# Deep Learning

| | |
|---|---|
| **Section Id :** | 64065357826 |
| **Section Number :** | 5 |
| **Section type :** | Online |
| **Mandatory or Optional :** | Mandatory |
| **Number of Questions :** | 17 |
| **Number of Questions to be attempted :** | 17 |
| **Section Marks :** | 50 |
| **Display Number Panel :** | Yes |
| **Section Negative Marks :** | 0 |
| **Group All Questions :** | No |
| **Enable Mark as Answered Mark for Review and Clear Response :** | Yes |
| **Maximum Instruction Time :** | 0 |
| **Sub-Section Number :** | 1 |
| **Sub-Section Id :** | 640653120743 |
| **Question Shuffling Allowed :** | No |
| **Is Section Default? :** | null |

**Question Number : 120 Question Id : 640653820784 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 0**

Question Label : Multiple Choice Question

**THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL : DEEP LEARNING (COMPUTER BASED EXAM)"**

**ARE YOU SURE YOU HAVE TO WRITE EXAM FOR THIS SUBJECT?**

**CROSS CHECK YOUR HALL TICKET TO CONFIRM THE SUBJECTS TO BE WRITTEN.**

**(IF IT IS NOT THE CORRECT SUBJECT, PLS CHECK THE SECTION AT THE TOP FOR THE SUBJECTS REGISTERED BY YOU)**

**Options :**

6406532754972. ✔ YES

6406532754973. ✖ NO

| | |
|---|---|
| **Sub-Section Number :** | 2 |
| **Sub-Section Id :** | 640653120744 |
| **Question Shuffling Allowed :** | Yes |
| **Is Section Default? :** | null |

**Question Number : 121 Question Id : 640653820785 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Multiple Choice Question

Consider a scenario where you have a dataset with overlapping classes (that is instances from different classes share similar or identical feature values), and you decide to train a perceptron model for classification.

**Assertion (A):** The perceptron model may struggle to classify instances accurately when classes overlap in the feature space.

**Reason (R):** The perceptron learning algorithm aims to find a linear decision boundary that separates the classes, and in the presence of overlapping classes, it may not be able to capture the underlying patterns effectively.

Select the correct option:

**Options :**

6406532754974. ✔ Both A and R are true, and R is the correct explanation of A.

6406532754975. ✖ Both A and R are true, but R is not the correct explanation of A.

6406532754976. ✖ A is true, but R is false.

6406532754977. ✖ A is false, but R is true.

**Question Number : 122 Question Id : 640653820787 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Multiple Choice Question

Consider a feedforward neural network with one hidden layer trained using backpropagation for a binary classification task. The network has the following architecture:

- Input layer with 15 neurons
- Hidden layer with 25 neurons
- Output layer with 1 neuron

During the backpropagation process, the derivative of the sigmoid activation function $\sigma(z)$ with respect to its argument $z$ is given by:

$$\sigma'(z) = \sigma(z) \cdot (1 - \sigma(z))$$

If the loss function used for binary classification is the binary cross-entropy loss, and the activation fuction at hidden layer and output layer is sigmoid. The output of the neural network is denoted as $\hat{y}$, and the true label is denoted as $y$, what is the expression for $\frac{\partial L}{\partial w_j}$, where $w_j$ represents the weights connecting the $j$th neuron of hidden layer to the output layer? Assume that the output of $j$th neuron of hidden layer is $h_j$ and no biases in the network.

**Options :**

6406532754980. ✖ $\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'(w_j h_j) \cdot h_j$

6406532754981. ✖ $$\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma' \left( \sum_{i=1}^{25} w_i \right) \cdot h_j$$

6406532754982. ✖ $$\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma'(h_j) \cdot h_j$$

6406532754983. ✔ $$\frac{\partial L}{\partial w_j} = -(y - \hat{y}) \cdot \sigma' \left( \sum_{i=1}^{25} w_i h_i \right) \cdot h_j$$

**Question Number : 123 Question Id : 640653820788 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Multiple Choice Question

In the context of mini-batch gradient descent, if doubling the size of the mini-batch makes your model take twice as many epochs to reach convergence, how does this affect the total number of parameter updates compared to using the original mini-batch size? Assume everything else remains constant.

**Options :**

6406532754984. ✖ The number of updates required is doubled.

6406532754985. ✖ The number of updates required is four times.

6406532754986. ✖ The number of updates required is halved.

6406532754987. ✔ The number of updates remains the same.

**Question Number : 124 Question Id : 640653820789 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Multiple Choice Question

Suppose you are working with a sparse dataset and using Momentum-based Gradient Descent (GD) and AdaGrad algorithms for optimization. Which of the following statements accurately describes the behavior of weight and bias term updates?

**Options :**

6406532754988. ✖ Bias vector will update frequently only in the case of the AdaGrad algorithm.

6406532754989. ✔ Weight vector will have very few updates in the case of Momentum-based GD.

6406532754990. ✖ Weight vector will have frequent updates in both cases.

6406532754991. ✖ Bias vector will have very few updates in the case of Momentum-based GD.

| | |
|---|---|
| **Sub-Section Number :** | 3 |
| **Sub-Section Id :** | 640653120745 |
| **Question Shuffling Allowed :** | Yes |
| **Is Section Default? :** | null |

**Question Number : 125 Question Id : 640653820786 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 2**

Question Label : Multiple Choice Question

State True or False. Sigmoid activation function helps mitigating vanishing gradient problem better than ReLU activation function.

**Options :**

6406532754978. ✔ FALSE

6406532754979. ✖ TRUE

**Question Number : 126 Question Id : 640653820802 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 2**

Question Label : Multiple Choice Question

How many weight matrices does a LSTM unit and GRU unit learns respectively during backpropagation?

**Options :**

6406532755010. ✖ 2,3

6406532755011. ✖ 3,4

6406532755012. ✖ 4,3

6406532755013. ✔ 8,6

**Question Number : 127 Question Id : 640653820803 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 2**

Question Label : Multiple Choice Question

Which statements about LSTM and GRU units in handling the problem of exploding/vanishing gradients are correct?

**Options :**

6406532755014. ✖ LSTM units are more susceptible to exploding gradients compared to GRU units due to their complex gating mechanisms.

6406532755015. ✖ The gating mechanisms in LSTM and GRU units implicitly handle exploding gradients without the need for additional techniques.

6406532755016. ✖ GRU units mitigate both vanishing and exploding gradient problems more effectively than LSTM units due to their simplified architecture.

6406532755017. ✔ Both LSTM and GRU units utilize gating mechanisms to address vanishing gradient problems, but GRU units typically converge faster due to their simpler design.

**Sub-Section Number :** 4

**Sub-Section Id :** 640653120746

**Question Shuffling Allowed :** Yes

**Is Section Default? :** null

**Question Number : 128 Question Id : 640653820790 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

Consider an ensemble consisting of 3 models, where each model has an individual error rate of 50% on the test set. Assuming a simple majority voting scheme, what is the probability that a given instance is misclassified by the ensemble?

Assume that the models are independent and their errors are uncorrelated.

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Range

**Text Areas :** PlainText

**Possible Answers :**

0.45 to 0.55

**Question Number : 129 Question Id : 640653820791 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

Given the true outputs for a dataset, y = [1,2,3,4], and the predictions from three different models (A, B, and C) trained on different subsets of the training data, as follows:

Model A predictions: A = [1.1,1.9,3.1,3.9]

Model B predictions: B = [0.9,2.1,2.9,4.1]

Model C predictions: C = [2,2,3,4]

Calculate the squared bias ($bias^2$) for these models evaluated in a regression task. Assume that each model uses the same parameters but is trained on different subsets of the training data. (Enter your answer up to 3 decimal places.)

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Range

**Text Areas :** PlainText

**Possible Answers :**

0.02 to 0.03

**Question Number : 130 Question Id : 640653820792 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

Consider a Convolutional Neural Network (CNN) architecture for image classification with the following layers:

1. Convolutional layer with 32 filters of size 3 × 3, with a stride of 1 and no padding.
2. Max pooling layer with a pool size of 2 × 2 and a stride of 2.
3. Convolutional layer with 64 filters of size 4 × 4, with a stride of 1 and no padding.
4. Max pooling layer with a pool size of 2 × 2 and a stride of 2.
5. Fully connected layer with 128 neurons.
6. Output layer with 10 neurons (for 10 classes) using softmax activation.

If the input image size is 64 × 64 × 3, and the network has no bias term, how many parameters are there in the CNN?

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Equal

**Text Areas :** PlainText

**Possible Answers :**

28256

**Question Number : 131 Question Id : 640653820793 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

Suppose you have a vocabulary of 10,000 unique words and you want to train a CBOW model with a window size of 2 (on each side) and with an embedding dimension of 300. How many parameters (weights) will the embedding layer have?

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Equal

**Text Areas :** PlainText

**Possible Answers :**

3000000

**Question Number : 132 Question Id : 640653820794 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

Assume that your CBOW model outputs a probability distribution over a vocabulary of 20,000 words for a given context. If the correct target word is word number 150, and the model's predicted probability for this word is 0.02, calculate the cross-entropy loss for this prediction.

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Range

**Text Areas :** PlainText

**Possible Answers :**

3.9 to 4.0

**Question Number : 133 Question Id : 640653820795 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

In a Skip-gram model with a window size of 2 (on each side), how many **unique** pairs of target and context words will be generated for the following sentence:

**‘The idea is to die young as late as possible’**

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Equal

**Text Areas :** PlainText

**Possible Answers :**

31

**Question Number : 134 Question Id : 640653820796 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

In a Skip-gram model with a vocabulary of 15000 words, if you choose a negative sampling rate of 10 negative samples for each positive sample, how many **total** samples will be generated for a sentence with 8 unique words and a window size of 3 (on each side)?

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Equal

**Text Areas :** PlainText

**Possible Answers :**

396

| | |
|---|---|
| **Sub-Section Number :** | 5 |
| **Sub-Section Id :** | 640653120747 |
| **Question Shuffling Allowed :** | No |
| **Is Section Default? :** | null |

**Question Id : 640653820797 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Question Numbers : (135 to 137)**

Question Label : Comprehension

Given a scenario where there are 5 possible letters represented as one-hot encoded vectors of length 5, the RNN employs the following formulas for the state vector and output at time step $t$:

$$s_t = \sigma(Ux_i + Ws_{t-1} + b)$$
$$\hat{y}_t = O(Vs_t + c)$$

Here, $\sigma$ and $O$ denote the sigmoid and softmax functions, respectively. Assume that $s_t \in \mathbb{R}^2$ and $y_t \in \mathbb{R}^5$

Based on the above data, answer the given subquestions.

**Sub questions**

**Question Number : 135 Question Id : 640653820798 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 2**

Question Label : Short Answer Question

With a total of 10 time steps ($T$ = 10, implying prediction for a word of length 9), what is the total count of parameters (including bias) within the network?

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Equal

**Text Areas :** PlainText

**Possible Answers :**

31

**Question Number : 136 Question Id : 640653820799 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Multiple Choice Question

If all the parameters (including bias) in the network are initialized to zero, what will be the predicted $\hat{y}_2$ at time step 2 (assume that indices start with 1) for the input $[0, 0, 1, 0, 0]^T$?

**Options :**

6406532755000. ✖ $[0, 0.5, 0.5, 0, 0]^T$

6406532755001. ✖ $[0.1, 0.1, 0.6, 0.1, 0.1]^T$

6406532755002. ✖ $[0, 0, 1, 0, 0]^T$

6406532755003. ✔ $[0.2, 0.2, 0.2, 0.2, 0.2]^T$

**Question Number : 137 Question Id : 640653820800 Question Type : SA Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3**

Question Label : Short Answer Question

If all the parameters (including bias) in the network are initialized to zero, what will be the total loss after 10 time steps (assume that indices start with 1) for the input $[0, 0, 1, 0, 0]^T$ ? assume the loss to be cross-entropy at each time step. Write your answer correct to two decimal places.

**Response Type :** Numeric

**Evaluation Required For SA :** Yes

**Show Word Count :** Yes

**Answers Type :** Range

**Text Areas :** PlainText

**Possible Answers :**

16 to 16.2

| | |
|---|---|
| **Sub-Section Number :** | 6 |
| **Sub-Section Id :** | 640653120748 |
| **Question Shuffling Allowed :** | Yes |
| **Is Section Default? :** | null |

**Question Number : 138 Question Id : 640653820801 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0**

**Correct Marks : 3 Max. Selectable Options : 0**

Question Label : Multiple Select Question

Select the correct statement regarding the vanishing and exploding gradient problem in RNN.

**Options :**

6406532755005. ✔ The vanishing gradient problem in RNNs occurs when the gradient approaches zero during backpropagation, hindering the training of long sequences.

6406532755006. ✔ The exploding gradient problem in RNNs occurs when the gradient grows uncontrollably during backpropagation, leading to numerical instability and difficulty in training.

6406532755007. ✖ The vanishing gradient problem in RNNs can be mitigated by using the Rectified Linear Unit (ReLU) activation function

6406532755008. ✔ The vanishing gradient problem in RNNs can be mitigated by using gradient