Unit 3 - Week 1

Assignment 1

The due date for submitting this assignment has passed. **Due on 2018-08-15, 23:59 IST.**

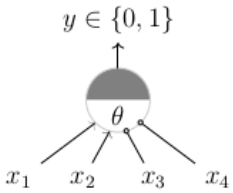
Assignment submitted on 2018-08-15, 23:09 IST

1) *1 point*

Recall that McCulloch Pitts (MP) neuron aggregates the inputs and takes a decision based on this aggregation. If the sum of all inputs is greater than the threshold (θ), then the output of MP neuron is 1, otherwise the output is 0. We say that a MP neuron implements a boolean function if the output of the MP neuron is consistent with the truth table of the boolean function. In other words, if for a given input configuration, the boolean function outputs 1 then the output of the neuron should also be 1. Similarly, if for a given input configuration, the boolean function outputs 0 then the output of the neuron should also be 0.
Consider the following boolean function:

$$(x_1 \text{ AND } x_2) \text{ AND } (!x_3 \text{ AND } !x_4)$$

The MP neuron for the above boolean function is as follows:



What should be the value of the threshold (θ) such that the MP neuron implements the above boolean function? (Note that the circle at the end of the input to the MP neuron indicates inhibitory input. If any inhibitory input is 1 the output will be 0.)

- $\theta = 1$
- $\theta = 2$
- $\theta = 3$
- $\theta = 4$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\theta = 2$

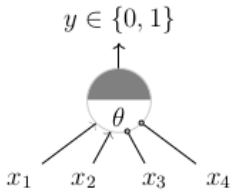
DOWNLOAD VIDEOS

2) 1 point

Keeping the concept discussed in question 1 in mind, consider the following boolean function:

$$(x_1 \text{ OR } x_2) \text{ AND } (!x_3 \text{ AND } !x_4)$$

The MP neuron for the above boolean function is as follows:



What should be the value of the threshold (θ) such that the MP neuron implements the above boolean function? (Note that the circle at the end of the input to the MP neuron indicates inhibitory input. If any inhibitory input is 1 the output will be 0.)

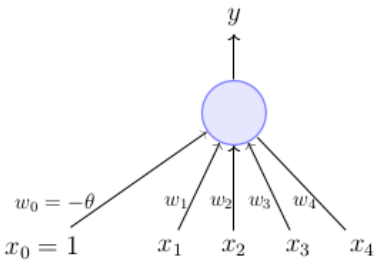
- $\theta = 1$
- $\theta = 2$
- $\theta = 3$
- $\theta = 4$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\theta = 1$

3) 1 point

Let us consider the movie example as discussed in this week's lecture. Suppose we want to predict whether a movie buff would like to watch a movie or not. Note that each movie is represented by a vector, $\mathbf{X} = [x_1\ x_2\ x_3\ x_4]$ and the description of each input (x_i) is mentioned in the figure below. Also, the weight assigned to each of these inputs (or features) is given by $\mathbf{W} = [w_1\ w_2\ w_3\ w_4]$ and the threshold is represented by the parameter θ.



$x_1 = popularity (between\ 1\ to\ 10)$
$x_2 = isGenreScifi (boolean)$
$x_3 = isDirectorNolan (boolean)$
$x_4 = imdbRating (between\ 0\ to\ 1)$

Now, consider the movie **Interstellar** has the feature vector $\mathbf{X} = [8\ \ 1\ \ 1\ \ 0.86]$; which means the movie has a *popularity* of 8 on a a scale of 10 and is a *Scifi* movie directed by *Nolan* with 0.86 as its *imdbRating*. Now consider a person who assigns the following weights to each of these inputs: $\mathbf{W} = [0.14\ \ 1\ \ 0.9\ \ 0.6]$. Further, suppose that $\theta = 2$. Based on the above information, what do you think will be his/her decision?

- Yes, (s)he will watch it.
- No, (s)he won't watch it.

Yes, the answer is correct.
Score: 1

Accepted Answers:
Yes, (s)he will watch it.

4) 1 point

Keeping the discussion of question 3 in mind, consider the movie **The Green Lantern** has the feature vector $\mathbf{X} = [5 \quad 1 \quad 0 \quad 0.53]$. Now consider a person who assigns the following weights to each of these inputs: $\mathbf{W} = [0.8 \quad 1 \quad 0.4 \quad 0.8]$. Further, suppose that $\theta = 7$.
Based on the above information, what do you think will be his/her decision?

- ○ Yes, (s)he will watch it.
- ◉ No, (s)he won't watch it.

Yes, the answer is correct.
Score: 1

Accepted Answers:
No, (s)he won't watch it.

5) 1 point

Consider a small training set with the following points in $\mathbb{R}^3$:

Index	Points $[x_0, x, y, z]$	Class
n_1	[1,0,0,0]	Class 0
p_1	[1,0,0,1]	Class 1
p_2	[1,0,1,0]	Class 1
p_3	[1,0,1,1]	Class 1
p_4	[1,1,0,0]	Class 1
p_5	[1,1,0,1]	Class 1
p_6	[1,1,1,0]	Class 1
p_7	[1,1,1,1]	Class 1

Note that there are 8 points which are divided into two classes, Class 0 and Class 1. We are interested in finding the plane which divides the input space into two classes. Starting with the weight vector, $\mathbf{w} = [0, 0, -1, 2]$, apply the perceptron algorithm by going over the points in the following order $[n_1, p_1, p_2, p_3, p_4, p_5, p_6, p_7]$. If needed, repeat in the same order till convergence. After the algorithm converges, what is the value of the weight vector?

- ○ $\mathbf{w} = [1, 1, 2, 3]$
- ○ $\mathbf{w} = [-1, 1, 1, 2]$
- ○ $\mathbf{w} = [-3, -2, -1, -1]$
- ◉ $\mathbf{w} = [-2, -1, -1, 1]$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mathbf{w} = [-1, 1, 1, 2]$

6) 1 point

A 2-dimensional dataset for 2 classes is given to you. Plot the data and comment wheth the 2 classes are linearly separable or not. Note that the top 500 rows are of Class and the rest 500 are of Class B. You can download the dataset.
Feel free to use any programming language/plotting tool of your choice. Once you plot th data, answer the following question: Is the data linearly separable ?

You can Download DATASET HERE

- ◉ True
- ○ False

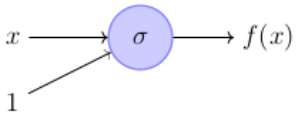
Yes, the answer is correct.
Score: 1

Accepted Answers:

True

7) *0 points*

Partial derivatives This question is not based on the material that we have covered so far. However, this is a part of the pre-requisites and will be required for the material that we will cover in the next class.
Consider the following function,



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

The value L is given by,

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

Here, x and y are constants and w and b are parameters that can be modified. In other words, L is a function of w and b.

Derive the partial derivatives, $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$ and choose the correct option.

- $\frac{\partial L}{\partial w} = (y - f(x))f(x)(1 - f(x))$
  $\frac{\partial L}{\partial b} = (y - f(x))f(x)(1 - f(x))x$
- $\frac{\partial L}{\partial w} = (y - f(x))(1 - f(x))x$
  $\frac{\partial L}{\partial b} = -(y - f(x))f(x)(1 - f(x))$
- $\frac{\partial L}{\partial w} = (y - f(x))f(x)(1 - f(x))x$
  $\frac{\partial L}{\partial b} = (y - f(x))f(x)(1 - f(x))$

Yes, the answer is correct.
Score: 0

Accepted Answers:

$\frac{\partial L}{\partial w} = (y - f(x))f(x)(1 - f(x))x$

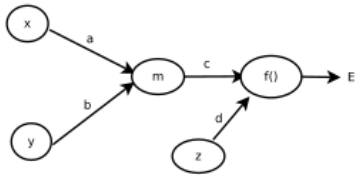
$\frac{\partial L}{\partial b} = (y - f(x))f(x)(1 - f(x))$

8) **1 point**

Consider the function E as given below,

$$E = g(x, y, z) = f(c(ax + by) + dz)$$

Represented as a graph, we have



Here x, y, z are inputs (constants) and a, b, c, d are parameters (variables). m is an intermediate computation and f is some differentiable function. Specifically, let us consider f to be the *tanh* function.

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

Note that here E is a function of a, b, c, d. Compute the following partial derivatives of E with respect to a i.e $\frac{\partial E}{\partial a}$, and choose the correct option.

- [x] $\frac{\partial E}{\partial a} = (1 - f(c(ax + by) + dz)^2)cx$
- [ ] $\frac{\partial E}{\partial a} = c(1 - f(c(ax + by) + dz)^2)$
- [ ] $\frac{\partial E}{\partial a} = (1 - f(c(ax + by) - dz)^2)cx$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\frac{\partial E}{\partial a} = (1 - f(c(ax + by) + dz)^2)cx$

9) **1 point**

Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial b}$ and choose the correct option.

- [ ] $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz))cy$
- [ ] $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)$
- [x] $\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)cy$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\frac{\partial E}{\partial b} = (1 - f(c(ax + by) + dz)^2)cy$

10) **1 point**

Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial c}$ and choose the correct option.

- [x] $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)(ax + by)$
- [ ] $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz))(ax + by)$
- [ ] $\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\frac{\partial E}{\partial c} = (1 - f(c(ax + by) + dz)^2)(ax + by)$

11) 1 point

Keeping the graph discussed in question 8 in mind, find $\frac{\partial E}{\partial d}$ and choose the correct option.

- $\frac{\partial E}{\partial d} = 2(1 - f(c(ax + by) + dz)^2)z$
- $\frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)z$
- $\frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$\frac{\partial E}{\partial d} = (1 - f(c(ax + by) + dz)^2)z$

Previous Page End

Unit 4 - Week 2

Assignment 2

The due date for submitting this assignment has passed.

Due on 2018-08-15, 23:59 IST.

Assignment submitted on 2018-08-15, 06:30 IST

1) **Cross Entropy Loss** *1 point*

In Lecture 3, we derived a formula to compute the partial derivative of the loss function with respect to the parameters of the model, *i.e.*, w and b. The loss function ($\mathscr{L}$) that we considered was the mean squared error loss. Let us assume there is only one point in the dataset, (x, y). We now define a new loss function known as the cross entropy loss function as follows,

$$\mathscr{L}(w, b) = -y * \log f(x)$$

where,

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

and w and b are the parameters of the model. Note that y is the true value given x whereas $f(x)$ is the output of the model given x as input. Derive an expression for the partial derivative of the cross-entropy loss function with respect to w and b and select the correct option from the options given below.

- $\nabla w = \frac{\partial \mathscr{L}(w,b)}{\partial w} = y * (1 - f(x)) * x$
  $\nabla b = \frac{\partial \mathscr{L}(w,b)}{\partial b} = y * (1 - f(x))$
- $\nabla w = \frac{\partial \mathscr{L}(w,b)}{\partial w} = -y * (1 - f(x))$
  $\nabla b = \frac{\partial \mathscr{L}(w,b)}{\partial b} = -y * (1 - f(x))$
- $\nabla w = \frac{\partial \mathscr{L}(w,b)}{\partial w} = -y * (1 - f(x)) * x$
  $\nabla b = \frac{\partial \mathscr{L}(w,b)}{\partial b} = -y * (1 - f(x))$

Yes, the answer is correct.
Score: 1

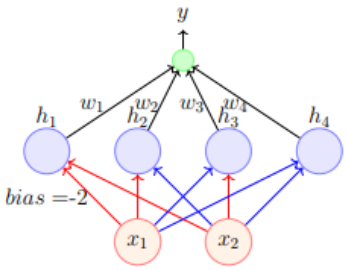
Accepted Answers:

$$\nabla w = \frac{\partial \mathscr{L}(w,b)}{\partial w} = -y * (1 - f(x)) * x$$
$$\nabla b = \frac{\partial \mathscr{L}(w,b)}{\partial b} = -y * (1 - f(x))$$

2) For this question let us assume True $= +1$ and False $= -1$. Consider the Multilayer Perceptron Network shown in the figure below with 2 inputs, x_1 and x_2 and 4 perceptrons in the hidden layer. The outputs of these 4 perceptrons are denoted by h_1, h_2, h_3, h_4. Each input is connected to all the 4 perceptrons with specific weights represented by red and blue edges in the figure below. The bias (w_0) of each perceptron is -2 (*i.e.*, each perceptron will fire only if the weighted sum of its input is ≥ 2). Each of these perceptrons is connected to an output perceptron by weights w_1, w_2, w_3 and w_4. The output of this perceptron (y) is the output of the network.

We have to find the weights w_1, w_2, w_3, w_4 such that this network represents the truth table of the $XNOR$ boolean function with two inputs.

1 point



red edge indicates $w = -1$
blue edge indicates $w = +1$

Under which of the following conditions will the above network behave as the $XNOR$ boolean function?

- $w_1 < w_0,\ w_2 \geq w_0,\ w_3 \geq w_0,\ w_4 < w_0$
- $w_1 = w_0,\ w_2 = w_0,\ w_3 = w_0,\ w_4 = w_0$
- $w_1 \geq w_0,\ w_2 < w_0,\ w_3 < w_0,\ w_4 \geq w_0$
- $w_1 \geq w_0,\ w_2 = w_0,\ w_3 = w_0,\ w_4 \geq w_0$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$w_1 \geq w_0,\ w_2 < w_0,\ w_3 < w_0,\ w_4 \geq w_0$

3) The logistic function is defined as follows,

1 point

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

where w and b are parameters.
What would happen if w increases?

- The slope of the logistic function increases
- The slope of the logistic function decreases
- The centre point of the logistic function moves to the left
- The centre point of the logistic function moves to the right

Yes, the answer is correct.
Score: 1

Accepted Answers:

The slope of the logistic function increases

4) Keeping in mind the logistic function defined in question 3, what would happen if b increases?

1 point

- The slope of the logistic function increases
- The slope of the logistic function decreases
- The centre point of the logistic function moves to the left
- The centre point of the logistic function moves to the right

No, the answer is incorrect.
Score: 0

Accepted Answers:
The centre point of the logistic function moves to the left

5) *3 points*

In this question you will implement the Gradient Descent algorithm on a toy 2-D dataset which consists of 40 data points. You can download the dataset from the following URL:

CLICK HERE TO DOWNLOAD

For this question you have to use the squared error loss function which is given as,

$$loss = \frac{1}{2}(\hat{y} - y)^2$$

where $\hat{y}$ is the output of your model (Refer slide 36 of Lecture 3).
Now given the following hyperparameter settings,

- learning rate = 0.01
- initial weight, w = 1
- initial bias, b = 1
- number of iterations = 100

Which of the following values is the closest to the value of loss that you get at the end of 100 iterations?

- loss = 0.028
- loss = 0.0
- loss = 1.28
- loss = 0.28

No, the answer is incorrect.
Score: 0

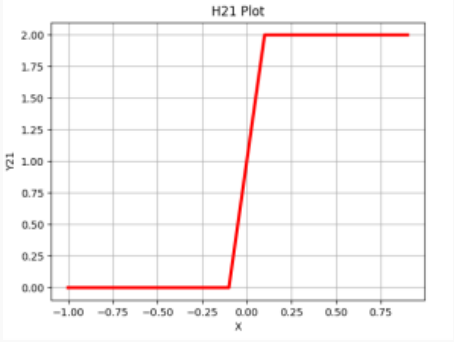
Accepted Answers:
loss = 0.028

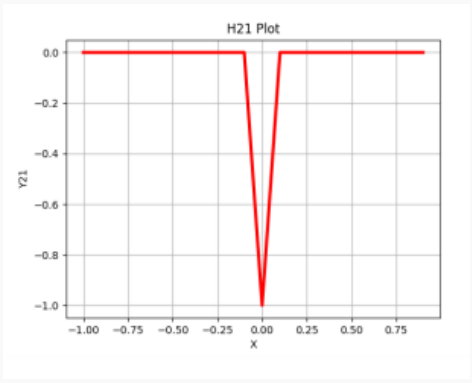
6) *1 point*

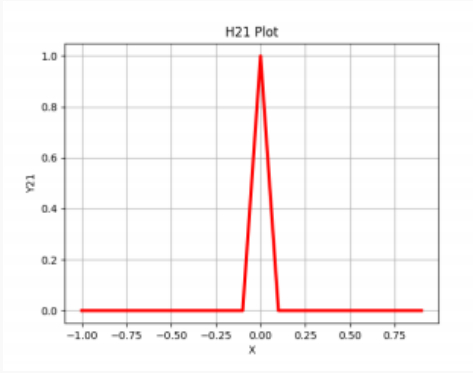
Consider the variable x and functions $h_{11}(x)$, $h_{12}(x)$ and $h_{21}(x)$ such that

$$h_{11}(x) = \frac{1}{1 + e^{-(400x+24)}}$$
$$h_{12}(x) = \frac{1}{1 + e^{-(400x-24)}}$$
$$h_{21}(x) = h_{11}(x) - h_{12}(x)$$

Plot the function $h_{21}(x)$ and choose the option which closely matches the shape of this function for $x \in (-1, 1)$

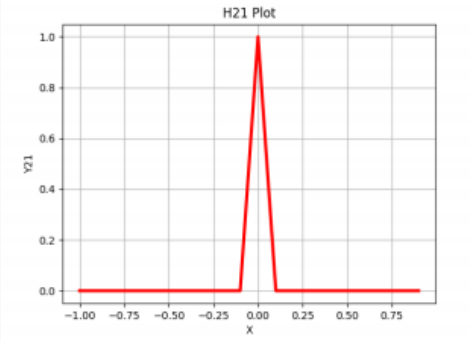






Yes, the answer is correct.
Score: 1

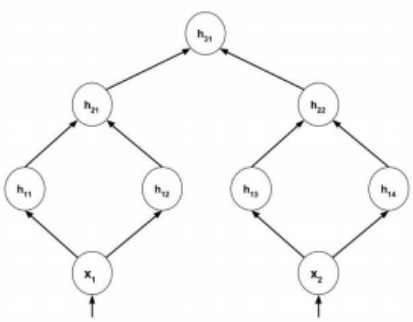
Accepted Answers:



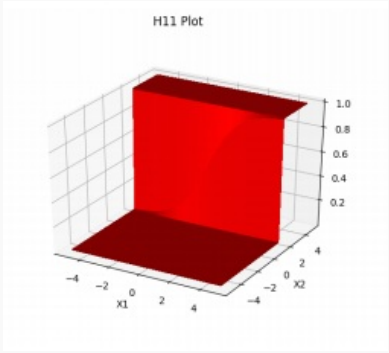
7) Now consider the variables x_1, x_2 and the following functions : *1 point*

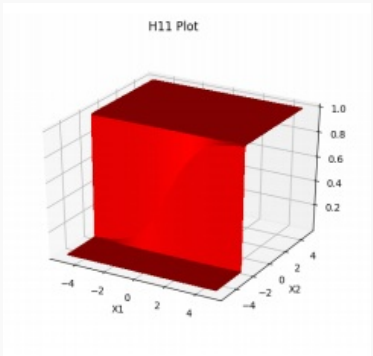
$$h_{11}(x_1, x_2) = \frac{1}{1 + e^{-(x_1 + 100x_2 + 200)}}$$
$$h_{12}(x_1, x_2) = \frac{1}{1 + e^{-(x_1 + 100x_2 - 200)}}$$
$$h_{13}(x_1, x_2) = \frac{1}{1 + e^{-(100x_1 + x_2 + 200)}}$$
$$h_{14}(x_1, x_2) = \frac{1}{1 + e^{-(100x_1 + x_2 - 200)}}$$
$$h_{21}(x_1, x_2) = h_{11}(x_1, x_2) - h_{12}(x_1, x_2)$$
$$h_{22}(x_1, x_2) = h_{13}(x_1, x_2) - h_{14}(x_1, x_2)$$
$$h_{31}(x_1, x_2) = h_{21}(x_1, x_2) + h_{22}(x_1, x_2)$$
$$f(x_1, x_2) = \frac{1}{1 + e^{-(50h_{31}(x) - 100)}}$$

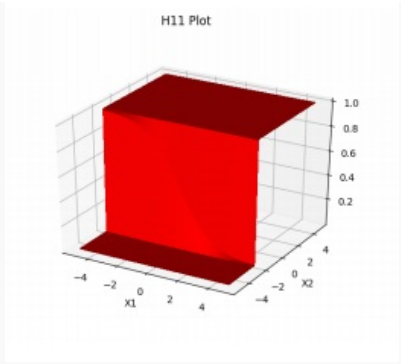
The above set of functions are summarized in the graph below.



Plot the function $h_{11}(x_1, x_2)$ and choose the option which closely matches the shape of this function for $x \in (-5, 5)$

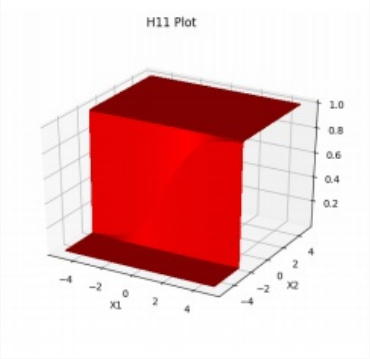






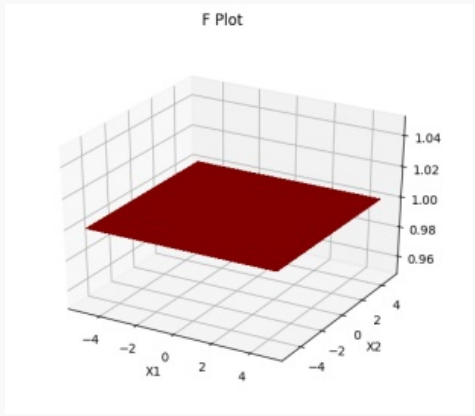
Yes, the answer is correct.
Score: 1

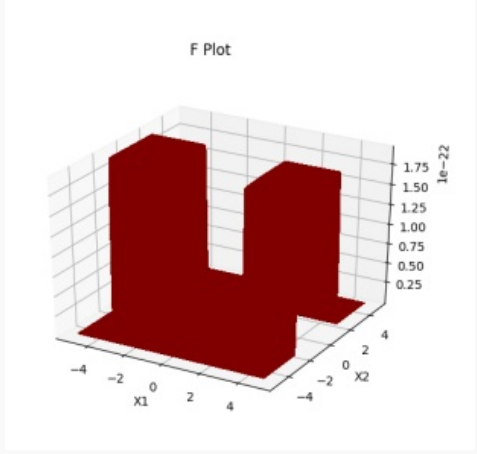
Accepted Answers:

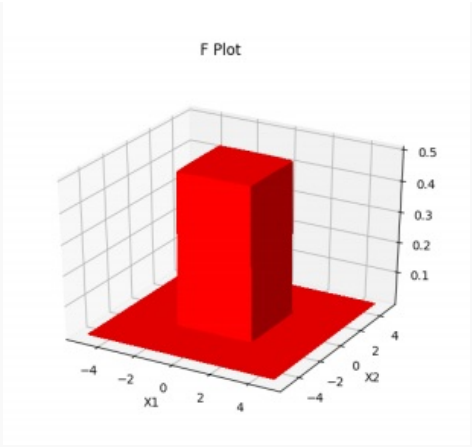


8) ***1 point***

Plot the function $f(x_1, x_2)$ as defined in question 7 and choose the option which closely matches the shape of this function for $x \in (-5, 5)$.

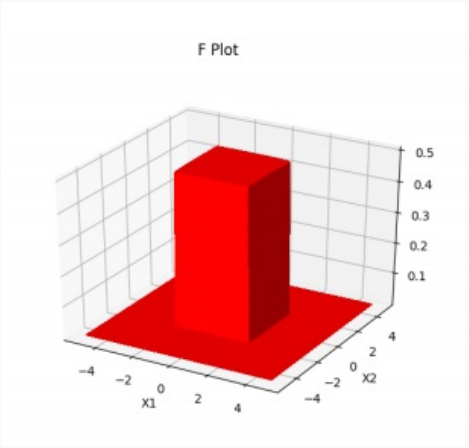






Yes, the answer is correct.
Score: 1

Accepted Answers:



9) *1 point*

Based on the plot, what is the maximum value of the function $f(x_1, x_2)$?

- 0.4
- 0.5
- 1

Yes, the answer is correct.
Score: 1

Accepted Answers:

0.5

10) What is the maximum value of the function $h_{31}(x_1, x_2)$? *1 point*

- 2
- 1.5
- 1

Yes, the answer is correct.
Score: 1

Accepted Answers:

2

Previous Page End

Unit 5 - Week 3

Assignment 3

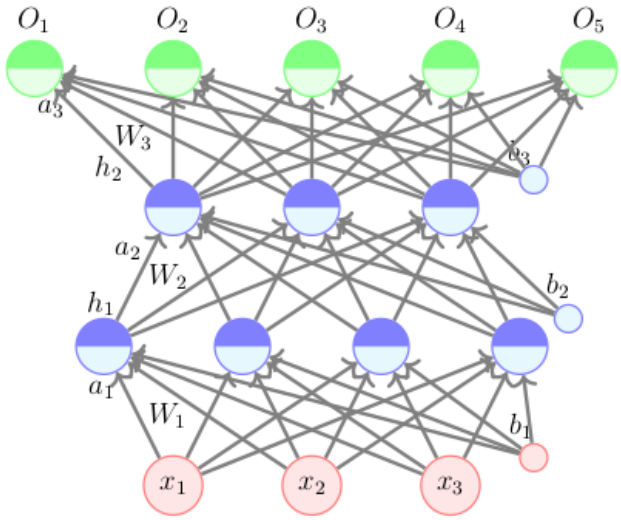
The due date for submitting this assignment has passed.

Due on 2018-09-05, 23:59 IST.

Assignment submitted on 2018-08-20, 22:42 IST

1) *1 point*

Consider the feedforward neural network in the figure shown below which has one input layer, two hidden layers and one output layer. The input x to this network $\in \mathbb{R}^3$ and the number of neurons in the two hidden layers and the output layer is 4,3,5 respectively. Each layer is fully connected to the next layer, *i.e.*, there is a weight connecting every neuron in layer i to every neuron in layer $i+1$. Further, every neuron in the hidden and output layers also has a bias connected to it. What is the total number of parameters in this feedforward neural network?



- 60
- 51
- 39
- 44

Yes, the answer is correct.
Score: 1

Accepted Answers:
51

2) *1 point*

Consider the function $f(\theta) = f(x, y, z) = x^2 + y^2 + z^2 - 8$. What is the gradient of this function at at $\theta = \{1, -1, 1\}$. Note that $\theta = [x, y, z]$ is a collection of all the parameters of this function.

- [2, -2, 2]
- [-2, 2, -2]
- [2, -2, -2]
- [-2, -2, 2]

Yes, the answer is correct.
Score: 1

Accepted Answers:
[2, -2, 2]

3) 1 point

Consider the function $f(\theta) = f(x, y, z) = x^2 + y^2 + z^2 - 8$. Suppose you start with $\theta_0 = \{1, -1, 1\}$ and run one step of gradient descent with the learning rate $\eta = 1$. What will be the updated value of θ?

- [1, -1, 1]
- [1, -1, -1]
- [-1, -1, 1]
- [-1, 1, -1]

Yes, the answer is correct.
Score: 1

Accepted Answers:
[-1, 1, -1]

4) 1 point

Consider the vector $a = [1.2, -2.5, 2.4, 3]$. What will the following lines of code do ?

```
import numpy as np

a = np.asarray([1.2, -2.5, 2.4, 3])
a = np.exp(a)
a = a/np.sum(a)
```

- It will compute the softmax of *a*
- It will compute the element-wise sigmoid of *a*
- It will compute the *e*-th power of each element of *a*

Yes, the answer is correct.
Score: 1

Accepted Answers:
It will compute the softmax of a

5) 1 point

Consider a box which contains 100 balls of which 30 are red, 50 are green and 20 are blue. Your friend peeps into the box and estimates the number of red, green and blue balls as 50, 25, 25. What is the cross entropy between the true distribution over the colors in the box and the distribution predicted by your friend?

- 1.5
- 1.0
- 1.7
- 2.3

Yes, the answer is correct.
Score: 1

Accepted Answers:
1.7

6) 1 point

Consider a vector $a \in \mathbb{R}^n$ and let $b \in \mathbb{R}^n$ be the output of the softmax function applied to this vector. The i-th entry of b is given by:

$$b_i = \frac{e^{a_i}}{\sum_{j=1}^{n} e^{a_j}}$$

Now suppose we introduce a parameter k such that

$$b_i = \frac{e^{k*a_i}}{\sum_{j=1}^{n} e^{k*a_j}}$$

Can b still represent a probability distribution ?

- True
- False

Yes, the answer is correct.
Score: 1

Accepted Answers:
True

7) *1 point*

Continuing the above question where $a \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ and we have a parameter k such that

$$b_i = \frac{e^{k*a_i}}{\sum_{j=1}^{n} e^{k*a_j}}$$

If $k = 1$ then we get the default softmax function. Now, for simplicity, let us assume that all elements of a are positive. Further, let us assume that the j-th element of a is the maximum/largest element of a. It should be obvious that the corresponding j-th element of b will also be the maximum/largest element of b. Now suppose we set $k = 2$, then what will happen to the j-th entry of b.

- ○ It will remain the same as the case when $k = 1$ because now $k = 2$ appears in the denominator also
- ◉ It will be greater than the case when $k = 1$
- ○ It will be lesser than the case when $k = 1$
- ○ Can't say as it will depend on the other values in a

Yes, the answer is correct.
Score: 1

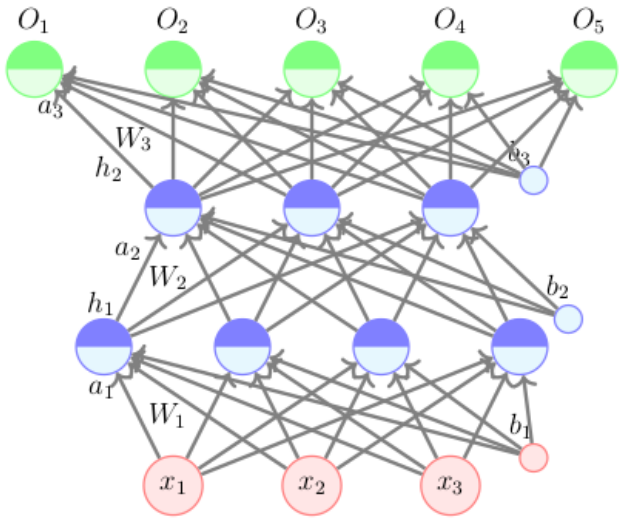
Accepted Answers:
It will be greater than the case when k = 1

8) 3 points

Consider the feedforward neural network in the figure shown below which has one input layer, two hidden layers and one output layer. The input x to this network $\in \mathbb{R}^3$ and the number of neurons in the two hidden layers and the output layer is 4,3,5 respectively. Each layer is fully connected to the next layer, *i.e.*, there is a weight connecting every neuron in layer i to every neuron in layer $i+1$. Also note that every neuron in the hidden and output layers has a bias connected to it. The activation function used in the two hidden layers is the logistic function as defined in the lecture and the output function is the softmax function. Now suppose that all the weights in layer 1 are 0.05, *i.e.*, each of the 3*4=12 elements of the matrix W_1 has a value 0.05. Similarly, let us assume that all the weights in layer 2 are 0.025, *i.e.*, each of the 4*3=12 elements of the matrix W_2 has a value 0.025. Also, let us assume that all the weights in layer 3 are 1.0, *i.e.*, each of the 3*5=15 elements of the matrix W_3 has a value 1. Finally, the bias vectors for the 3 layers are as follows:

$$b_1 = [0.1, 0.2, 0.3, 0.4]$$
$$b_2 = [5.2, 3.2, 4.3]$$
$$b_3 = [0.2, 0.45, 0.75, 0.55, 0.95]$$

Now, suppose we feed the input $x = [1.5, 2.5, 3]$ to this network, what will be the value of O_3 (i.e., the value output by the third neuron in the output layer).



- 0.132
- 0.189
- 0.229
- 0.753

Yes, the answer is correct.
Score: 3

Accepted Answers:
0.229

Previous Page End

Unit 6 - Week 4

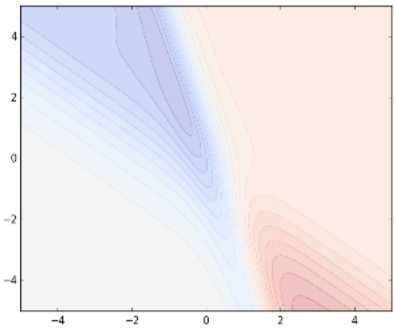
Assignment 4

The due date for submitting this assignment has passed.

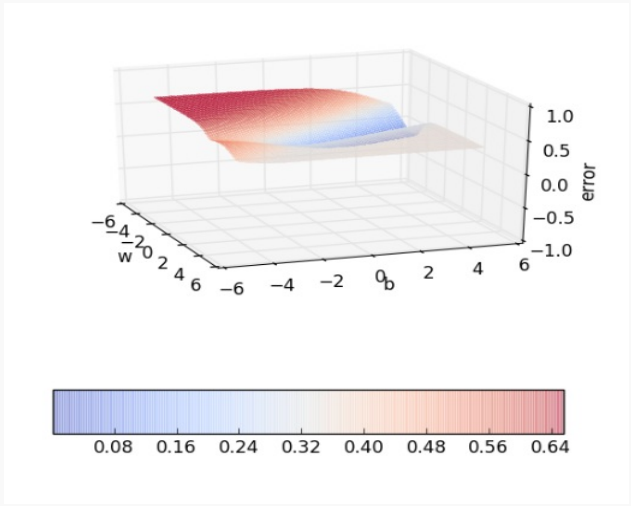
Due on 2018-09-05, 23:59 IST.

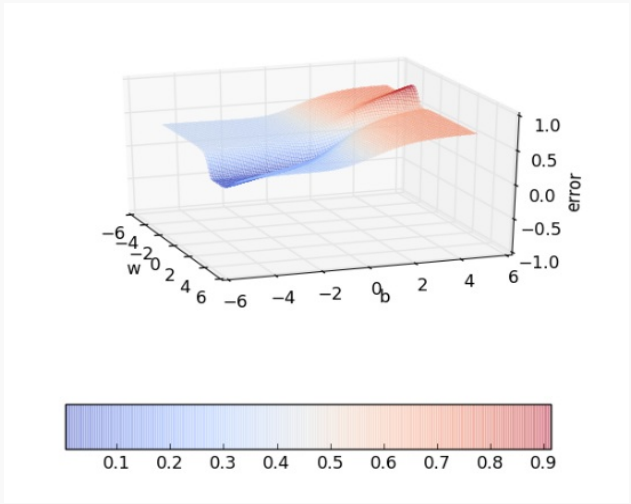
Assignment submitted on 2018-09-05, 22:25 IST

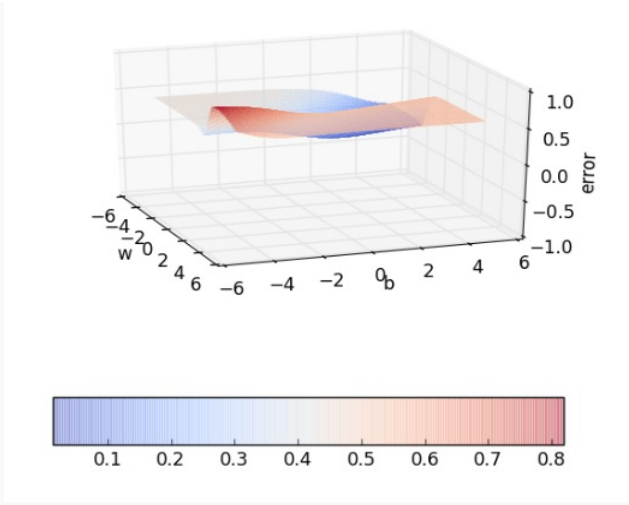
1) Consider the following contour map plotted in 2d. *1 point*



Note that in the above 2d plot the horizontal axis corresponds to the parameter w and the vertical axis corresponds to the parameter b. Which of the 3d plots below corresponds to the 2d plot shown in the figure above (please see carefully which axis corresponds to w and which to b in the 3d plot).

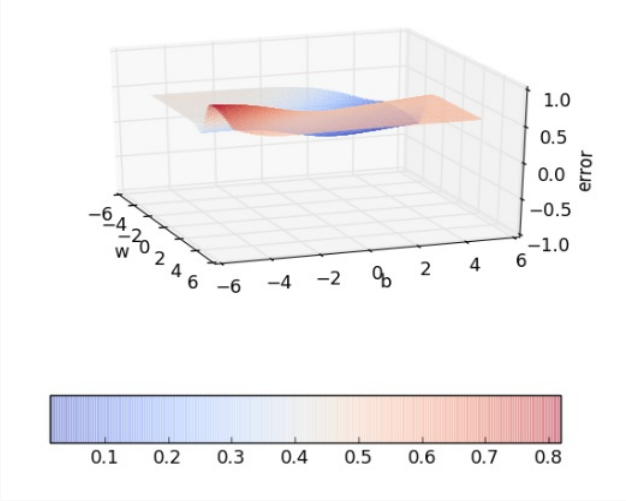






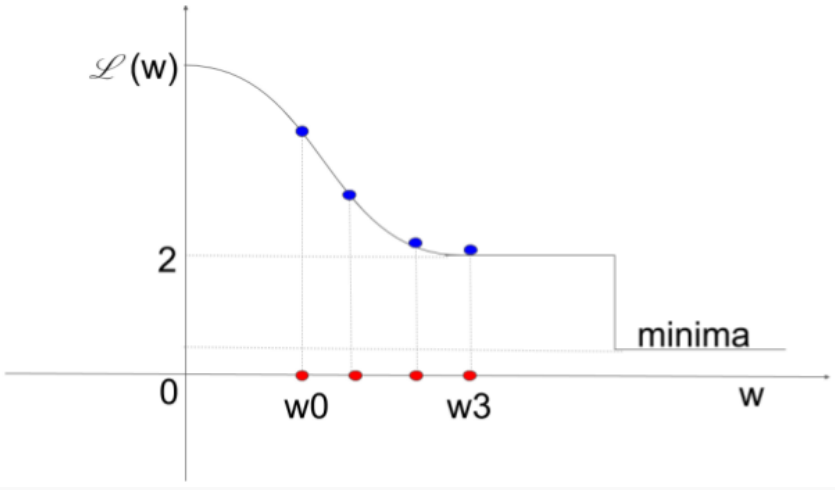
Yes, the answer is correct.
Score: 1

Accepted Answers:



2) *1 point*

Consider the loss function $\mathscr{L}(w)$ as shown in the figure below. You are interested in finding the minima of this function *i.e.*, the value(s) of w for which the function will take its lowest value. To do so you run gradient descent starting with a random value w_0 (the leftmost red dot in the figure). After running, three steps of gradient descent you have the updated value of w as w_3. The red dots in the figure show the value of w at each step and the blue dots show the corresponding value of the loss function $\mathscr{L}(w)$. Now, what will happen if you run the 4th step of gradient descent, *i.e.*, if you try to update the value of w using the gradient descent update rule. Assume that the learning rate is 1.



- the value of w will increase ($i.e.$, $w_4 > w_3$)
- the value of w will remain the same ($i.e.$, $w_4 = w_3$)

- the value of w will decrease $(i.e.,\ w_4 < w_3)$

Yes, the answer is correct.
Score: 1

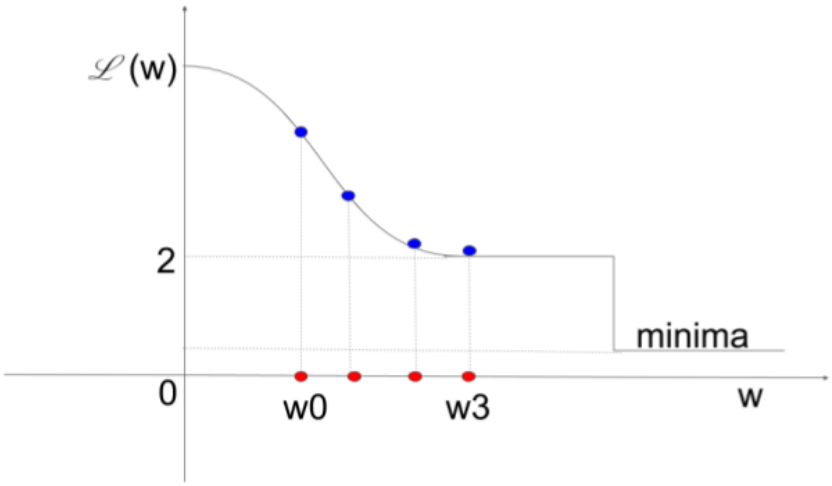
Accepted Answers:
the value of w will remain the same $(i.e.,\ w_4 = w_3)$

3) *1 point*

Continuing the previous question and referring to the same figure again, suppose instead of gradient descent you ran 3 iterations of momentum based gradient descent resulting in the value w_3 as shown in the figure. Note that the update rule of momentum based gradient descent is:

$$update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - update_t$$

Assume that the learning rate is 1 and the momentum parameter $\gamma > 0$. Now, what will happen if you run the 4th step of gradient descent, *i.e.*, if you try to update the value of w using the update rule of momentum based gradient descent.



- the value of w will increase $(i.e.,\ w_4 > w_3)$
- the value of w will remain the same $(i.e.,\ w_4 = w_3)$
- the value of w will decrease $(i.e.,\ w_4 < w_3)$

Yes, the answer is correct.
Score: 1

Accepted Answers:
the value of w will increase $(i.e.,\ w_4 > w_3)$

4) *1 point*

Suppose we choose a model $f(x) = \sigma(wx + b)$ which has two parameters w, b. Further, assume that we are trying to learn the parameters of this model using 200 training points. If we use mini-batch gradient descent with a batch size of 10 then how many times will each parameter get updated in one epoch.

- 10
- 20
- 100
- 200

Yes, the answer is correct.
Score: 1

Accepted Answers:
20

5) 1 point

Note that the update rule for momentum based gradient descent is given by

$$update_t = \gamma \cdot update_{t-1} + \eta \nabla w_t$$
$$w_{t+1} = w_t - update_t$$

Let $\eta = 1$ and $\gamma = 0.9$ and ∇w_1 be the derivative computed at the first time step. you run momentum based gradient descent for 10 iterations then what fraction of ∇ will be a part of $update_{10}$

- $0.9\nabla w_1$
- $\frac{1}{0.9}\nabla w_1$
- $\frac{0.9}{10-1}\nabla w_1$
- $(0.9)^{(10-1)}\nabla w_1$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$(0.9)^{(10-1)}\nabla w_1$

6) 1 point

We saw the following update rule for Adam :

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

$\hat{m}_t, \hat{v}_t$ are the bias corrected values of m_t, v_t. Suppose, instead of using the above equation for m_t we use the following equation where $0 \leq \alpha_1 \leq 1$ and $0 \leq \beta_1 \leq 1$

$$m_t = \frac{\alpha_1}{\beta_1} * m_{t-1} + \frac{(\beta_1 - \alpha_1)}{\beta_1} * \nabla w_t$$

then what would the bias corrected value of m_t be ?

- $\hat{m}_t = \frac{m_t}{\alpha_1^t - \beta_1^t}$
- $\hat{m}_t = \frac{\alpha_1^t m_t}{1-\beta_1^t}$
- $\hat{m}_t = \frac{\alpha_1^t m_t}{\alpha_1^t - \beta_1^t}$
- $\hat{m}_t = \frac{\alpha_1^t m_t}{1-\beta_1^t}$
- $\hat{m}_t = \frac{\beta_1^t m_t}{\beta_1^t - \alpha_1^t}$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\hat{m}_t = \frac{\beta_1^t m_t}{\beta_1^t - \alpha_1^t}$

7) In this question you will implement the Adam algorithm on toy 2-D dataset which consists of 40 data points, *i.e.*, 40 (x,y) pairs. *0 points*

For this question you have to use the squared error loss function which is given as,

$$loss = \frac{1}{2}(\hat{y} - y)^2$$

where $\hat{y}$ is the output of your model given by:

$$\hat{y} = \frac{1}{1 + e^{-(wx+b)}}$$

Now given the following hyperparameter settings,

- learning rate = 0.01
- initial weight, w = 1
- initial bias, b = 1
- number of iterations = 100
- $\beta_1 = 0.9$
- $\beta_2 = 0.99$

What is the value of the loss at the end of 100 iterations?

You can download the dataset using the following link :

CLICK HERE TO DOWNLOAD DATA

- $\text{loss} = 0.058$
- $\text{loss} = 0.0$
- $\text{loss} = 1.58$
- $\text{loss} = 0.58$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\text{loss} = 0.058$

Previous Page End

Unit 7 - Week 5

Assignment 5

The due date for submitting this assignment has passed. **Due on 2018-09-12, 23:59 IST.**

Assignment submitted on 2018-09-07, 19:48 IST

1) Consider the following matrix: *1 point*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

Which of the following vectors is not an eigenvector of this matrix ?

- $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$ ◉

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}$

2) *0 points*

Consider a square matrix $A \in \mathbb{R}^{3\times 3}$ such that $A^T = A$. My friend told me that the following three vectors are the eigenvectors of this matrix A:

$$x = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, z = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}$$

Is my friend telling the truth ?

- Yes
- No

- [x] Can’t say without knowing all the elements of A
- [ ] Yes, only if all the diagonal elements of A are 1

No, the answer is incorrect.
Score: 0

Accepted Answers:
No

3) Consider the following matrix: *1 point*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

What can you say about the series $x, Ax, A^2xA^3x, \ldots$?

- [x] It will diverge (explode)
- [ ] It will converge (vanish)
- [ ] It will reach a steady state
- [ ] Can’t say without knowing all the elements of x

Yes, the answer is correct.
Score: 1

Accepted Answers:
It will diverge (explode)

4) Which of the following sets of vectors **does not** form a valid basis in $\mathbb{R}^3$ *1 point*

- [ ] $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$
- [ ] $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \\ 5 \end{bmatrix}$
- [x] $\begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$
- [ ] $\begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 \\ 7 \\ 4 \end{bmatrix}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$$\begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 6 \\ 5 \\ 4 \end{bmatrix}$$

5) **0 points**

Consider the matrix A:

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{bmatrix}$$

Now consider the following optimization problem:

$$\min_x \; x^T A x$$
$$\text{s.t } \|x\| = 1$$

Which of the following vectors is a solution to the above minimization problem?

- $\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$
- $\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$
- None of the above

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$

6) **1 point**

Consider a row stochastic matrix $M \in \mathbb{R}^3$. The sum of the elements of each row of this matrix is 1. Is the vector $x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$ an eigenvector of this matrix?

- Yes
- No
- Can’t say without knowing the elements of A
- Yes, only if each row represents a uniform distribution

Yes, the answer is correct.
Score: 1

Accepted Answers:

Yes

7) 1 point

Consider a set of points $x_1, x_2, ..., x_m \in \mathbb{R}^2$ represented using the standard basis $x = \begin{bmatrix}1 & 0\end{bmatrix}$ and $y = \begin{bmatrix}0 & 1\end{bmatrix}$. Let $X \in \mathbb{R}^{m \times 2}$ be a matrix such that $x_1, x_2, ..., x_m$ are the rows of this matrix. Using PCA, we want to represent this data using a new basis. To do so, we find the eigenvectors of $X^T X$, which happen to be $u_1 = \begin{bmatrix}\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$ and $u_2 = \begin{bmatrix}\frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{bmatrix}$. Now suppose, we want to represent one of the m points, say $x_i = \begin{bmatrix}2.1 & 2.4\end{bmatrix}$ using only u_1 (*i.e.*, we want to represent the data using fewer dimensions then what would be the squared error in reconstructing x_i using only u_1 ?

- (•) 0.045
- ( ) 0.030
- ( ) 0.015
- ( ) 0

Yes, the answer is correct.
Score: 1

Accepted Answers:
0.045

Previous Page End

Unit 8 - Week 6

Assignment 6

The due date for submitting this assignment has passed. **Due on 2018-09-12, 23:59 IST.**

Assignment submitted on 2018-09-11, 21:29 IST

1) *1 point*

Consider an autoencoder which has one input layer, one hidden layer and one output layer. There is a weight connecting every neuron in the input layer to every neuron in the hidden layer and similarly there is a weight connecting every neuron in the hidden layer to every neuron in the output layer. There are no bias parameters in the network. Now, if the input to the network is $x \in \mathbb{R}^{10}$ and the total number of weights in the network is 160, then what kind of autoencoder is this ?

- Overcomplete autoencoder
- Undercomplete autoencoder
- Can't say because the size of the hidden layer in not known

Yes, the answer is correct.
Score: 1

Accepted Answers:
Undercomplete autoencoder

2) *1 point*

In the lecture we saw that an autoencoder is equivalent to PCA if we use a linear encoder, a linear decoder, the squared error loss function and standardize the elements of the input matrix in a certain way. Keeping this in mind, what would the following matrix look like after its elements are standardized.

$$\begin{bmatrix} 1 & 6 & 12 \\ 3 & 12 & 24 \\ 5 & 18 & 72 \\ 7 & 48 & 36 \end{bmatrix}$$

- $\begin{bmatrix} -3 & -15 & -24 \\ -1 & -9 & -12 \\ 1 & -3 & 36 \\ 3 & 27 & 0 \end{bmatrix}$
- $\begin{bmatrix} 0.5 & 3 & 6 \\ 1.5 & 6 & 12 \\ 2.5 & 9 & 36 \\ 3.5 & 24 & 18 \end{bmatrix}$
- $\begin{bmatrix} -1.5 & -7.5 & -12 \\ -0.5 & -4.5 & -6 \\ 0.5 & -1.5 & 18 \\ 1.5 & 13.5 & 0 \end{bmatrix}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$$\begin{bmatrix} -1.5 & -7.5 & -12 \\ -0.5 & -4.5 & -6 \\ 0.5 & -1.5 & 18 \\ 1.5 & 13.5 & 0 \end{bmatrix}$$

3) *1 point*

Consider a dataset containing m black and white images *i.e.*, every pixel in the image is either black or white. Further, assume that each image contains k pixels and x_{ij} denotes the j-th pixel of the i-th image. Suppose you want to train an autoencoder using this data such that the autoencoder takes an image x as input and reconstructs it as $\hat{x}$. Which of the following is the most appropriate loss function to use in this case?

- $\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij} - \hat{x}_{ij})$
- $\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij} - \hat{x}_{ij})^2$
- $-\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij}\log\hat{x}_{ij} - (1-x_{ij})\log(1-\hat{x}_{ij}))$
- $-\sum_{i=1}^{m}\log\hat{x}_{ij}$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$$-\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij}\log\hat{x}_{ij} - (1-x_{ij})\log(1-\hat{x}_{ij}))$$

4) *1 point*

We saw that we can use L2 regularization along with the reconstruction loss to regularize autoencoders. The total loss function then contains two parts

$$\sum_{i=1}^{m}\sum_{j=1}^{k}(x_{ij} - \hat{x}_{ij})^2 + \lambda||\theta||^2$$

where θ is a collection of all the parameters of the network .The second term, *i.e.*, $\lambda||\theta||^2$ is the L2-regularization loss and the first term is the reconstruction loss. Suppose you use the same training data containing m points and train the autoencoder using different values of λ. For example, you train one autoencoder using $\lambda = 1$, another using $\lambda = 10$ and another using $\lambda = 100$. What do you expect will happen to the reconstruction loss as you increase the value of λ. Assume that the data is a complex real world dataset (for example, images uploaded on facebook, instagram, *etc*).

- the reconstruction loss will be unaffected as λ only gets multiplied by $||\theta||^2$
- the reconstruction loss will increase as λ increases
- the reconstruction loss will decrease as λ increases

Yes, the answer is correct.
Score: 1

Accepted Answers:

the reconstruction loss will increase as λ increases

5) **3 points**

We have trained two autoencoders on a small dataset containing 1000 training points where each point $\in \mathbb{R}^{50}$. Each of these autoencoders has one input layer, one hidden layer and one output layer. Each autoencoder has the same number of neurons in the hidden layer and uses the following encoder and decoder functions.

$$h = \frac{1}{1 + e^{-(W_e x + b)})}$$
$$x = \frac{1}{1 + e^{-(W_d h + c)})}$$

where, $x \in \mathbb{R}^{50}$, $W_e \in \mathbb{R}^{20 \times 50}$ is the weight matrix for the input layer ($i.e.$, it contains the weights connecting the input layer to the hidden layer). $b \in R^{20}$ contains the biases for the hidden layer (one per neuron in the hidden layer). Similarly, $W_d \in \mathbb{R}^{50 \times 20}$ is the weight matrix for the output layer ($i.e.$, it contains the weights connecting the hidden layer to the output layer). Finally $c \in R^{50}$ contains the biases for the hidden layer (one per neuron in the hidden layer). We have trained the two autoencoders using two different algorithms and hence we have different values for W_e, W_d, b and c. You can download the weights learned by the two networks and load them using the following code:

```
# Load packages
import numpy as np

# Load weights for autoencoder 1
parameters1 = np.load('autoencoder1.npy')
W_e_1 = parameters1[0]
b_1   = parameters1[1]
W_d_1 = parameters1[2]
c_1   = parameters1[3]

# Load weights for autoencoder 2
parameters2 = np.load('autoencoder2.npy')
W_e_2 = parameters2[0]
b_2   = parameters2[1]
W_d_2 = parameters2[2]
c_2   = parameters2[3]
```

You can download the training data and the weights from this URL :

CLICK HERE TO DOWNLOAD THE TRAINING DATA AND WEIGHTS

You need to find out which autoencoder gives a smaller squared error loss on the training data.

- (●) the loss of autoencoder1 $<$ the loss of autoencoder2
- ( ) the loss of autoencoder1 $>$ the loss of autoencoder2
- ( ) the loss of autoencoder2 $=$ the loss of autoencoder1

Yes, the answer is correct.
Score: 3

Accepted Answers:

the loss of autoencoder1 $<$ the loss of autoencoder2

6) 1 point

This question is in the context of contractive autoencoders. Consider an autoencoder where the input $x \in \mathbb{R}^3$ and the hidden layer $h \in \mathbb{R}^2$. Suppose, we use the following encoder function:

$$h = Wx + b$$

where, $x \in \mathbb{R}^3, h \in \mathbb{R}^2, W \in \mathbb{R}^{2\times3}, b \in \mathbb{R}^2$. Further, let us assume that the network has the following parameters:

$$W = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

The ij-th entry of the matrix W, *i.e.*, the element in the i-th row and j-th column is the weight connecting the j-th input to the i-th neuron in the hidden layer. Which of the following matrices is the Jacobian matrix for this autoencoder (refer to the definition of the Jacobian matrix as discussed in the lecture).

- $\begin{bmatrix} 2 & 5 \\ 3 & 4 \\ 2 & 7 \end{bmatrix}$
- $\begin{bmatrix} 2 & 3 & 2 \\ 5 & 4 & 7 \end{bmatrix}$
- $\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 4 \end{bmatrix}$
- $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \end{bmatrix}$ (selected)

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 4 \end{bmatrix}$

Previous Page End

Unit 9 - Week 7

Assignment 7

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-09-19, 23:59 IST.

1) A model having a high bias with very few parameters will *1 point*

- underfit the training data
- overfit the training data

No, the answer is incorrect.
Score: 0

Accepted Answers:
underfit the training data

2) A model having a high variance with a large number of parameters will *1 point*

- have a low expected mean square error
- have a high expected mean square error

No, the answer is incorrect.
Score: 0

Accepted Answers:
have a high expected mean square error

3) *1 point*

Is the following statements true: A complex model will be more sensitive to changes training data than a simple model.

- True
- False

No, the answer is incorrect.
Score: 0

Accepted Answers:
True

4) *1 point*

Consider a model which has only 3 parameters, w_1, w_2, w_3 or more compactly θ $[w_1, w_2, w_3]$. Let $\mathscr{L}(\theta)$ be the loss function and $\Omega(\theta)$ be the regularizer. For example we use L2 regularization than $\Omega(\theta) = ||\theta||^2$. The corresponding gradient descent upd rule for w_1 is given by

$$w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - 2\eta\alpha w_1$$

and we will have a similar equation for w_2 and w_3 also. Suppose, instead of L2 regul ization we use L4 regularization then what would the update rule for w_1 be

- $w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - \eta\alpha w_1$
- $w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - \eta\alpha$

- $w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - 3\eta\alpha w_1^2$
- $w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - 4\eta\alpha w_1^3$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$w_1 = w_1 - \eta \frac{\partial}{\partial w_1} \mathscr{L}(w_1) - 4\eta\alpha w_1^3$

5) *3 points*

Suppose you are training a simple neural network as defined by the equations below:

$$y = f(x) = \frac{1}{1 + e^{-(w^T x)}}$$

where $x \in \mathbb{R}^{40}, w \in \mathbb{R}^{40}$. The dataset is divided into a training set and a validation set. We have trained the network for 10 epochs and the weights after each epoch are downloadable. You can load these weights using the following code:

```
import numpy as np
# Before epoch 0, load this
w = np.load('weights_after_epoch_0.npy')
# Before epoch 1, load this
w = np.load('weights_after_epoch_1.npy')
# .. and so on till
# Before epoch 9, load this
w = np.load('weights_after_epoch_9.npy')
```

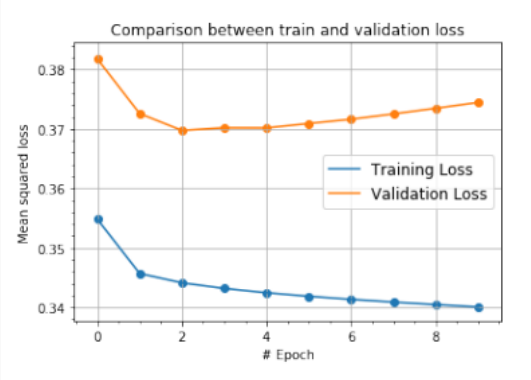
You need to compute the training and validation error after each epoch and say which of the following plots correctly shows the training and validation error. We have used mean squared loss function while training the network. You can calculate the loss for each individual data-point using the definition of "error" as defined in the code snippets (Slide 36/62 in Week 2 study material) and the total loss after every epoch is calculated by the taking the mean of these individual losses.

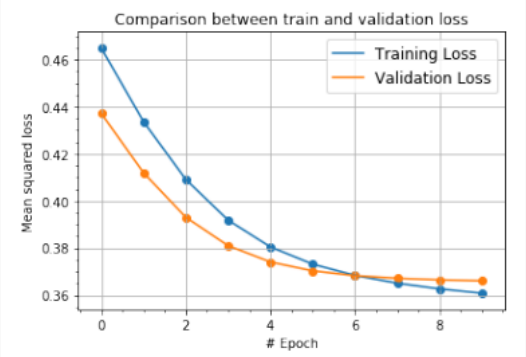
You can download the data using the url:

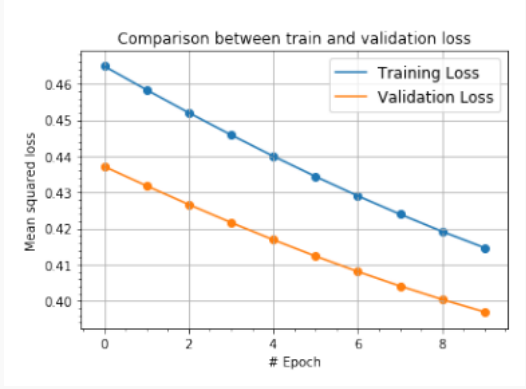
https://drive.google.com/open?id=1gw26BEHSuwx3Enp300Rr1uOrshrhHRA8

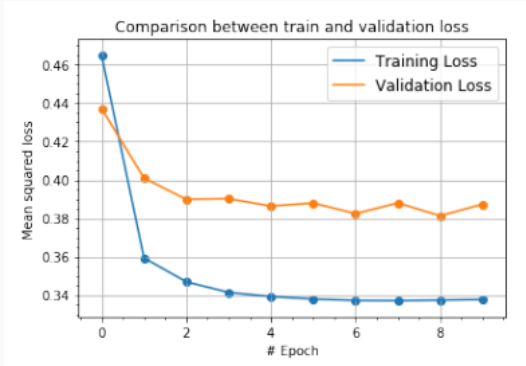
You can download the weights after every epoch using the url:

https://drive.google.com/open?id=1-yXdSiYviVtpPnEp0V6lR_ZNmlPpKkGA

- 

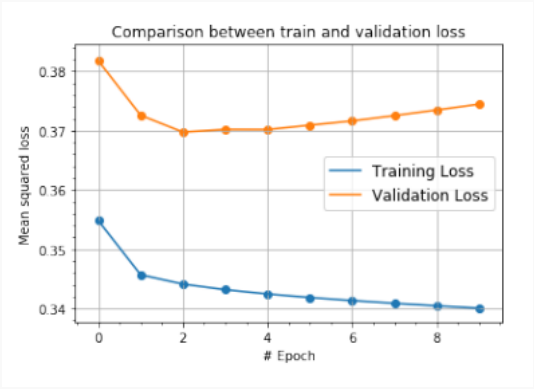
- 





No, the answer is incorrect.
Score: 0

Accepted Answers:



6) *1 point*

Continuing the previous question, if you decide to use early stopping with a patience 3 epochs then which model will you pick ?(Note that we number the epochs from epo 0 to epoch 9)

- The model saved after epoch 0
- The model saved after epoch 2
- The model saved after epoch 4
- The model saved after epoch 6

No, the answer is incorrect.
Score: 0

Accepted Answers:

The model saved after epoch 4

Previous Page End

Unit 10 - Week 8

Assignment 8

The due date for submitting this assignment has passed.

Due on 2018-09-26, 22:59 IST.

Assignment submitted on 2018-09-21, 17:10 IST

1) *1 point*

Let $X \in \mathbb{R}$ be a random variable with mean μ_x and variance σ_x^2. Similarly, let $Y \in \mathbb{R}$ be a random variable with mean μ_y and variance σ_y^2. If X and Y are independent random variables, then E[XY] = ?

- (●) $\mu_x \cdot \mu_y$
- ( ) μ_x
- ( ) μ_y
- ( ) $\sigma_x \cdot \sigma_y$

Yes, the answer is correct.
Score: 1
Accepted Answers:
$\mu_x \cdot \mu_y$

2) *1 point*

Continuing the above question, Var(XY) = ?

- (●) $\sigma_x^2\sigma_y^2$
- ( ) $\sigma_x^2\sigma_y^2 + \sigma_x^2\mu_y^2 + \sigma_y^2\mu_x^2$
- ( ) $\mu_x^2\mu_y^2$
- ( ) $\sigma_x\mu_x + \sigma_y\mu_y$

No, the answer is incorrect.
Score: 0
Accepted Answers:
$\sigma_x^2\sigma_y^2 + \sigma_x^2\mu_y^2 + \sigma_y^2\mu_x^2$

3) *1 point*

Is ReLU activation function differentiable at origin?

- ( ) True
- (●) False

Yes, the answer is correct.
Score: 1
Accepted Answers:
False

4) *1 point*

Is ReLU activation function a non-linear activation function?

- (●) True
- ( ) False

Yes, the answer is correct.
Score: 1
Accepted Answers:
True

5) 1 point

The maxout activation function is given by $f(x) = max(w_1^T x + b_1, w_2^T x + b_2)$. You obtain the ReLU function from this by setting:

- $b_1 = b_2 = 0$
- $w_1 = b_1 = b_2 = 0$
- $w_1 = w_2 = 0$

Yes, the answer is correct.
Score: 1
Accepted Answers:
$w_1 = b_1 = b_2 = 0$

6) 1 point

The symmetry breaking problem occurs only if all the weights in a layer are initiali to zero.

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
False

7) 1 point

The batch normalization layer does not introduce any new parameters.

- True
- False

No, the answer is incorrect.
Score: 0
Accepted Answers:
False

8) 1 point

A deep neural network with linear activation functions throughout and no bias paran ers is equivalent to a shallow neural network

- True
- False

Yes, the answer is correct.
Score: 1
Accepted Answers:
True

9) 1 point

We can use backpropagation to train a deep neural network only if all the hidden layers in the network have the same activation function.

- True
- False

Yes, the answer is correct.
Score: 1
Accepted Answers:
False

10) 1 point

Consider the functions $f(x) = \frac{1}{1+e^{-x}}$ and $g(x) = tanh(x)$. Which of the following statements is true ?

- $g(x) = 2 * f(x) - 1$
- $g(x) = 2 * f(2x) - 1$
- $g(x) = f(\frac{x}{2}) - 0.5$
- $g(x) = f(\frac{x}{2}) - 1$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$g(x) = 2 * f(2x) - 1$

11) *1 point*

Suppose there are 16 neurons in layer1 and 64 neurons in layer 2. The weight matrix connecting the neurons in layer 1 to layer 2 will be $W \in \mathbb{R}^{16 \times 64}$. Further, assume that you are using the logistic activation function in all the hidden layers of your network. If you are using Xavier initialization to initialize the weights of this layer then you will draw the weights from a unit Gaussian and multiply them by ?

- $\frac{1}{4}$
- $\frac{1}{8}$
- $\frac{1}{16}$
- $\frac{1}{64}$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\frac{1}{4}$

12) *1 point*

Continuing the above question but this time assuming that you are using the ReLU activation function in all the hidden layers of your network. Now, if you are using He initialization to initialize the weights of the layer mantioned in the previous question then you will draw the weights from a unit Gaussian and multiply them by ?

- $\frac{1}{32}$
- $\frac{1}{8}$
- $\frac{1}{4}$
- $\frac{1}{2}$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\frac{1}{2}$

Previous Page End

Unit 11 - Week 9

Assignment 9

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-10-03, 23:59 IST.

1) *1 point*

Consider the following corpus: "human machine interface for computer applicatio user opinion of computer system response time. user interface management syste system engineering for improved response time". What is the size of the vocabulary the above corpus ? (You can ignore punctuation.)

- 13
- 15
- 14
- 16

No, the answer is incorrect.
Score: 0

Accepted Answers:
15

2) *1 point*

Let $count(w, c)$ be the number of times the words w and c appear together in the corp $(i.e.,$ occur within a window of few words around each other). Further, let $count(w)$ a $count(c)$ be the total number of times the word w and c appear in the corpus respectiv and let N be the total number of words in the corpus. The PMI between w and then given by:

- $\log \frac{count(w,c)*count(w)}{N*count(c)}$
- $\log \frac{count(w,c)*count(c)}{N*count(w)}$
- $\log \frac{count(w,c)*N}{count(w)*count(c)}$
- $\log \frac{count(w)*count(c)}{count(w,c)*N}$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\log \frac{count(w,c)*N}{count(w)*count(c)}$

3) *1 point*

The SVD of a matrix X is given by $X = U\Sigma V^\top$ where U contains the eigen vectors

- X
- $XX^\top$
- $X^\top$
- $X^\top X$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$XX^\top$

4) 1 point

Let X be the co-occurrence matrix such that the (i, j)-th entry of X captures the P between the i-th and j-th word in the corpus. Every row of X corresponds to t representation of the i-th word in the corpus. Suppose each row of X is normalized (i the L_2 norm of each row is 1) then the (i, j)-th entry of $XX^\top$ captures the:

- PMI between word i and word j
- euclidean distance between word i and word j
- probability that word i co-occurs with word j
- cosine similarity between word i and word j

No, the answer is incorrect.
Score: 0

Accepted Answers:
cosine similarity between word i and word j

5) 1 point

Let the co-occurrence matrix $X \in \mathbb{R}^{m \times n}$ ($i.e.$, there are m words and n context word Once we do a k-rank approximation of X using SVD, we take $W_{word} = U\Sigma$ as the mat containing the representations of the words. What are the dimensions of W_{word} ?

- $m \times n$
- $n \times k$
- $m \times k$
- $k \times m$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$m \times k$

6) 1 point

At the input layer of continuous bag of words model, we multiply a one-hot vector $x \in \mathbb{R}^{|V|}$ with the parameter matrix $\mathbf{W} \in \mathbb{R}^{k \times |V|}$. What does each column of $\mathbf{W}$ correspond to ?

- the representation of the i-th word in the vocabulary
- the i-th eigen vector of the co-occurrence matrix

No, the answer is incorrect.
Score: 0

Accepted Answers:
the representation of the i-th word in the vocabulary

7) 1 point

Consider the word w and a word c which appears before it. For example, w could be the word *barks* and c could be the word *dog*. Let v_w and u_c be the representations of w and c respectively. Further, assume that you are training the bag-of-words model using $n = 1$ ($i.e.$, you are training the model to predict the next word given the current word). The loss function used in the continuous bag-of-words model ensures that:

- v_w and u_c are orthogonal to each other
- v_w and u_c are similar to each other

- does not guarantee anything about v_w and u_c (after all, in the above example, why should the algorithm care about the relation between the representations of *dog* and *barks*. It should rather be interested in the relation between the representations of {*dog* and *cat*} or {*barks* and *howls*})

No, the answer is incorrect.
Score: 0

Accepted Answers:
v_w and u_c are similar to each other

8) *1 point*

Consider the word w and a word c which appears before it. For example, w could be the word *barks* and c could be the word *dog*. Let v_w and u_c be the representations of w and c respectively. Further, assume that you are training the bag-of-words model using $n = 1$ (*i.e.*, you are training the model to predict the next word given the current word). Let $\hat{y}$ be the output of the model (*i.e.*, $\hat{y}$ is the probability distribution over all words in the vocabulary). In particular, $\hat{y}_w$ is the probability assigned by the model to the word w. If you are using gradient descent to train the model and η is the learning rate then the update rule for v_w is given by:

- $v_w = v_w + \eta\ u_c(1 - \hat{y}_w)$
- $v_w = v_w + \eta\ \hat{y}_w(1 - u_c)$
- $v_w = v_w - \eta\ u_c(1 - \hat{y}_w)$
- $v_w = v_w - \eta\ \hat{y}_w(1 - u_c)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$v_w = v_w + \eta\ u_c(1 - \hat{y}_w)$

Previous Page End

Unit 12 - Week 10

Assignment 10

The due date for submitting this assignment has passed.

Due on 2018-10-10, 23:59 IST.

Assignment submitted on 2018-10-10, 23:49 IST

1) *1 point*

Consider an image $I \in \mathbb{R}^{p \times q}$ and a kernel $K \in \mathbb{R}^{m \times n}$. Suppose we apply this kerne on the pixel I_{ij} of the image such that this pixel is placed at the center of the image Which of the following equations represents the resulting output S_{ij} of this convolutio operation?

- $S_{ij} = \sum_{a=-\lfloor \frac{m}{2} \rfloor}^{\lfloor \frac{m}{2} \rfloor} \sum_{b=-\lfloor \frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} I_{i-a,j-b} K_{a,b}$
- $S_{ij} = \sum_{a=\lfloor -\frac{m}{2} \rfloor}^{\lfloor \frac{m}{2} \rfloor} \sum_{b=\lfloor -\frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} I_{i-a,j-b} K_{\frac{m}{2}+a,\frac{n}{2}+b}$
- $S_{ij} = \sum_{a=0}^{m-1} \sum_{b=0}^{n-1} I_{i+a,j+b} K_{a,b}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$S_{ij} = \sum_{a=\lfloor -\frac{m}{2} \rfloor}^{\lfloor \frac{m}{2} \rfloor} \sum_{b=\lfloor -\frac{n}{2} \rfloor}^{\lfloor \frac{n}{2} \rfloor} I_{i-a,j-b} K_{\frac{m}{2}+a,\frac{n}{2}+b}$

2) *1 point*

What is the dimension of the output when a 7×7 kernel is applied to a 9×9 imag with stride $S = 1$ and no padding ?

- 3 x 3
- 5 x 5
- 2 x 2

Yes, the answer is correct.
Score: 1

Accepted Answers:

3 x 3

3) *1 point*

Given a $128 \times 128 \times 3$ image and 6 filters of size $9 \times 9 \times 3$, what will be the dimensio of the output volume when the stride $S = 1$ and the padding $P = 2$?

- 124 x 124 x 6
- 119 x 119 x 6
- 121 x 121 x 6

Yes, the answer is correct.
Score: 1

Accepted Answers:

124 x 124 x 6

4) *1 point*

Consider the following Convolutional neural network where all the convolution filters ar of size 3×3. For all the convolution layers, the stride $S = 1$ and padding $P = 1$:

- **CONV1**: convolutional layer which takes an image of size $28 \times 28 \times 1$ as input and produces 64 outputs (64 filters of size $3 \times 3 \times 1$)
- **POOL1**: 2×2 max-pooling layer
- **CONV2**: convolutional layer with 64 inputs, 128 outputs (128 filters of size $3 \times$ $\times 64$)
- **POOL2**: 2×2 max-pooling layer
- **CONV3**: convolutional layer with 128 inputs, 256 outputs
- **CONV4**: convolutional layer with 256 inputs, 256 outputs
- **POOL3**: 2×2 max-pooling layer
- **FC1**: fully connected layer with 1024 outputs.

What is the number of parameters in the FC1 layer?

- ○ 256 * 1024
- ○ 4096 * 1024
- ◉ 1024 * 1024

No, the answer is incorrect.
Score: 0

Accepted Answers:
*4096 * 1024*

5) *1 point*

Consider the problem where we have a Feed-forward Neural Network (FNN) with layers such that it takes a m-dimensional vector as input and produces an a n-dimensiona vector as output. Is it possible to represent this FNN as a Convolutional Neural Networ (CNN)?

- ◉ No, it is not possible.
- ○ Yes, it is possible, by choosing *m* filters of size *n* x 1.
- ○ Yes, it is possible, by choosing *n* filters of size *m* x 1.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Yes, it is possible, by choosing n filters of size m x 1.

6) Is the max-pooling layer differentiable? *0 points*

- ○ True
- ◉ False

No, the answer is incorrect.
Score: 0

Accepted Answers:
True

7) *1 point*

In the context of Deep Art, if $V \in \mathcal{R}^{32 \times (128 \times 128)}$ is the activation at a layer, then whicl of the following captures the style of the image at that layer?

- ○ V^T
- ◉ $V^T V$
- ○ $V^{-1} V$

Yes, the answer is correct.
Score: 1

Accepted Answers:
$V^T V$

8) *1 point*

Consider an image $\in \mathbb{R}^{28\times28}$ which is stored as a numpy array of size 28×28. You ca download the numpy array from here :

CLICK HERE TO DOWNLOAD THE FILES

What will be the result of applying the following kernel to this image ?

$$\begin{bmatrix} 0.053 & 0.110 & 0.054 \\ 0.111 & 0.225 & 0.111 \\ 0.054 & 0.110 & 0.053 \end{bmatrix}$$

- it will result in a sharpened image
- it will result in a blurred image
- it will result in an image which contains all the edges in the original image
- it will result in a black and white version of the original image

Yes, the answer is correct.
Score: 1

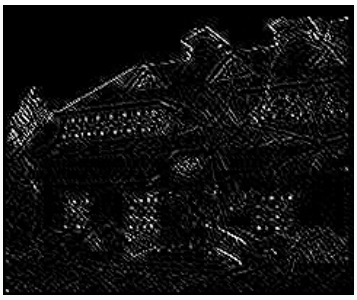
Accepted Answers:
it will result in a blurred image

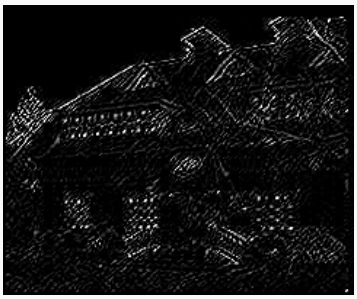
9) Consider the image shown below: *1 point*

Figure 1: House Image

What will be the result of applying the following kernel to the above image ?

$$\begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix}$$





Yes, the answer is correct.
Score: 1

Accepted Answers:

Previous Page

End

Unit 13 - Week 11

Assignment 11

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-10-17, 23:59 IST.

1) *1 point*

What is the difference between backpropagation algorithm and backpropagation through time (BPTT) algorithm ?

- There is no difference.
- Unlike backpropagation, in BPTT we add the gradients for corresponding weight for each time step.
- Unlike backpropagation, in BPTT we subtract the gradients for corresponding weight for each time step.

No, the answer is incorrect.
Score: 0

Accepted Answers:

Unlike backpropagation, in BPTT we add the gradients for corresponding weight for each time step.

2) *1 point*

What approach is taken to deal with the problem of Exploding Gradients in Recurrent Neural Networks?

- Gradient clipping
- Using modified architectures like LSTMs and GRUs
- Using dropout

No, the answer is incorrect.
Score: 0

Accepted Answers:

Gradient clipping

3) *1 point*

In the context of the state equations of LSTM, we have seen that $h_t = o_t \odot \sigma(s_t)$ where $h_t, o_t, s_t \in \mathbb{R}^n$. What is the derivative of h_t w.r.t. s_t?

- Vector
- Tensor
- Matrix

No, the answer is incorrect.
Score: 0

Accepted Answers:

Matrix

4) *1 point*

Continuing the previous question, how many non-zero entries does the derivative of h_t w.r.t. s_t have?

- No non-zero entries
- n
- $n^2 - n$

No, the answer is incorrect.
Score: 0

Accepted Answers:
n

5) In the context of LSTMs, the gradient of $\mathcal{L}_t(\theta)$ w.r.t θ_i vanishes when *1 point*

- the gradients flowing through at least one path from $\mathcal{L}_t(\theta)$ to θ_i vanishes.
- the gradients flowing through each and every path from $\mathcal{L}_t(\theta)$ to θ_i vanishes.

No, the answer is incorrect.
Score: 0

Accepted Answers:
the gradients flowing through each and every path from $\mathcal{L}_t(\theta)$ to θ_i vanishes.

6) *1 point*

Which of the following options represent the full set of equations for GRU gates where s_t represents the state of the GRU and h_t refers to the intermediate output?

- $o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$

  $i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$
- $o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$

  $i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$

$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$

7) *1 point*

Consider a GRU where the input $x \in \mathbb{R}^m$ and the state of GRU $s \in \mathbb{R}^n$ at any time step t. What is the total number of parameters in this GRU ?

- $n^2 + nm + 2n$
- $3 \times (n^2 + nm + n)$

- $n + 3 \times (n^2 + nm + n)$
- $4 \times (n^2 + nm + n)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$3 \times (n^2 + nm + n)$

8) Consider the following statements in the context of LSTMs: *1 point*

1. During forward propagation, the gates control the flow of information.
2. During backward propagation, the gates control the flow of gradients.

Which of the following option is correct ?

- Statement 1 is True and Statement 2 is False.
- Statement 2 is True and Statement 1 is False.
- Both are False.
- Both are True.

No, the answer is incorrect.
Score: 0

Accepted Answers:
Both are True.

9) *1 point*

Consider the RNN with the following equations:

$$s_t = \sigma(Ux + Ws_{t-1} + b)$$
$$y_t = \mathcal{O}(Vs_t + c)$$

where s_t is the state of the network at timestep t and the parameters W,U,V,b,c are shared across timesteps. The loss $\mathscr{L}_t(\theta)$ is defined as :

$$\mathscr{L}_t(\theta) = -\log(y_{tc})$$

where y_{tc} is the predicted probability of true output at time-step t. Given the above RNN, find $\frac{\partial \mathscr{L}_t(\theta)}{\partial s_t}$ at $t = 4$.

- $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -\frac{\mathcal{O}(Vs_4+c)}{\mathcal{O}'(Vs_4+c)}$
- $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\frac{\mathcal{O}(Vs_4+c)}{\mathcal{O}'(Vs_4+c)}$
- $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\frac{\mathcal{O}'(Vs_4+c)}{\mathcal{O}(Vs_4+c)}$
- $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\mathcal{O}'(Vs_4 + c)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\frac{\mathcal{O}'(Vs_4+c)}{\mathcal{O}(Vs_4+c)}$

10) *1 point*

Considering the same RNN setup as defined in the previous question, find $\frac{\partial \mathscr{L}(\theta)}{\partial V}$.

- $\frac{\partial \mathcal{L}(\theta)}{\partial V} = -s_t \frac{\mathcal{O}(Vs_t+c)}{\mathcal{O}'(Vs_t+c)}$
- $\frac{\partial \mathcal{L}(\theta)}{\partial V} = -s_t \frac{\mathcal{O}'(Vs_t+c)}{\mathcal{O}(Vs_t+c)}$
- $\frac{\partial \mathcal{L}(\theta)}{\partial V} = \sum_{t=1}^{T} -s_t \frac{\mathcal{O}(Vs_t+c)}{\mathcal{O}'(Vs_t+c)}$
- $\frac{\partial \mathcal{L}(\theta)}{\partial V} = \sum_{t=1}^{T} -s_t \frac{\mathcal{O}'(Vs_t+c)}{\mathcal{O}(Vs_t+c)}$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\frac{\partial \mathcal{L}(\theta)}{\partial V} = \sum_{t=1}^{T} -s_t \frac{\mathcal{O}'(Vs_t+c)}{\mathcal{O}(Vs_t+c)}$

Previous Page End

Unit 14 - Week 12

Assignment 12

The due date for submitting this assignment has passed.
As per our records you have not submitted this assignment.

Due on 2018-10-24, 23:59 IST.

1) *1 point*

Consider the task of Video QA where given a video and the question (example, "What is the person in the video doing?") the task is to generate an answer (example, "Walking."). Assume all videos are of the same length T and the answers contain a single word picked from a fixed vocabulary. We can model this task using the encoder-decoder framework as shown below:

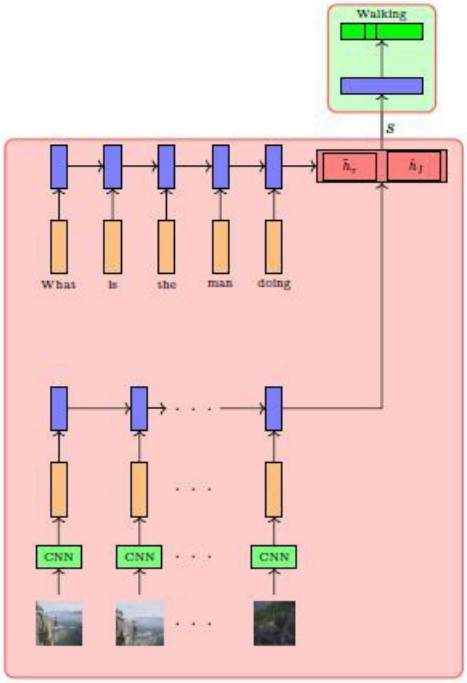


Figure 1: Video Question Answering

- **Task:** Video Question Answering
- **Data:** $\{x_i = \{video, q\}_i,\ y_i = Answer_i\}_{i=1}^{N}$
- **Model:**
  - **Encoder:**

    $\hat{h}_t =$_____, $\tilde{h}_t =$_____

    $s = [\tilde{h}_T; \hat{h}_T]$
  - **Decoder:**
- **Loss:**

  $P(y|q, Video) =$_____

  $\mathscr{L}(\theta) =$_____
- **Algorithm:** Gradient descent with backpropagation

Given the above model, what will be a natural choice for the encoder, or what will be a natural choice for $\hat{h}_t$ and $\tilde{h}_t$, where $\hat{h}_t$ represents the video encoding while $\tilde{h}_t$ represents the question encoding.

- $\hat{h}_t = CNN(Video_{it}),\ \tilde{h}_t = RNN(\tilde{h}_{t-1}, q_{it})$
- $\hat{h}_t = RNN(\hat{h}_{t-1}, Video_{it}),\ \tilde{h}_t = RNN(\tilde{h}_{t-1}, q_{it})$
- $\hat{h}_t = RNN(\hat{h}_{t-1}, CNN(Video_{it})),\ \tilde{h}_t = RNN(\tilde{h}_{t-1}, q_{it})$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\hat{h}_t = RNN(\hat{h}_{t-1}, CNN(Video_{it})),\ \tilde{h}_t = RNN(\tilde{h}_{t-1}, q_{it})$

2) In the Video Question Answering task defined in Question 1, what will be the equation of the decoder? *1 point*

- $P(y|q, Video) = sigmoid(Vs + b)$
- $P(y|q, Video) = softmax(Vs + b)$
- $P(y|q, Video) = Vs + b$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$P(y|q, Video) = softmax(Vs + b)$

3) In the Video Question Answering task defined in Question 1, what will be an appropriate loss function ? *1 point*

- $\mathscr{L}(\theta) = -\log P(y = \ell|video)$
- $\mathscr{L}(\theta) = -\log P(y = \ell|video, q)$
- $\sum_{t=1}^{T} \mathscr{L}_t(\theta) = -\sum_{t=1}^{T} \log P(y_t = \ell_t|y_1^{t-1}, video, q)$

No, the answer is incorrect.
Score: 0

Accepted Answers:
$\mathscr{L}(\theta) = -\log P(y = \ell|video, q)$

4) *1 point*

Consider the task of Video Captioning where you want to generate a textual description given a video. For example, in the following example, we want to generate the caption "A man is walking on a rope." Assume all videos are of same length T and the caption generated for each video is of length J. We can model this task using an encoder, decoder and attention mechanism as shown below:

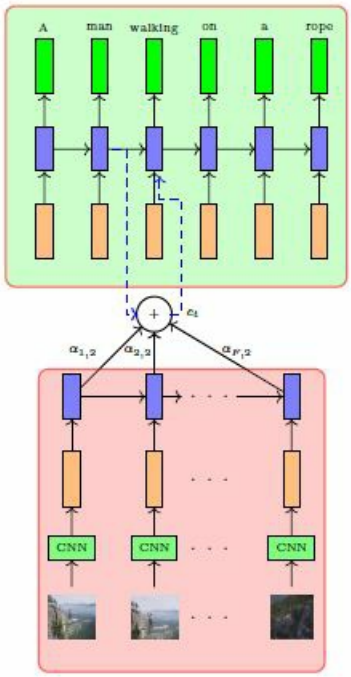


Figure 2: Video Captioning

- **Task:** Video Caption Generation (With attention)
- **Data:** $\{x_i = video_i,\ y_i = desc_i\}_{i=1}^{N}$
- **Model:**
  - **Encoder:**

$$h_t = \_\_\_\_\_$$
$$s_0 = h_T$$

  - **Decoder:**

$$e_{jt} = V_{attn}^T tanh(U_{attn} h_j + W_{attn} s_t)$$
$$\alpha_{jt} = softmax(e_{jt})$$
$$c_t = \_\_\_\_\_$$
$$s_t = \_\_\_\_\_$$
$$l_t = softmax(V s_t + b)$$

- **Loss:**

$$\sum_{t=1}^{T} \mathscr{L}_t(\theta) = -\sum_{t=1}^{T} \log P(y_t = \ell_t | y_1^{t-1}, x)$$

- **Algorithm:** Gradient descent with backpropagation

What will be the encoder equation for this task, *i.e.*, what will be h_t?

- $h_t = RNN(h_{t-1}, CNN(f_{attn}(x_{it})))$
- $h_t = RNN(h_{t-1}, f_{attn}(CNN(x_{it})))$
- $h_t = RNN(h_{t-1}, CNN(x_{it}))$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$h_t = RNN(h_{t-1}, CNN(x_{it}))$

5) In the context of the Video Captioning task defined in Question 4, we have seen the equation, *1 point*

$$e_{jt} = V_{attn}^T tanh(U_{attn} h_j + W_{attn} s_t)$$

where e_{jt} is a _____ which tells us how much attention should be given to the j-th input frame at time step t.

- Scalar
- Vector
- Matrix
- Tensor

No, the answer is incorrect.
Score: 0

Accepted Answers:

Scalar

6) In the context of the Video Captioning task defined in Question 4, what will be the equation to calculate c_t which is the context being passed to the decoder at timestep t? *1 point*

- $c_t = \alpha_{jt} s_j$, for $t = 1$ to T
- $c_t = \alpha_{jt} h_j$, for $t = 1$ to T
- $c_t = \sum_{t=1}^{T} \alpha_{jt} s_j$
- $c_t = \sum_{t=1}^{T} \alpha_{jt} h_j$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$c_t = \sum_{t=1}^{T} \alpha_{jt} h_j$

7) In the context of the Video Captioning task defined in Question 4, what will be the equation to calculate s_t which is the hidden state of the decoder at time step t? *1 point*

- $s_t = \text{RNN}(s_t, [e(\hat{y}_t), c_{t-1}])$
- $s_t = \text{RNN}(s_{t-1}, [e(\hat{y}_{t-1}), c_{t-1}])$
- $s_t = \text{RNN}(s_{t-1}, [e(\hat{y}_{t-1}), c_t])$
- $s_t = \text{RNN}(s_t, [e(\hat{y}_t), c_t])$

No, the answer is incorrect.
Score: 0

Accepted Answers:

$s_t = \text{RNN}(s_{t-1}, [e(\hat{y}_{t-1}), c_t])$

8) Consider the output of the 4^{th} convolutional layer of VGGNet network given in Slide 50 of Lecture 15, which is a $28 \times 28 \times 512$ size feature map. If we were to use this model as an encoder and then introduce an attention mechanism then how many locations will the model have to learn to attend to? ? *1 point*

- 196 locations
- 512 locations
- 784 locations

No, the answer is incorrect.
Score: 0

Accepted Answers:

784 locations