

EECS545 Lecture 14 Quiz Solutions

1. What is the main function of the attention mechanism in deep learning?
 - (a) To reduce overfitting in the neural network
 - (b) To focus on the most important parts of the input sequence
 - (c) To speed up the training of the neural network
 - (d) To regularize the weights in the neural network

Solution: (b).

2. What are the benefits of using a attention over an recurrent layers for sequential tasks?
 - (a) Attention is less memory intensive.
 - (b) Attention is more parallelizable during training
 - (c) Attention is more parallelizable for generating a new sequence during test time
 - (d) Attention doesn't suffer from vanishing gradients
 - (e) Attention can directly use inputs from long sequences without memory

Solution: (b) and (e).

In attention during training (b), we can encode the entire sequence length in one pass, whereas RNN requires T passes, where T is the sequence length. However, when inferring a new sequence (c) of some length T , both RNNs and attention have to do T passes.

3. In a transformer model, what is the purpose of the positional encoding?
 - (a) To add information about the position of each token in the input sequence
 - (b) To reduce the dimensionality of the input sequence
 - (c) To compute the attention weights between each pair of tokens in the input sequence
 - (d) To aggregate the information from each token in the input sequence into a single vector

Solution: (a).

4. Path length is the number of layers an input has to go through, before it is output. An RNN has a maximum path length of $O(n)$, where n is the sequence length. This is because the first input goes through n RNN calls then “exits” through the last sequence output. What is the maximum path length of a self-attention layer?

- (a) $O(1)$
- (b) $O(n)$
- (c) $O(n^2)$
- (d) $O(n^3)$

Solution: $O(1)$. Each input in the sequence goes through the same key-query-value matrix product once in self-attention.

5. In a transformer model, why is it important to use causal attention masks in the decoder?

- (a) To prevent the model from overfitting to the training data.
- (b) To allow the model to attend to all positions in the input sequence.
- (c) To avoid introducing future information into the decoding process.
- (d) To increase the model's capacity to handle long input sequences.

Solution: (c). For an input x_t , A causal attention mask masks out all values from x_{t+1}, x_{t+2}, \dots . So, when encoding x_t , the transformer cannot use any information from the “future”.