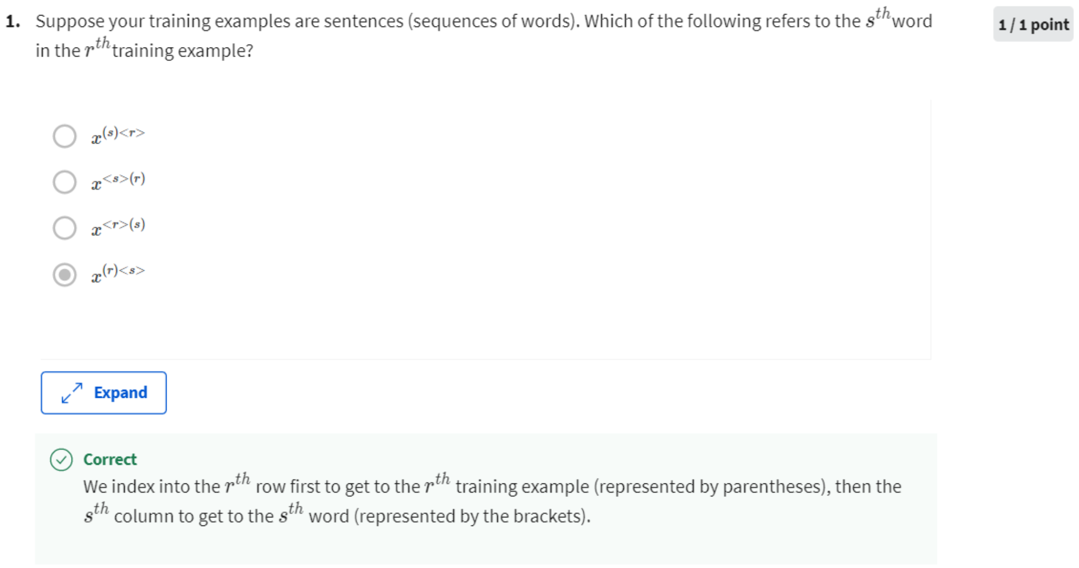
Recurrent Neural Networks



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

1 / 1 point

○ $x^{(s)<r>}$

○ $x^{<s>(r)}$

○ $x^{<r>(s)}$

● $x^{(r)<s>}$

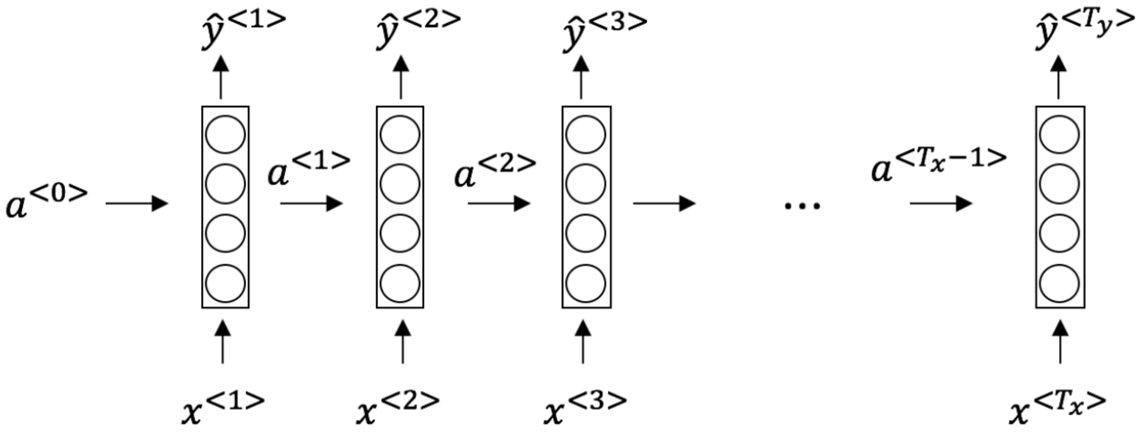
Expand

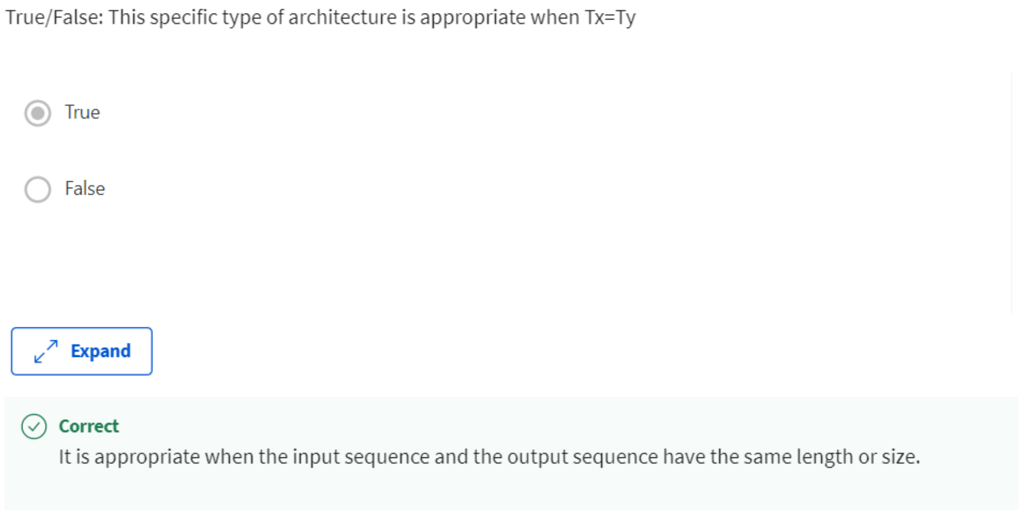
Correct

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

2.
Question 2

Consider this RNN:





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

● True

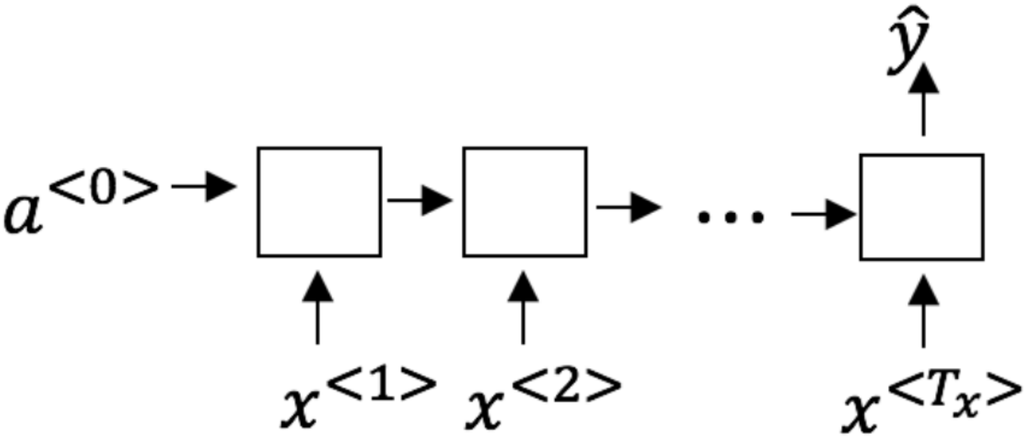
○ False

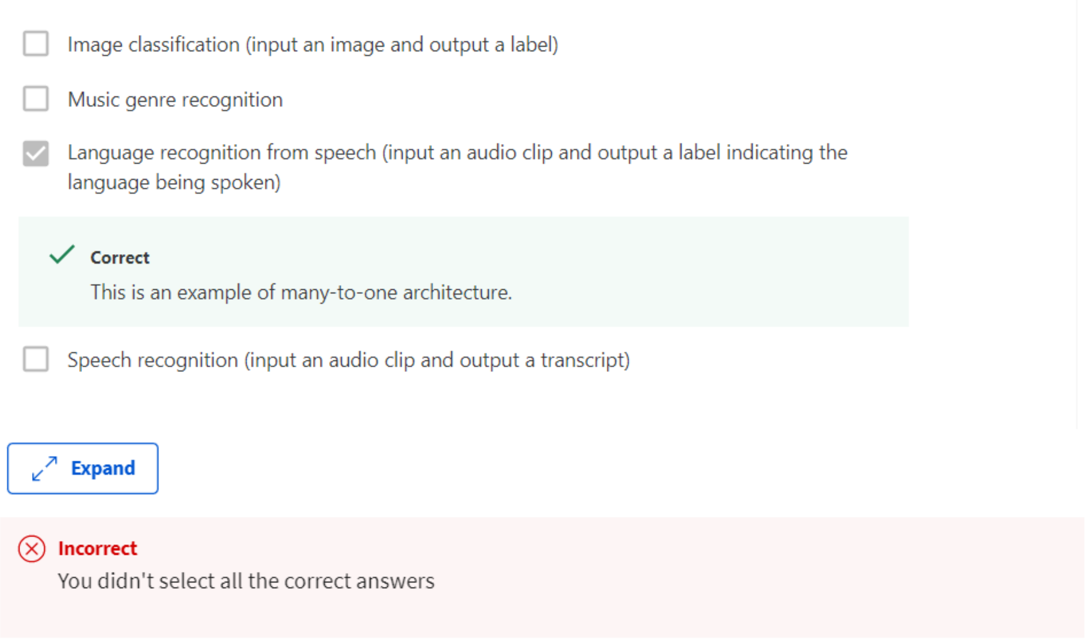
✓ Correct

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

3.
Question 3

To which of these tasks would you apply a many-to-one RNN architecture?





□ Image classification (input an image and output a label)

□ Music genre recognition

☑ Language recognition from speech (input an audio clip and output a label indicating the language being spoken)

✓ **Correct**

This is an example of many-to-one architecture.

□ Speech recognition (input an audio clip and output a transcript)

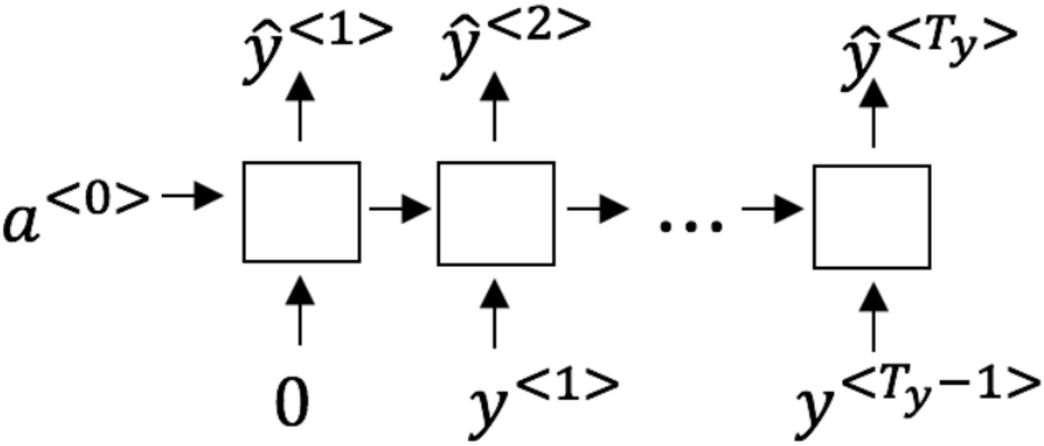
ⓧ **Incorrect**

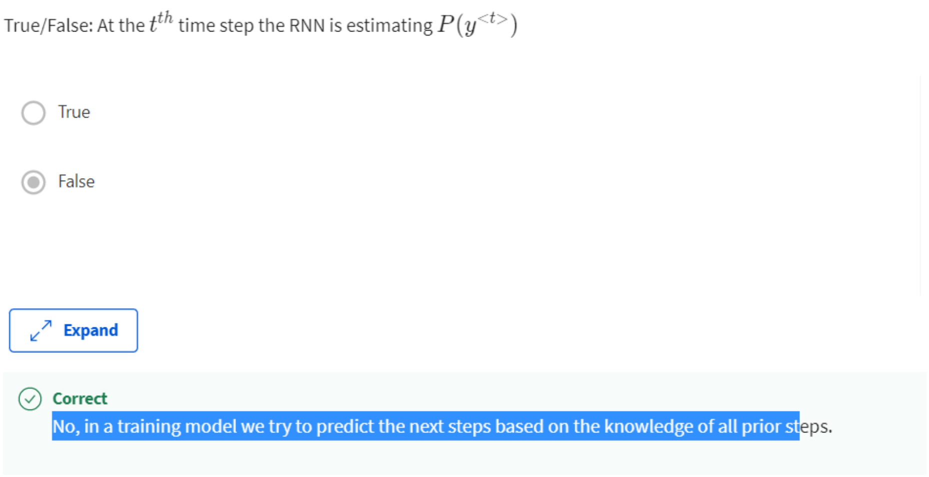
You didn't select all the correct answers

4.

Question 4

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





True/False: At the t^{th} time step the RNN is estimating $P(y^{<t>})$

○ True

● False

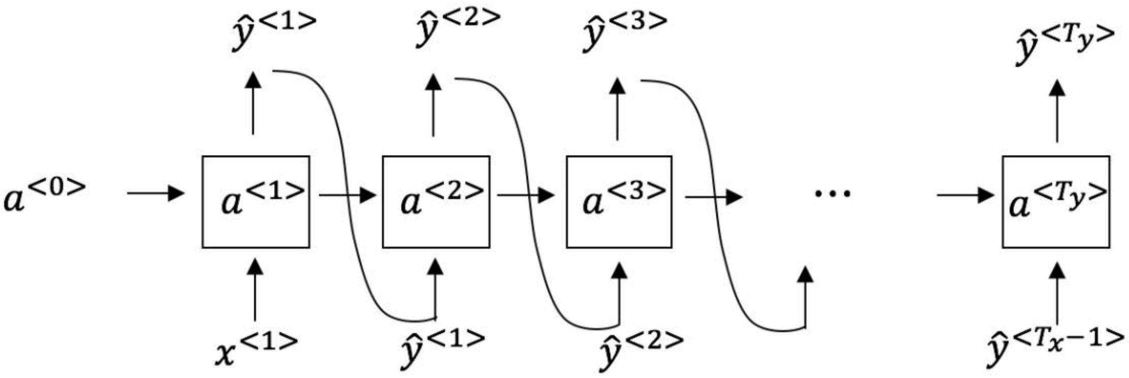
Expand

✓ Correct

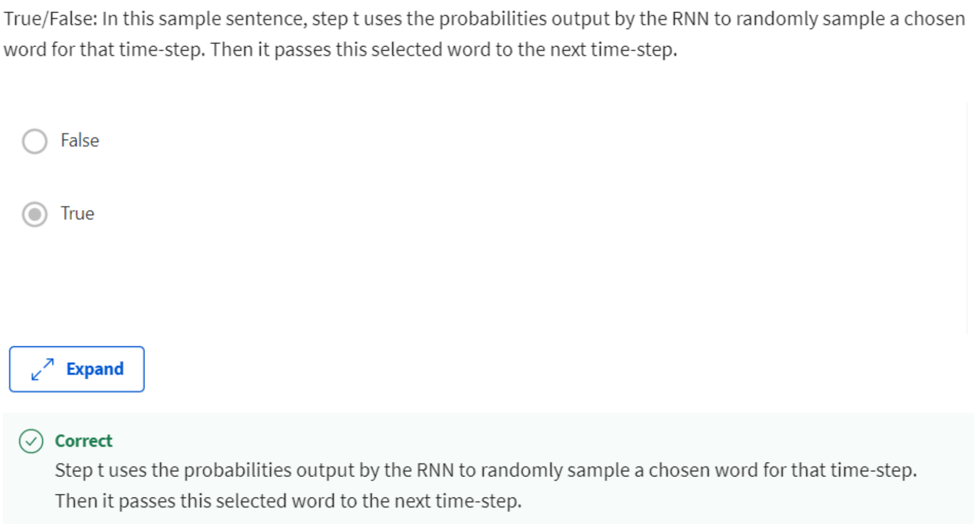
No, in a training model we try to predict the next steps based on the knowledge of all prior steps.

5.
Question 5

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 randomly sample a chosen word for that time-step. Then it passes this selected word to the next time-step.



True/False: In this sample sentence, step t uses the probabilities output by the RNN to randomly sample a chosen word for that time-step. Then it passes this selected word to the next time-step.

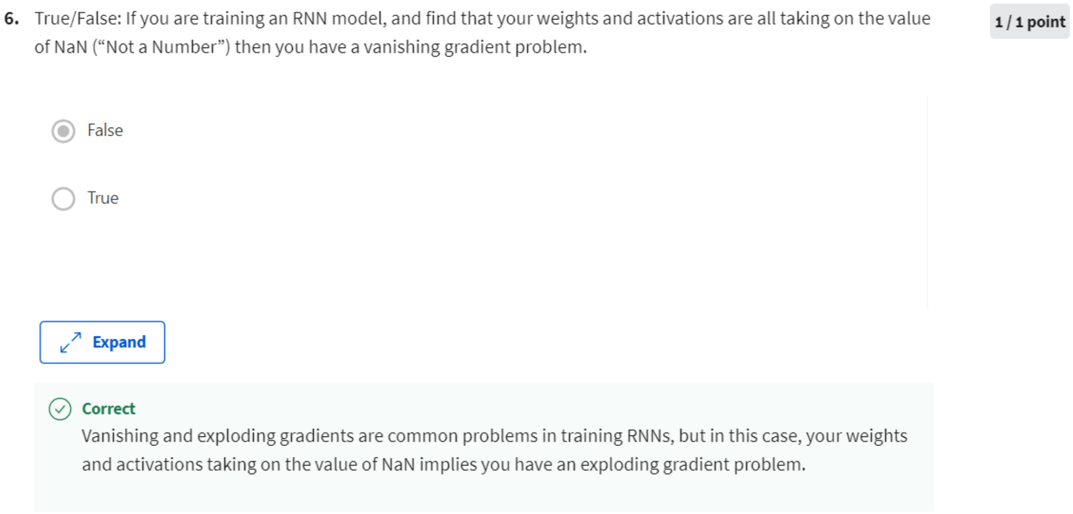
○ False

● True

Expand

✓ Correct

Step t uses the probabilities output by the RNN to randomly sample a chosen word for that time-step. Then it passes this selected word to the next time-step.

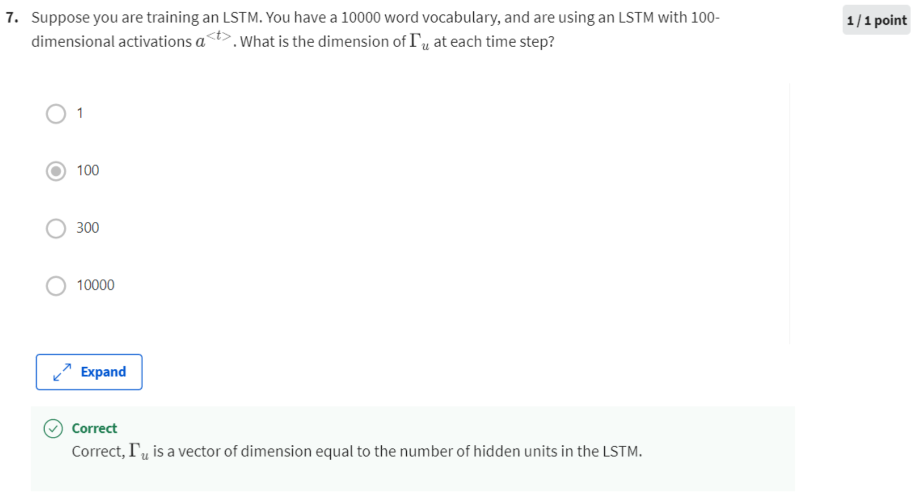


6. True/False: If you are training an RNN model, and find that your weights and activations are all taking on the value of NaN ("Not a Number") then you have a vanishing gradient problem.

● False

○ True

✓ Correct
Vanishing and exploding gradients are common problems in training RNNs, but in this case, your weights and activations taking on the value of NaN implies you have an 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?

○ 1

● 100

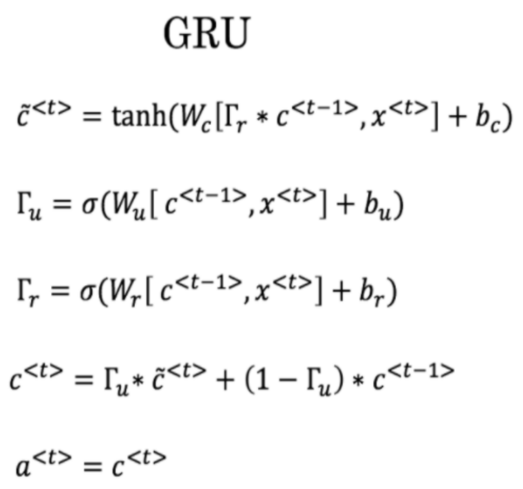
○ 300

○ 10000

✓ Correct
Correct, Γ_u is a vector of dimension equal to the number of hidden units in the LSTM.

8.
Question 8

Here are the update equations for the GRU.



GRU

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

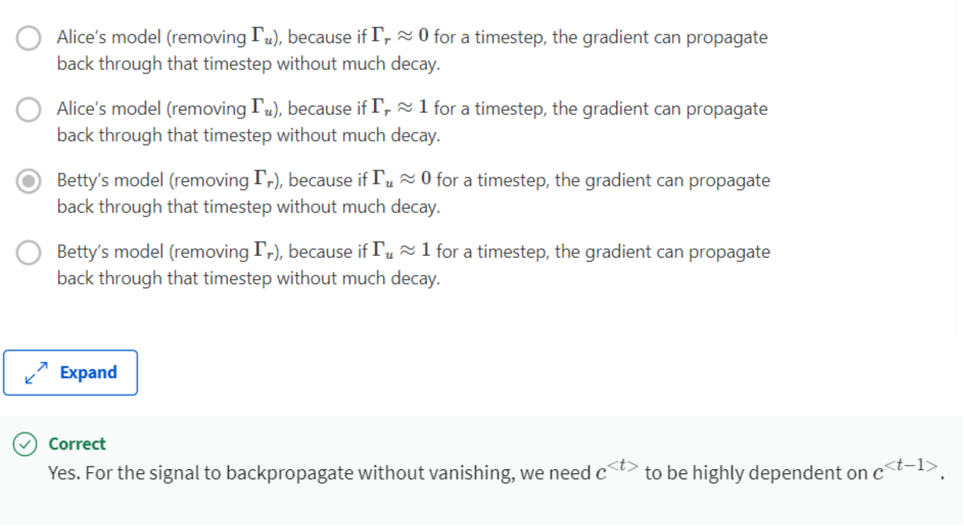
$$\Gamma_u = \sigma(W_u[\, c^{<t-1>}, x^{<t>}] + b_u)$$

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

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

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

Alice proposes to simplify the GRU by always removing the Γ�Γu. I.e., setting Γ�Γu = 0. Betty proposes to simplify the GRU by removing the Γ�Γr. I. e., setting Γ�Γr = 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.

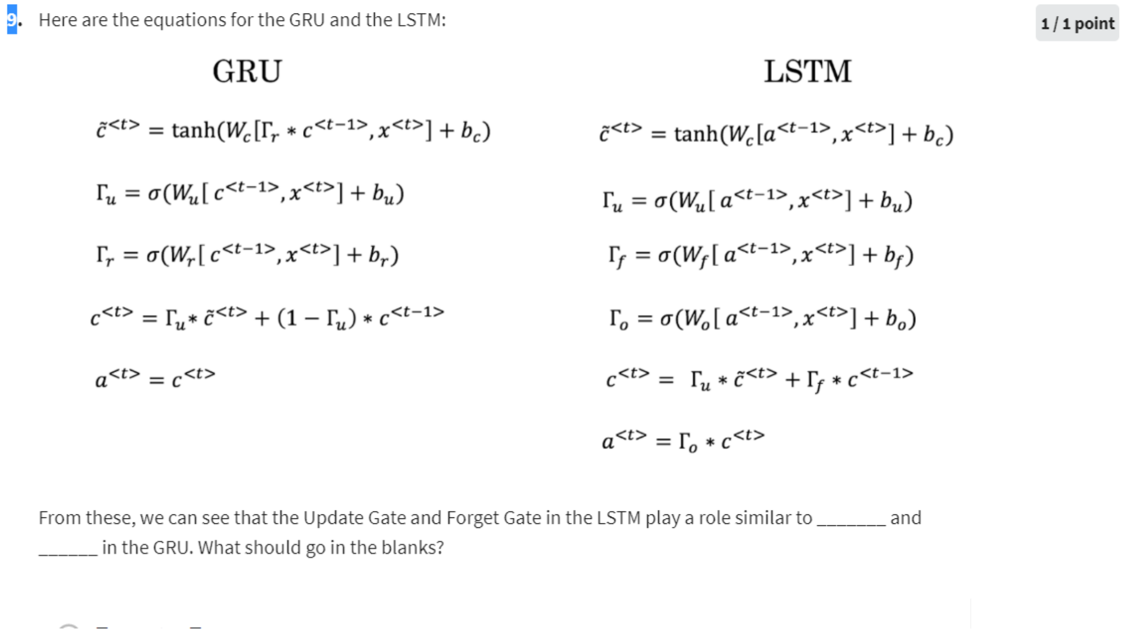
○ Alice's model (removing Γ_u), because if $\Gamma_r \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay.

● 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.

Correct

Yes. For the signal to backpropagate without vanishing, we need $c^{<t>}$ to be highly dependent on $c^{<t-1>}$.

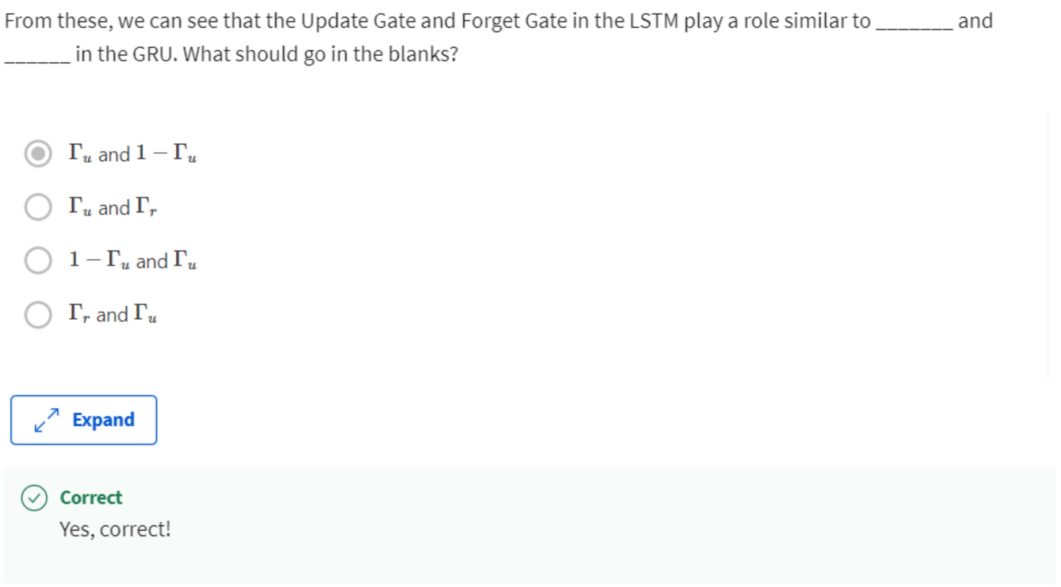


9. Here are the equations for the GRU and the LSTM:

1 / 1 point

GRU	LSTM
$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$	$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$
$\Gamma_u = \sigma(W_u[\, c^{<t-1>}, x^{<t>}] + b_u)$	$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$
$\Gamma_r = \sigma(W_r[\, c^{<t-1>}, x^{<t>}] + b_r)$	$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$
$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$	$\Gamma_o = \sigma(W_o[\, a^{<t-1>}, x^{<t>}] + b_o)$
$a^{<t>} = c^{<t>}$	$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$
	$a^{<t>} = \Gamma_o * c^{<t>}$

From these, we can see that the Update Gate and Forget Gate in the LSTM play a role similar to _______ and ______ in the GRU. What should go in the blanks?



From these, we can see that the Update Gate and Forget Gate in the LSTM play a role similar to _______ and ______ in the GRU. What should go in the blanks?

● Γ_u and $1 - \Gamma_u$

○ Γ_u and Γ_r

○ $1 - \Gamma_u$ and Γ_u

○ Γ_r and Γ_u

Expand

✓ Correct

Yes, correct!

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 $x \rightarrow y$. 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.

○ Unidirectional RNN, because the value of $y^{<t>}$ depends only on $x^{<t>}$, and not other days' weather.

○ Bidirectional RNN, because this allows backpropagation to compute more accurate gradients.

Expand

✓ **Correct**