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?

0 / 1 point

- $x^{<k>(l)}$
- $x^{(l)<k>}$
- $x^{(k)<l>}$
- $x^{<l>(k)}$

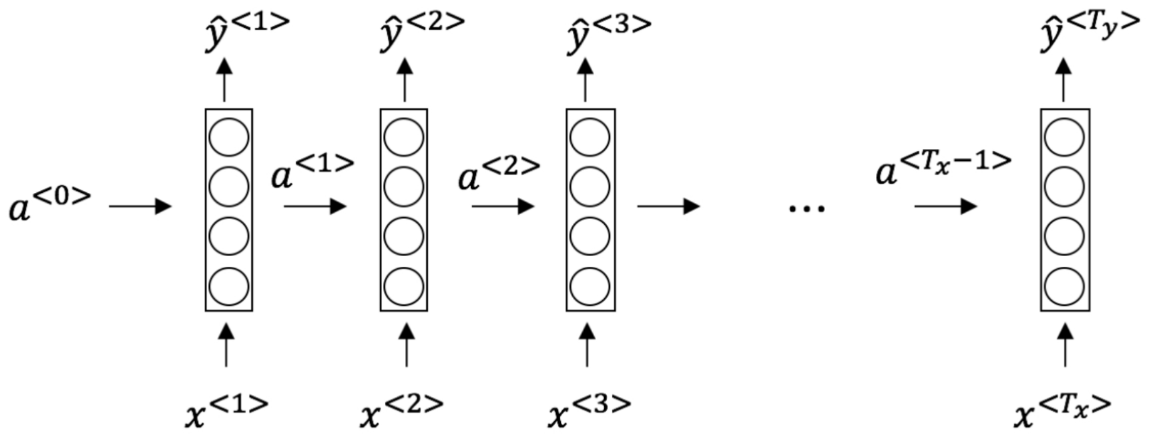
Expand

Incorrect

The parentheses represent the training example and the brackets represent the word. You should choose the training example and then the word.

2. Consider this RNN:

1 / 1 point



True/False: This specific type of architecture is appropriate when Tx=Ty

- False
- True

Expand

Correct

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

3. Select the two tasks combination that could be addressed by a many-to-one RNN model architecture from the following:

0 / 1 point

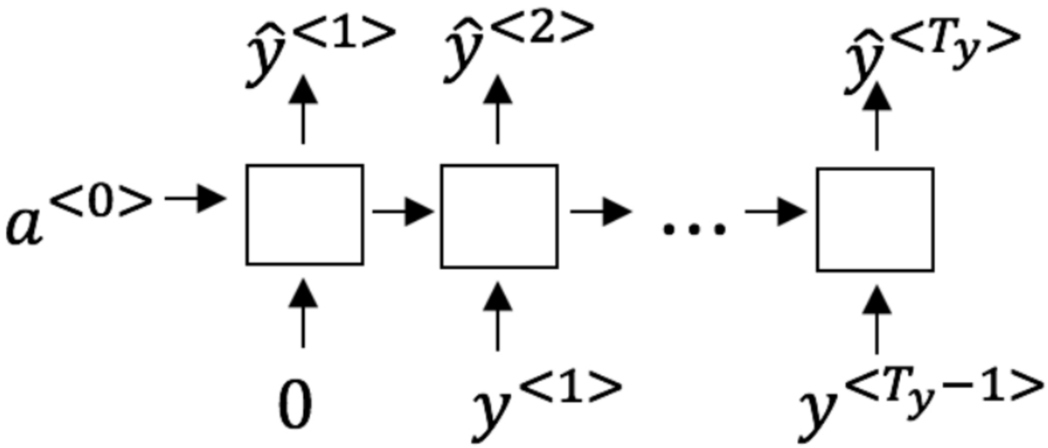
- **Task 1:** Gender recognition from audio. **Task 2:** Image classification.
- **Task 1:** Speech recognition. **Task 2:** Gender recognition from audio.
- **Task 1:** Image classification. **Task 2:** Sentiment classification.
- **Task 1:** Gender recognition from audio. **Task 2:** Movie review (positive/negative) classification.

Expand

Incorrect
Speech recognition is an example of many-to-many recognition.

4. Using this as the training model below, answer the following:

1 / 1 point



True/False: At the t^{th} time step the RNN is estimating $P(y^{<t>} \mid y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$

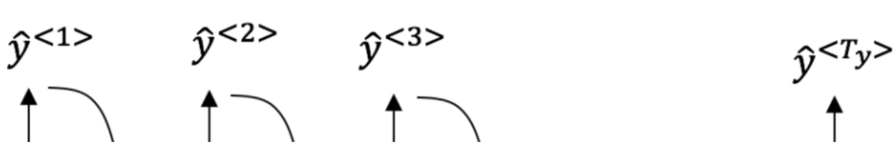
- True
- False

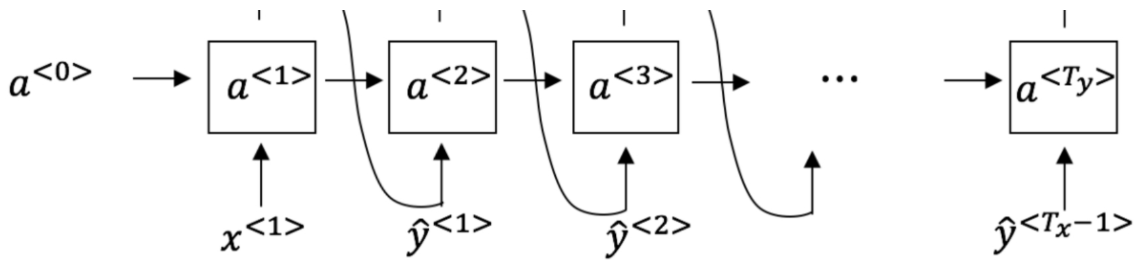
Expand

Correct
Yes, in a training model we try to predict the next step based on knowledge of all prior steps.

5. You have finished training a language model RNN and are using it to sample random sentences, as follows:

1 / 1 point





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.

- True
- False

Expand

Correct
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. 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?

1 / 1 point

- Vanishing gradient problem.
- Exploding gradient problem.
- The model used the ReLU activation function to compute g(z), where z is too large.
- The model used the Sigmoid activation function to compute g(z), where z is too large.

Expand

Correct

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?

1 / 1 point

- 1
- 100
- 300
- 10000

Expand

Correct

Correct, $$\Gamma_u$$ is a vector of dimension equal to the number of hidden units in the LSTM.

8. Here are the update equations for the GRU. **1 / 1 point**

GRU

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[\, c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[\, c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

Alice proposes to simplify the GRU by always removing the Γ_u. I.e., setting Γ_u = 0. Betty proposes to simplify the GRU by removing the Γ_r. I. e., setting Γ_r = 1 always. Which of these models is more likely to work without vanishing gradient problems even when trained on very long input sequences?

- ○ Alice's model (removing Γ_u), because if $\Gamma_r \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ○ Alice's model (removing Γ_u), because if $\Gamma_r \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ◉ Betty's model (removing Γ_r), because if $\Gamma_u \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ○ Betty's model (removing Γ_r), because if $\Gamma_u \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay.

Expand

Correct

Yes. 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 role similar to 1- Γu and Γu. **1 / 1 point**

GRU

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

LSTM

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[\,c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[\,c^{<t-1>}, x^{<t>}] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$$a^{<t>} = c^{<t>}$$

$$\Gamma_u = \sigma(W_u[\,a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[\,a^{<t-1>}, x^{<t>}] + b_f)$$

$$\Gamma_o = \sigma(W_o[\,a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * c^{<t>}$$

- ○ True
- ◉ False

Expand

Correct
Instead of using Γu to compute 1 - Γu, LSTM uses 2 gates (Γu and Γf) to compute the final value of the hidden state. So, Γf is used instead of 1 - Γu.

10. Your 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 mood, which you represent as $y^{<1>}, \ldots, y^{<365>}$. You'd like to build a model to map from *x*→*y*. Should you use a Unidirectional RNN or Bidirectional RNN for this problem?

1 / 1 point

- ◉ Unidirectional RNN, because the value of $y^{<t>}$ depends only on $x^{<1>}, \ldots, x^{<t>}$, but not on $x^{<1>}, \ldots, x^{<365>}$.
- ○ Bidirectional RNN, because this allows the prediction of mood on day t to take into account more information.
- ○ Bidirectional RNN, because this allows backpropagation to compute more accurate gradients.
- ○ Unidirectional RNN, because the value of $y^{<t>}$ depends only on $x^{<t>}$, and not other days' weather.

Expand

Correct