\infty) = 0$, the paddings' attention is ignored.
This process is necessary, because paddings are manually added to make the batch data equal length. Thus, it should not be included in the encoding information.

(g) Please report the model's corpus BLEU Score. It should be larger than 10.

Solution:
The model's corpus BLEU Score is 12.33.

(g) In class, we learned about dot product attention, multiplicative attention, and additive attention.

i. Explain one advantage and one disadvantage of dot product attention compared to multiplicative attention.

Solution:
One advantage of dot product attention is that it can be calculated very easily as it doesn't involve any additional variables. One disadvantage is that it requires the s_t and h_i to have the same dimension. It only returns a scalar, thus it contains less information. The multiplicative attention however, doesn't require s_t and h_i to have the same dimension. It's also has more representation than dot product attention. But, it's more costly in high dimension and it involves a new W parameter.

ii. Explain one advantage and one disadvantage of additive attention compared to multiplicative attention.

Solution:
Additive attention has the similar computational complexity as the multiplicative attention, but it has better computational efficiency in high dimensions. However, additive attention has more

new parameters than multiplicative attention (W1 and W2). Also, dimension is another hyperparameter.

Question 2: Analyzing NMT Systems (30 points

(a) Why might it be important to model our Cherokee-to-English NMT problem at the subword-level vs. the whole word-level?

Solution:
Because Cherokee is a polysynthetic language, meaning that words are composed of many morphemes that each have independent meanings. If we use the whole word-level, as the embedding matrix is stationary, it's hard for us to deal with new words. However, if we use the subword-level, then we can embed the new words through its morphemes.

(b) Character-level and subword embeddings are often smaller than whole word embeddings. In 1-2 sentences, explain one reason why this might be the case.

Solution:
The number of characters and subwords is smaller than the number of whole words. Therefore, the embedding matrix of character-level and subword-level is smaller than the whole word-level embedding matrix.

(c) How does multilingual training help in improving NMT performance with low-resource languages?

Solution:
The multilingual training learns shared representations for linguistically similar languages without the need for external constraints, validating long-standing intuitions and empirical results that exploit these similarities. And these learnable shared representations can help to improve the NMT performance with low-resource languages.

(d) Here we present three examples of errors we found in the outputs of our NMT model (which is the same as the one you just trained). The errors are underlined in the NMT translation sentence. For each example of a source sentence, reference (i.e., 'gold') English translation, and NMT (i.e., 'model') English translation, please:
1. Provide possible reason(s) why the model may have made the error (either due to a specific linguistic construct or a specific model limitation).
2. Describe one possible way we might alter the NMT system to fix the observed error. There are more than one possible fixes for an error. For example, it could be tweaking the size of the hidden layers or changing the attention mechanism.

Solution:

i. **Source Translation:** Yona utsesdo ustiyegv anitsilvsgi digvtanv uwoduisdei.
Reference Translation: Fern had a crown of daisies in her hair.
NMT Translation: Fern had <u>her hair</u> with her hair.
Possible reasons for the error: Not enough training data showing the translation to a crown of daisies.
Possible ways to fix: Add more related training data.

ii. **Source Translation:** Ulihelisdi nigalisda.
Reference Translation: She is very excited.

NMT Translation: It's joy.
Possible reasons for the error: Model limitation. It didn't capture the this type of sentence structure.
Possible ways to fix: Increase the size of hidden units. Add more related training data.

iii. **Source Translation:** Tsesdi hana yitsadawoesdi usdi atsadi!
Reference Translation: Don't swim there, Littlefish!
NMT Translations: Don't know how a small fish!
Possible reasons for the error: The model didn't understand that Littlefish is a specific linguistic construct.
Possible ways to fix: Update the embedding matrix.

(e) Now it is time to explore the outputs of the model that you have trained! The test-set translations your model produced in question 1-i should be located in outputs/test outputs.txt.

i. Find a line where the predicted translation is correct for a long (4 or 5 word) sequence of words. Check the training target file (English); does the training file contain that string (almost) verbatim? If so or if not, what does this say about what the MT system learned to do?

Solution:
NMT Translation: But if the demons casteth the word of God, ye shall not see the kingdom of God come unto you.
Reference Translation: But if I by the finger of God cast out demons, then is the kingdom of God come upon you.
The training file contains the string verbatim. This indicates that the MT system learned the translation of the existing word combination in the training file.

ii. Find a line where the predicted translation starts off correct for a long (4 or 5 word) sequence of words, but then diverges (where the latter part of the sentence seems totally unrelated). What does this say about the model's decoding behavior?

Solution:
NMT Translation: And the wall of the city twelve twelve, and the breadth of the twelve tribes: and the twelve tribes of the twelve, and the names of the twelve.
Reference Translation: And the wall of the city had twelve foundations, and on them twelve names of the twelve apostles of the Lamb.
When decoding, the model can not cover all the inputs, and is not good at dealing with rare word combinations.

(f) BLEU score is the most commonly used automatic evaluation metric for NMT systems. It is usually calculated across the entire test set, but here we will consider BLEU defined for a single example. Suppose we have a source sentence s, a set of k reference translations $r_1, ..., r_k$, and a candidate translation c.

i Please consider this example from Spanish. Please compute the BLEU scores for c_1 and c_2. Let $\lambda_i = 0.5$ for $i \in 1, 2$ and $\lambda_i = 0$ for $i \in 3, 4$. Which of the two NMT translations is considered the better translation according to the BLEU Score? Do you agree that it is the better translation?

Solution:
For c_1: the love can always do:
$p_1 = \frac{0+1+1+1+0}{5} = 0.6$
$p_2 = \frac{0+1+1+0}{4} = 0.5$

$BP = 1$
$BLEU = \exp(0.5 \cdot \ln 0.6 + 0.5 \cdot \ln 0.5) = 0.548$

For c_2: love can make anything possible:
$p_1 = \frac{1+1+0+1+1}{5} = 0.8$
$p_2 = \frac{1+0+0+1}{4} = 0.5$
$BP = 1$
$BLEU = \exp(0.5 \cdot \ln 0.8 + 0.5 \cdot \ln 0.5) = 0.632$

The c_2 translation is better according to the BLEU Score. I agree that it's a better translation.

ii Our hard drive was corrupted and we lost Reference Translation r_2. Please recompute BLEU scores for c_1 and c_2, this time with respect to r_1 only. Which of the two NMT translations now receives the higher BLEU score? Do you agree that it is the better translation?

Solution:
For c_1: the love can always do:
$p_1 = \frac{0+1+1+1+0}{5} = 0.6$
$p_2 = \frac{0+1+1+0}{4} = 0.5$
$BP = \exp(1 - 6/5) = 0.8187$
$BLEU = BP \cdot \exp(0.5 \cdot \ln 0.6 + 0.5 \cdot \ln 0.5) = 0.4484$

For c_2: love can make anything possible:
$p_1 = \frac{1+1+0+0+0}{5} = 0.4$
$p_2 = \frac{1+0+0+0}{4} = 0.25$
$BP = 0.8187$
$BLEU = BP \cdot \exp(0.5 \cdot \ln 0.4 + 0.5 \cdot \ln 0.25) = 0.2589$

The c_1 translation is better according to the BLEU Score. I don't agree that it is the better translation.

iii Due to data availability, NMT systems are often evaluated with respect to only a single reference translation. Please explain (in a few sentences) why this may be problematic.

Solution:
As the last exercise shows, if we only have a single reference translation, it may restrict our NMT translation performance. Even if the translation result is better, the BLEU result may be lower due to the limited reference translation.

iv List two advantages and two disadvantages of BLEU, compared to human evaluation, as an evaluation metric for Machine Translation.

Solution:
Advantages: The BLEU Score is faster and easier than human evaluation. It is also language independent.
Disadvantages: Scoring is inflexible and depends largely on the reference translation. Can't evaluate advanced translations.