CSE 493 G1/ 599 G1
Deep Learning
Spring 2024 Quiz 2

SOLUTIONS

Feb 2, 2024

Full Name: ______________________________

UW Net ID: ______________________________

Question		Score
True/False	(5 pts)	
Multiple Choice	(8 pts)	
Short Answer	(9 pts)	
Total	(22 pts)	

Welcome to the CSE 493 G1 Quiz 2!

- The exam is 20 min and is **double-sided**.
- No electronic devices are allowed.

I understand and agree to uphold the University of Washington Honor Code during this exam.

Signature: ______________________ Date: ______________

Good luck!

Mean: 16.4
Median: 18.0
Stdev: 4.54

This page is left blank for scratch work only. DO NOT write your answers here.

1 True / False (5 points) - Recommended 5 Minutes

Fill in the circle next to True or False, or fill in neither. Fill it in completely like this: ●. No explanations are required.

Scoring: Correct answer is worth 2 points. To discourage guessing, incorrect answers are worth -1 points. Leaving a question blank will give 0 points.

1.1 ReLU (Rectified Linear Unit) activation functions can suffer from the problem of dead neurons.

○ True

○ False

SOLUTION:
True, In a ReLU activation, if the input is less than zero, the output is zero. This can lead to situations where neurons stop responding to inputs altogether, known as dead ReLU problem.

1.2 One downside of layer normalization is that it works differently during training and testing, leading to a common source of errors.

○ True

○ False

SOLUTION:
False. Layer normalization maintains consistent behavior between training and testing, as it computes normalization statistics from the inputs to each layer for each training case, unlike batch normalization which behaves differently in these phases.

1.3 Batch normalization is applied to the input of each layer before the activation function.

○ True

○ False

SOLUTION:
True, Batch normalization is typically applied to the inputs of a layer, normalizing them before they are passed through the activation function.

Grading Note: This question was poorly worded. Here, we were alluding to the definition of "layer" from the assignments' `FullyConnectedNet`, which is as follows: `affine - [batch/layer norm] - relu - [dropout]`. Assuming this definition, the question extrapolates that BN is usually applied immediately before `relu`.

1.4 Dropout regularization can be applied only to fully connected layers in a neural network.

○ True

○ False

SOLUTION:
False, While the lecture slides specifically mention an example of dropout in a 3-layer network, they do not restrict its application to fully connected layers only. Dropout can be applied to different types of layers in a neural network. People have tried using dropout for conv layers and recurrent layers.

The key insight students needed to answer this question correctly was that there's no inherent restriction in the design of dropout that limits us to fully connected layers.

1.5 Saliency maps in convolutional neural networks (CNNs) are primarily used to reduce the dimensionality of the output layer for easier interpretation and visualization.

◯ True

◯ False

SOLUTION:

False, This question is relevant to the topic of saliency maps as discussed in the lecture, and the statement is false because saliency maps are actually used to highlight the areas of an input image that are most relevant to the network's decision, rather than for reducing dimensionality of the output layer.

2 Multiple Choices (8 points) - Recommended 6 Minutes

Fill in the circle next to the letter(s) of your choice (like this: ●). No explanations are required. **Choose ALL options that apply.**

Each question is worth 4 points and the answer may contain one or more options. Selecting all of the correct options and none of the incorrect options will get full credits. For questions with multiple correct options, each incorrect or missing selection gets a 2-point deduction (up to 4 points).

2.1 During the training of a deep neural network using Stochastic Gradient Descent (SGD), you notice that the model's performance on the validation set starts declining after a certain number of epochs, even though the loss on the training set is still going down and the performance continues to improve. What would you do to improve the model's training (Treat each option as independent of each other)? Select ALL options that apply.

- ◯ A: Decrease the regularization parameter λ.
- ◯ B: Reduce the number of layers in the neural network.
- ◯ C: Increase the number of neurons on each layer of the neural network.
- ◯ D: Increase the dropout rate.
- ◯ E: Use more powerful GPUs.

SOLUTION:
B, D

This problem is to test the students' understanding of model overfitting.

A is False. The regularization term is to help prevent overfitting, so it should increase.

B is True. Reducing the number of layers of the neural network would reduce the model complexity and further reduce the risk of overfitting.

C is False. Opposite of B.

D is True because dropout is a technique used to prevent overfitting by randomly setting a fraction of input units to 0 at each update during training time.

E is False. Irrelevant to overfitting.

2.2 Which of the following would you consider to be valid activation functions (elementwise non-linearities) to train a neural net in practice?

- ◯ A: $f(x) = max(0, x)$
- ◯ B: $f(x) = 0.8x + 1$
- ◯ C: $f(x) = \begin{cases} \min(x, .1x) & \mid \quad x >= 0 \\ \min(x, .1x) & \mid \quad x < 0 \end{cases}$
- ◯ D: $f(x) = \begin{cases} \max(x, .1x) & \mid \quad x >= 0 \\ \min(x, .1x) & \mid \quad x < 0 \end{cases}$
- ◯ E: $f(x) = \begin{cases} \max(x, .111x) & \mid \quad x >= 0 \\ \min(x, .111x) & \mid \quad x < 0 \end{cases}$

SOLUTION:
A, C

A is True. This is the Rectified Linear Unit (ReLU) function. It's a widely used activation function in neural networks due to its simplicity and efficiency. It outputs the input directly if it is positive, otherwise, it outputs zero.

B is False. This function is a linear function. Linear functions are not typically used as activation functions in neural networks because they do not introduce non-linearity. Without non-linearity, stacking multiple layers of neurons would not be more powerful than using a single layer, as the composition of linear functions is still a linear function.

C is True. This is very similar to the inverse of a Leaky ReLU function. It works because it introduces non-linearity. See the following Desmos link: http://tinyurl.com/cse493g1-24wi-quiz2-c.

D is False. This is a linear function. See the following Desmos link: http://tinyurl.com/cse493g1-24wi-quiz2-d

E is True. This is a linear function. See the following Desmos link: http://tinyurl.com/cse493g1-24wi-quiz2-e

3 Short Answers (9 points) - Recommended 9 Minutes

Please make sure to write your answer only in the provided space.

1. (2D Convolution) (9 points) We have introduced Convolutional Neural Networks (CNNs) and the concept of 2D convolution in our lecture. Convolutional operations play a pivotal role in the field of deep learning, especially in the processing and understanding of image data. These operations help in extracting important features from the input data by applying filters that capture spatial hierarchies and patterns.

   In this question, you are provided with an input matrix I and a filter F, represented as:

$$I = \begin{array}{|c|c|c|c|} \hline 0 & 2 & 1 & 2 \\ \hline 0 & 3 & 2 & 2 \\ \hline 0 & 2 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 \\ \hline \end{array} \qquad F = \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array}$$

   In this question, you will apply the filter F to the input matrix I using 2D convolution operations under different padding and stride configurations. Padding adds a border of zeros to the input matrix to control the spatial size of the output feature map. Stride determines the step size the filter takes as it slides over the input matrix.

   Your task is to calculate the output matrices for the following configurations. For each configuration, write your answer in the grid with the right size. **Note: If the grid is too small, you can draw your own on the right empty space of each part.**

   (a) (2 points) Apply F on I with **Padding = 0, Stride = 1**

   (b) (2 points) Apply F on I with **Padding = 0, Stride = 2**

   (c) (2 points) Apply max pool on I with 2×2 filter and **Stride =2**

(d) (3 points) Apply F on I with **Padding = 1, Stride = 1**

SOLUTION:

(a) Apply F on I with **Padding = 0, Stride = 1**

3	4	3
2	4	3
0	3	1

(b) Apply F on I with **Padding = 0, Stride = 2**

3	3
0	1

(c) Apply max pool on I with 2×2 filter and **Stride =2**

3	2
2	1

(d) Apply F on I with **Padding = 1, Stride = 1**

0	2	1	2	0
0	3	4	3	2
0	2	4	3	2
1	0	3	1	1
0	1	0	1	0