Sequence Models

Week 1: Recurrent Neural Networks

1 Version 1: Suppose your training examples are sentences (sequences of words). Which of the following refers to the l^{th} word in the k^{th} training example?

Answer: $x^{(k)<l>}$.

Comment: We index into the k^{th} row first to get to the k^{th} training example (represented by parentheses), then the l^{th} column to get to the l^{th} word (represented by the brackets).

1 Version 2: Suppose your training examples are sentences (sequences of words). Which of the following refers to the jth word in the ith training example?

Answer: $x^{(i)}$.

2 Version 1: Consider this RNN: True/False: This specific type of architecture is appropriate when Tx = Ty.

Answer: True.

Comment: It is appropriate when the input sequence and the output sequence have the same length or size.

2 Version 2: Consider this RNN: This specific type of architecture is appropriate when:

Answer: Tx = Ty.

3 To which of these tasks would you apply a many-to-one RNN architecture? (Check all that apply).

Answer: Sentiment classification (input a piece of text and output a 0/1 to denote positive or negative sentiment); Gender recognition from speech (input an audio clip and output a label indicating the speaker's gender)

4 Using this as the training model below, answer the following: True/False: At the l^{th} time step the RNN is estimating $P(y^{<t>}|y^{<1>}, \ldots, y^{<t-1>})$

Answer: True.

Comment: In a training model we try to predict the next step based on knowledge of all prior steps.

5 Version 1: You have finished training a language model RNN and are using it to sample random sentences, as follows: True/False: In this sample sentence, step t uses the probabilities output by the RNN to pick the highest probability word for that time-step. Then it passes the ground-truth word from the training set to the next time-step.

Answer: False.

Comment: The probabilities output by the RNN are not used to pick the highest probability word and the ground-truth word from the training set is not the input to the next time-step.

6 Version 2: You have finished training a language model RNN and are using it to sample random sentences, as follows: What are you doing at each time step t?

Answer: (i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as y. (ii) Then pass this selected word to the next time-step.

6 You are training an RNN model, and find that your weights and activations are all taking on the value of NaN ("Not a Number"). Which of these is the most likely cause of this problem?

Answer: Exploding gradient problem.

7 Suppose you are training an LSTM. You have a 10000 word vocabulary, and are using an LSTM with 100-dimensional activations $a^{<t>}$. What is the dimension of Γ_u at each time step?

Answer: 100.

Comment: Γ_u is a vector of dimension equal to the number of hidden unites in the LSTM.

8 True/False: In order to simplify the GRU without vanishing gradient problems even when training on very long sequences you should remove the Γ_r i.e., setting $\Gamma_r = 1$ always.

Answer: True.

Comments: If $\Gamma_u \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay. For the signal to backpropagate without vanishing, we need $c^{<t>}$ to be highly dependent on $c^{<t-1>}$.

9 True/False: Using the equations for the GRU and LSTM below the Update Gate and Forget Gate in the LSTM play a different role to Γ_u and $1 - \Gamma_u$.

Answer: True.

Comment: Instead of using Γ_u to compute $1-\Gamma_u$, LSTM uses 2 gates (Γ_u and Γ_f) to compute the final value of the hidden state. So, Γ_f is used instead of $1 - \Gamma_u$.

10 Version 1: True/False: You would use unidirectional RNN if you were building a model map to show how your mood is heavily dependent on the current and past few days' weather.

Answer: True.

Comment: Your mood is contingent on the current and past few days' weather, not on the current, past, AND future days' weather.

10 Version 2: You have a pet dog whose mood is heavily dependent on the current and past few days' weather. You've collected data for the past 365 days on the weather, which you represent as a sequence as $x^{<1>}, \ldots, x^{<365>}$. You've also collected data on your dog's mood, which you represent as $y^{<1>}, \ldots, y^{<365>}$. You'd like to build a model to map from x to y. Should you use a Unidirectional RNN or Bidirectional RNN for this problem?

Answer: Unidirectional RNN, because the value of y depends only on $x^{<1>}, \ldots X$ but not on $x^{<t+1>}, \ldots, x^{<365>}$.