Sequence Models & Attention Mechanism

Congratulations! You passed!

Grade received 100%

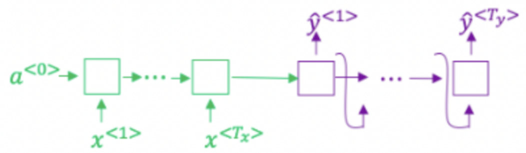
Latest Submission Grade 100%

To pass 80% or higher

Go to next item

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

1 / 1 point



True/False: This model is a “conditional language model” in the sense that the decoder portion (shown in purple) is modeling the probability of the output sentence y given the input sentence x.

- False
- True

Expand

Correct
The encoder-decoder model for machine translation models the probability of the output sentence y conditioned on the input sentence x.

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

- [x] Beam search will use up more memory.

✓ Correct

- [ ] Beam search will converge after fewer steps.
- [x] Beam search will run more slowly.

✓ Correct

- [x] Beam search will generally find better solutions (i.e. do a better job maximizing $P(y \mid x)$)

✓ 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

- ( ) True
- (•) False

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 $P(y \mid x)$.

1 / 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?

- ○ 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.
- ○ 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.
- ○ 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 search algorithm rather than to the RNN.

Expand

Correct

$P(y^* \mid x) > P(\hat{y} \mid x)$ 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, $P(y^* \mid x) > P(\hat{y} \mid x)$. This suggests you should not focus your attention on improving the search algorithm.

1 / 1 point

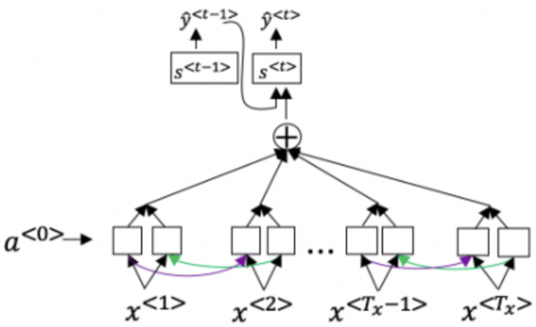
- (●) False
- ( ) True

Expand

Correct

$P(y^* \mid x) > P(\hat{y} \mid x)$ indicates the error should be attributed to the search algorithm rather than to the RNN.

6. Consider the attention model for machine translation. **1 / 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.

- [x] $\sum_{t'} \alpha^{<t,t'>} = 1$ (Note the summation is over t'.)

✓ **Correct**

- [ ] $\sum_{t} \alpha^{<t,t'>} = 1$ Timed quizzes e the summation is over 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.)

✓ **Correct**

- [ ] 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.)

Expand

Correct
Great, you got all the right 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**

We can replace $s^{<t-1>}$ with $s^{<t>}$ as an input to this neural network because $s^{<t>}$ is independent of $\alpha^{<t,t'>}$ and $e^{<t,t'>}$.

- ( ) True
- (•) False

Expand

Correct
We can't replace $s^{<t-1>}$ with $s^{<t>}$ as an input to this neural network. This is because $s^{<t>}$ depends on $\alpha^{<t,t'>}$ which in turn depends on and $e^{<t,t'>}$; so at the time we need to evaluate this network, we haven't computed $s^{<t>}$.

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

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

9.

1 / 1 point

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?

aaa_aaaaaa_______rr_dddddddd______v_aaaaaa_rrrr_______kk

- ( ) aa rd var k
- (●) aardvark
- ( ) ardvark
- ( ) aaaaaaaaarrddddddddvaaaaaarrrrkk

Expand

Correct

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.

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

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.
- (selected) Someone has just finished saying the trigger word at time <i>t</i>.

Expand

Correct

Target labels indicate whether or not a trigger word has been said.