Sub-Section Number : 1

Sub-Section Id : 640653230613

Question Shuffling Allowed : No

Question Number : 103 Question Id : 6406531425035 Question Type : MCQ

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 :

6406534763347. ✓

YES

6406534763348. ✗

NO

Sub-Section Number : 2

Sub-Section Id : 640653230614

Question Shuffling Allowed : Yes

Question Number : 104 Question Id : 6406531425036 Question Type : MCQ

Correct Marks : 2

Question Label : Multiple Choice Question

Consider the following two statements regarding model performance:

Statement 1: A model achieving zero training loss is guaranteed to perform well on unseen data.

Statement 2: Incorporating a regularization term in the loss function may lead to higher training loss but lower generalization error.

Which of the following options is correct?

Options :

6406534763349. ✗

Both Statement 1 and Statement 2 are true.

6406534763350. ✗

Statement 1 is true, but Statement 2 is false.

6406534763351. ✗

None of these.

6406534763352. ✓

Statement 1 is false, but Statement 2 is true.

Question Number : 105 Question Id : 6406531425037 Question Type : MCQ

Correct Marks : 2

Question Label : Multiple Choice Question

How does unsupervised layerwise pretraining help in alleviating the vanishing gradient problem?

Options :

6406534763353. ✓

It allows the network to learn a better representation of the data in each layer, which leads to better-initialized weights for subsequent supervised training.

6406534763354. ✖

It adds skip connections to the network, which are then removed before the supervised training.

6406534763355. ✖

It replaces the sigmoid functions with ReLU functions during the pretraining phase.

6406534763356. ✖

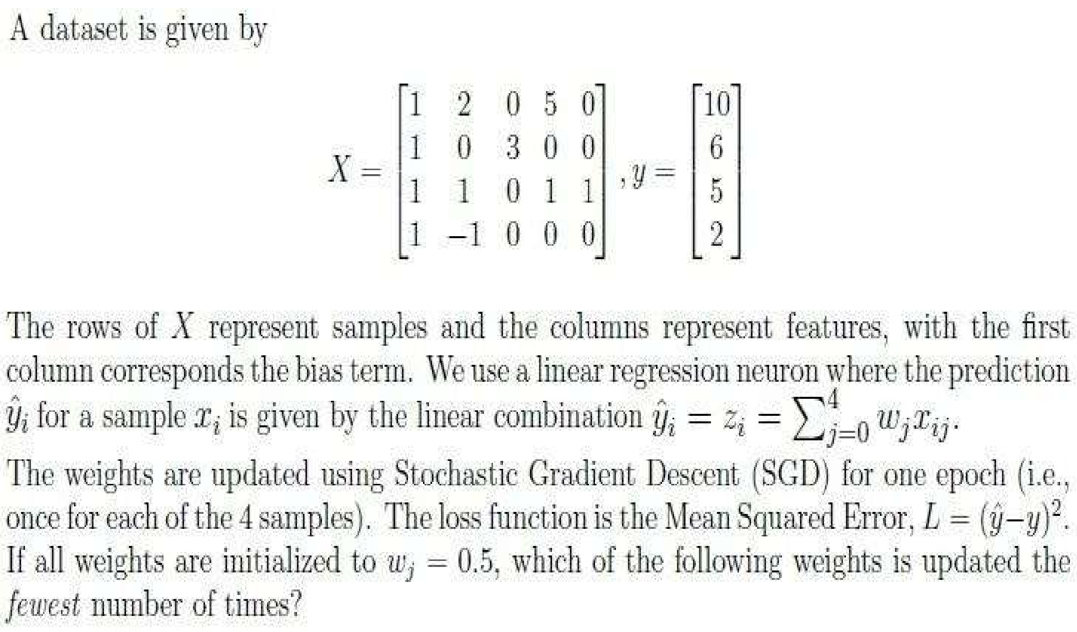
It regularizes the network’s weights, making them smaller and less likely to cause the gradients to explode.

Sub-Section Number :	3
Sub-Section Id :	640653230615
Question Shuffling Allowed :	Yes

Question Number : 106 Question Id : 6406531425038 Question Type : MSQ

Correct Marks : 2 Max. Selectable Options : 0

Question Label : Multiple Select Question



A dataset is given by

$$X = \begin{bmatrix} 1 & 2 & 0 & 5 & 0 \\ 1 & 0 & 3 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & 0 \end{bmatrix}, y = \begin{bmatrix} 10 \\ 6 \\ 5 \\ 2 \end{bmatrix}$$

The rows of X represent samples and the columns represent features, with the first column corresponds the bias term. We use a linear regression neuron where the prediction $\hat{y}_i$ for a sample x_i is given by the linear combination $\hat{y}_i = z_i = \sum_{j=0}^{4} w_j x_{ij}$.

The weights are updated using Stochastic Gradient Descent (SGD) for one epoch (i.e., once for each of the 4 samples). The loss function is the Mean Squared Error, $L = (\hat{y} - y)^2$. If all weights are initialized to $w_j = 0.5$, which of the following weights is updated the *fewest* number of times?

Options :

6406534763357. ✗
w0

6406534763358. ✗
w1

6406534763359. ✓
w2

6406534763360. ✗
w3

6406534763361. ✓
w4

Sub-Section Number : 4

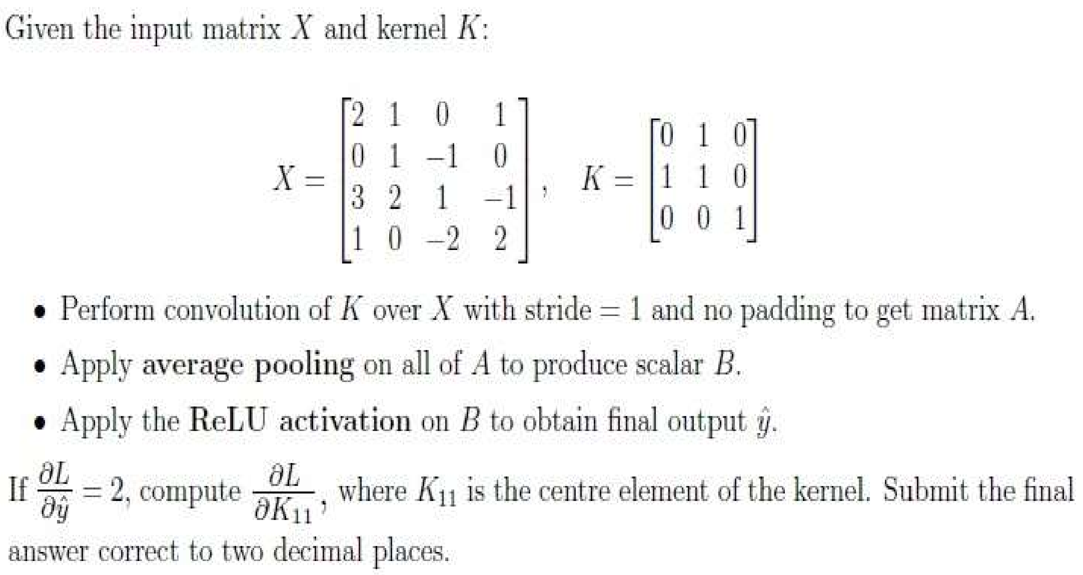
Sub-Section Id : 640653230616

Question Shuffling Allowed : Yes

Question Number : 107 Question Id : 6406531425039 Question Type : SA

Correct Marks : 3

Question Label : Short Answer Question



Given the input matrix X and kernel K:

$$X = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 3 & 2 & 1 & -1 \\ 1 & 0 & -2 & 2 \end{bmatrix}, \quad K = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Perform convolution of K over X with stride $= 1$ and no padding to get matrix A.
- Apply **average pooling** on all of A to produce scalar B.
- Apply the **ReLU activation** on B to obtain final output $\hat{y}$.

If $\frac{\partial L}{\partial \hat{y}} = 2$, compute $\frac{\partial L}{\partial K_{11}}$, where K_{11} is the centre element of the kernel. Submit the final 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 :

1 to 2

Sub-Section Number : 5

Sub-Section Id : 640653230617

Question Shuffling Allowed : No

Question Id : 6406531425040 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix

Question Numbers : (108 to 110)

Question Label : Comprehension

Suppose you are given three encoder hidden states at time t:

$\text{Score}(s_{t-1}, h_1)$

$= \begin{bmatrix}1 & 1\end{bmatrix} \tanh\left(I\begin{pmatrix}1\\2\end{pmatrix} + I\begin{pmatrix}1\\0\end{pmatrix}\right)$

$= \begin{bmatrix}1 & 1\end{bmatrix} \tanh\begin{pmatrix}2\\2\end{pmatrix}$

$= \begin{bmatrix}1 & 1\end{bmatrix}\begin{pmatrix}0.964\\0.964\end{pmatrix}$

$\text{Sum}(c_t) = 1.355$ $\quad c_t = \begin{pmatrix}0.706\\0.649\end{pmatrix}$

$$h_j = h_1 = 0.349\left(\begin{bmatrix}1\\0\end{bmatrix}\right), + 0.292\left(h_2 = \begin{bmatrix}0\\1\end{bmatrix}\right), + 0.357\left(h_3 = \begin{bmatrix}1\\1\end{bmatrix}\right)$$

Homework

The previous decoder hidden state is:

$e_{11} = 1.928$ $\quad e_{12} = 1.75$ $\quad e_{13} = 1.95$

Softmax

$\alpha_{11} = 0.349$ $\quad \alpha_{12} = 0.292$ $\quad \alpha_{13} = 0.357$

$$s_{t-1} = \begin{bmatrix}1\\2\end{bmatrix}$$

$\begin{bmatrix}1 & 1\end{bmatrix} \tanh\begin{pmatrix}1\\3\end{pmatrix}$ $\quad \begin{bmatrix}1 & 1\end{bmatrix} \tanh\begin{pmatrix}2\\3\end{pmatrix}$

1.75 $\quad 1.95$

Given the attention score function:

$$\text{score}(s_{t-1}, h_j) = V_{att}^{\top} \tanh\left(U_{att}s_{t-1} + W_{att}h_j\right)$$

where the hyperbolic tangent function is defined as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$

where

$$V_{att} = \begin{bmatrix}1\\1\end{bmatrix}, \quad U_{att} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, \quad W_{att} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$$

Based on the above data, answer the given subquestions.

Sub questions

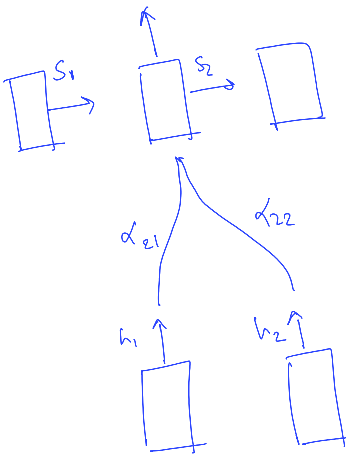
Question Number : 108 Question Id : 6406531425041 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question

Compute the attention score for hidden state h_1 using the given function.

Submit the final answer correct to two decimal places.



$\alpha_{21} = f(s_1, h_1)$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

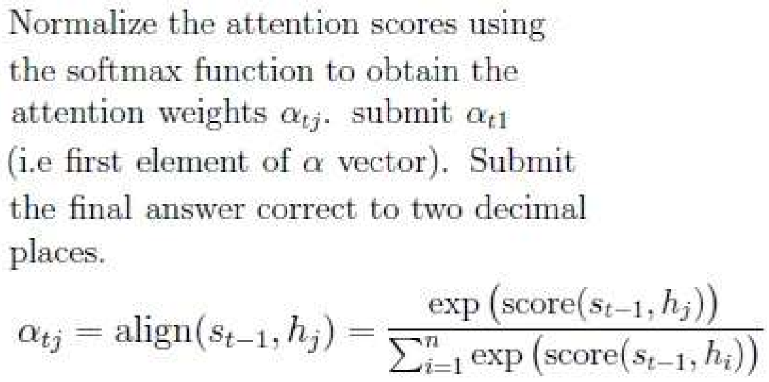
Possible Answers :

1.89 to 1.96 1,928

Question Number : 109 Question Id : 6406531425042 Question Type : SA

Correct Marks : 4

Question Label : Short Answer Question



Normalize the attention scores using the softmax function to obtain the attention weights α_{tj}. submit α_{t1} (i.e first element of α vector). Submit the final answer correct to two decimal places.

$$\alpha_{tj} = \text{align}(s_{t-1}, h_j) = \frac{\exp\left(\text{score}(s_{t-1}, h_j)\right)}{\sum_{i=1}^{n} \exp\left(\text{score}(s_{t-1}, h_i)\right)}$$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

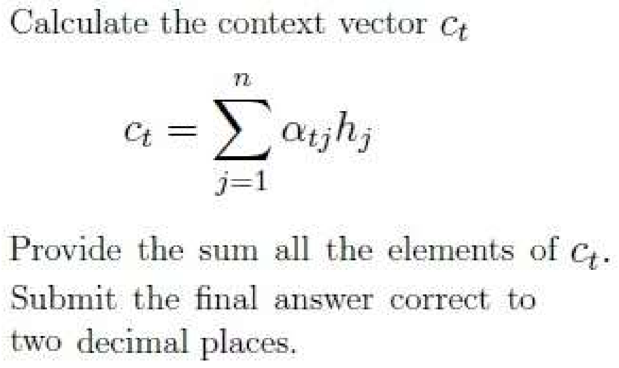
Possible Answers :

0.30 to 0.38 0.35

Question Number : 110 Question Id : 6406531425043 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question



Calculate the context vector c_t

$$c_t = \sum_{j=1}^{n} \alpha_{tj} h_j$$

Provide the sum all the elements of c_t. Submit the final 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 :

1.32 to 1.40 1,355

Sub-Section Number : 6

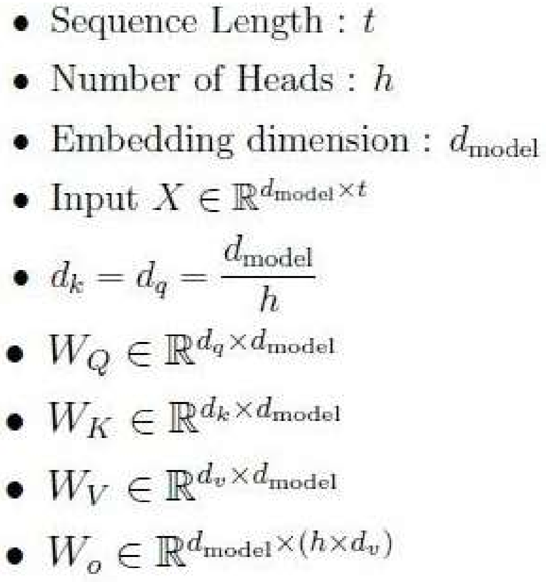
Sub-Section Id : 640653230618

Question Shuffling Allowed : No

Question Id : 6406531425044 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix

Question Numbers : (111 to 113)

Question Label : Comprehension



- Sequence Length : t
- Number of Heads : h
- Embedding dimension : d_{model}
- Input $X \in \mathbb{R}^{d_{\text{model}} \times t}$
- $d_k = d_q = \dfrac{d_{\text{model}}}{h}$
- $W_Q \in \mathbb{R}^{d_q \times d_{\text{model}}}$
- $W_K \in \mathbb{R}^{d_k \times d_{\text{model}}}$
- $W_V \in \mathbb{R}^{d_v \times d_{\text{model}}}$
- $W_o \in \mathbb{R}^{d_{\text{model}} \times (h \times d_v)}$

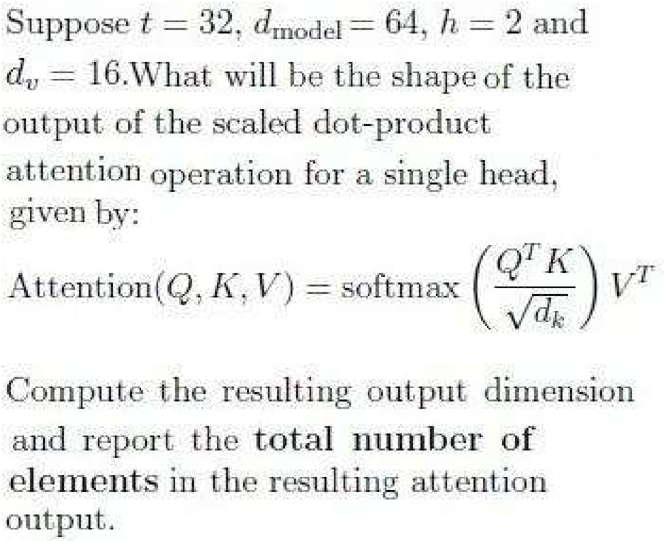
Based on the above data, answer the given subquestions.

Sub questions

Question Number : 111 Question Id : 6406531425045 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question



Suppose $t = 32$, $d_{\text{model}} = 64$, $h = 2$ and $d_v = 16$.What will be the shape of the output of the scaled dot-product attention operation for a single head, given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q^T K}{\sqrt{d_k}}\right) V^T$$

Compute the resulting output dimension and report the **total number of elements** in the resulting attention output.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

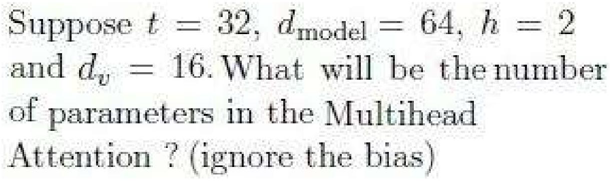
Possible Answers :

512

Question Number : 112 Question Id : 6406531425046 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question



Suppose $t = 32$, $d_{\text{model}} = 64$, $h = 2$ and $d_v = 16$. What will be the number of parameters in the Multihead Attention ? (ignore the bias)

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

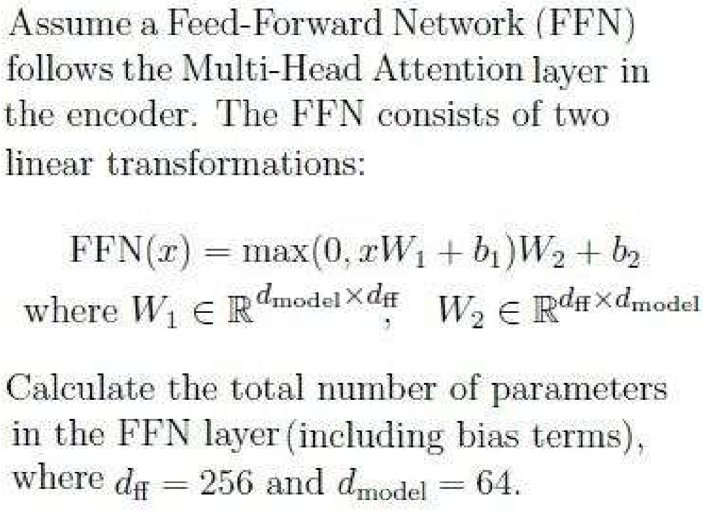
Possible Answers :

12288

Question Number : 113 Question Id : 6406531425047 Question Type : SA

Correct Marks : 1

Question Label : Short Answer Question



Assume a Feed-Forward Network (FFN) follows the Multi-Head Attention layer in the encoder. The FFN consists of two linear transformations:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

where $W_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$, $W_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$

Calculate the total number of parameters in the FFN layer (including bias terms), where $d_{\text{ff}} = 256$ and $d_{\text{model}} = 64$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

33088

Sub-Section Number : 7

Sub-Section Id : 640653230619

Question Shuffling Allowed : No

Question Id : 6406531425048 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix

Question Numbers : (114 to 116)

Question Label : Comprehension

Consider a CBOW model for learning word embeddings. The vocabulary is made up of three words, {good, bad, ugly}. *W* and *C* are the matrices that contain the word and context embeddings respectively. The columns in each matrix correspond to the embeddings. Both matrices are of shape 2 x 3:

$$W = \left[\begin{array}{ccc} \text{good} & \text{bad} & \text{ugly} \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{array}\right], \; C = \left[\begin{array}{ccc} \text{good} & \text{bad} & \text{ugly} \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{array}\right]$$

The context window is 1, meaning, the next word is predicted using just the current word as context. Recall that we use softmax to make predictions at the output.

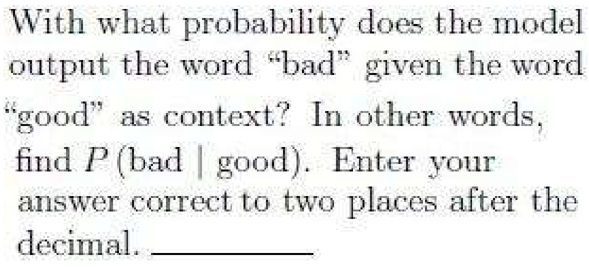
Based on the above data, answer the given subquestions.

Sub questions

Question Number : 114 Question Id : 6406531425049 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question



With what probability does the model output the word "bad" given the word "good" as context? In other words, find $P\,(\text{bad} \mid \text{good})$. Enter your answer correct to two places after the decimal. ____________

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

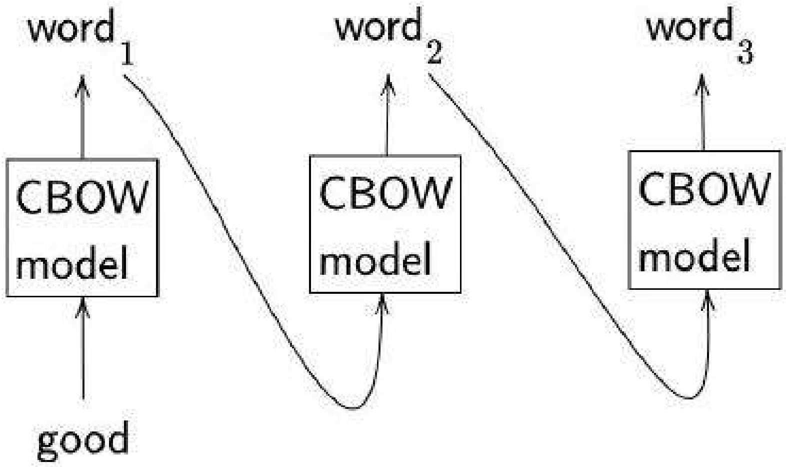
0.62 to 0.72

Question Number : 115 Question Id : 6406531425050 Question Type : MCQ

Correct Marks : 2

Question Label : Multiple Choice Question

The CBOW model is now used to generate a "sentence" or a string of words. First we pass the word "good" and retain the word with highest probability as the output, say word$_1$, which is in turn passed as input to the model. If the model is run this way for exactly three time steps, what is the sentence that it outputs? Note that the sentence here is "word$_1$ word$_2$ word$_3$".



Options :

6406534763370. ✓

bad ugly good

6406534763371. ✗

good bad ugly

6406534763372. ✗

bad bad bad

6406534763373. ✗

good good good

6406534763374. ✗

bad ugly bad

Question Number : 116 Question Id : 6406531425051 Question Type : MCQ

Correct Marks : 3

Question Label : Multiple Choice Question

Now consider updating the word embeddings using the sample "good good". The first "good" in the string is used as context and the second "good" as the true label. Use cross entropy as the loss function and run one iteration of gradient descent with η = 1 starting with the existing values for the embeddings.

Find the updated word embedding for "good" and choose the most appropriate option from below. Note that you have to compute the updated word embedding for "good" and not its context embedding.

Options :

6406534763375. ✓

(1.76, -0.24)

6406534763376. ✖
(1.24, -0.76)

6406534763377. ✖
(1.76, 1.76)

6406534763378. ✖
(1.76, -1.76)

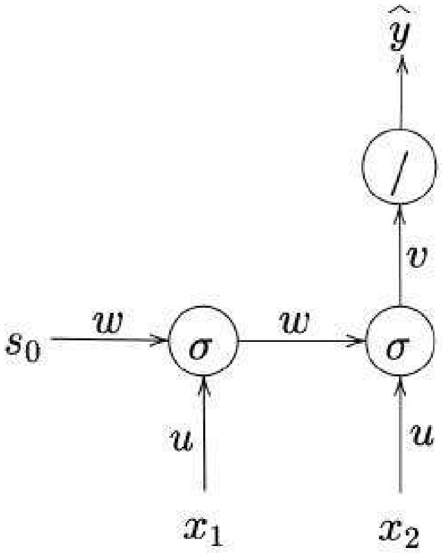
6406534763379. ✖
(1.24, -1.24)

Question Id : 6406531425052 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix

Question Numbers : (117 to 120)

Question Label : Comprehension

Consider a toy-RNN for a regression problem with input $\in \mathbb{R}$ and just one neuron in the hidden layer unrolled in time for two time steps.



The hidden layer neuron is a sigmoid neuron and the output layer neuron is linear.

- w is the weight corresponding to the recurrent connection
- u is the weight of the connection between the input and the hidden layer neuron
- v is the weight of the connection between the hidden layer neuron and the output neuron
- s_0 is the hidden state at time step $t = 0$. s_1 and s_2 are the outputs (activations) of the hidden layer neurons at time step $t = 1$ and $t = 2$ respectively.

Ignore biases everywhere. The loss is squared error and given by $L(y, \hat{y}) = \frac{1}{2} \cdot (\hat{y} - y)^2$.

Numerical Data

- $w = 1, \; u = 1, \; v = 2$
- $s_0 = 0$
- $x_1 = -\log_e 3, \; x_2 = \frac{-1}{4}$ and $y = 0$

Based on the above data, answer the given subquestions.

Sub questions

Question Number : 117 Question Id : 6406531425053 Question Type : SA
Correct Marks : 1

Question Label : Short Answer Question

Find s_1

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.25

Question Number : 118 Question Id : 6406531425054 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question

Find s_2.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.5

Question Number : 119 Question Id : 6406531425055 Question Type : SA

Correct Marks : 1

Question Label : Short Answer Question

Find $\hat{y}$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1

Question Number : 120 Question Id : 6406531425056 Question Type : SA

Correct Marks : 3

Question Label : Short Answer Question

Find $\frac{\partial L}{\partial w}$.

Hint: Expressing $\hat{y}$ as a function of u, v, w, and then computing the partial derivative might help.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

0.125

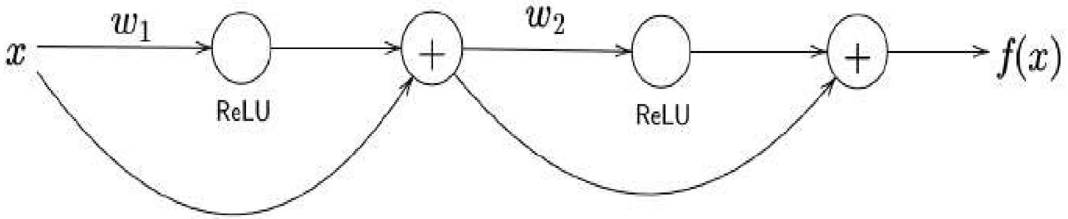
Sub-Section Number :	8
Sub-Section Id :	640653230620
Question Shuffling Allowed :	No

Question Id : 6406531425057 Question Type : COMPREHENSION Sub Question Shuffling Allowed : No Group Comprehension Questions : No Question Pattern Type : NonMatrix

Question Numbers : (121 to 122)

Question Label : Comprehension

Consider a setup where we use what are called skip connections. The operation $\oplus$ adds the output activation of the ReLU neuron before it and the input to that ReLU neuron. For example, the first $\oplus$ would output $x + \max(0, w_1 x)$. Ignore biases.



Based on the above data, answer the given subquestions.

Sub questions

Question Number : 121 Question Id : 6406531425058 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question

If $w_1 = 1, w_2 = -5$, compute $f(1)$.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

2

Question Number : 122 Question Id : 6406531425059 Question Type : SA

Correct Marks : 2

Question Label : Short Answer Question

Find $\dfrac{\partial f}{\partial w_1}$ and evaluate it at $w_1 = 1$,
$w_2 = -5, x = 1.$

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

1

Programming in C

Section Id :	64065399887
Section Number :	6
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	25
Number of Questions to be attempted :	25
Section Marks :	100
Display Number Panel :	Yes
Section Negative Marks :	0
Group All Questions :	No
Enable Mark as Answered Mark for Review and Clear Response :	No
Section Maximum Duration :	0
Section Minimum Duration :	0
Section Time In :	Minutes
Maximum Instruction Time :	0
Sub-Section Number :	1
Sub-Section Id :	640653230621
Question Shuffling Allowed :	No

Question Number : 123 Question Id : 6406531425060 Question Type : MCQ

Correct Marks : 0

Question Label : Multiple Choice Question

THIS IS QUESTION PAPER FOR THE SUBJECT "DEGREE LEVEL : PROGRAMMING IN C (COMPUTER BASED EXAM)"