JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3EXAMINATION- 2025

B.Tech-VII Semester (CSE/IT)

COURSE CODE (CREDITS):19B1WCI738 (3) MAX. MARKS: 35

COURSE NAME: INTRODUCTION TO DEEP LEARNING

COURSE INSTRUCTORS: VANI SHARMA MAX. TIME: 2 Hours

Note: (a) All questions are compulsory.

(b) The candidate is allowed to make Suitable numeric assumptions wherever required

for solving problems

(c) Use of calculator is allowed

Q.No	Question	CO	Marks
Q1	Suppose we choose a model $f(x) = \sigma(wx + b)$ which has two parameters w, b. Further, assume that we are trying to learn the parameters of this model using 200 training points. If we use mini-batch gradient descent with a batch size of 10, then how many times will each parameter get updated in one epoch?	[1]	[2]
Q2	(a) Consider the following Convolutional neural network where all the convolution filters are of size 3×3. For all the convolution layers, the stride $S = 1$ and padding $P = 1$: • CONV1:convolutional layer which takes an image of size $28 \times 28 \times 1$ as input, and produces 64 outputs (64 filters of size $3 \times 3 \times 1$) • POOL1: 2×2 max-pooling layer • CONV2: convolutional layer with 64 inputs, 128 outputs (128 filters of size $3 \times 3 \times 64$) • POOL2: 2×2 max-pooling layer • CONV3: convolutional layer with 128 inputs, 256 outputs • CONV4: convolutional layer with 256 inputs, 256 outputs • POOL3: 2×2 max-pooling layer • FC1: fully connected layer with 1024 outputs What is the number of parameters in the FC1 layer? (b) Does pooling make the model more or less sensitive to small changes in the input images? Why? By small changes, we mean moving the input images to the left or right, rotating them slightly etc.	[3]	[5+2]
Q3	(a) Suppose that we need to develop an RNN model for sentiment classification. The input to the model is a sentence composed of five words and the output is the sentiments (positive or negative). Assume	[3]	[2+6]

	that each word is represented as a vector of length 100 × 1 and the output labels are one-hot encoded. Further, the state vector s_t is initialized with all zeros of size 30 × 1. How many parameters (including bias) are there in the network? (b) Consider a simple RNN used for sentiment analysis. Write the equations for the parameter updates using Backpropagation Through Time (BPTT) for at least three time steps.		
Q4	A convolution operation is performed over an input gray scale image of size 3×3 (represented as matrix X) with a filter of size 2×2 representing its weight matrix w_1 and bias b_1 that results in the next layer feature map z_1. Then after the ReLU, Maxpooling and flatten the 1-Dimensional flatten vector is fed to a single, perceptron. At last, the sigmoid activation function is applied to make a binary classification, and the loss (L) is computed as the binary cross entropy. Assume that during the back propagation the derivative of loss with respect to z_1 is known of already computed. Write a gradient descent back propagation solution to update the w_1 and b_1 trainable parameters in the above CNN architecture.	[3]	[7]
Q5	Describe the architecture of a GRU with a neat and clean diagram. Discuss the significance of all the gates used, along with the mathematical equations that lead to the calculation of the reset gate, update gate, candidate activation, hidden state, and final output. Does the GRU help address the issue of poor long-term memory in standard RNNs?	[4]	[5]
Q6	Describe the concept of cross attention in the transformer architecture. How does masked multi-head attention prevent information leakage during model training?	[3]	[6]