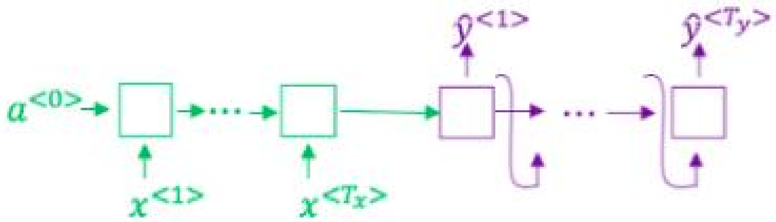
1. Consider using this encoder-decoder model for machine translation.

1 / 1 point



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

- ○ True
- ◉ False

Expand

Correct

The encoder-decoder model for machine translation models the probability of the output sentence y conditioned on the input sentence x. The encoder portion is shown in green, while the decoder portion is shown in purple.

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

✓ Correct

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

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

1 / 1 point

- (selected) True
- False

Expand

Correct

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,

$P(\hat{y} \mid x) = 1.95*10^{-7}$

$P(y^* \mid x) = 3.42*10^{-9}$

True/False: Trying a different network architecture could help correct this example.

- ○ False
- ◉ True

Expand

Correct

$P(y^* \mid x) < P(\hat{y} \mid x)$ indicates the error should be attributed to the RNN rather than to the search algorithm. If the RNN model is at fault, then a deeper layer of analysis could help to figure out if you should add regularization, get more training data, or try a different network architecture.

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

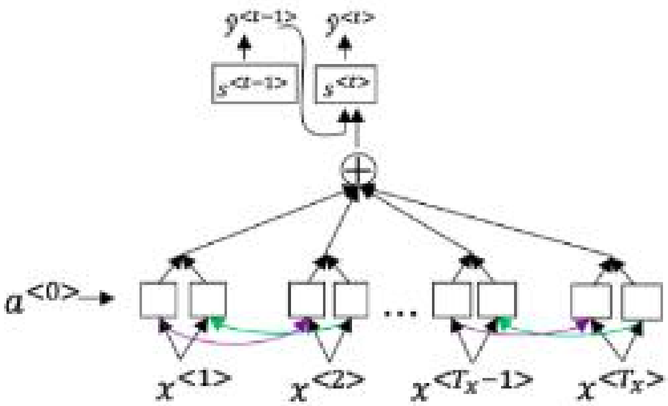
- [ ] False.
- [x] True.

Expand

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.

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

✓ Correct

Correct! If we sum over $\alpha^{<t,t'>}$ for all t' (the formulation can be seen in the image), the numerator will be equal to the denominator, therefore, $\sum_{t'} \alpha^{<t,t'>} = 1$.

- [ ] 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.)**
- [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

Correct! $\alpha^{<t,t'>}$ is equal to the amount of attention $y^{<t>}$ should pay to $a^{<t'>}$. So, if a value of $a^{<t'>}$ is highly relevant to $y^{<t>}$, then the attention coefficient $\alpha^{<t,t'>}$ should be larger. Note the difference between a (activation) and α (attention coefficient).

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

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 greatest advantage when:

1 / 1 point

- [ ] The input sequence length T_x is small.
- [x] The input sequence length T_x is large.

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 / 1 point

__c_oo_o_kk___b_ooooo__oo__kkk

- cokbook
- cookbook
- cook book
- coookkbooooooookkk

10. In trigger word detection, $x^{<t>}$ is:

1 / 1 point

- (selected) Features of the audio (such as spectrogram features) 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.
- Whether the trigger word is being said at time t.