Deep Learning

Section Id :	64065349323
Section Number :	4
Section type :	Online

Mandatory or Optional :	Mandatory
Number of Questions :	15
Number of Questions to be attempted :	15
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 :	640653103740
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 82 Question Id : 640653699284 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 :

6406532335107. ✔ YES

6406532335108. ✖ NO

Sub-Section Number : 2

Sub-Section Id : 640653103741

Question Shuffling Allowed : Yes

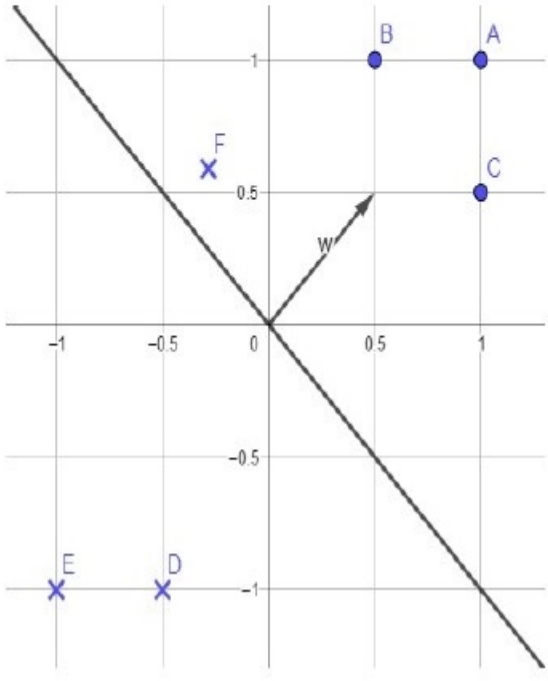
Is Section Default? : null

Question Number : 83 Question Id : 640653699285 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

The diagram below shows a decision boundary of a perceptron. The weight vector w is perpendicular to the decision boundary initially. The current decision boundary misclassifies some data points. After running the perceptron learning algorithm for a few iterations, the weight gets updated and the corresponding decision boundary separates all the data points correctly. Then the angle between the updated weight vector and its corresponding decision boundary



Options :

6406532335109. ✖ reduces

6406532335110. ✔ remains 90^{o}

6406532335111. ✖ increases

6406532335112. ✖ insufficient information

Sub-Section Number : 3

Sub-Section Id : 640653103742

Question Shuffling Allowed : Yes

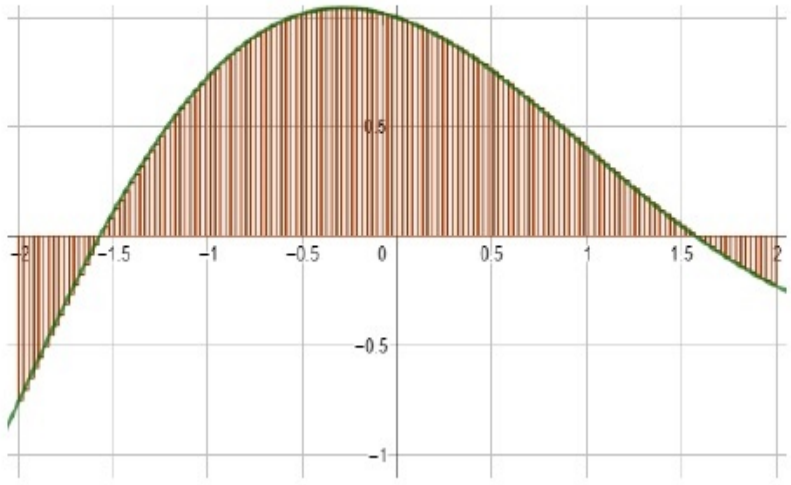
Is Section Default? : null

Question Number : 84 Question Id : 640653699287 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 1

Question Label : Multiple Choice Question

The function $f(x)$ shown below is approximated using 100 tower functions. What is the minimum number of neurons required to construct the network that approximates the function?



Options :

6406532335117. ✖ 200

6406532335118. ✖ 100

6406532335119. ✖ 301

6406532335120. ✔ 201

6406532335121. ✖ 300

Sub-Section Number : 4

Sub-Section Id : 640653103743

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 85 Question Id : 640653699286 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

Consider a sigmoid function

$$f(x) = \frac{1}{1 + e^{-(wx+b)}}$$

Suppose that w is restricted to take only negative values ($w < 0$). Suppose further that we define the steepness of the curve as absolute value of the slope. Then, select all the correct statements about the function

Options :

6406532335113. ✖ Increasing the value of *b* shifts the sigmoid function to the left (i.e., towards negative infinity)

6406532335114. ✔ Increasing the value of *b* shifts the sigmoid function to the right (i.e., towards positive infinity)

6406532335115. ✖ Increasing the value of *w* increases the steepness of the sigmoid function

6406532335116. ✔ Increasing the value of *w* decreases the steepness of the sigmoid function

Question Number : 86 Question Id : 640653699292 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

Suppose that a neural network has millions of parameters (weights and biases). A team decides to use an optimization algorithm with a learning rate scheme that is **local** to each parameter in the

network. Moreover, the learning rate changes in each iteration such that it should decrease on the steep surface and increase on the gentle surface. Which of the following optimization algorithms satisfy the team's requirements?

Options :

6406532335128. ✖ GD with an exponentially decaying learning rate scheduler

6406532335129. ✖ AdaGrad

6406532335130. ✔ AdaM

6406532335131. ✔ NADAM

6406532335132. ✔ RMSProp

6406532335133. ✖ SGD with line search

Question Number : 87 Question Id : 640653699305 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

Suppose we divide the available training samples into mini batches of size 32 to train a model with mini-batch gradient descent. Assume that we have 33 different machines to train the model. One out of 33 machines acts as a master machine. The actual weight update happens in the master machine. The master machine can send one sample for the rest of the machines along with a copy of the model in its current state to compute the gradients. We call this entire set-up parallelization. Which of the following deep learning architectures can be trained in parallel then?

Options :

6406532335155. ✔ Fully connected Feed forword neural network

6406532335156. ✔ Convolutional Neural network

6406532335157. ✔ Reccurent Neural Network

6406532335158. ✔ Transformers

Sub-Section Number : 5

Sub-Section Id : 640653103744

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 88 Question Id : 640653699293 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Select all correct statements about the bias and variance of deep learning models.

Options :

6406532335134. ✔ A model with high bias produces high training error and high validation error

6406532335135. ✖ A model with high variance produces high training error and low validation error

6406532335136. ✔ A model with high variance produces low training error and high validation error

6406532335137. ✔ Bias and variance can be both made zero if the underlying distribution of the samples are known and also the samples are not corrupted by noise

Question Number : 89 Question Id : 640653699294 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

Suppose that we use Dropout scheme to regularize the deep learning model. Assume that the dropout probability $p = 0.75$ for all the layers in the model and the model is trained using a Stochastic Gradient Descent (SGD) optimizer. Choose the correct statements about Dropout regularization.

Options :

6406532335138. ✖ Dropout reduces (removes) 25 percent of weights from the model during training and inference

6406532335139. ✖ For each iteration, 75% of neurons are dropped randomly from the model

6406532335140. ✔ For each iteration, 25% of neurons are dropped randomly from the model

6406532335141. ✔ Each weight gets multiplied by a factor p during inference.

6406532335142. ✖ Each weight gets multiplied by a factor $1 - p$ during inference.

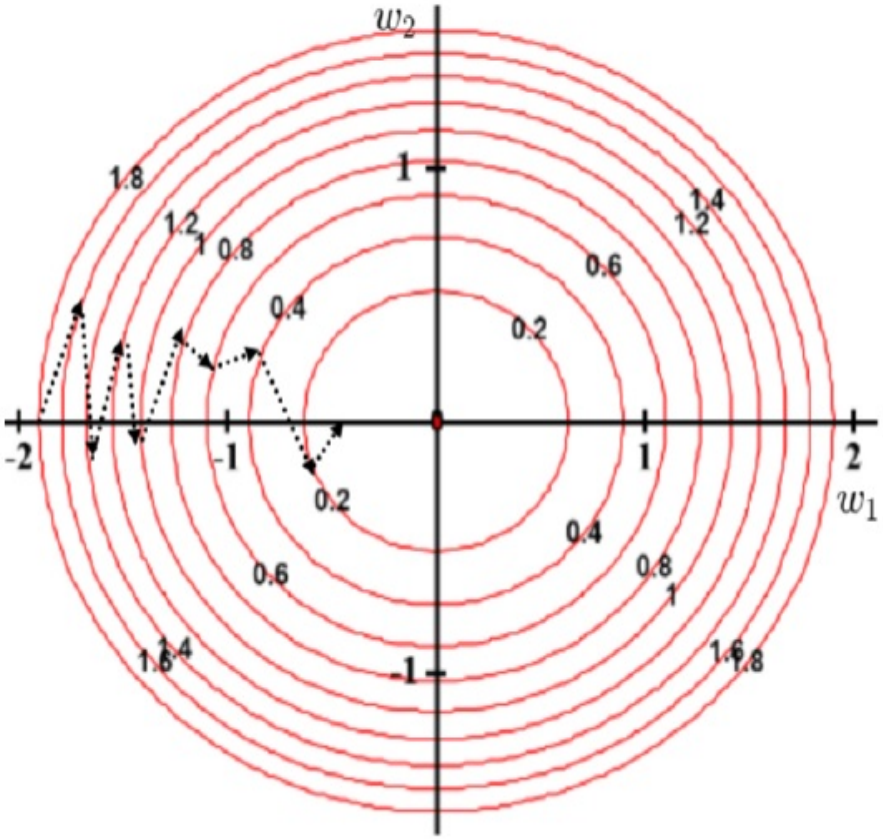
Sub-Section Number :	6
Sub-Section Id :	640653103745
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 90 Question Id : 640653699291 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 5 Max. Selectable Options : 0

Question Label : Multiple Select Question

Consider the contours of a loss surface as shown in the figure below. The parameters (w_1, w_2) were initialized to $(w_1 = -2, w_2 = 0)$. Suppose we run an optimization algorithm (not necessarily gradient-based) for a few iterations. The dotted arrows in the figure show the trajectory of updated parameters in each iteration.



Choose the correct statements

Options :

6406532335124. ✔ The loss surface is convex (with a global minimum)

6406532335125. ✖ The loss value oscillates over iterations

6406532335126. ✔ The loss value decreases consistently over iterations

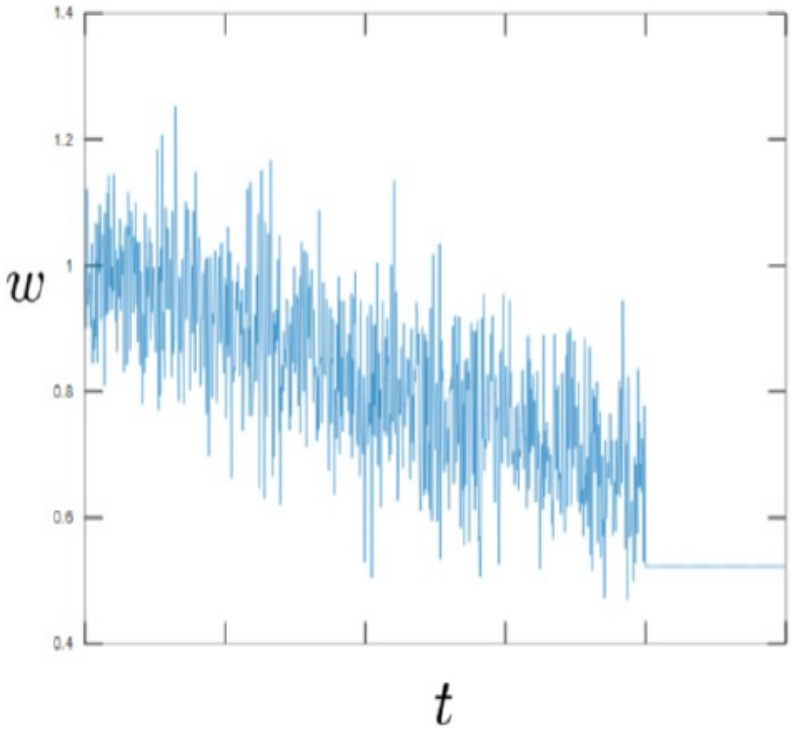
6406532335127. ✖ The loss value increases consistently

Question Number : 91 Question Id : 640653699295 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 5 Max. Selectable Options : 0

Question Label : Multiple Select Question

Suppose we have a fully connected neural network with three hidden layers containing 10 neurons each. All the neurons in the network use the standard ReLU activation function. Suppose we use cross entropy loss with Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ to update the parameters. The figure below shows the change in the value of one of the weights w in the first hidden layer over the entire training duration.



Based on the given information and from the figure, it is implied that.

Options :

6406532335143. ✖ definitely, the loss value at the end of the training is zero

6406532335144. ✔ The neuron in the hidden layer might be experiencing a vanishing gradient problem

6406532335145. ✔ The loss could have converged to the local minimum after a finite number of iterations

6406532335146. ✔ The loss value at the end of the training may not necessarily be zero

Sub-Section Number :	7
Sub-Section Id :	640653103746
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653699288 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 : (92 to 93)

Question Label : Comprehension

Consider a fully connected feed forward neural network with 3 hidden layers. The weight matrix W_1 connecting the input layer to the first hidden layer is of shape 20×150, similarly, the shape of other weight matrices are as follows $W_2 : 150 \times 100$, $W_3 : 100 \times 10$, and the weight W_4 connecting the final hidden layer and the output layer is of shape 10×3. The network solves the multi-class classification problem by using the cross-entropy loss function. Moreover, the labels are one hot encoded.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 92 Question Id : 640653699289 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

How many neurons are there in the network? Every neuron in the network has bias associated with it?

Note: A neuron is a computation unit that takes in some inputs and produces an output.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

263

Question Number : 93 Question Id : 640653699290 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

How many parameters (including bias) does the network have to learn? Assume dropout regularization is applied.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

19293

Question Id : 640653699297 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 : (94 to 95)

Question Label : Comprehension

Construct a vocabulary V from the following text corpus

- How much wood could a woodchuck chuck
- If a woodchuck could chuck wood
- as much wood as a woodchuck could chuck

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 94 Question Id : 640653699298 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

What is the size of the vocabulary $|V|$?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

9

Question Number : 95 Question Id : 640653699299 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

Suppose we consider the three words (wood,woodchuck, much). Assume we use one-hot encoded vector representation for all these words. The statement that,“The cosine similarity between the pair (wood,woodchuck) is greater than the pair (wood,much)” is

Options :

6406532335149. ✖ True

6406532335150. ✔ False

Sub-Section Number :	8
Sub-Section Id :	640653103747
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653699301 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 : (96 to 98)

Question Label : Comprehension

Suppose that we need to develop an RNN model for sentiment classification tasks. The input to the model is a **sentence composed of 10 words** and the output is the sentiment (positive or negative). Assume 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 and the prediction $\hat{y}_t$ are computed as follows

$$s_t = \sigma(Ux_t + Ws_{t-1} + b)$$

$$\hat{y}_t = \mathcal{O}(Vs_t + c)$$

The state vector s_t is initialized with all zeros of size 50×1.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 96 Question Id : 640653699302 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

How many parameters (including bias) are there in the network?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

7652

Question Number : 97 Question Id : 640653699303 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

For the given input sentence containing 10 words, how many sequential time steps does RNN take to make a final prediction?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

10

Question Number : 98 Question Id : 640653699304 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 we train the model using the BPTT algorithm for 200 iterations. In each iteration, we feed the input sentence and make a prediction, compute the loss, back-propagate through time and update the parameter. How many times does the parameter matrix *W* get updated over 200 iterations?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

200

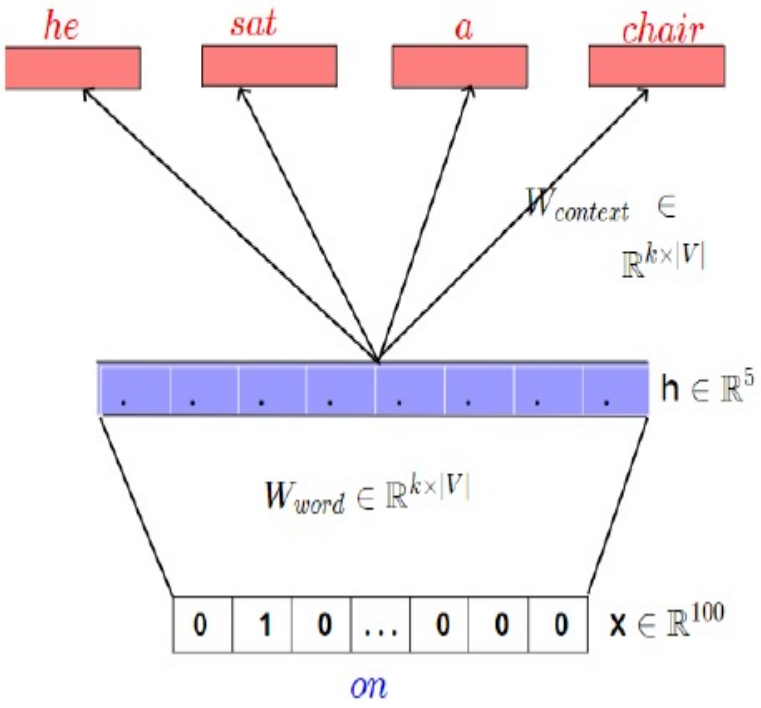
Sub-Section Number :	9
Sub-Section Id :	640653103748
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 99 Question Id : 640653699300 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 skip-gram model shown below. Each word in the vocabulary is represented as one-hot vector of size 100×1. The embedding dimension h is 5×1. Enter the number of parameters (exclude bias) in the entire network



Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2500

Sub-Section Number :	10
Sub-Section Id :	640653103749
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 100 Question Id : 640653699296 Question Type : SA Calculator : None

Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 5

Question Label : Short Answer Question

Consider an intermediate feature map H obtained after applying convolution operation on the input X using kernel F.

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

Apply the max-pooling operation with stride $s = 2$ and no padding ($p = 0$) and store the resultant output in matrix H_m. The prediction $\hat{y}$ is simply the sum of elements in H_m. Suppose $\frac{\partial L}{\partial \hat{y}} = 1.5$ (that is, the gradient of loss with respect to the prediction). What is the gradient $\frac{\partial L}{\partial H_{10}}$ where H_{10} is the element at the 1-st row and 0-th column? If you think the given info is insufficient, enter -1 as the answer.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1.5