NO SPACES, TABS, DOTS, BRACKETS OR EXTRANEOUS CHARACTERS.

Answer format: X9,Y9,Z9,W9

Response Type : Alphanumeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Answers Case Sensitive : No

Text Areas : PlainText

Possible Answers :

NIL

Deep Learning

Section Id :	64065341412
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 :	64065388807
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Number : 74 Question Id : 640653614051 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 :

6406532049909. ✔ YES

6406532049910. ✖ NO

Sub-Section Number :	2
Sub-Section Id :	64065388808
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 75 Question Id : 640653614052 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 of the following threshold θ value of MP neuron implements AND Boolean function denoted by $f(x)$? Assume that the number of inputs x_i to the neuron is seven and the neuron does not have any inhibitory inputs.

$$f(x) = \begin{cases} 1, & \text{if } \sum_{i=0}^{6} x_i > \theta \\ 0, & otherwise \end{cases}$$

Options :

6406532049911. ✖ 0

6406532049912. ✖ -6

6406532049913. ✔ 6

6406532049914. ✖ 7

6406532049915. ✖ -7

Question Number : 76 Question Id : 640653614071 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

Consider the statement “the attention mechanism in RNN based Encoder- Decoder architecture helps the decoder to understand the context of words in a given sentence”. The statement is

Options :

6406532049952. ✔ True

6406532049953. ✖ False

Sub-Section Number :	3
Sub-Section Id :	64065388809
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 77 Question Id : 640653614053 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 MP neuron takes in 7 Boolean inputs $(x_0, \cdots, x_6)$ and produces the Boolean output y. Assume none of the inputs is inhibitory. Select all true statements

Options :

6406532049916. ✔ There are 2^{2^7} possible Boolean functions

6406532049917. ✖ There are 2^7 possible Boolean functions

6406532049918. ✔ The function $y = min(x_0, \cdots, x_6)$ is linearly separable

6406532049919. ✖ The function $y = min(x_0, \cdots, x_6)$ is not linearly separable

Question Number : 78 Question Id : 640653614062 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 we have a deep Feed Forward Fully Connected Neural Network. The network is observed to have a high variance. Then, which of the following techniques regularize the parameter of the network to reduce the high variance ?

Options :

6406532049934. ✔ Adding *L2* norm of weights to the loss function

6406532049935. ✔ Adding a noise to the input samples

6406532049936. ✔ Adding a noise to the output prediction

6406532049937. ✔ Adding more samples to the dataset by augmenting existing samples using some augmentation techniques

6406532049938. ✖ Dropping hidden layers in a neural network randomly during training

Sub-Section Number :	4
Sub-Section Id :	64065388810
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 79 Question Id : 640653614054 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

The logistic sigmoid function is defined as follows,

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

The parameters are initialized to $w = 1$ $b = 1$. Suppose the loss in defined as

$$L = \frac{1}{2}(f(x) - y)^2$$

where y is the true value. Compute the gradient of b for the following sample $x = 0.2, y = 0$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.024 to 0.029

Question Number : 80 Question Id : 640653614059 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 training set that contains 10 samples to train a neural network. Further, mini-batch GD algorithm has been chosen to update the parameters of the network with a batch size of 2. Suppose that we use an exponentially decaying learning rate scheme $\eta_t = 2\exp(-\frac{t}{4})$ and train the model for 2 epochs. What will be the value of the learning rate η_t at the end of the training? Assume, t starts from zero.
Enter the answer to 3 decimal points (that is, if your answer is -0.12145, then enter it as -0.121)

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.19 to 0.23

Question Number : 81 Question Id : 640653614063 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 input image of size $256 \times 256 \times 3$, where 3 is the number of channels. Suppose we apply a set of convolution kernels on the input image that generates the output feature maps of size $248 \times 248 \times 32$. How many parameters (including bias) do the kernels have? Assume stride $(s = 1)$ and padding $p = 1$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

11648

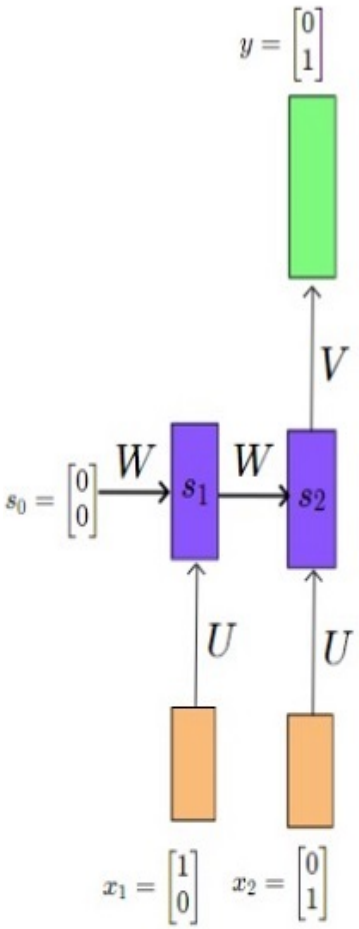
Question Number : 82 Question Id : 640653614070 Question Type : SA Calculator : None

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

Correct Marks : 3

Question Label : Short Answer Question

Consider a simple RNN for a binary sequence classification problem.



Suppose the weight matrices U, V, W are initialized as follows

$$W = U = V = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

The state vector s_t is computed as follows

$$s_t = tanh(Ux_t + Ws_{t-1})$$

What is the loss value? Use cross entropy loss with natural logarithm.

Note:In all your calculations, consider only the first two decimal places of any number (such as inputs, intermediate results..). That is, if the number is -1.0234, take it as -1.02.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.12 to 0.26

Sub-Section Number :	5
Sub-Section Id :	64065388811
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653614055 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 : (83 to 85)

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 : 83 Question Id : 640653614056 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 produce 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 : 84 Question Id : 640653614057 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 learnable parameters (including bias) does the network have? 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 Number : 85 Question Id : 640653614058 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

For the given neural network configuration, we can replace the output layer with softmax activation by logistic sigmoid and still use cross entropy loss. The statement is

Options :

6406532049923. ✖ True

6406532049924. ✔ False

Question Id : 640653614064 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 : (86 to 87)

Question Label : Comprehension

Consider a sentence inside the quote “I may be wrong, and you may be right, and by an effort, we may get nearer to the truth”

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 86 Question Id : 640653614065 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 :

16

Question Number : 87 Question Id : 640653614066 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 all words in the vocabulary are represented using one-hot-encoded vector of size $|V|$. Then compute the ordered pair-wise (that is, Cartesian product of $V \times V$) cosine similarity between word representations and enter their sum.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

16

Sub-Section Number :	6
Sub-Section Id :	64065388812
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 88 Question Id : 640653614060 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 5

Question Label : Multiple Choice Question

The update rule for the ADAM (Adaptive Moments) optimization algorithm is given below,

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla w_t$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla w_t)^2$$
$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}}\hat{m}_t$$

Here, $0 \leq \beta_1 < 1$ and $0 \leq \beta_2 < 1$ and t starts from zero (that is, $t = 0, 1, 2, \cdots$). Both m_t and v_t are initialized to zero. However, the update rule uses the bias corrected version of m_t and v_t. Which of the following is the bias corrected version of m_t?

Helper:

$$m_t = (1-\beta_1)\sum_{\tau=0}^{t} \beta_1^{t-\tau}\nabla w_\tau$$

and assume that $E[\nabla w_\tau] = E[\nabla w] \quad \forall \tau$, if required.

Options :

6406532049926. ✖ $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$

6406532049927. ✔ $\hat{m}_t = \frac{m_t}{1-\beta_1^{t+1}}$

6406532049928. ✖ $\hat{m}_t = m_t$

6406532049929. ✖ $\hat{m}_t = \frac{m_t}{1-t\beta_1^t}$

Sub-Section Number :	7
Sub-Section Id :	64065388813
Question Shuffling Allowed :	Yes
Is Section Default? :	null

Question Number : 89 Question Id : 640653614061 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 target variable $y = f(x) + \epsilon$ that is related to x, where ϵ is a random variable (noise) distributed normally. About 1000 points are sampled from the true function $f(x)$ to form a training set. Suppose that a prediction model $\hat{f}(x)$ is sufficiently complex in that $f(x) \subset \hat{f}(x)$. Then, the statement that the training error is lower bounded by σ^2 (that is, the variance of the noise) is

Options :

6406532049930. ✖ True, due to the presence of noise in the target

6406532049931. ✖ True, due to the high variance of the prediction model

6406532049932. ✔ False, due to the high variance of the prediction model

6406532049933. ✖ False, due to zero mean of the noise added to the target

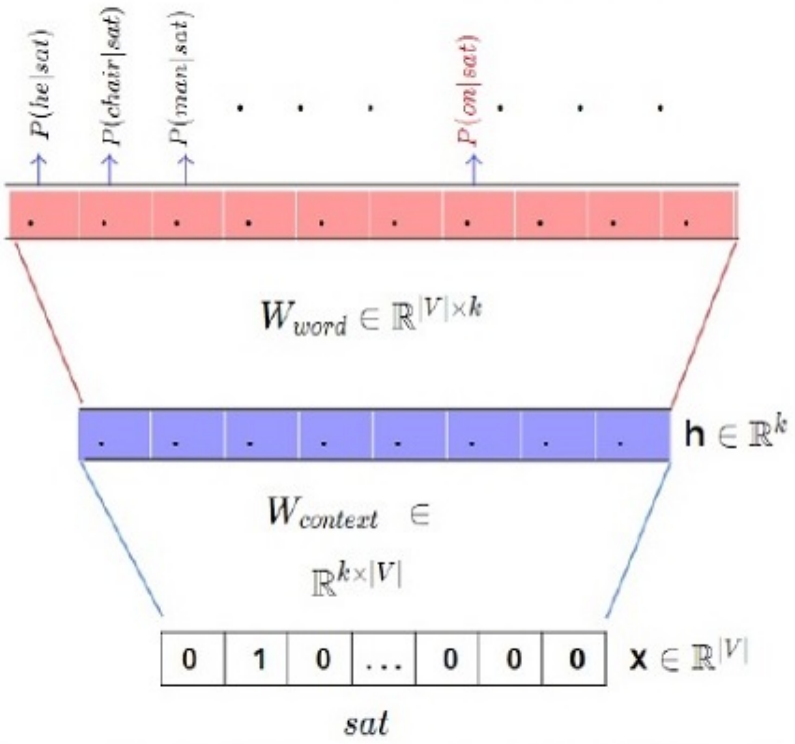
Sub-Section Number :	8
Sub-Section Id :	64065388814
Question Shuffling Allowed :	No
Is Section Default? :	null

Question Id : 640653614067 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 : (90 to 91)

Question Label : Comprehension

Consider a model shown below that learns the distributed vector representation of words by learning to predict the target word v_w given the context word u_c. Here, v_w and u_c are the vector representation of target word at index w of the output vocabulary and context word at index c of the input vocabulary, respectively.



In the diagram, $|V|$ denotes the size of the vocabulary, $W_{context}$ and W_{word} are learnable parameters. The vector representation of all context words are arranged as columns of $W_{context}$ and the vector representation for all target words are arranged as row vectors in W_{word}. The parameters are initialized randomly. The input x is one-hot-representation of a word in the input vocabulary. Assume that the size of both input and output vocabulary are equal.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 90 Question Id : 640653614068 Question Type : MCQ Is Question Mandatory : No Calculator : None Response Time : N.A Think Time : N.A Minimum Instruction Time : 0

Correct Marks : 5

Question Label : Multiple Choice Question

Suppose the input is x (one hot representation of context word) and the corresponding label is y (one hot representation of target word). The quantities $h, u_c, \hat{y}$ are computed as follows,

$$h = u_c = W_{context}\ x, \quad z = W_{word}\ u_c \quad \hat{y} = softmax(z)$$

Choose the expression that the model has to minimize using cross entropy loss (Assume natural logarithm where required).

Options :

6406532049942. ✔ $-v_w u_c + \log\left(\sum_{w' \in V} exp(v_{w'} u_c)\right)$

6406532049943. ✖ $-u_c v_w^T + \log\left(\sum_{w' \in V} exp(u_c v_{w'}^T)\right)$

6406532049944. ✖ $-u_c v_w^T - \log\left(\sum_{w' \in V} exp(u_c v_{w'}^T)\right)$

6406532049945. ✖ $-v_w^T u_c - \log\left(\sum_{w' \in V} exp(v_{w'}^T u_c)\right)$

Question Number : 91 Question Id : 640653614069 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 compute the gradients and update the representations with $\eta = 1$. Choose the correct statement(s)

Options :

6406532049946. ✔ Suppose the model predicts target word v_w with probability score of 1 (that is, $\hat{y_w} = 1$). Then no elements in v_w will get modified after one iteration (i.e, parameter update).

6406532049947. ✔ Suppose the model predicts target word v_w with probability score of 1 (that is, $\hat{y_w} = 1$). Then no elements in $v_{w'}, (w' \neq w)$ will get modified after one iteration (i.e, parameter update).

6406532049948. ✔ Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of v_w will be modified as $v_w = v_w + 0.5u_c^T$

6406532049949. ✔ Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of $v_{w'}$ will be modified as $v_{w'} = v_{w'} - \hat{y}_{w'}u_c^T$

6406532049950. ✖ Suppose the model predicts target word v_w with probability score of 0.5 (that is, $\hat{y_w} = 0.5$). Then the elements of $v_{w'}$ will be modified as $v_{w'} = v_{w'} + \hat{y}_{w'}u_c^T$