Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

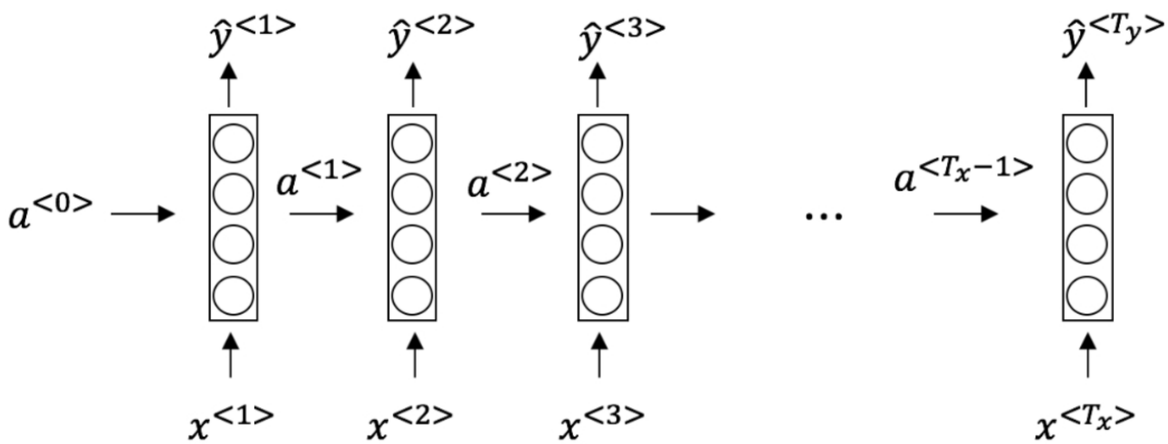
Next item →

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? **1 / 1 point**

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

Correct
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).

2. Consider this RNN: **1 / 1 point**

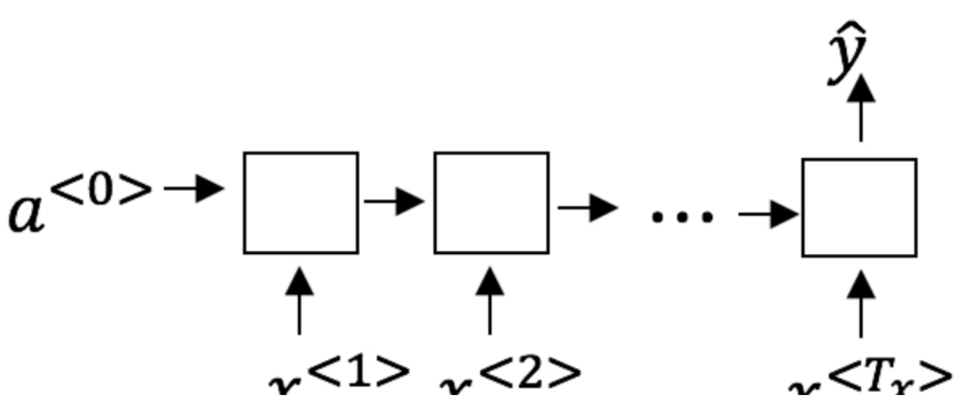


This specific type of architecture is appropriate when:

- ● $T_x = T_y$
- ○ $T_x < T_y$
- ○ $T_x > T_y$
- ○ $T_x = 1$

Correct
It is appropriate when every input should have an output.

3. To which of these tasks would you apply a many-to-one RNN architecture? (Check all that apply). **1 / 1 point**



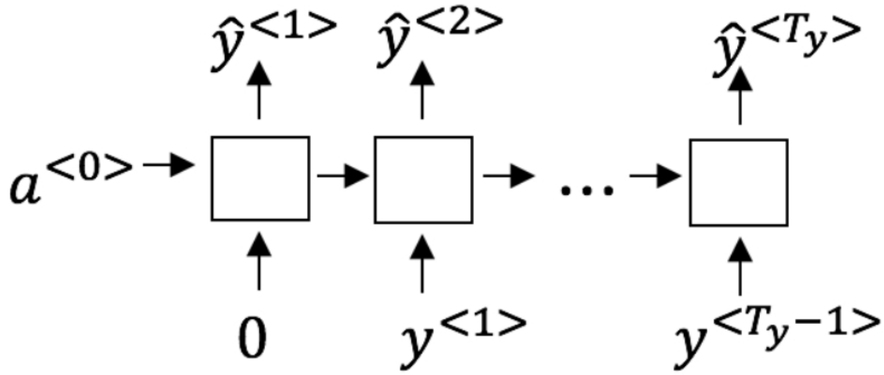
- [ ] Speech recognition (input an audio clip and output a transcript)
- [x] Sentiment classification (input a piece of text and output a 0/1 to denote positive or negative sentiment)

> **Correct**
> Correct!

- [ ] Image classification (input an image and output a label)
- [x] Gender recognition from speech (input an audio clip and output a label indicating the speaker's gender)

> **Correct**
> Correct!

4. Using this as the training model below, answer the following: **1 / 1 point**

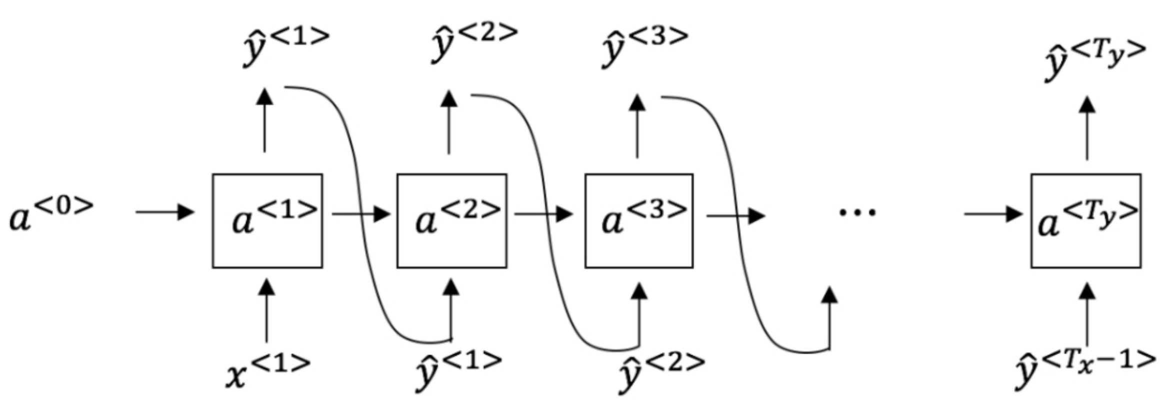


True/False: At the t^{th} time step the RNN is estimating $P(y^{<t>} \mid y^{<1>}, y^{<2>}, \ldots, y^{<t-1>})$

- ( ) False
- (•) True

> **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
- [x] False

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

- [x] True
- [ ] False

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 an 80000 word vocabulary, and are using an LSTM with 800-dimensional activations $a^{<t>}$. What is the dimension of Γ_u at each time step?

1 / 1 point

- [ ] 100
- [ ] 80000
- [x] 800
- [ ] 8

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

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?

- (●) Γ_u and $1 - \Gamma_u$
- ( ) Γ_u and Γ_r
- ( ) $1 - \Gamma_u$ and Γ_u
- ( ) Γ_r and Γ_u

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**

- (●) True
- ( ) False

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