Back Sequence Models & Attention Mechanism
Graded Assignment • 50 min

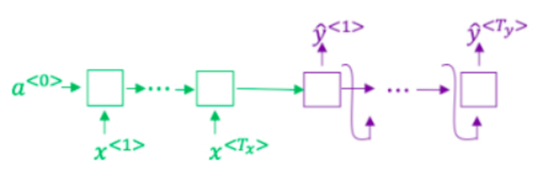
English Due Nov 10, 11:59 PM IST

Your grade: 100%

Your latest: **100%** • Your highest: **100%** • 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.

- ○ True
- ◉ False

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 converge after fewer steps.
- ☑ Beam search will run more slowly.

  ✓ **Correct**

- ☑ Beam search will use up more memory.

  ✓ **Correct**

- ☑ 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

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 $P(y \mid x)$. **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 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.
- ◉ Yes, because

  $$P(y^* \mid x) > P(\hat{y} \mid x)$$

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.

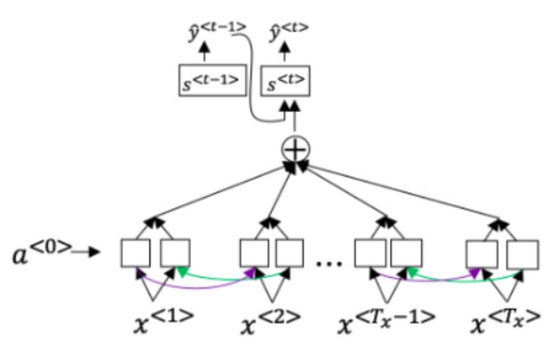
5. Continuing the example from Q4, suppose you work on your algorithm for a few more weeks, and now find that for the vast majority of examples on which your algorithm makes a mistake, $P(y^* \mid x) > P(\hat{y} \mid x)$. This suggests you should focus your attention on improving the RNN. **1 / 1 point**

- ○ True
- ◉ False

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.

- ☑ $\sum_{t'} \alpha^{<t,t'>} = 1$ (**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.**)

  ✓ **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. $\alpha^{<t,t'>}$ should be larger. Note the difference between a (activation) and α (attention coefficient).

- ☐ We expect

  $$\alpha^{<t,t'>}$$

  to be generally larger for values of

  $$a^{<t>}$$

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**

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

- ☑ $e^{<t,t'>}$

  ✓ **Correct**

  $$s^{<t>}$$

  depends on

  $$\alpha^{<t,t'>}$$

  which in turn depends on

  $$e^{<t,t'>}$$

  .

- ☐ $$s^{<t+1>}$$

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/AMS-Regular.js

Expand

Correct
Great, you got all the right answers.

8. The attention model performs the same as the encoder-decoder model, no matter the sentence length. **1 / 1 point**

- ◉ False
- ○ True

Expand

Correct
The performance of the encoder-decoder model declines as the amount of words increases. The attention model has the greatest advantage when 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

- ○ cokbok
- ◉ cookbook
- ○ coookkbooooooookkk
- ○ cook book

Expand

Correct

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

- ○ Only one word has been stated.
- ○ There is exactly one trigger word.
- ○ The total time that the trigger word detection algorithm has been running is 1.
- ◉ 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.