JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

MOOC End Term Examination- 2026

B.Tech-VII Semester (CSE/IT)

COURSE CODE (CREDITS): 18B2WCI711 MAX. MARKS: 100

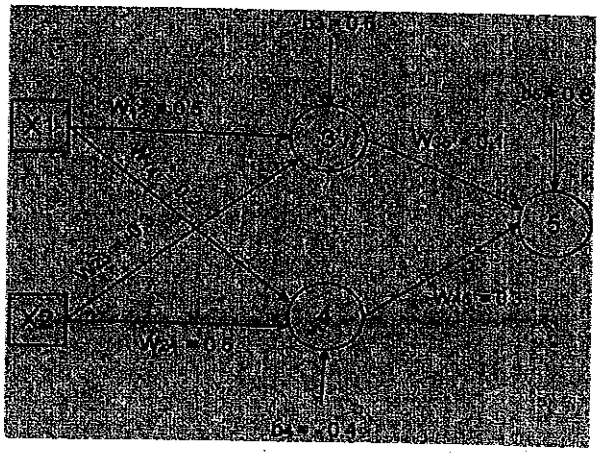
COURSE NAME: DEEP LEARNING

COURSE INSTRUCTORS: VANI SHARMA MAX. TIME: 3 Hours

Note: (a) All questions are compulsory.

(b)Marks are indicated against each question in square brackets.

(c) Use of Calculator is allowed

Q.No	Question	Marks
Q1	Discuss the limitations of self-attention that led to the development of multi-head attention and explain how multi-head attention overcomes these limitations by allowing the model to jointly attend to information from different representation subspaces.	[10]
Q2	a) Describe the architecture of 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 hidden state, and final hidden-state. b) Does the GRU help address the issue of vanishing gradients and poor long-term memory in standard RNNs?	[5+5]
Q3	Consider the following ANN model with the back-propagation algorithm. The weights and biases are given in the figure. The network uses the sigmoid function as the activation function. Using the given information, compute the output of each neuron during the forward pass. Then, calculate the error term for each neuron using the back-propagation algorithm and update the parameters for one iteration. Given: X1=1, X2=1, target T=0 and learning rate (η)=0.1. 	[10]

Q4	You are designing a Convolutional Neural Network (CNN) for image classification. Each input image is an RGB photograph with spatial dimensions of 96 × 96, giving an input volume of 96 × 96 × 3. The first two convolutional layers of this network are configured as follows: **Layer 1**: • Filters: 12, each of size 3 × 3 • Stride: 1 | Padding: 1 **Layer 2**: • Filters: 24, each of size 5 × 5 • Stride: 2 | Padding: 2 a) Determine the output volume (height × width × depth) produced by each convolutional layer. b) For each of the two convolutional layers, calculate the total number of trainable parameters. c) A 2 × 2 max pooling operation with stride = 2 (no padding) is applied to the output of Layer 2. Compute the resulting output volume dimensions and briefly state what information is retained versus discarded by this operation. d) Consider the stride and padding values currently set in Layer 2 (stride = 2, padding = 2). Discuss what would happen to the spatial resolution of the output, and the computational cost of the layer, if the stride were reduced to 1 and the padding were reduced to 0. In your answer, refer to how these two hyper-parameters jointly govern the trade-off between preserving spatial detail and controlling the number of operations performed.	[15]
Q5	What will happen if the bias terms of every neuron in a neural network are initialized to the following values before training begins? For each case, describe the effect on the forward pass, the gradient flow during back propagation, and the final learned behavior of the network: a) All biases set to a large positive constant b) All biases set to a large negative constant c) All biases initialized to zero	[10]

Q6	a) Explain how L1 and L2 regularization differ in the way they penalize large weights in a deep neural network. Write the modified loss functions for both, and explain geometrically why L1 regularization tends to produce sparse weight vectors while L2 regularization tends to produce small but non-zero weights. How does each method reduce overfitting? b) Explain Batch Normalization as a regularization technique for deep neural networks. Discuss how batch normalization reduces internal covariate shift and acts as an implicit regularizer. c) Discuss the following optimization-related concepts in the context of training deep neural networks: • **Learning Rate Scheduling** — Explain the motivation behind decaying the learning rate during training. • **AdaGrad Optimization Algorithm** — Write the complete update equations for AdaGrad. Explain how it adapts the learning rate individually for each parameter, why it is well-suited for sparse data, and what its main limitation is in the context of long training runs.	[15]
Q7	a) Explain the role of the cost function in logistic regression and describe how it differs from the cost function used in linear regression. Provide the mathematical formulation for both. b) Derive the steps involved in updating the parameters of a multiple linear regression model using the Batch Gradient Descent algorithm. c) Write the cost function minimized in Locally Weighted Regression (LWR) and explain the significance of the weights. Show that when $\tau \to \infty$, Locally Weighted Regression reduces to Ordinary Linear Regression.	[5+10+5]
Q8	A convolution operation is performed over an input gray scale image of size (represented as matrix X) with a filter of size representing its weight matrix and bias that results in the next layer feature map . 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 is known of already computed. Write a gradient descent back propagation solution to update the trainable parameters in the above CNN architecture.	[10]