Student's Name:____________________ | ID:__________________

	Course: CSAL4333- Introduction to Deep Learning
	Program: BSCS
	Sections: H2, H4, H6
	Instructors: Dr. M. Umair, Syed Nisar Ali
	Total Marks: 30
	Time Allowed: 1 Hours 30 Minutes

Solution-Mid-Term Examination-Spring25

CHECK LIST (TO BE FILLED BY STUDENT)

SN	Instructions	Checked
1	The front page of my answer script is fully filled and signed by me and the invigilator.	
2	Questions are attempted in sequence.	
3	Answers involving mathematical calculation show clear mathematical steps.	
4	I have used the right words and jargon to explain my answers.	

Question 1 – Short Answers (3 to 5 lines) [20 marks]

5 marks each

i. What is the difference between deep learning and traditional machine learning?

Deep learning is a subset of machine learning that differs primarily in how it handles feature extraction, data requirements, and model complexity. Traditional machine learning relies on manual feature engineering, where domain expertise is needed to select relevant features. In contrast, deep learning automatically extracts features from raw data using multi-layer neural networks. Deep learning models generally require large amounts of data and significant computational resources, such as GPUs, to perform effectively, whereas traditional machine learning models can perform well on smaller datasets with less computation. Additionally, traditional models like decision trees or SVMs are usually more interpretable, while deep learning models are often considered black boxes due to their complexity.

ii. What are the roles of activation functions in a neural network?

Activation functions in a neural network introduce non-linearity into the model, allowing it to learn complex patterns and relationships in the data. Without activation functions, a neural network composed of only linear operations would behave like a single-layer linear model, regardless of the number of layers, and thus fail to capture non-linear dependencies.

iii. What is overfitting in deep learning, and how can it be prevented?

Overfitting in deep learning occurs when a model learns the training data too well, capturing noise and irrelevant patterns, which leads to poor generalization on unseen data. This typically happens when the model is too complex relative to the amount of training data. To prevent overfitting, techniques such as regularization (L1/L2), dropout, early stopping, and data augmentation can be used. Additionally, using more training data or simpler models can help improve generalization and reduce overfitting.

iv. What is the function of backpropagation in training a neural network?

Backpropagation is the algorithm used to train neural networks by updating the model's weights to minimize the loss function. It works by computing the gradient of the loss with respect to each weight through the chain rule, moving backward from the output layer to the input layer. These gradients are then used by an optimization algorithm, typically stochastic gradient descent (SGD) or its variants, to adjust the weights and reduce the error. Backpropagation ensures that the network learns which parameters contribute most to the prediction error and updates them accordingly.

Question 2	**[20 marks]**

In the Leaky ReLU activation function, the negative slope parameter α controls the amount of "leak" for negative inputs.

i. 8 marks
ii. 4 marks
iii. 4 marks
iv. 4 marks

Discuss (i) how varying the value of α (e.g., very small like 0.01 vs. relatively large like 1.0) affects the behavior of the activation function, (ii) the gradient flow during backpropagation, and (iii) the overall training dynamics of a deep neural network.

(iv) What trade-offs should be considered when choosing α?

(i) Behavior of the activation function:
In Leaky ReLU, the value of α determines the slope for negative inputs. A small α (e.g., 0.01) means the function outputs a small negative value for negative inputs, allowing a slight flow of information. A large α (e.g., 1.0) makes the function nearly linear across the entire input range, behaving more like a linear function which leads towards a convex behavior (single minima) in loss function.

(ii) Gradient flow during backpropagation:
With a small α, gradients for negative inputs are small but non-zero, which helps avoid the "dying ReLU" problem where neurons become inactive. However, very small gradients can still slow learning. A larger α allows stronger gradient flow for negative inputs, which can improve training speed and mitigate vanishing gradients.

(iii) Overall training dynamics:
Smaller α values maintain non-linearity and help the network learn complex representations, but may slow convergence if negative gradients are too weak. Larger α speeds up learning by maintaining stronger gradient flow, but excessive linearity may reduce the network's capacity to model non-linear functions effectively.

(iv) Trade-offs when choosing α:
Choosing α involves balancing between preserving non-linearity and ensuring sufficient gradient flow. A very small α risks slow learning or partial neuron inactivation, while a large α can reduce model expressiveness. Typically, α is chosen empirically (e.g., 0.01 or 0.1), ensuring a good compromise between learning stability and model flexibility.

Question 3	[10 marks]
Given the 3×3 feature map below, apply sigmoid and ReLU activation functions to each element and produce the resulting matrices:	5 marks each

$$\begin{bmatrix} 2 & -1 & 3 \\ -2 & 0 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

ReLU

$$\begin{bmatrix} 2 & 0 & 3 \\ 0 & 0 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

Sigmoid

$$\begin{bmatrix} 0.8808 & 0.2689 & 0.9526 \\ 0.1192 & 0.5 & 0.7311 \\ 0.7311 & 0.8808 & 0.9820 \end{bmatrix}$$