Problem Set 4

CS 4277: Deep Learning

Name (print clearly): ______________________________ Section: (e.g., 01) ________

Signature: ____

Student account username (e.g., msmith3): ______________________

Signing signifies that you agree to comply with the **Academic Honor Code** and course policies stated in the syllabus.

Choose one of these two options for turn-in:

1. Print this document, write or answers, scan your finished homework to a PDF, name the PDF `cs4277-ps4-<your-student-account-username>.pdf`, e.g., `cs4277-ps4-msmith3.pdf` and submit the PDF to the assignment on D2L.

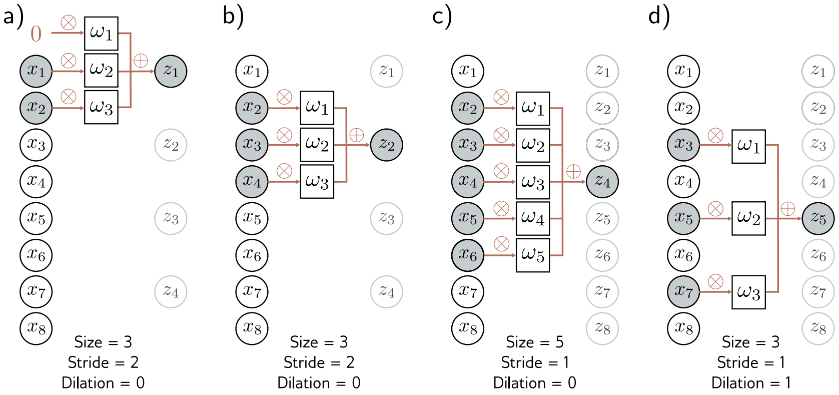
2. While viewing this document in your web browser, in the address bar change `.pdf` to `.tex`, save the LaTeX source as a text file, add your answers in appropriate LaTeX markup in the appropriate spaces, compile to a PDF named as in the instructions above, and submit the PDF file to the assignment on D2L.

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	12	12	12	12	12	12	12	12	12	108
Score:										

1. (12 points) Problem 10.1 Show that the operation in Equation 10.3 (reproduced below) is equivariant with respect to translation.

$$z_i = \omega_1 x_{i-1} + \omega_2 x_i + \omega_3 x_{i+1} \tag{10.3}$$

2. (12 points) Problem 10.3 Write out the equation for the 1D dilated convolution with a kernel size of three and a dilation rate of two, as pictured in Figure 10.3d (reproduced below).



a)
0
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
ω_1 ω_2 ω_3
z_1 z_2 z_3 z_4
Size = 3
Stride = 2
Dilation = 0
b)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
ω_1 ω_2 ω_3
z_1 z_2 z_3 z_4
Size = 3
Stride = 2
Dilation = 0
c)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
ω_1 ω_2 ω_3 ω_4 ω_5
z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8
Size = 5
Stride = 1
Dilation = 0
d)
x_1 x_2 x_3 x_4 x_5 x_6 x_7 x_8
ω_1 ω_2 ω_3
z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8
Size = 3
Stride = 1
Dilation = 1

3. (12 points) Problem 10.8 Consider a 1D convolutional network where the input has three channels. The first hidden layer is computed using a kernel size of three and has four channels. The second hidden layer is computed using a kernel size of five and has ten channels. How many biases and how many weights are needed for each of these two convolutional layers?

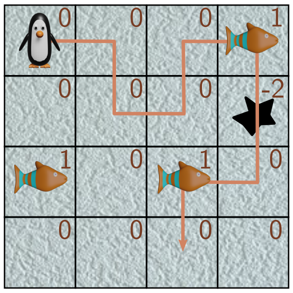
4. (12 points) Problem 10.9 A network consists of three 1D convolutional layers. At each layer, a zero-padded convolution with kernel size three, stride one, and dilation one is applied. What size is the receptive field of the hidden units in the third layer?

5. (12 points) Problem 12.1 Consider a self-attention mechanism that processes N inputs of length D to produce N outputs of the same size. How many weights and biases are used to compute the queries, keys, and values? Assume that all three quantities are also of length D. How many attention weights $\boldsymbol{a}[\cdot, \cdot]$ will there be? How many weights and biases would there be in a fully connected shallow network relating all DN inputs to all DN outputs?

6. (12 points) Problem 12.5 Why is implementation more efficient if the values, queries, and keys in each of the H heads each have dimension $\frac{D}{H}$ where D is the original dimension of the data?

7. (12 points) Problem 12.7 Consider adding a new token to a precomputed masked self-attention mechanism with N tokens. Describe the extra computations that must be done to incorporate this new token with their Big-Os assuming embeddings of length D.

8. (12 points) Problem 19.1 Figure 19.18 (reproduced below) shows a single trajectory through an example Markov reward process. Calculate the return for 1st, 2nd, 7th and 8th steps in the trajectory given that the discount factor γ is 0.9.



9. (12 points) Problem 19.4 The Boltzmann policy strikes a balance between exploration and exploitation by basing the action probabilities $\pi(a|s)$ on the current state-action reward function $q(s,a)$:

$$\pi(s|a) = \frac{e^{q(s,a)/\tau}}{\sum_{a'} e^{q(s,a')/\tau}}$$

Explain how the temperature parameter τ can be varied to prioritize exploration or exploitation.