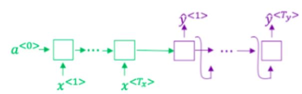
1. Consider using this encoder-decoder model for machine translation. 1 point



True/False: This model is a "conditional language model" in the sense that the encoder portion (shown in green) is modeling the probability of the input sentence x.

- [ ] True
- [x] False

2. In beam search, if you increase the beam width B, which of the following would you expect to be true? Check all that apply. 1 point

- [x] Beam search will generally find better solutions (i.e. do a better job maximizing $P(y \mid x)$)
- [ ] Beam search will converge after fewer steps.
- [x] Beam search will run more slowly.
- [x] Beam search will use up more memory.

3. True/False: In machine translation, if we carry out beam search using sentence normalization, the algorithm will tend to output overly short translations. 1 point

- [x] False
- [ ] True

4. Suppose you are building a speech recognition system, which uses an RNN model to map from audio clip x to a text transcript y. Your algorithm uses beam search to try to find the value of y that maximizes $P(y \mid x)$. 1 point

On a dev set example, given an input audio clip, your algorithm outputs the transcript $\hat{y} =$ "I'm building an A Eye system in Silly con Valley.", whereas a human gives a much superior transcript $y^* =$ "I'm building an AI system in Silicon Valley."

According to your model,

$P(\hat{y} \mid x) = 7.21*10^{-8}$

$P(y^* \mid x) = 1.09*10^{-7}$

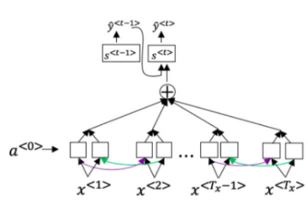
Would you expect increasing the beam width B to help correct this example?

- [ ] No, because $P(y^* \mid x) > P(\hat{y} \mid x)$ indicates the error should be attributed to the RNN rather than to the search algorithm.
- [ ] Yes, because $P(y^* \mid x) > P(\hat{y} \mid x)$ indicates the error should be attributed to the RNN rather than to the search algorithm.
- [ ] Yes, because $P(y^* \mid x) > P(\hat{y} \mid x)$ indicates the error should be attributed to the search algorithm rather than to the RNN.
- [x] No, because $P(y^* \mid x) > P(\hat{y} \mid x)$indicates the error should be attributed to the search algorithm rather than the RNN.

5. Continuing the example from Q4, suppose you work on your algorithm for a few more weeks, and now find that for the vast majority of examples on which your algorithm makes a mistake, $P(y^* \mid x) > P(\hat{y} \mid x)$. This suggests you should focus your attention on improving the search algorithm. 1 point

- [ ] False.
- [x] True.

6. Consider the attention model for machine translation. 1 point



Further, here is the formula for $\alpha^{<t,t'>}$.

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$

Which of the following statements about $\alpha^{<t,t'>}$ are true? Check all that apply.

- [ ] $\sum_{t'} \alpha^{<t,t'>} = 1$. (**Note the summation is over t'.**)
- [ ] $\sum_{t'} \alpha^{<t,t'>} = 0$. (**Note the summation is over t'.**)
- [x] $\alpha^{<t,t'>}$ is equal to the amount of attention $y^{<t>}$ should pay to $a^{<t'>}$
- [x] We expect $\alpha^{<t,t'>}$ to be generally larger for values of $a^{<t'>}$ that are highly relevant to the value the network should output for $y^{<t'>}$. (**Note the indices in the superscripts.**)

7. The network learns where to "pay attention" by learning the values $e^{<t,t'>}$, which are computed using a small neural network: 1 point

Which of the following does $s^{<t>}$ depend on? Select all that apply.

- [x] $\alpha^{<t,t'>}$
- [x] $e^{<t,t'>}$
- [ ] s^t is independent of $\alpha^{<t,t'>}$ and $e^{<t,t'>}$.
- [ ] $s^{<t+1>}$

8. The attention mechanism was primarily introduced to solve the information bottleneck problem in standard encoder-decoder models. In which situation is this mechanism **less critical** for achieving good performance? — 1 point

   - ( ) The input sequence length T_x is large.
   - (•) The input sequence length T_x is small.

9. Under the CTC model, identical repeated characters not separated by the "blank" character (_) are collapsed. Under the CTC model, what does the following string collapse to? — 1 point

   __c_oo_o_kk___b_ooooo__oo__kkk

   - ( ) cokbok
   - ( ) cook book
   - ( ) cooookkbooooooookkk
   - (•) cookbook

10. In trigger word detection, if the target label for $x^{<t>}$ is 1: — 1 point

   - ( ) The total time that the trigger word detection algorithm has been running is 1.
   - ( ) Only one word has been stated.
   - ( ) There is exactly one trigger word.
   - (•) Someone has just finished saying the trigger word at time *t*.