## Congratulations! You passed!

**Grade received** 90% **Latest Submission Grade** 90% **To pass** 80% or higher

Go to next item

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

**1 / 1 point**

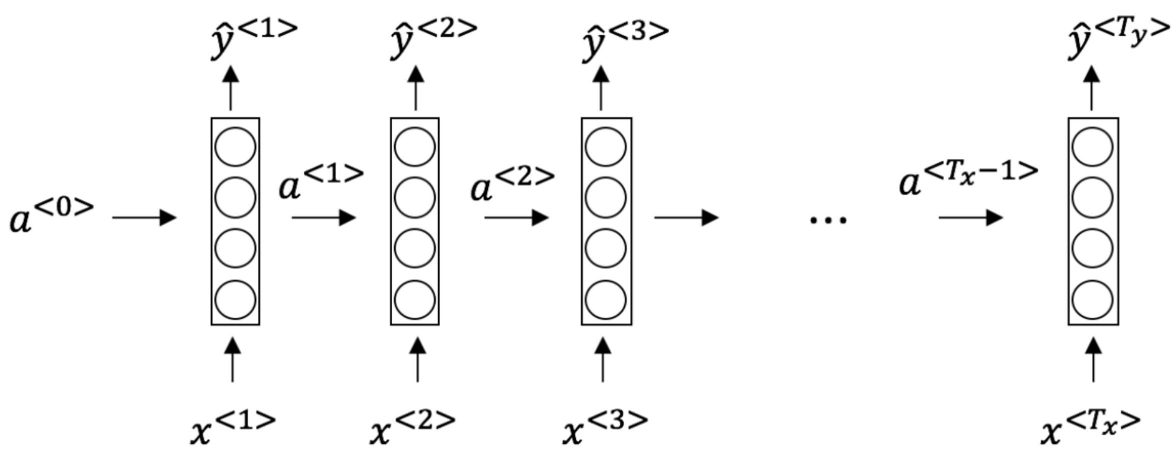
- (•) $x^{(i)<j>}$
- ( ) $x^{<i>(j)}$
- ( ) $x^{(j)<i>}$
- ( ) $x^{<j>(i)}$

Expand

**Correct**
We index into the $$i^{th}$$ row first to get the $$i^{th}$$ training example (represented by parentheses), then the $$j^{th}$$ column to get the $$j^{th}$$ word (represented by the brackets).

**2.** Consider this RNN:

**1 / 1 point**




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

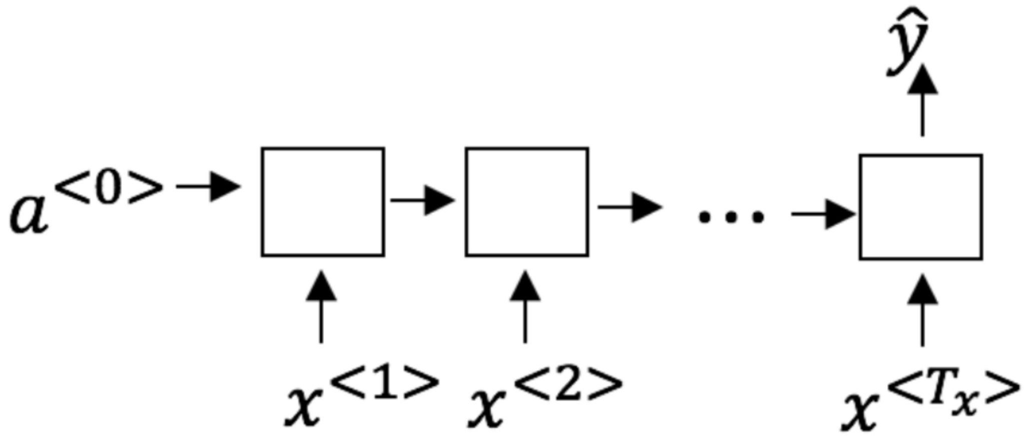
- (•) True
- ( ) False

Expand

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

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

**0 / 1 point**




- [ ] Image classification (input an image and output a label)
- [ ] Music genre recognition
- [x] 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.

- [x] Speech recognition (input an audio clip and output a transcript)

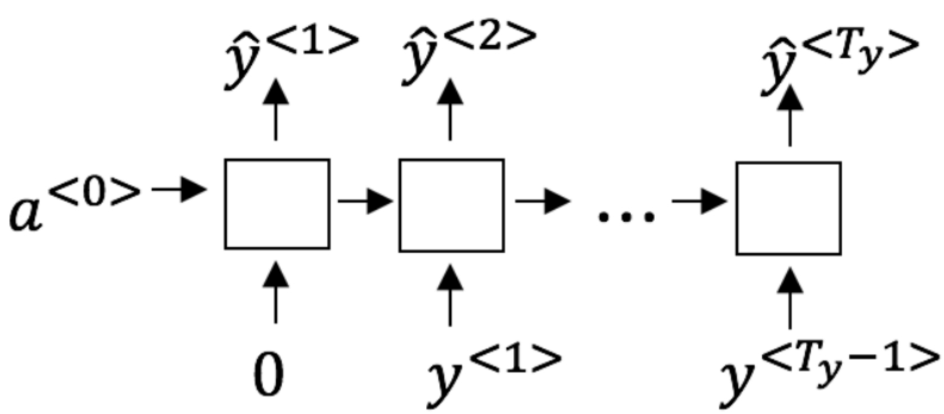
! **This should not be selected**
This is an example of many-to-many architecture.

Expand

**Incorrect**
You didn't select all the correct answers

**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 $P(y^{<t>})$

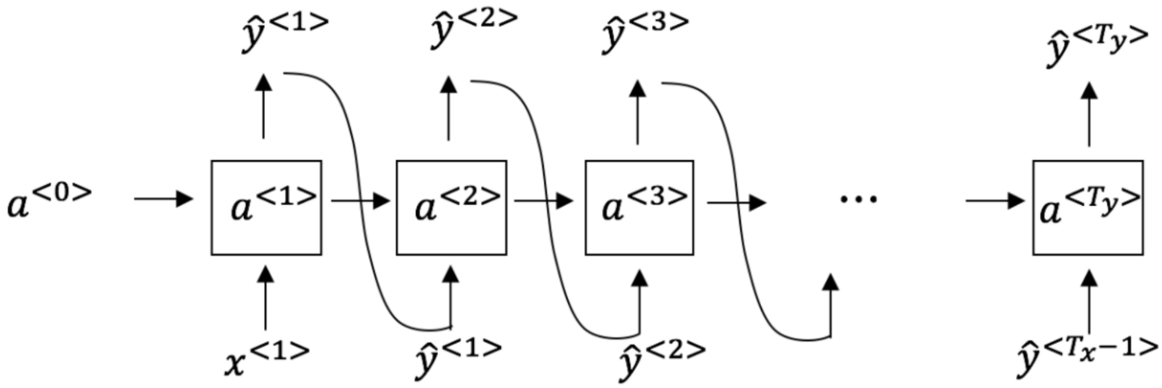
- ( ) True
- (•) False

Expand

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

**5.** 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.

- (●) False
- ( ) True

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.** 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 an exploding gradient problem.

1 / 1 point

- (●) True
- ( ) False

Expand

Correct
Correct! Exploding gradients happen when large error gradients accumulate and result in very large updates to the NN model weights during training. These weights can become too large and cause an overflow, identified as NaN.

**7.** Suppose you are training an LSTM. You have a 50000 word vocabulary, and are using an LSTM with 500-dimensional activations $a^{<t>}$. What is the dimension of $\Gamma_u$ at each time step?

1 / 1 point

- ( ) 50000

- ◉ 500
- ○ 200
- ○ 5

Expand

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

**8.** Sarah proposes to simplify the GRU by always removing the Γu. I.e., setting Γu = 0. Ashely proposes to simplify the GRU by removing the Γr. I. e., setting Γr= 1 always. Which of these models is more likely to work without vanishing gradient problems even when trained on very long input sequences?

**1 / 1 point**

GRU

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

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

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

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

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

- ○ Ashely's model (removing $\Gamma_r$), because if $\Gamma_u \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ○ Sarah's model (removing $\Gamma_u$), because if $\Gamma_r \approx 1$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ◉ Ashely's model (removing $\Gamma_r$), because if $\Gamma_u \approx 0$ for a timestep, the gradient can propagate back through that timestep without much decay.
- ○ Sarah's model (removing $\Gamma_u$), because if $\Gamma_r \approx 0$ 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.** Here are the equations for the GRU and the LSTM:

**1 / 1 point**

GRU

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

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

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

LSTM

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

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

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

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

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

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

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

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

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?

- (selected) $\Gamma_u$ and $1 - \Gamma_u$
- $\Gamma_u$ and $\Gamma_r$
- $1 - \Gamma_u$ and $\Gamma_u$
- $\Gamma_r$ and $\Gamma_u$

Expand

Correct
Yes, correct!

**10.** 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.

1 / 1 point

- (selected) True
- False

Expand

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