← Back **Sequence Models & Attention Mechanism**
Graded Assignment • 50 min

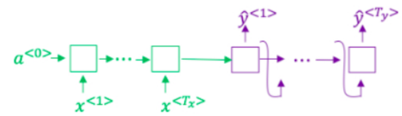
English **Due** Dec 23, 10:29 AM IST

Your grade: 97.50%

Your latest: **97.50%** • Your highest: **97.50%** • To pass you need at least 80%. We keep your highest score.

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

Correct

1 / 1 point

2. In beam search, if you increase the beam width B, which of the following would you expect to be true?

- ( ) Beam search will use up less memory.
- ( ) Beam search will converge after fewer steps.
- ( ) Beam search will run more quickly.
- (●) Beam search will generally find better solutions (i.e. do a better job maximizing *P(y|x)*).

Correct
As the beam width increases, beam search runs more slowly, uses up more memory, and converges after more steps, but generally finds better solutions.

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

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,

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

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

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

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

Correct

5. Continuing the example from Q4, suppose you work on your algorithm for a few more weeks, and now find **1 / 1 point**

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.

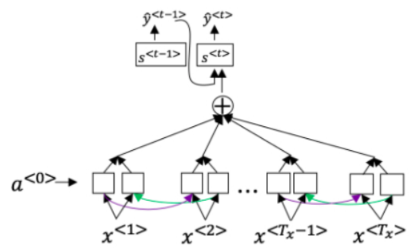
1 / 1 point

- [ ] False.
- [x] True.

Correct

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.

- [ ] $\sum_{t'} \alpha^{<t,t'>} = 0$
- [ ] $\sum_{t'} \alpha^{<t,t'>} = -1$
- [x] $\alpha^{<t,t'>}$ is equal to the amount of attention $y^{<t>}$ should pay to $a^{< t' >}$

Correct
Correct! $\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

- [ ] 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:

0.75 / 1 point

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

- [ ] $e^{<t,t'>}$
- [x] $\alpha^{<t,t'>}$

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

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

You didn't select all the correct 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 large.
- [x] The input sequence length T_x is small.

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.

1 / 1 point

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?

aaa_aaaaaa________rr_ddddddddd_______v_aaaaaa_rrrr________kk

- ( ) aa rd var k
- ( ) ardvark
- (•) aardvark
- ( ) aaaaaaaaarrdddddddddvaaaaaarrrrkk

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, $x^{<t>}$ is:

1 / 1 point

- (•) Features of the audio (such as spectrogram features) at time t.
- ( ) Whether the trigger word is being said at time t.
- ( ) The t-th input word, represented as either a one-hot vector or a word embedding.
- ( ) Whether someone has just finished saying the trigger word at time t.

Correct