Deep Learning

Section Id :	64065349293
Section Number :	4
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	14

Number of Questions to be attempted :	14
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 :	640653103496
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 80 Question Id : 640653698497 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 :

6406532332767. ✔ YES

6406532332768. ✖ NO

Sub-Section Number :	2
Sub-Section Id :	640653103497

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 81 Question Id : 640653698498 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

Which of the following threshold θ values of MP neuron implements OR Boolean function denoted by $f(\mathrm{x})$? Assume that the number of inputs x_i to the neuron is five and the neuron does not have any inhibitory inputs.

$$f(\mathrm{x}) = \begin{cases} 1, & \text{if } \sum_{i=0}^{4} x_i > \theta \\ 0, & otherwise \end{cases}$$

Options :

6406532332769. ✖ 4

6406532332770. ✖ 3

6406532332771. ✖ 2

6406532332772. ✔ 0

6406532332773. ✖ 1

Sub-Section Number : 3

Sub-Section Id : 640653103498

Question Shuffling Allowed : Yes

Is Section Default? : null

Question Number : 82 Question Id : 640653698499 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 two data points $\mathbf{x_1} = \begin{bmatrix} a \\ a \end{bmatrix}$ and $\mathbf{x_2} = -1 * \mathbf{x_1}$, where $a > 0$. The data point $\mathbf{x_1}$ belongs to positive class (denoted as 1) and the datapoint $\mathbf{x_2}$ belongs to negative class (denoted by 0). Suppose that the perceptron learning algorithm is used to find the decision boundary that separates these data points with the following rule,

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{w^T x} \geq 0, \\ 0 & \mathbf{w^T x} < 0 \end{cases}$$

The algorithm checks $\mathbf{x_1}$ in the first iteration and $\mathbf{x_2}$ in the second iteration and so on. How many times the weights get updated until convergence (That is, the algorithm classifies both the points correctly)? The weights do not include bias.

Assume the weights are initialized to zero

Options :

6406532332774. ✔ 1

6406532332775. ✖ 2

6406532332776. ✖ 4

6406532332777. ✖ It oscillates and never converges

Sub-Section Number :	4
Sub-Section Id :	640653103499
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 83 Question Id : 640653698500 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 sigmoid neuron. Suppose that input $x = 10$, output $y = 1$ and the parameters, $w = 0.1, b = 0.1$. The loss function is $L = \frac{1}{2}(y - \hat{y})^2$. Update the parameters once using GD with η=1. Enter the sum of updated parameter values. (that is, if $w = 0.01$ and $b = 0.09$ after updating them, then you need to enter the sum 0.1)

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.68 to 0.72

Question Number : 84 Question Id : 640653698501 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 neural network with three hidden layers and one output layer. The hidden layers contain 100 neurons each. Suppose we have a square image of size 30×30 containing either a cat (class-1) or a dog (class-2). The neural network is designed to recognize it by outputting a probability distribution over two classes using softmax activation at the output layer. The input image is flattened into an array of size 900. Assume that all neurons in the network have bias associated with them and use the sigmoid activation function in the hidden layers. How many parameters 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 :

110502

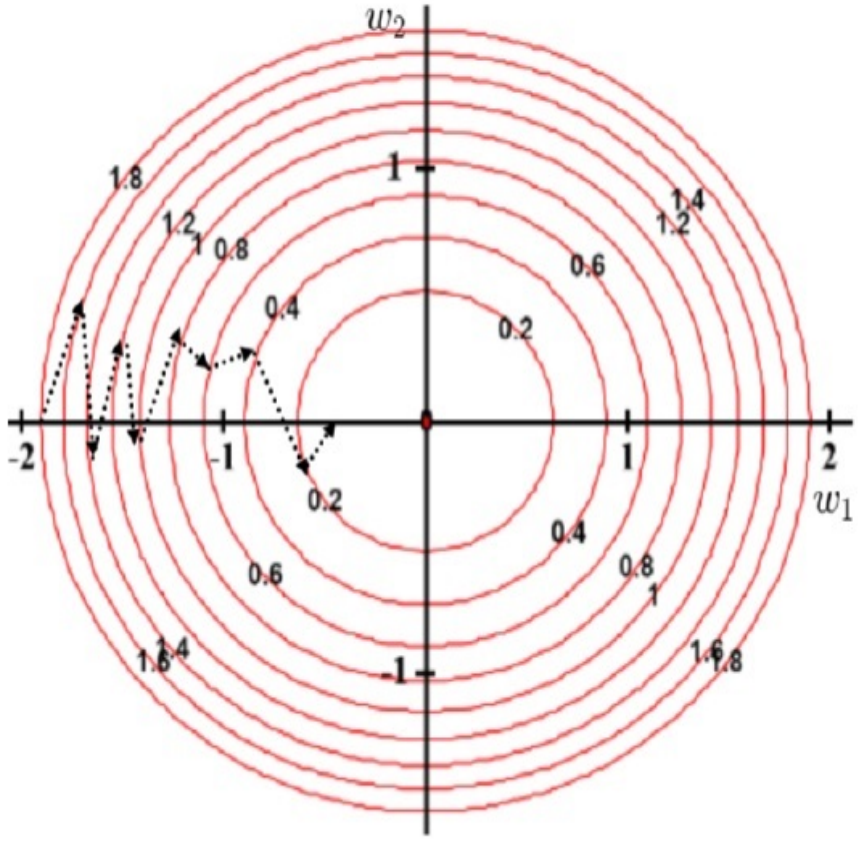
Sub-Section Number :	5
Sub-Section Id :	640653103500
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 85 Question Id : 640653698502 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 that are inferred from the figure

Options :

6406532332780. ✔ The loss surface is convex (with a global minimum)

6406532332781. ✖ The loss value oscillates over iterations

6406532332782. ✔ The loss value decreases consistently over iterations

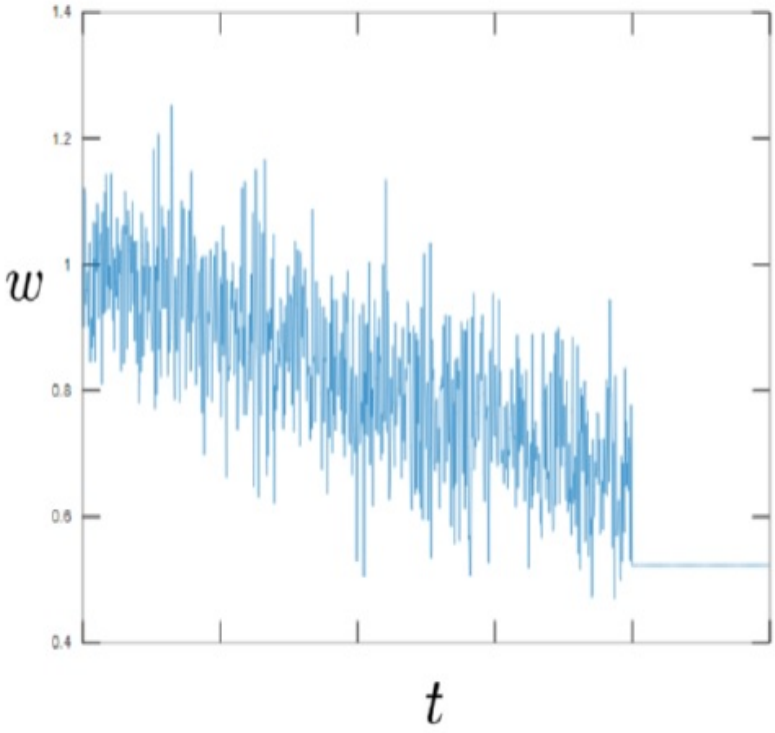
6406532332783. ✖ The loss value increases consistently

Question Number : 86 Question Id : 640653698505 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 :

6406532332797. ✖ Definitely, the loss value at the end of the training is zero

6406532332798. ✔ The neuron in the hidden layer might be experiencing a vanishing gradient problem

6406532332799. ✔ The loss could have converged to the local minimum after a finite number of iterations

6406532332800. ✔ The loss value at the end of the training may not necessarily be zero

Sub-Section Number :	6
Sub-Section Id :	640653103501
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 87 Question Id : 640653698503 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 :

6406532332784. ✖ GD with an exponentially decaying learning rate scheduler

6406532332785. ✖ AdaGrad

6406532332786. ✔ AdaM

6406532332787. ✔ NADAM

6406532332788. ✔ RMSProp

6406532332789. ✖ SGD with line search

Question Number : 88 Question Id : 640653698517 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 :

6406532332813. ✔ Fully connected Feed forword neural network

6406532332814. ✔ Convolutional Neural network

6406532332815. ✔ Reccurent Neural Network

6406532332816. ✔ Transformers

Sub-Section Number :	7
Sub-Section Id :	640653103502
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 89 Question Id : 640653698504 Question Type : MSQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 4 Max. Selectable Options : 0

Question Label : Multiple Select Question

The diagram below shows contours of a loss function $\hat{L}(\theta)$ where, $\hat{L}(\theta) =$

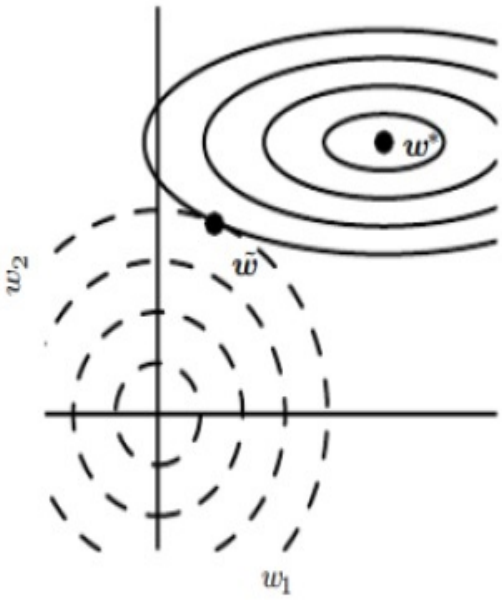


Figure 1: Contours

$L(\theta) + \alpha\Omega(\theta)$. Here, $L(\theta)$ is an un-regularied loss function and $\Omega(\theta)$ is a regularization term. Suppose we use L_2 regularization, then choose the correct statements from the following statements (with respect to this diagram)

Options :

6406532332790. ✔ The dotted circles represent contours of regularization term $\Omega(\theta)$

6406532332791. ✖ The dotted circles represent contours of loss term $L(\theta)$

6406532332792. ✔ The solid circles represent the contours of loss term $L(\theta)$

6406532332793. ✖ The solid circles represent the contours of regularization term $\Omega(\theta)$

6406532332794. ✔ Increasing the value of α make the parameters sparse

6406532332795. ✖ Decreasing the value of α makes the parameter sparse

6406532332796. ✔ The L_2 regularization is independent of the input samples used to train the network

Sub-Section Number :	8
Sub-Section Id :	640653103503
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 90 Question Id : 640653698506 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 $\frac{\partial L}{\partial \hat{y}} = 2.5$ (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 :

2.5

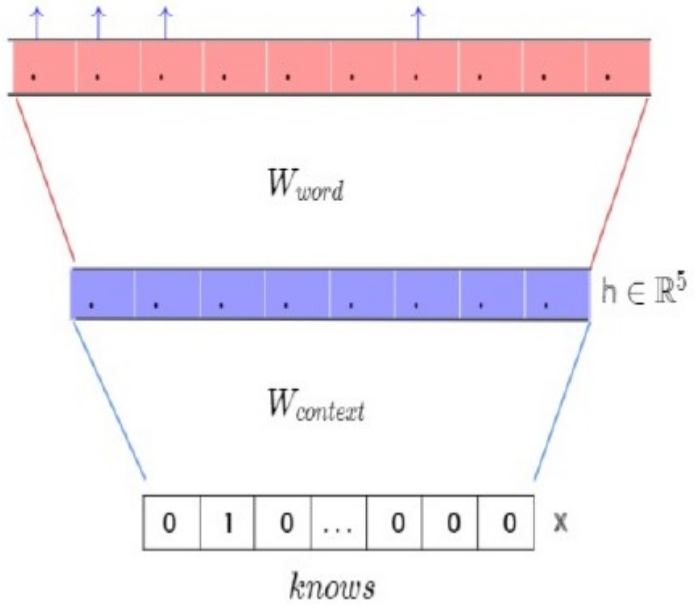
Sub-Section Number :	9
Sub-Section Id :	640653103504
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653698510 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 : (91 to 92)

Question Label : Comprehension

Suppose we use the CBOW (Continuous Bag of words) model shown below to find a distributed vector representation for all the words in the vocabulary. The size of the vocabulary $|V|$ is 150 and all the words in the vocabulary are one hot encoded (as a **column vector**) and fed as input to the network. The output layer uses softmax to produce the probability score for each word given the context. Here, $\mathbf{W}_{word}$ and $\mathbf{W}_{context}$ are weight matrices.



Based on the above data, answer the given subquestions.

Sub questions

Question Number : 91 Question Id : 640653698511 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 number of parameters are there in the network (excluding bias)?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1500

Question Number : 92 Question Id : 640653698512 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 word representation for the 2-nd word in the vocabulary corresponds to

Options :

6406532332806. ✔ 2-nd column of $\mathbf{W}_{\mathbf{context}}$

6406532332807. ✖ 2-nd row of $\mathbf{W}_{\mathbf{context}}$

6406532332808. ✖ 2-nd column of $\mathbf{W}_{\mathbf{word}}$

6406532332809. ✖ 2-nd row of $\mathbf{W}_{\mathbf{word}}$

Sub-Section Number :	10
Sub-Section Id :	640653103505
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653698507 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 : (93 to 94)

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 : 93 Question Id : 640653698508 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 : 94 Question Id : 640653698509 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 :

6406532332803. ✖ True

6406532332804. ✔ False

Sub-Section Number : 11

Sub-Section Id : 640653103506

Question Shuffling Allowed : No

Is Section Default? : null

Question Id : 640653698513 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 : (95 to 97)

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 15 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 : 95 Question Id : 640653698514 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 : 96 Question Id : 640653698515 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 15 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 :

15

Question Number : 97 Question Id : 640653698516 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 100 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 100 iterations?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

100