0.1 Calculating Output of Attention Layers (15 points)

1. (4 points) Consider a self-attention layer on an input sequence of two row vectors $X = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$. Given a key weight matrix W_k, query weight matrix W_q and value weight matrix W_v:

$$W_k = \begin{bmatrix} -2 \\ -1 \end{bmatrix}, W_q = \begin{bmatrix} 2\sqrt{2} \\ \sqrt{2} \end{bmatrix}, W_v = \begin{bmatrix} -0.1 \\ 0.3 \end{bmatrix},$$

find the output sequence.

.

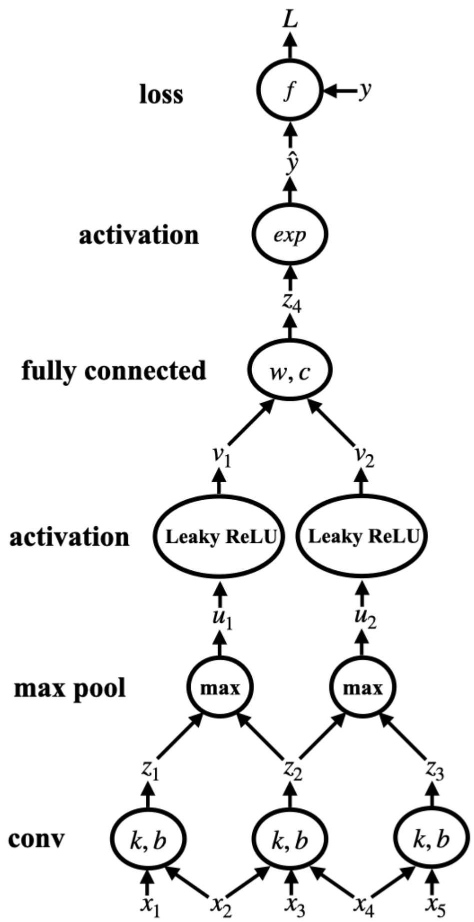
2. (4 points) How would you compute the output sequence for input $X = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$, given the answer to the above problem?

3. (4 points) Now consider incorporating a positional encoding with values $P = \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix}$ for the first two positions. Assuming we add the positional encoding to the input, what is the output sequence for the input from part (a) and the same key weights, query weights, and value weights?

4. (3 points) Does your answer to part (b) change in the case of a positional encoding? Explain why or why not.

0.2 Funky 1D ConvNet Backpropagation (16 points)

Consider the following 1-dimensional convolutional neural network, where all variables are scalars:



$$L = \frac{1}{2}(y - \hat{y})^2$$

$$\hat{y} = e^{z_4}$$

$$z_4 = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + c$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \max(0.1u_1, u_1) \\ \max(0.1u_2, u_2) \end{bmatrix}$$

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} \max(z_1, z_2) \\ \max(z_2, z_3) \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} k_1 & k_2 & k_3 & 0 & 0 \\ 0 & k_1 & k_2 & k_3 & 0 \\ 0 & 0 & k_1 & k_2 & k_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix}$$

Figure 1: Computation graph of a 1D ConvNet.

1. (2 point) List the parameters in this network.

2. (3 points) Determine the following:

$$\frac{\partial L}{\partial z_4} =$$

$$\frac{\partial L}{\partial v_1} =$$

$$\frac{\partial L}{\partial v_2} =$$

3. (3 points) Given the gradients of the loss L with respect to u_1 and u_2, derive the gradients of the loss with respect to z_1, z_2, and z_3. More precisely, given

$$\frac{\partial L}{\partial u_1} = \delta_1 \qquad \frac{\partial L}{\partial u_2} = \delta_2,$$

determine the following:

$$\frac{\partial L}{\partial z_1} =$$

$$\frac{\partial L}{\partial z_2} =$$

$$\frac{\partial L}{\partial z_3} =$$

4. (4 points) Given the gradients of the loss L with respect to z_1, z_2, z_3, derive the gradients of the loss with respect to k_1, k_2, k_3, and b. More precisely, given

$$\frac{\partial L}{\partial z_1} = \delta_1 \qquad \frac{\partial L}{\partial z_2} = \delta_2 \qquad \frac{\partial L}{\partial z_3} = \delta_3,$$

determine the following:

$$\frac{\partial L}{\partial k_1} =$$

$$\frac{\partial L}{\partial k_2} =$$

$$\frac{\partial L}{\partial k_3} =$$

$$\frac{\partial L}{\partial b} =$$

5. (4 points) Suppose that we know the exact numeric values of some intermediate variables/derivatives in the computation graph:

$$z_1 = 1 \quad z_2 = -2 \quad z_3 = -3 \qquad w_1 = 5 \quad w_2 = 10 \quad c = 3 \qquad \frac{\partial L}{\partial z_4} = 1.$$

Given these values, what are the numeric values of z_4, $\frac{\partial L}{\partial z_1}$, $\frac{\partial L}{\partial z_2}$, $\frac{\partial L}{\partial z_3}$?

(Your answers should be numbers rather than expressions containing variables).

$z_4 =$

$$\frac{\partial L}{\partial z_1} =$$

$$\frac{\partial L}{\partial z_2} =$$

$$\frac{\partial L}{\partial z_3} =$$