复查测验提交: Homework 2

用户	生物医学工程学院 吉泓光
课程	自然语言处理
测试	Homework 2
已开始	24-3-27 下午1:14
已提交	24-4-1 下午10:11
截止日期	24-4-2 下午11:59
状态	已完成
尝试分数	得 130 分，满分 150 分
已用时间	128 小时 56 分钟
显示的结果	所有答案, 已提交的答案, 正确答案

问题 1

得 10 分，满分 10 分

Bidirectional LSTM provide an easier way to learn long distance dependencies, so it usually perform better than vanilla RNNs in language modeling tasks.

所选答案: 错

答案: 对

错

问题 2

得 10 分，满分 10 分

It is safe to compare the perplexities of language models with different vocabularies.

所选答案: 错

答案: 对

错

问题 3

得 10 分，满分 10 分

Four students are training a simple language model. The vocabulary size

is 33279. Below are the perplexities of language models reported by these students. Who is absolutely lying?

所选答案: ✓ Macha: 0.198248938724835

答案: Ben: 124853942.3429384

Saoirse: 965.0860734119312

Tomm: 1.0

✓ Macha: 0.198248938724835

问题 4

得 10 分，满分 10 分

We can solve the vanishing gradient problem in vanilla RNN by multiply the gradient with a constant scalar in when the gradient is small.

所选答案: ✓ 错

答案: 对

✓ 错

问题 5

得 10 分，满分 10 分

Compared to n-gram language models, fixed window neural language models won't suffer from the data sparsity problem, because:

所选答案: ✓ Fixed window neural language models use embeddings to represent words instead of the words themselves.

答案: Neural language models always have enough data do train.

✓ Fixed window neural language models use embeddings to represent words instead of the words themselves.

Fixed window neural language models do not care about the word order while n-gram language models do consider the word order.

Fixed window neural language models have less parameters than n-gram language models.

问题 6

得 0 分，满分 10 分

Select all correct statements

所选答案: ✓ A. The attention module is fast to compute because it can be easily vectorized and is parallelizable.

B.
Self-attention means the "queries", "keys" and "values" are all from the same sequence of vectors.

C.
Multi-head self-attention takes self-attention over the original input vectors multiple times and concatenate the attention output together.

D.
[a, b, c, d], [d, b, a, c] are two sequences of input vectors, after feeding them into a multi-head self-attention layer without a positional encoding, the attention output for the input vector b are the same.

答案：A.
The attention module is fast to compute because it can be easily vectorized and is parallelizable.

B.
Self-attention means the "queries", "keys" and "values" are all from the same sequence of vectors.

C.
Multi-head self-attention takes self-attention over the original input vectors multiple times and concatenate the attention output together.

D.
[a, b, c, d], [d, b, a, c] are two sequences of input vectors, after feeding them into a multi-head self-attention layer without a positional encoding, the attention output for the input vector b are the same.

问题 7

得 10 分，满分 10 分

Select all correct statements about the **RNN language model with attention** covered in the lecture. Suppose we are training the language model on a sentence of length L, and we are at time step t and predicting the (t+1)-th token.

所选答案：Keys and values are hidden states from time step 1 to t.
Query vector is the hidden state of the t-th token

答案：Keys and values are hidden states from time step 1 to t.
Keys and values are hidden states from time step 1 to t-1.
Keys and values are hidden states from time step 1 to L.
Query vector is the hidden state of the t-th token
Query vector is the hidden state of the (t+1)-th token

问题 8

得 10 分，满分 10 分

Select all correct statements

所选答案：A. RNNs take O(sequence length) steps for distant word pairs to interact.
C. The attention scores are calculated by the "queries" and "keys".
D. The final attention output is a weighted sum of "values".

答案：A. RNNs take O(sequence length) steps for distant word pairs to interact.
B. The attention scores are calculated by the "keys" and "values".
C. The attention scores are calculated by the "queries" and "keys".
D. The final attention output is a weighted sum of "values".
E.

Attention takes $\Theta(\log_2(\text{sequence length}))$ steps for distant word pairs to interact.

问题 9

得 10 分，满分 10 分

Select the correct statement

所选答案：
Not all linear attention mechanisms guarantee their attention to be a valid distribution.

答案：
A dot product attention has complexity that scales linearly with the length of input.

In dot product attention, we compute a distribution of **query** vectors that a **key** vector focuses on.

Not all linear attention mechanisms guarantee their attention to be a valid distribution.

In linear attention, by limiting the attention matrix according to some pre-defined patterns, the complexity of attention is reduced.

问题 10

得 0 分，满分 0 分

Select all correct statements

所选答案：
B.
Adding a feed-forward layer with nonlinearities improves the model's ability and allows the neural network model more complex functions

答案：
A.
Position embedding is used to encode the word-order and word-position information. A randomly initialized position embedding is invalid because it cannot encode the word-order information.

B.
Adding a feed-forward layer with nonlinearities improves the model's ability and allows the neural network model more complex functions

C.
The residual connection smoothes the loss landscape and makes training easier, it also introduces more model parameters to improve the model's expressive power.

D.
Layer normalization shifts the "mean" and "standard deviation" of the previous layer output to zero.

问题 11

得 10 分，满分 10 分

BERT is finetuned using MLM (masked language modeling) and NSP (next sentence prediction) and then pretrained on downstream tasks such as text classification.

所选答案：
错

答案：
对

错

问题 12

得 0 分，满分 10 分

Select all correct statements

所选答案：

The difference between MEMM and CRF is that the former is locally normalized and suffers from the label bias issue, while the latter is globally normalized.

In CRF training, both max-margin loss and negative log-likelihood loss involve the use of the forward algorithm to compute the partition function.

Neural CRFs use neural networks (e.g., BiLSTMs, Transformers) to compute CRF potential scores.

答案：

The difference between MEMM and CRF is that the former is locally normalized and suffers from the label bias issue, while the latter is globally normalized.

MEMMs prefer states with higher number of transitions, thus suffer from the label bias issue.

In CRF training, both max-margin loss and negative log-likelihood loss involve the use of the forward algorithm to compute the partition function.

Neural CRFs use neural networks (e.g., BiLSTMs, Transformers) to compute CRF potential scores.

问题 13

得 10 分，满分 10 分

Select all correct statements

所选答案：

1.
The Baum-Welch algorithm is a special case of the EM algorithm, and can be used for unsupervised learning of HMM parameters.

3.
The forward-backward algorithm has the same time and space complexity as the Viterbi algorithm.

4.
We can use the "count and normalize" strategy to train an HMM in both supervised and unsupervised manners. The difference is that, in supervised learning, "count" is the actual counts, whereas in unsupervised learning, "count" is the expected counts.

答案：

1.
The Baum-Welch algorithm is a special case of the EM algorithm, and can be used for unsupervised learning of HMM parameters.

2.
If we use the Baum-Welch algorithm to train an HMM, it will finally converge to a global optimum after running a sufficient number of iterations.

3.
The forward-backward algorithm has the same time and space complexity as the Viterbi algorithm.

4.
We can use the "count and normalize" strategy to train an HMM in both supervised and unsupervised manners. The difference is that, in supervised learning, "count" is the actual counts, whereas in unsupervised learning, "count" is the expected counts.

问题 14

得 10 分，满分 10 分

Beam search can always find the optimal translation in the decoding process of neural machine translation.

所选答案：✔ 错

答案：对

✔ 错

问题 15

得 10 分，满分 10 分

Layernorm makes training easier by normalizing the mean and derivation of hidden states to zero.

所选答案：✔ 错

答案：对

✔ 错

问题 16

得 10 分，满分 10 分

Select all correct options.

所选答案：✔ If we remove the cross-attention in a transformer, the decoder will always output the same sequence when using greedy decoding.

答案：When training an encoder-decoder seq2seq model, the encoder and decoder must be trained separately.

The future tokens should be masked in the decoder of a non-autoregressive model to avoid leaking the answer.

Beam search is guaranteed to decode a sentence with a higher score than greedy decode

✔ If we remove the cross-attention in a transformer, the decoder will always output the same sequence when using greedy decoding.

2024年5月29日 星期三 下午07时31分26秒 CST

确定