به نام خدا

درس مبانی پردازش زبان و گفتار

تمرین سری چهارم

مدرس درس:
جناب آقای دکتر اعتمادی

تهیه شده توسط:
الناز رضایی ۹۸۴۱۱۳۸۷

تاریخ ارسال: ۱۴۰۲/۰۲/۲۷

سوال ۱ :

پاسخ (g)

In the decoding process, we generate masks based on the length of each sentence and incorporate them into the decoding procedure. Within the `step()` function, after calculating the attention scores e_t, we apply a masking operation to eliminate the influence of padding. Specifically, we identify the padded positions (marked as 1s in the mask) and replace their corresponding attention scores in e_t with negative infinity.

When we subsequently pass e_t through the softmax function to obtain the attention distribution α_t, the exponential term of negative infinity results in a probability of zero for the padded positions. Consequently, the attention mechanism effectively ignores the paddings during the decoding process.

This masking step is necessary because the padding tokens are not part of the original data and are solely introduced to achieve equal length within the batch. It would be nonsensical for the decoder to attend to non-existent words in the encoding. Therefore, by masking out the padded positions and assigning them a probability of zero, we ensure that the decoder focuses only on the relevant and meaningful parts of the input sentences.

پاسخ (h)

corpus BLEU Score: 19.7539239421

پاسخ (i)

i . Advantage: Dot product attention requires fewer computational operations compared to multiplicative attention, making it faster and

more efficient to compute. This efficiency is especially beneficial when dealing with large-scale models or when performing attention operations in real-time applications.

Disdvantage: A disadvantage of dot product attention is that it does not capture complex interactions between the query and key vectors. The dot product operation measures the similarity between the query and key vectors based on their raw values, without explicitly considering any learned weights or transformations. As a result, dot product attention may struggle to model more intricate relationships between the query and key vectors, potentially limiting its representational capacity in certain scenarios.

ii . Advantage: One advantage of additive attention compared to multiplicative attention is its ability to capture more complex relationships between the input and output sequences. In additive attention, the attention weights are computed by passing the concatenation of the encoder and decoder states through a feed-forward neural network. This additional non-linearity allows the model to learn more intricate patterns and dependencies, making it suitable for tasks where the relationships are not easily captured by a simple dot product.

Disdvantage: The additional feed-forward neural network introduces more parameters and computational operations, which can make training and inference slower. In cases where efficiency is crucial, such as in real-time applications or large-scale models, the increased computational cost of additive attention can be a limiting factor. Multiplicative attention, on the other hand, offers a simpler and more computationally efficient approach by directly calculating attention scores using a

dot product or similarity measure between the encoder and decoder states.

سوال ۲:

پاسخ (a)

By incorporating a Conv1D layer, we can extract contextual features from a sentence, enabling us to capture the nuanced meanings that arise from the combination of Mandarin Chinese characters. This approach is particularly valuable as it allows us to discern and differentiate the intricate features associated with complex Chinese words.

پاسخ (b)

i .

Error: using singular format of culprits.

Reason: the model may have difficulty handling pluralization in English.

Solution: improve the model's understanding of noun number agreement by providing more training examples with plural nouns and their corresponding translations.

ii .

Error: repeating the phrase "resources have been exhausted."

Reason: the model might struggle with generating diverse and varied output, leading to repetitive phrases.

Solution: implementing a coverage mechanism to ensure that the

model attends to different parts of the input sentence when generating translations.

iii .

Error: incorrect translation for ”a national mourning today”. (the model translated ”today’s day”)

Reason: lack contextual knowledge of cultural events and their corresponding translations.

Solution: augment the training data with more examples that involve cultural events and their accurate translations. Additionally, providing contextual information during training, such as information about national holidays, can help the model generate more accurate translations.

iv .

Error: incorrect translation for ”“ act not, err not”. (the model translated ”it’s not wrong”)

Reason: the model may struggle with accurately translating idiomatic expressions or capturing the nuanced meaning of the phrase.

Solution: enhance the model’s understanding of idiomatic expressions by including more examples of similar expressions during training. Additionally, incorporating a specific module or attention mechanism that focuses on preserving the idiomatic nature of certain phrases can help improve the accuracy of such translations.

پاسخ (e)

i . c_1 :

$$P_1 = \frac{4}{9} \quad , \quad P_2 = \frac{3}{8}$$

$$len(c_1) = 9 \quad , \quad len(r_1) = 11$$

$$BP(c_1) = e^{1-\frac{11}{9}} = e^{\frac{-2}{9}}$$

$$BLEU(c_1) = e^{\frac{-2}{9}}.e^{\frac{1}{2}\cdot\frac{4}{9}+\frac{1}{2}\cdot\frac{3}{8}} = 1.2$$

c_2 :

$$P_1 = 1 \quad , \quad P_2 = \frac{3}{5}$$

$$len(c_2) = 6 \quad , \quad len(r_2) = 6$$

$$BP(c_2) = 1$$

$$BLEU(c_2) = 1.e^{\frac{1}{2}.1+\frac{1}{2}\cdot\frac{3}{5}} = 2.2$$

According to BLEU scores, c2 score is higher than c1, but it has a weaker performance than c1.

ii . c_1 :

$$P_1 = \frac{4}{9} \quad , \quad P_2 = \frac{3}{8}$$

$$BP(c_1) = 1$$

$$BLEU(c_1) = 1.e^{\frac{1}{2}\cdot\frac{4}{9}+\frac{1}{2}\cdot\frac{3}{8}} = 1.5$$

c_2 :

$$P_1 = \frac{1}{2} \quad , \quad P_2 = \frac{1}{5}$$

$$BP(c_2) = 1$$

$$BLEU(c_2) = 1.e^{\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{5}} = 1.4$$

According to BLEU scores, c1 score is higher than c2, and it has a stronger performance than c2.

iii . Evaluating NMT systems with respect to only a single reference translation can be problematic because it doesn't capture the full range of possible translations for a given source sentence. Different reference translations can have variations in wording, sentence structure, and word order, and evaluating against only one reference may favor translations that align more closely with that particular reference, even if other valid translations exist.

The BLEU score metric, when computed across multiple reference translations, aims to mitigate this problem by considering the modified n-gram precision of the candidate translation with respect to all references. By incorporating multiple references, BLEU encourages translations that capture common patterns across the references while penalizing translations that overly rely on specific phrasing from a single reference. This helps provide a more comprehensive evaluation of the translation quality.

iv . Advantages:

(a) Automation: BLEU can be computed automatically without the need for human judges, making it a faster and more cost-effective evaluation method.

(b) Quantitative Measure: BLEU provides a numerical score that allows for easy comparison and tracking of translation quality

across different systems, models, or iterations.

Disdvantages:

(a) Limited Scope: BLEU primarily focuses on n-gram precision and does not capture other aspects of translation quality, such as semantic accuracy, fluency, or grammatical correctness. It may not reflect the overall translation quality comprehensively.

(b) Insensitivity to Synonyms and Paraphrasing: BLEU heavily relies on exact word matches and does not consider synonymous or paraphrased expressions. Translations that convey the same meaning but use different words may receive lower BLEU scores.