Sequence Models

Week 3: Sequence models and Attention Mechanism

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

Answer: False.

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.

Answer: Beam search will use up more memory; Beam search will run more slowly; Beam search will generally find better solutions (i.e. do a better job maximizing $P(y|x)$).

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

Answer: True.

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 $P(y|x)$. On a dev set example, given an input audio clip, your algorithm outputs the transcript $\hat{y}$ = I am building an A Eye system in Silly con Valley, whereas a human gives a much superior transcript y^* = I am building an AI system in Silicon Valley. According to your model, $P(y^{|}x) = 1.09 \times 10^{-7}$, $P(y^*|x) = 7.21 \times 10^{-8}$. Would you expect increasing the beam width B to help correct this example?

Answer: Yes, because $P(y^*|x) > P(\hat{y}|x)$ indicates the error should be attributed to the search algorithm rather than to the RNN.

Comment: $P(y^*|x) > P(\hat{y}|x)$ 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.

Answer: True.

6 Consider the attention model for machine translation. Which of the following statements about $\alpha^{<t,t^{'}>}$ are true? Check all that apply.

Answer: $\sum_{t'} \alpha^{<t,t^{'}>} = 1$ Note the summation is over $t^{'}$; W 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.

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

Answer: False.

Comment: We cannot 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 The attention model performs the same as the encoder-decoder model, no matter the sentence length.

Answer: False.

Comment: 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?

Answer: aardvark.

Comment: letter separated by - only repeat once. 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:

Answer: Features of the audio (such as spectrogram features) at time t.