CS577 Deep Learning Midterm Examination

Illinois Institute of Technology

Name: ____________________________________

Illinois Tech A #: ____________________________________

Instructions:

1. You have **75 minutes** to complete the examination.
2. Write your answers on an empty piece of paper. After finishing, you need to take a picture for your answer (better to use a PDF scanner APP, such as Adobe Scan.) And then upload it to BlackBoard.
3. This exam is open book. You may bring in your homework, class notes and textbooks to help you. Internet browser searches are **NOT** allowed.
4. You may use a calculator. You may not share a calculator with anyone.

In recognition of and in the spirit of the Illinois Institute of Technology Honor Code, I certify that each student will neither give nor receive unpermitted aid on this examination after starting the exam.

Question 1: Convolutional Neural Networks

(a) [8 pts] Given an $5 \times 5 \times 2$ input I, calculate the result of the convolution with $3 \times 3 \times 2$ filter K and bias = 1, with stride = 2 and no padding.

I :

2	2	0	1	1
2	1	0	1	0
1	2	0	2	1
2	0	1	0	0
1	0	0	1	0

1	1	0	2	2
1	0	2	1	0
1	2	2	1	2
0	1	1	0	1
0	0	1	2	2

K :

-1	-1	0
-1	-1	1
0	1	1

1	1	1
1	0	-1
-1	-1	-1

bias: 1

(b) [4 pts] Calculate the result of applying ReLU to the answer of (a).

(c) [8 pts] Given an input I, calculate the result of the max pooling with a 2×2 filter, stride = 2.

I :

17	14	21	0	30	32
1	38	16	33	34	23
12	7	26	18	5	28
22	25	11	40	15	19
13	36	24	48	3	20
47	29	4	6	42	31

(d) [20 pts] A CNN has the architecture given in the following table. For each layer, calculate the number of parameters and the size of corresponding feature maps. The notation follows the convention as below:

- CONV-K-N-L denotes a convolutional layer with N filters, each of the size $K \times K$, SAME [1] padding and stride = L.
- POOL-K-L indicates a pooling layer with $K \times K$ filter and stride = L.
- FC-N stands for a fully-connected layer with N neurons.

Layer	Feature map shape	Number of parameters
INPUT	$32 \times 32 \times 3$	0
CONV-5-16-2	(1)	(2)
POOL-2-2	(3)	(4)
CONV-3-32-1	(5)	(6)
POOL-2-2	(7)	(8)
Flatten	/	/
FC-10	(9)	(10)

[1] SAME padding means adding zeros to the edges of the input feature maps so that the output feature maps have the same spatial dimensions as the input feature maps

Question 2: RNN

(a) [20 pts] Consider a simple RNN with a linear activate function for one output defined by these equations, for $t = 1, ..., T$:

$$s_t = c(w_x \times x_t + w_{rec} \times s_{t-1}), \text{ and } y = c(w_h \times s_T).$$

The below picture illustrates a simple RNN.

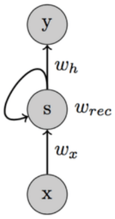

Figure 1: The simple Recurrent Neural Network.

Assume $w_x = 4$, $w_{rec} = 2$, $w_h = 3$, $s_0 = 0$, $c = 2$ and $T = 3$, please calculate the output y for the sample input x = (1, 2, 2).

(b) [5 pts] (Multiple Choice Question. There is only one correct choice for each question.)

What is the primary difference between RNNs (Recurrent Neural Networks) and traditional feedforward neural networks?

(i) RNNs use convolutional layers.

(ii) RNNs have more hidden layers.

(iii) RNNs have loops that allow for information to be passed between time steps.

(iv) RNNs are only suitable for text data.

Your Answer:

(c) [5 pts] (Multiple Choice Question. There is only one correct choice for each question.)

What is the most common issue when RNNs process long sequence data?

(i) They typically overfit.

(ii) They can't recognize patterns in sequences.

(iii) They may experience gradient explosion or gradient vanishing issues.

(iv) Their outputs are always binary.

Your Answer:

(d) [5 pts] (Multiple Choice Question. There is only one correct choice for each question.)

Which of the following statements about the bidirectional RNNs is true?

(i) They process sequences from the last element to the first only.

(ii) They have two layers of RNNs, one processing the sequence from start to end and the other from end to start.

(iii) They only return the final state from the backward pass.

(iv) They are designed originally for image data.

Your Answer:

(e) [5 pts] (Multiple Choice Question. There is only one correct choice for each question.)

Given the following statements, which one is uniquely true for LSTMs?

(i) Utilize skip connections to bypass certain layers.

(ii) Employ gating mechanisms to regulate the flow of information through memory cells.

(iii) Use attention mechanisms to weight input features.

(iv) Process input data using filter kernels.

Your Answer:

Question 3: Self-Attention Calculation

(a) [8 pts] Given the following input sequence with 3 vectors and 4 elements for each vector, and the self-attention mechanism parameters:

Input Sequence: [[4, 2, 1, 5], [2, 1, 4, 2], [4, 7, 0, 2]]

Parameters:

Query Weight:

$$\begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}$$

Key Weight:

$$\begin{bmatrix} 0.4 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}$$

Value Weight:

$$\begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0 & 0.9 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.3 \end{bmatrix}$$

(Hint: Attention$(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V$, where $Q = XW^Q, K = XW^K, V = XW^V$, X is the input, W^i is the weight matrix, d_k is the scale factor.)

Calculate the Query vector, Key vector and Value vector given the above input and parameters.

(i) Show your calculation process for the first vector [4, 2, 1, 5].

(ii) Show the result for the whole input sequence.

(b) [4 pts] Based on your calculations above

(i) Calculate the raw attention scores (you don't need to apply the scale factor or the softmax function.) for the second vector of the input. Show your calculation process and result.

(ii) Then you apply the softmax function to this raw attention scores to get the attention score (you don't need to apply the scale factor). Your result can contain the exponential term.

(c) [8 pts] Calculate the final representation of the second vector after this self-attention mechanism, using the Q vector, K vector, V vector and the attention value you calculated above. For the attention score, you can just use the raw attention scores. Show your calculation process and the result.