EAI6080 Week 5 Practice (OPTIONAL for Extra Credits)

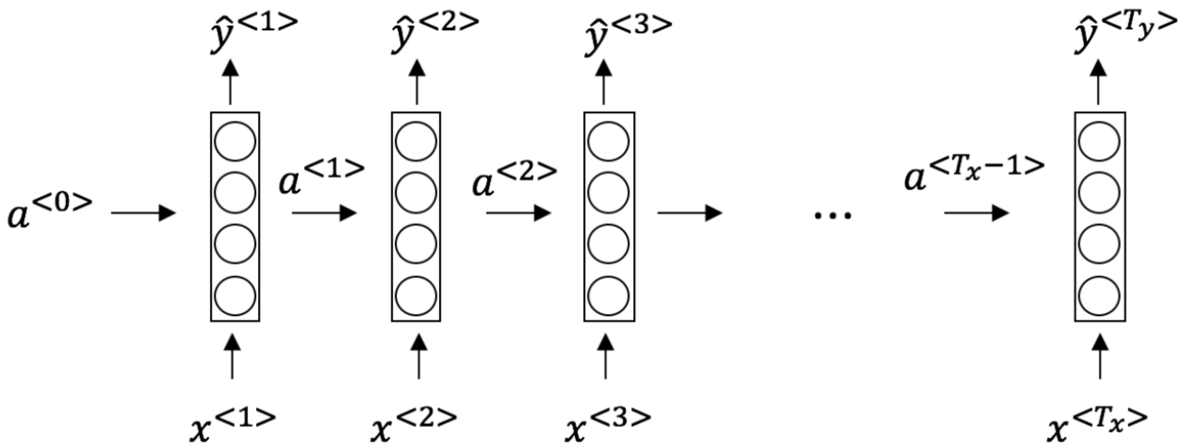
NAME:

Date:

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?

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

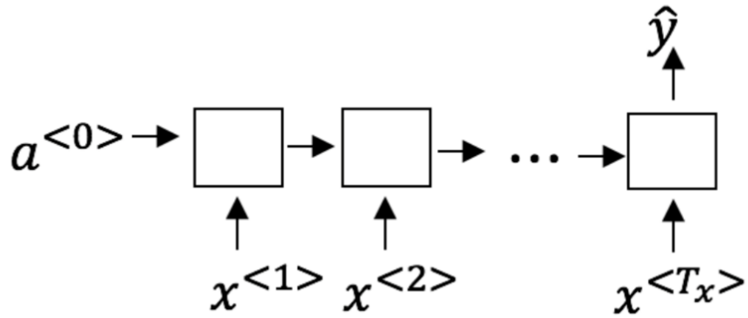
2. Consider this RNN:



This specific type of architecture is appropriate when:

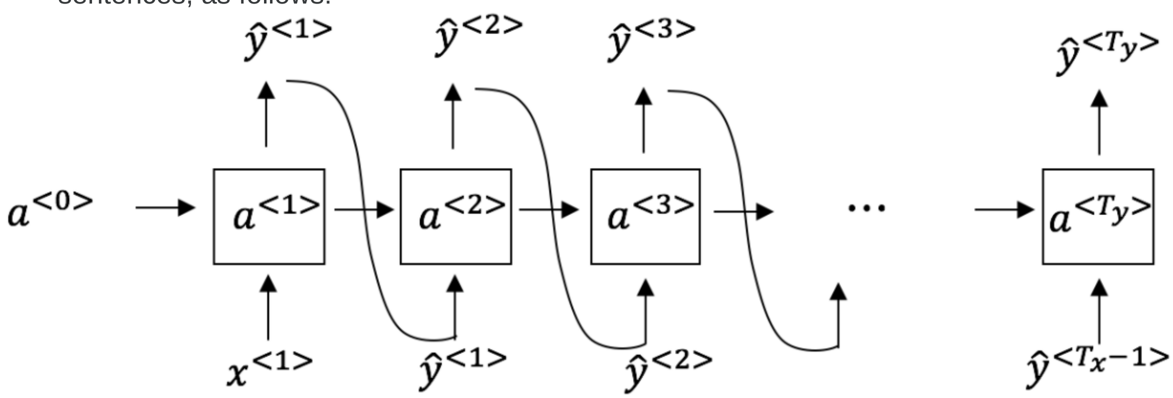
- ◉ $T_x = T_y$
- ○ $T_x < T_y$
- ○ $T_x > T_y$
- ○ $T_x = 1$

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



- [ ] 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)
- [ ] 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)

4. 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 tt?

○ (i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as y^$<t>$. (ii) Then pass the ground-truth word from the training set to the next time-step.

○ (i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as y^$<t>$. (ii) Then pass the ground-truth word from the training set to the next time-step.

○ (i) Use the probabilities output by the RNN to pick the highest probability word for that time-step as y^$<t>$. (ii) Then pass this selected word to the next time-step.

○ **(i) Use the probabilities output by the RNN to randomly sample a chosen word for that time-step as y^$<t>$. (ii) Then pass this selected word to the next time-step. TRUE**

5. You are training an RNN, 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?

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

6. Here're 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_u. I.e., setting $\Gamma_u = 1$. Betty proposes to simplify the GRU by removing the Γ_r. I. e., setting $\Gamma_r = 1$ 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. TRUE**
- 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.

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

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?

- (•) **Γ_u and 1-1−Γ_u** **TRUE**
- ( ) Γ_u and Γ_r
- ( ) 1-Γ_u and Γ_u
- ( ) Γ_r and Γ_u