Practical No. 3

Jan Dziuba

April 9, 2023

1 Part 1

g)

Mask sets weights corresponding to padding tokens to $-\infty$. After softmax those weights are 0. We use masks because we want attention vector to be only dependent on words in the sentence.

i)

Corpus BLEU: 11.934234823654876

j)

Please provide one possible advantage and disadvantage of each attention mechanism.

1. dot product attention

   Dot product attention is computationally efficient, but it is more sensitive to the scale of the input data. If the input vectors have a large magnitude, the dot product may produce very large or very small values, leading to numerical instability issues. In such cases, scaling the input vectors becomes necessary.

2. multiplicative attention

   Slower than dot product attention, but usually gives better results.

3. additive attention

   Even slower than multiplicative attention, but provides better results for larger dimensions.

2 Part 2

a)

1. Bad grammar

   (en) Or you simplify it .. let me write this .. you simplify it.

   (pl) Albo się to upraszcza

   (pred) Albo to się uprości – napiszę to trochę uprościć.

   This sentence is quite confusing, so I am not surprised by imprecise translation from the model.

   I am however surprised by proposed translation, which is very short.

   I would translate the sentence as

   "Albo to upraszczasz .. daj mi to napisać .. upraszczasz to."

   Anyway more training, better model, better data should help with bad grammar.

2. Model translating only part of the sentence

   (en) it is a times sign like that or we could write it as a dot

   (pl) znak razy w ten sposób albo możemy to zapisać jako kropka

   (pred) to jest a razy <unk>

   Because model is translating each line independently, it loses context.

   Giving model more context and changing the model to be able to handle larger context (for example transformer architecture) should help with this error.

3. Wrong word.

   (en) That is going to approach 0.

   (pl) To będzie dążyć do 0.

   (pred) To będzie równe 0.

   More robust model and more training data could help with this problem.

4. Literal translation.

   (en) not quite infinity.

   (pl) może nie do końca nieskończoność.

   (pred) nie dość nieskończoność.

More robust model and more training data could help with this problem.

5. Repeating words.

   (en) I am now going to find some fish for my next

   (pl) Chcę znaleźć obrazki ryb do mojego kolejnego

   (pred) Zamierzam teraz znaleźć mi mi mi

   This error is called text degeneration and is common in NLP.

   One approach to solving it would be to replace beam search with top-k or top-p sampling.

   Another would be to penalize for repeating words or sequences of words.

b)

Corpus BLEU: 7.114279307554786

Difference is large, because model will not be able to translate directly words that are not covered by our vocabulary, and will have to predict them using sentence context.
The drop will be big because Polish grammar differs from English.
Polish is also an inflected language, so the model needs more data to know all the forms the word can exist in.
To avoid this issue we would need either more data, or model that can better understand inflection rules.

c)

1. c_1

$$p_1 = \frac{1+1+1+1}{1+1+1+1} = \frac{4}{4} = 1, \quad p_2 = \frac{1+0+1}{1+1+1} = \frac{2}{3}$$

$$c = 4, \quad r^* = 5, \quad BP = \exp\left(-\frac{1}{4}\right)$$

$$BLEU = \exp\left(-\frac{1}{4}\right) \cdot \exp\left(\frac{1}{2}(\ln(1) + \ln\left(\frac{2}{3}\right))\right) \approx 0.636$$

c_2

$$p_1 = \frac{1+1+1+1+1+0+0+1}{1+1+1+1+1+1+1+1} = \frac{6}{8}, \quad p_2 = \frac{1+1+1+1+0+0+0}{1+1+1+1+1+1+1} = \frac{4}{7}$$

$$c = 8, \quad r^* = 7, \quad BP = 1$$

$$BLEU = 1 \cdot \exp\left(\frac{1}{2}(\ln\left(\frac{3}{4}\right) + \ln\left(\frac{4}{7}\right))\right) \approx 0.655$$

According to the BLUE score, c_2 is considered the better translation. In my opinion, the better translation is c_1 because while c_1 has slightly different meaning than r_1, c_2 contains both 'to oznacza' and 'jest' which is grammatically incorrect, it also contains 'zbiór' instead of 'podzbiór'.

2. c_1

$$p_1 = \frac{1+1+1+1}{1+1+1+1} = \frac{4}{4} = 1, \quad p_2 = \frac{1+0+1}{1+1+1} = \frac{2}{3}$$

$$c = 4, \quad r^* = 5, \quad BP = \exp\left(-\frac{1}{4}\right)$$

$$BLEU = \exp\left(-\frac{1}{4}\right) \cdot \exp\left(\frac{1}{2}(\ln(1) + \ln\left(\frac{2}{3}\right))\right) \approx 0.636$$

For c_2:

$$p_1 = \frac{3}{8}, \quad p_2 = \frac{1}{7}, \quad c = 8, \quad r^* = 5, \quad BP = 1$$

$$BLEU = 1 \cdot \exp\left(\frac{1}{2}(\ln\left(\frac{3}{8}\right) + \ln\left(\frac{1}{7}\right))\right) \approx 0.231$$

Now c_1 has higher BLUE score. As before I believe that c_1 is a better translation.

3. One sentence can often be translated in many ways. It is especially true with larger sentences.

   This means that correct translation can potentially get a low BLEU score.

4. Advantages:

   - It's quick to calculate and easy to understand.
   - It's the most popular metric for machine translation, so it's easy to compare model performance.

   Disadvantages:

   - Needs high quality reference translations.
   - It does not take into account synonymous words.