

✓ **Congratulations! You passed!**

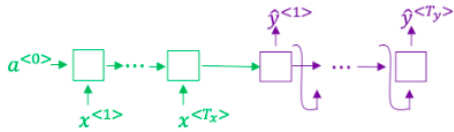
Grade received 90% Latest Submission Grade 90% To pass 80% or higher

Retake the assignment in  
23h 52m

Go to next  
item

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

1 / 1 point



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  $x$ .

☒ False

☐ True

↗ Expand

✓ Correct

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 / 1 point

☒ Beam search will run more slowly.

✓ Correct

☐ Beam search will converge after fewer steps.

☒ Beam search will generally find better solutions (i.e. do a better job maximizing  $P(y | x)$ )

✓ Correct

☒ Beam search will use up more memory.

✓ Correct

↗ Expand

✓ Correct

Great, you got all the right answers.

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 / 1 point

☒ False

☐ True

Expand

Correct

In machine translation, if we carry out beam search without using sentence normalization, the algorithm will tend to output overly short translations.

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  $P(y | x)$ .

1 / 1 point

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

According to your model,

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

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

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

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

Expand

Correct

$P(y^* | x) > P(\hat{y} | x)$  indicates the error should be attributed to the search algorithm rather than to the RNN. Increasing the beam width will generally allow beam search to find better solutions.

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,  $P(y^* | x) > P(\hat{y} | x)$ . This suggests you should focus your attention on improving the search algorithm.

1 / 1 point

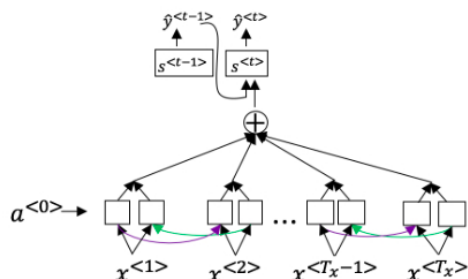
- ☒ True.
- ☐ False.

Expand

Correct

6. Consider the attention model for machine translation.

0 / 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.

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

✓ Correct

We expect  $\alpha^{<t,t'>}$  to be larger for activation values that are highly relevant to the value the network should output for  $y^{<t>}$ .

- ☐  $\sum_{t'} \alpha^{<t,t'>} = 0$ . (Note the summation is over  $t'$ .)

↗ Expand

✗ Incorrect

You didn't select all the correct answers

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

1 / 1 point

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

- ☐  $s^t$  is independent of  $\alpha^{<t,t'>}$  and  $e^{<t,t'>}$ .
- ☒  $\alpha^{<t,t'>}$

✓ Correct

$s^{<t>}$  depends on  $\alpha^{<t,t'>}$  which in turn depends on  $e^{<t,t'>}$ .

- ☐  $s^{<t+1>}$
- ☒  $e^{<t,t'>}$

✓ Correct

$s^{<t>}$  depends on  $\alpha^{<t,t'>}$  which in turn depends on  $e^{<t,t'>}$ .

↗ Expand

✓ Correct

Great, you got all the right answers.

8. Compared to the encoder-decoder model shown in Question 1 of this quiz (which does not use an attention mechanism), we expect the attention model to have the least advantage when:

1 / 1 point

- ☒ The input sequence length  $T_x$  is small.
- ☐ The input sequence length  $T_x$  is large.

↗ Expand

The encoder-decoder model works quite well with short sentences. The true advantage for the attention model occurs when the input sentence is large.

1 / 1 point

Expand

The basic rule for the CTC cost function is to collapse repeated characters not separated by "blank". If a character is repeated, but separated by a "blank", it is included in the string.

1 / 1 point

Expand

$x^{\langle t \rangle}$  represents the features of the audio at time  $t$ .