Deep Learning

Section Id :	64065364133
Section Number :	5
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	25
Number of Questions to be attempted :	25
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 :	No
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	640653134103
Question Shuffling Allowed :	No

Question Number : 94 Question Id : 640653904136 Question Type : MCQ Calculator : Yes

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 :

6406533044517. ✓ YES

6406533044518. ✖ NO

Question Number : 95 Question Id : 640653904137 Question Type : MCQ Calculator : Yes

Correct Marks : 0

Question Label : Multiple Choice Question

Note: If the base of log is not mentioned, use log base e.

Options :

6406533044519. ✓ Instructions has been mentioned above.

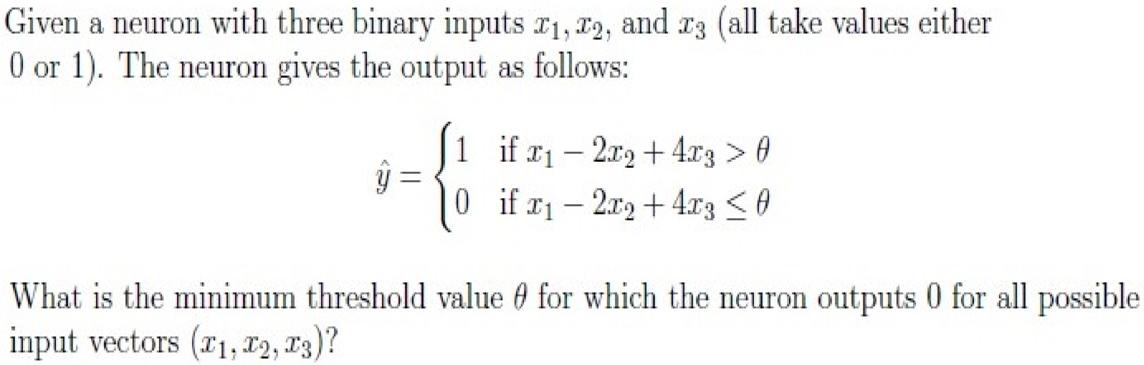
6406533044520. ✖ This Instructions is just for a reference & not for an evaluation.

Sub-Section Number :	2
Sub-Section Id :	640653134104
Question Shuffling Allowed :	Yes

Question Number : 96 Question Id : 640653904138 Question Type : SA Calculator : None

Correct Marks : 2

Question Label : Short Answer Question



Given a neuron with three binary inputs x_1, x_2, and x_3 (all take values either 0 or 1). The neuron gives the output as follows:

$$\hat{y} = \begin{cases} 1 & \text{if } x_1 - 2x_2 + 4x_3 > \theta \\ 0 & \text{if } x_1 - 2x_2 + 4x_3 \leq \theta \end{cases}$$

What is the minimum threshold value θ for which the neuron outputs 0 for all possible input vectors (x_1, x_2, x_3)?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

5

Question Number : 97 Question Id : 640653904141 Question Type : SA Calculator : None

Correct Marks : 2

Question Label : Short Answer Question

Consider a dataset with 150 samples and a batch size of 15. If each minibatch iteration contributes an average loss of 0.4, what will be the total loss after 15 epochs?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

60

Question Number : 98 Question Id : 640653904149 Question Type : SA Calculator : None

Correct Marks : 2

Question Label : Short Answer Question

If you use hierarchical softmax with a binary tree where each leaf node represents a word in the vocabulary, and the vocabulary size (V) is 16000, how many binary classifiers are needed?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

15998 to 16001

Question Number : 99 Question Id : 640653904150 Question Type : SA Calculator : None

Correct Marks : 2

Question Label : Short Answer Question

In a Skip-gram model with a vocabulary size $V = 100$, an embedding dimension $D = 10$, and a window size of 3 (on each side), using negative sampling with 5 negative samples per positive sample, what is the total number of parameters in the model?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

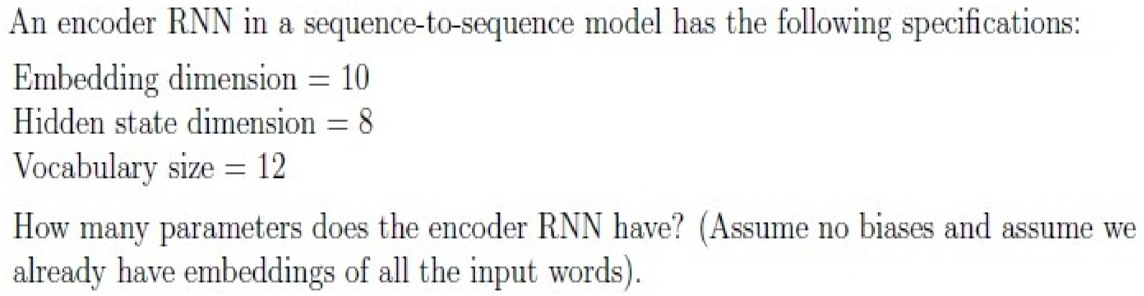
Text Areas : PlainText

Possible Answers :

2000

Question Number : 100 Question Id : 640653904159 Question Type : SA Calculator : None
Correct Marks : 2

Question Label : Short Answer Question



An encoder RNN in a sequence-to-sequence model has the following specifications:

Embedding dimension = 10
Hidden state dimension = 8
Vocabulary size = 12

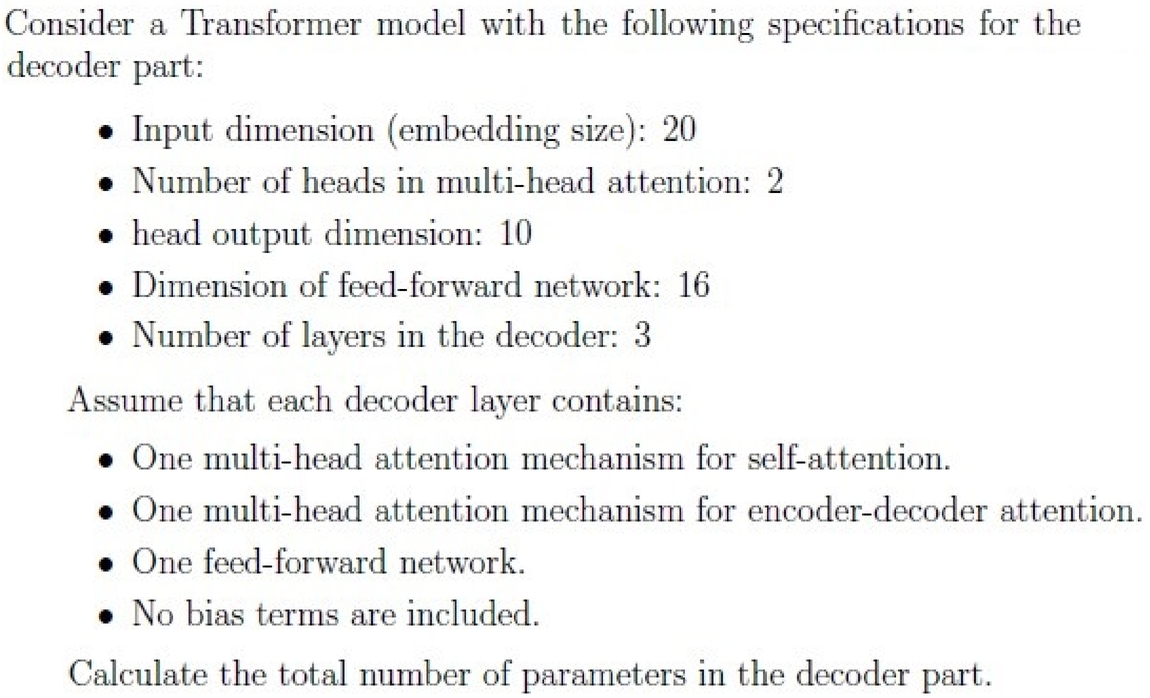
How many parameters does the encoder RNN have? (Assume no biases and assume we already have embeddings of all the input words).

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes
Answers Type : Equal
Text Areas : PlainText
Possible Answers :

176

Question Number : 101 Question Id : 640653904163 Question Type : SA Calculator : None
Correct Marks : 2

Question Label : Short Answer Question



Consider a Transformer model with the following specifications for the decoder part:

- Input dimension (embedding size): 20
- Number of heads in multi-head attention: 2
- head output dimension: 10
- Dimension of feed-forward network: 16
- Number of layers in the decoder: 3

Assume that each decoder layer contains:

- One multi-head attention mechanism for self-attention.
- One multi-head attention mechanism for encoder-decoder attention.
- One feed-forward network.
- No bias terms are included.

Calculate the total number of parameters in the decoder part.

Response Type : Numeric
Evaluation Required For SA : Yes
Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

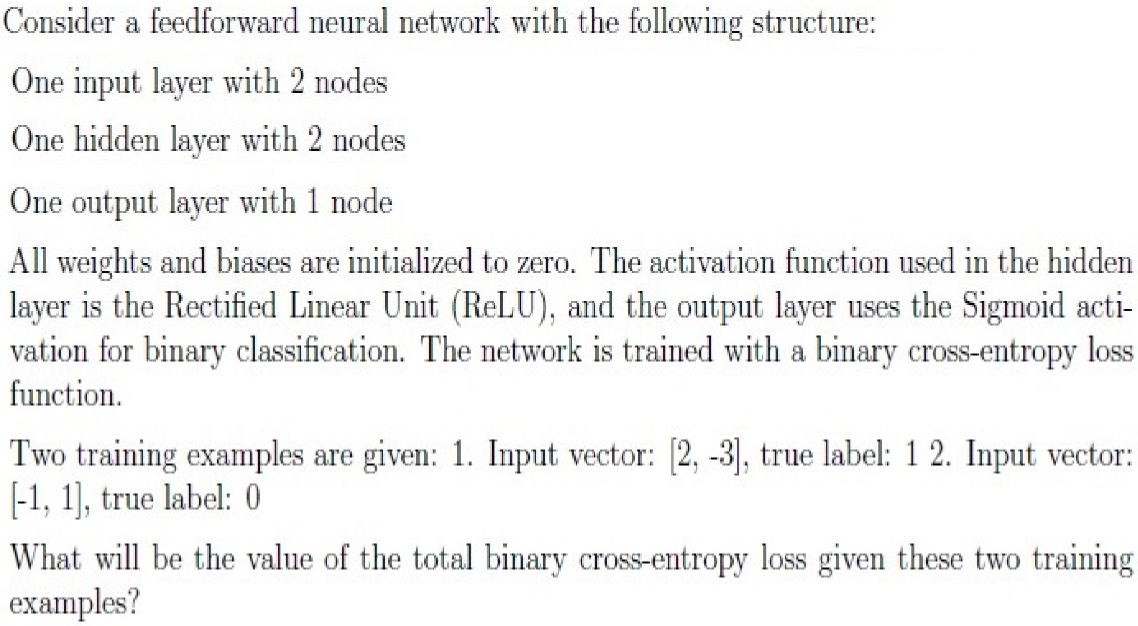
3680

Sub-Section Number :	3
Sub-Section Id :	640653134105
Question Shuffling Allowed :	Yes

Question Number : 102 Question Id : 640653904139 Question Type : SA Calculator : None

Correct Marks : 3

Question Label : Short Answer Question



Consider a feedforward neural network with the following structure:

One input layer with 2 nodes

One hidden layer with 2 nodes

One output layer with 1 node

All weights and biases are initialized to zero. The activation function used in the hidden layer is the Rectified Linear Unit (ReLU), and the output layer uses the Sigmoid activation for binary classification. The network is trained with a binary cross-entropy loss function.

Two training examples are given: 1. Input vector: [2, -3], true label: 1 2. Input vector: [-1, 1], true label: 0

What will be the value of the total binary cross-entropy loss given these two training examples?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

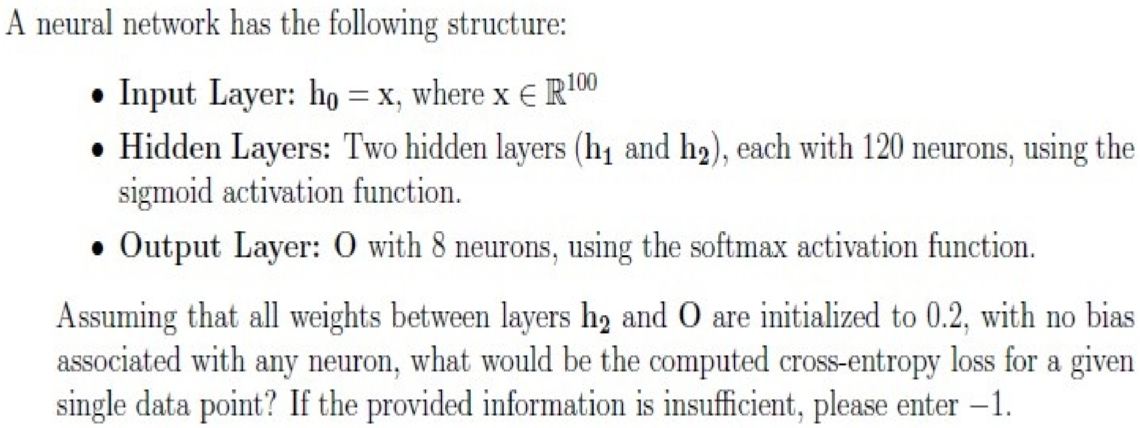
Possible Answers :

1 to 1.5

Question Number : 103 Question Id : 640653904140 Question Type : SA Calculator : None

Correct Marks : 3

Question Label : Short Answer Question



A neural network has the following structure:

- **Input Layer:** $\mathbf{h_0} = \mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{100}$
- **Hidden Layers:** Two hidden layers ($\mathbf{h_1}$ and $\mathbf{h_2}$), each with 120 neurons, using the sigmoid activation function.
- **Output Layer:** O with 8 neurons, using the softmax activation function.

Assuming that all weights between layers $\mathbf{h_2}$ and O are initialized to 0.2, with no bias associated with any neuron, what would be the computed cross-entropy loss for a given single data point? If the provided information is insufficient, please enter −1.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

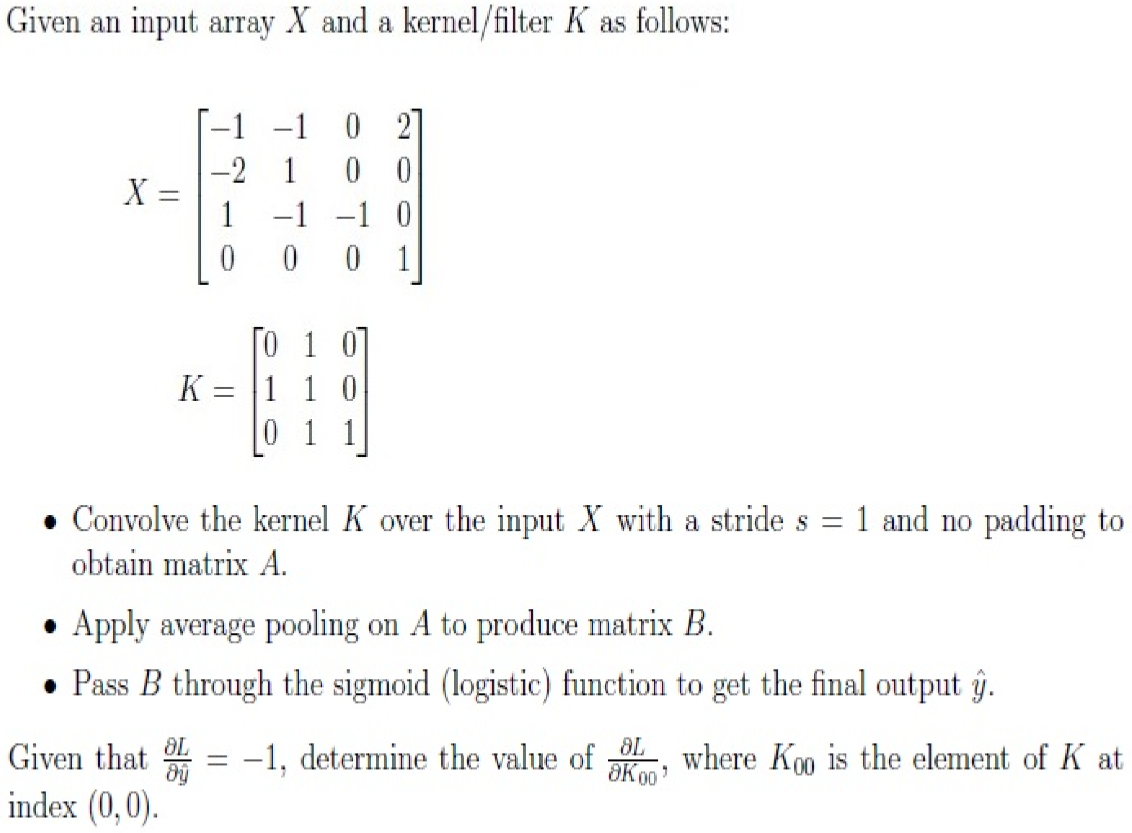
Possible Answers :

2 to 2.2

Question Number : 104 Question Id : 640653904148 Question Type : SA Calculator : None

Correct Marks : 3

Question Label : Short Answer Question



Given an input array X and a kernel/filter K as follows:

$$X = \begin{bmatrix} -1 & -1 & 0 & 2 \\ -2 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$K = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

- Convolve the kernel K over the input X with a stride $s = 1$ and no padding to obtain matrix A.
- Apply average pooling on A to produce matrix B.
- Pass B through the sigmoid (logistic) function to get the final output $\hat{y}$.

Given that $\frac{\partial L}{\partial \hat{y}} = -1$, determine the value of $\frac{\partial L}{\partial K_{00}}$, where K_{00} is the element of K at index $(0, 0)$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

0.02 to 0.22

Sub-Section Number :	4
Sub-Section Id :	640653134106
Question Shuffling Allowed :	Yes

Question Number : 105 Question Id : 640653904147 Question Type : SA Calculator : None

Correct Marks : 1

Question Label : Short Answer Question

What is the derivative of the ReLU activation function at x = 10?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1

Sub-Section Number :	5
Sub-Section Id :	640653134107
Question Shuffling Allowed :	Yes

Question Number : 106 Question Id : 640653904142 Question Type : MCQ Calculator : Yes

Correct Marks : 1

Question Label : Multiple Choice Question

In terms of convergence speed, which gradient descent method can show the most rapid progress initially but may suffer from high variance in updates?

Options :

6406533044525. ✖ Batch Gradient Descent

6406533044526. ✔ Stochastic Gradient Descent

6406533044527. ✖ Mini-batch Gradient Descent

6406533044528. ✖ None of these

Question Number : 107 Question Id : 640653904143 Question Type : MCQ Calculator : Yes

Correct Marks : 1

Question Label : Multiple Choice Question

How does the use of early stopping in training a neural network affect the model’s performance on unseen data?

Options :

6406533044529. ✔ It usually leads to better performance on unseen data by preventing overfitting

6406533044530. ✗ It generally worsens the performance on unseen data by halting training too early

6406533044531. ✗ It does not affect the performance on unseen data

6406533044532. ✗ It increases the risk of overfitting by allowing more epochs of training

Sub-Section Number :	6
Sub-Section Id :	640653134108
Question Shuffling Allowed :	Yes

Question Number : 108 Question Id : 640653904144 Question Type : MCQ Calculator : Yes

Correct Marks : 2

Question Label : Multiple Choice Question

Which of the following statements is/are not true with respect to a dropout rate of 0.2?

Options :

6406533044533. ✓ The exact number of neurons dropped in each iteration will always be exactly 20%.

6406533044534. ✗ The exact number of neurons dropped and retained can vary slightly from one iteration to another due to the probabilistic nature of dropout.

6406533044535. ✗ Each neuron has a 20% chance of being dropped during any given training iteration.

6406533044536. ✗ Over many training iterations, the average percentage of retained neurons will approximate 80%.

Question Number : 109 Question Id : 640653904145 Question Type : MCQ Calculator : Yes

Correct Marks : 2

Question Label : Multiple Choice Question

In the context of unsupervised pretraining of artificial neural networks, which of the following statements accurately describes the role and benefits of using unsupervised pretraining techniques for initializing a neural network?

Options :

6406533044537. ✗ Unsupervised pretraining methods help in identifying patterns in unlabeled data, which can be used to initialize weights and reduce the risk of overfitting in the subsequent supervised training phase.

6406533044538. ✗ The primary purpose of unsupervised pretraining is to generate synthetic data that can be used to expand the training dataset for the neural network, leading to more robust performance.

6406533044539. ✓ Unsupervised pretraining enables the network to learn a hierarchical representation of data, which can be fine-tuned with supervised learning, enhancing the model’s generalization capabilities.

6406533044540. ✗ Using unsupervised pretraining techniques ensures that the neural network can skip the initial training phase, directly achieving high accuracy on test data without further training.

Question Number : 110 Question Id : 640653904146 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

What are the maximum values of the derivatives of sigmoid and tanh?

Options :

6406533044541. ✗ 1, 1

6406533044542. ✗ 0.5, 0.5

6406533044543. ✗ 0, 0.5

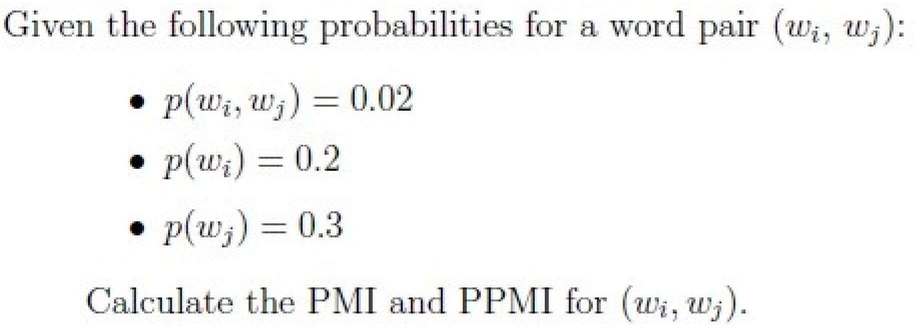
6406533044544. ✗ 0.5, 0

6406533044545. ✓ 0.25, 1

6406533044546. ✗ 0.25, 0.5

Question Number : 111 Question Id : 640653904151 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question



Given the following probabilities for a word pair (w_i, w_j):

- $p(w_i, w_j) = 0.02$
- $p(w_i) = 0.2$
- $p(w_j) = 0.3$

Calculate the PMI and PPMI for (w_i, w_j).

Options :

6406533044551. ✓ -0.477, 0

6406533044552. ✗ -0.301, 0

6406533044553. ✗ 0.301, 0

6406533044554. ✗ 0.477, 0.477

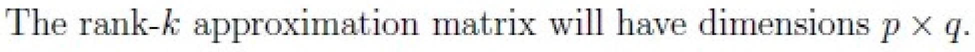
6406533044555. ✗ -0.477, 0.477

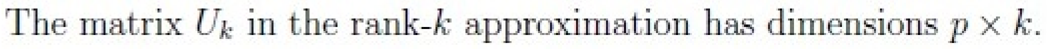
Question Number : 112 Question Id : 640653904152 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

Given a matrix *A* with dimensions *p* × *q*, which of the following statements is NOT true regarding the rank-*k* approximation of *A* obtained through Singular Value Decomposition (SVD)?

Options :

6406533044556. ✗ 
The rank-k approximation matrix will have dimensions $p \times q$.

6406533044557. ✗ 
The matrix U_k in the rank-k approximation has dimensions $p \times k$.

6406533044558. ✖ The matrix Σ_k in the rank-k approximation has dimensions $k \times k$.

6406533044559. ✓ The rank-k approximation matrix will have dimensions $k \times k$.

Question Number : 113 Question Id : 640653904157 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

Consider an encoder-decoder model trained with a batch size of 64. Each input sequence has a length of 12 tokens, and each output sequence has a length of 18 tokens. How many computational steps do the encoder and decoder take per batch respectively during training?

Options :

6406533044569. ✓ 768, 1152

6406533044570. ✖ 1152, 768

6406533044571. ✖ 12, 18

6406533044572. ✖ 18, 12

6406533044573. ✖ 12, 1

6406533044574. ✖ 1, 18

Question Number : 114 Question Id : 640653904158 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

In an encoder-decoder model, what is the significance of the context vector?

Options :

6406533044575. ✖ It stores the hidden states of the decoder.

6406533044576. ✓ It summarizes the input sequence information to be used by the decoder.

6406533044577. ✖ It acts as the final output of the decoder.

6406533044578. ✖ It initiates the hidden states of the encoder.

6406533044579. ✖ It contains the parameters of the attention mechanism.

Question Number : 115 Question Id : 640653904160 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

Given the attention weights $\alpha_{t,1}$ = 0.3, $\alpha_{t,2}$ = 0.4, $\alpha_{t,3}$ = 0.3 and the corresponding encoder hidden states h_1 = [2, 1, 0], h_2 = [1, 2, 1], h_3 = [0, 1, 2], calculate the context vector c_t.

Options :

6406533044581. ✖ [0.7, 1.4, 0.9]

6406533044582. ✓ [1.1, 1.6, 1.3]

6406533044583. ✖ [0.6, 1.3, 0.9]

6406533044584. ✖ [1.1, 1.4, 1.1]

Question Number : 116 Question Id : 640653904161 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

In the Transformer model, what is the purpose of the multi-head attention mechanism?

Options :

6406533044585. ✓ To allow the model to focus on different parts of the input sequence using different sets of attention weights.

6406533044586. ✖ To average the attention weights across multiple heads for more stable training.

6406533044587. ✖ To reduce the dimensionality of the input sequence before applying attention.

6406533044588. ✖ To apply attention in parallel across multiple layers of the Transformer model.

Sub-Section Number :	7
Sub-Section Id :	640653134109
Question Shuffling Allowed :	Yes

Question Number : 117 Question Id : 640653904162 Question Type : MSQ Calculator : Yes
Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question

In the context of the Transformer model’s encoder-decoder architecture, which of the following statements are correct?

Options :

6406533044589. ✓ The encoder processes the entire input sequence at once at a particular time step, and its output serves as the context for the decoder during generation.

6406533044590. ✓ The multi-head attention mechanism in the decoder allows the model to focus on different parts of the encoder’s output while generating the sequence.

6406533044591. ✖ The decoder applies self-attention over its entire sequence of inputs without any restrictions, allowing it to consider all future tokens at once.

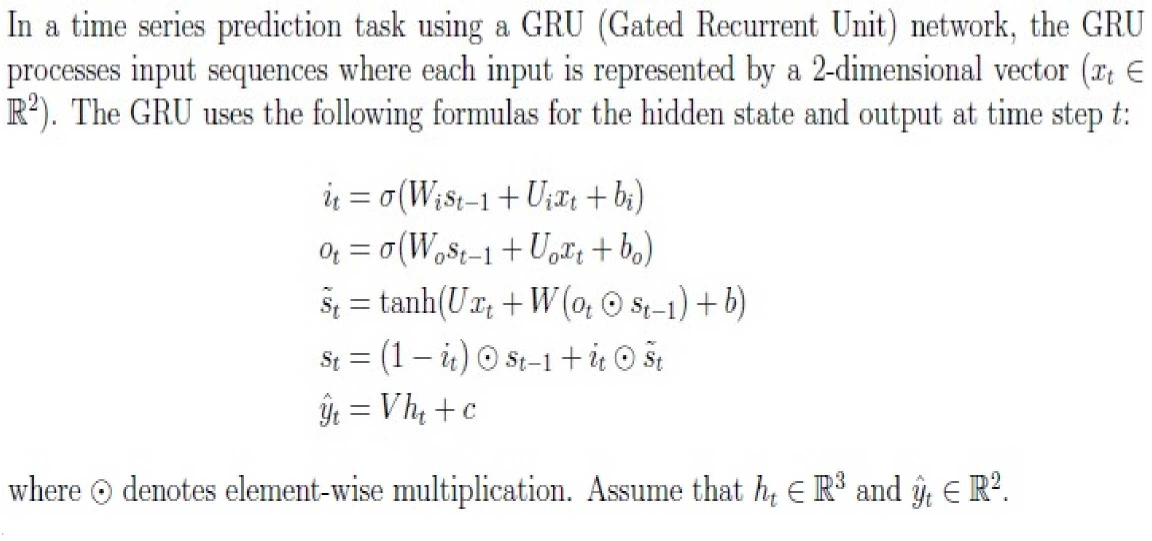
6406533044592. ✓ The decoder’s self-attention mechanism includes a masking component to prevent attending to future positions, ensuring the model generates outputs one step at a time.

Sub-Section Number :	8
Sub-Section Id :	640653134110
Question Shuffling Allowed :	No

Question Id : 640653904153 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix Calculator : None

Question Numbers : (118 to 120)

Question Label : Comprehension



In a time series prediction task using a GRU (Gated Recurrent Unit) network, the GRU processes input sequences where each input is represented by a 2-dimensional vector ($x_t \in \mathbb{R}^2$). The GRU uses the following formulas for the hidden state and output at time step t:

$$
\begin{aligned}
i_t &= \sigma(W_i s_{t-1} + U_i x_t + b_i) \\
o_t &= \sigma(W_o s_{t-1} + U_o x_t + b_o) \\
\tilde{s}_t &= \tanh(U x_t + W(o_t \odot s_{t-1}) + b) \\
s_t &= (1 - i_t) \odot s_{t-1} + i_t \odot \tilde{s}_t \\
\hat{y}_t &= V h_t + c
\end{aligned}
$$

where $\odot$ denotes element-wise multiplication. Assume that $h_t \in \mathbb{R}^3$ and $\hat{y}_t \in \mathbb{R}^2$.

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 118 Question Id : 640653904154 Question Type : SA Calculator : None

Correct Marks : 2

Question Label : Short Answer Question

Given that the GRU processes sequences of length 6 (*T* = 6), what is the total number of parameters (including biases) in the network?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

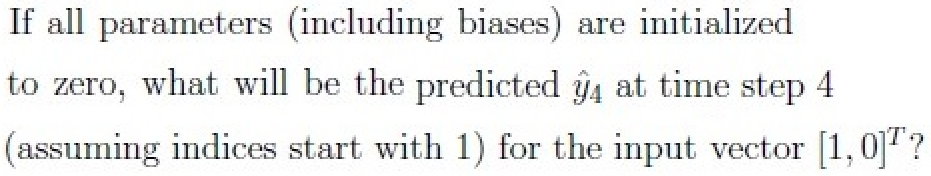
Possible Answers :

113

Question Number : 119 Question Id : 640653904155 Question Type : MCQ Calculator : Yes

Correct Marks : 2

Question Label : Multiple Choice Question



If all parameters (including biases) are initialized to zero, what will be the predicted $\hat{y}_4$ at time step 4 (assuming indices start with 1) for the input vector $[1, 0]^T$?

Options :

6406533044561. ✓ $[0.5, 0.5]^T$

6406533044562. ✖ $[0.3, 0.7]^T$

6406533044563. ✖ $[0.4, 0.6]^T$

6406533044564. ✗ $[0.25, 0.75]^T$

Question Number : 120 Question Id : 640653904156 Question Type : MCQ Calculator : Yes
Correct Marks : 2

Question Label : Multiple Choice Question

If the GRU is modified such that $\hat{y}_t \in \mathbb{R}^4$, and
all parameters are initialized randomly, what will be
the dimensionality of the weight matrix V?

Options :

6406533044565. ✗ 3 × 2

6406533044566. ✓ 3 × 4

6406533044567. ✗ 4 × 3

6406533044568. ✗ 2 × 4