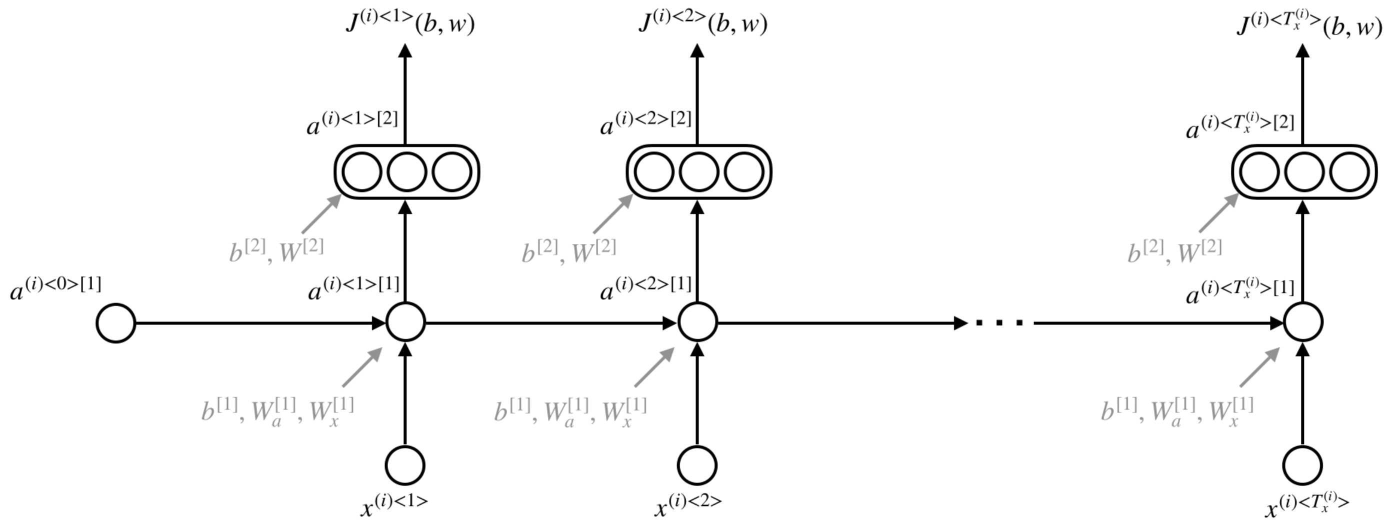
HW5 Written Part

Due 5pm Tuesday April 28, 2020

What is your name?

Problem 1: RNN Activation Functions

Consider a RNN with a single hidden recurrent layer as in the diagram below.



(a) For simplicity, suppose that there is only one unit in the recurrent layer, that all inputs $x^{(i)<1>}, \dots, x^{(i)<T>}$ are 0, that $b^{[1]} = 0$, and that $W_a^{[1]} = [1]$. For parts i and ii below, suppose a sigmoid activation is used for the recurrent layer:

$$a^{(i)<t>} = \frac{\exp\left(z^{(i)<t>}\right)}{1 + \exp\left(z^{(i)<t>}\right)} \text{ where}$$
$$z^{(i)<t>} = b^{[1]} + W_a^{[1]} a^{(i)<t-1>} + W_x^{[1]} x^{(i)<t>}$$

(continued on next page)

i. Find $\frac{\partial a^{(i)<t>}}{\partial a^{(i)<t-1>}}$. If you prefer, you can do this in two steps, first finding $\frac{\partial a^{(i)<t>}}{\partial z^{(i)<t>}}$.

ii. Recall that the recurrent layer is initialized with $a^{(i)<0>} = 0$. For long sequences (imagine $T \to \infty$, though you don't need to take a formal limit), will this network tend to suffer from vanishing gradients, exploding gradients, or neither? You can justify your answer in a sentence or two.

(b) For simplicity, suppose that there is only one unit in the recurrent layer, that all inputs $x^{(i)<1>}, \ldots, x^{(i)<T>}$ are 0, that $b^{[1]} = 0$, and that $W_a^{[1]} = [1]$. For parts i and ii below, suppose a sigmoid activation is used for the recurrent layer:

$$\begin{aligned} a^{(i)<t>} &= \frac{\exp\left(2z^{(i)<t>}\right) - 1}{\exp\left(2z^{(i)<t>}\right) + 1} \text{ where} \\ z^{(i)<t>} &= b^{[1]} + W_a^{[1]} a^{(i)<t-1>} + W_x^{[1]} x^{(i)<t>} \end{aligned}$$

i. Find $\frac{\partial a^{(i)<t>}}{\partial a^{(i)<t-1>}}$. If you prefer, you can do this in two steps, first finding $\frac{\partial a^{(i)<t>}}{\partial z^{(i)<t>}}$.

ii. Recall that the recurrent layer is initialized with $a^{(i)<0>} = 0$. For long sequences (imagine $T \to \infty$, though you don't need to take a formal limit), will this network tend to suffer from vanishing gradients, exploding gradients, or neither? You can justify your answer in a sentence or two.

Problem 2: CNN dimensions

Consider a CNN architecture with the following specification:

- Input layer: an image of shape $128 \times 128 \times 3$
- Convolutional layer: 10 filters each of shape $3 \times 3 \times 3$, padding 0, stride 1
- Max pooling layer: 2×2 window, padding 0, stride 2

(a) How many parameters are there in each layer?

- Convolutional layer:
- Max pooling layer:

(b) What are the dimensions (shape) of the activation outputs for each layer?

- Convolutional layer:
- Max pooling layer:

(c) Suppose you wanted to use "same" padding so that the convolutional layer's output volume was the same as the input layer volume. How much padding would you use?

(d) What are the width and height of the effective receptive field of one unit in the output from the max pooling layer?

Problem 3: 1D Convolutions

Suppose we want to do convolutions for a one-dimensional input of length 5 and filter width of 3:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ x_4^{(i)} \\ x_5^{(i)} \end{bmatrix} \qquad f = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}$$

The activation output from this convolutional layer will be calculated as follows, where the *relu* function is applied elementwise:

$$a^{(i)} = relu\left(\begin{bmatrix} f_1 x_1^{(i)} + f_2 x_2^{(i)} + f_3 x_3^{(i)} \\ f_1 x_2^{(i)} + f_2 x_3^{(i)} + f_3 x_4^{(i)} \\ f_1 x_3^{(i)} + f_2 x_4^{(i)} + f_3 x_5^{(i)} \end{bmatrix}\right)$$

(a) Show how the argument to the *relu* function above could be obtained as a matrix product of a matrix F involving the filter and the column vector $x^{(i)}$. Your goal is to fill in values in the matrix F below (I did not set it up to be to scale):

$$\begin{bmatrix} f_1 x_1^{(i)} + f_2 x_2^{(i)} + f_3 x_3^{(i)} \\ f_1 x_2^{(i)} + f_2 x_3^{(i)} + f_3 x_4^{(i)} \\ f_1 x_3^{(i)} + f_2 x_4^{(i)} + f_3 x_5^{(i)} \end{bmatrix} = \begin{bmatrix} \qquad\qquad\qquad\qquad\qquad\qquad \\ \\ \\ \\ \end{bmatrix} \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ x_4^{(i)} \\ x_5^{(i)} \end{bmatrix}$$

(b) One of the quantities needed for gradient descent would be $\frac{\partial a^{(i)}}{\partial f_1}$. Show how this could be calculated, assuming that $f_1 x_1^{(i)} > 0$, $f_1 x_2^{(i)} > 0$, and $f_1 x_3^{(i)} < 0$. Your answer will be a column vector of length 3.