CENG 506 Deep Learning
Exercises, 27th of March 2023

Q1. True or False? Write T or F to the space given at the beginning of each sentence.

____ Zero centering and scale normalization are two essential steps of data preprocessing

____ Sigmoid neuron's activation saturates at either 0 or 1, using tanh function solves the saturation problem.

____ Using ReLU solves the vanishing gradient problem.

____ More hidden layers in neural networks enable us to model/compute more complex functions.

____ Ideally all weights are initialized to be the same (with a value close to zero, e.g. 0.001).

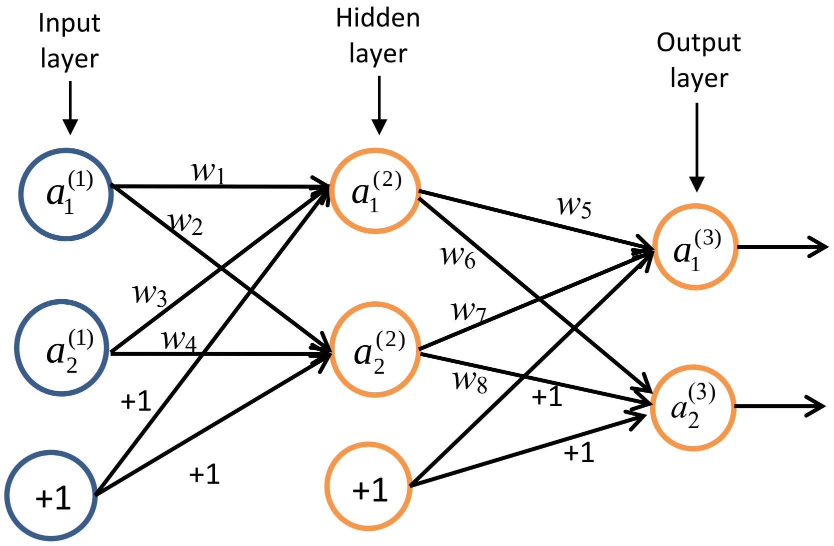
____ Sigmoid outputs are always positive. When the input to a neuron (x) is all positive, it causes weight updates to be in the same direction (all positive or all negative).

Answer: T F T T F T

Q2. Suppose this is an image classification task. There is only one animal in an image. You are asked to predict the animal in a given image. There are five classes: dog, cat, bear, alligator, zebra. Ground truth labels for samples are organized as one-hot vectors, i.e. y=(1,0,0,0,0) and y=(0,0,1,0,0) are the ground truth vectors for dog and bear respectively. After some training, for an image of bear, model predicts with a normalized probability vector y=(0.15,0.15,0.4,0.2,0.1), what is the value of the softmax loss?

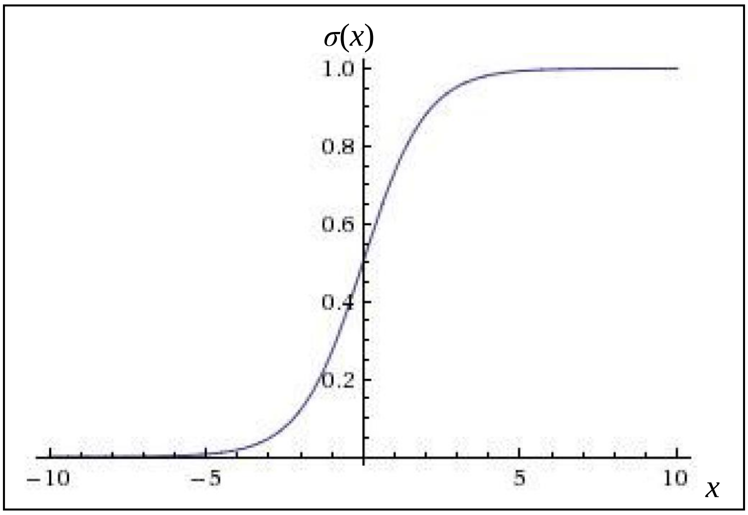
Ans: - log (0.4) = 0.9

Q3.



For the above neural network, hidden and output layer neurons use sigmoid function as the activation function.

Sigmoid function produces the values shown in the plot given on the right.



a) Take $a_1^{(1)}$=0.5, $a_2^{(1)}$=0.2, w_1=0.8, w_2=0.2, w_3=0.75, w_4=0.1, w_5=0.4, w_6=0.2, w_7=0.5, w_8=0.1 and apply forward propagation. Please read the sigmoid output from the plot given above (select approximate value by eye, don't worry too much).

b) For this input, the desired output (ground truth) is 0.1 and 0.5 for $a_1^{(3)}$ and $a_2^{(3)}$ respectively. With this ground truth value, we would like to update w_5 with $w_5 = w_5 - \alpha \frac{\partial L}{\partial w_5}$ where α is the learning rate. Compute $\partial L/\partial w_5$.

c) Compute $\partial L/\partial w_1$

Answer:

a) $a_1^{(2)}$= σ (0.5*0.8+0.2*0.75+1) = σ (1.55) =0.75
$a_2^{(2)}$= σ (0.5*0.2+0.2*0.1+1) = σ (1.12) =0.7
$\boldsymbol{a_1^{(3)}}$= σ (0.75*0.4+0.7*0.5+1) = σ (1.65) = **0.75**
$\boldsymbol{a_2^{(3)}}$= σ (0.75*0.2+0.7*0.1+1) = σ (1.22) = **0.7**

b) $\frac{\partial L}{\partial a_1^{(3)}}$ = - (target-out) = - (0.1-0.75) = 0.65

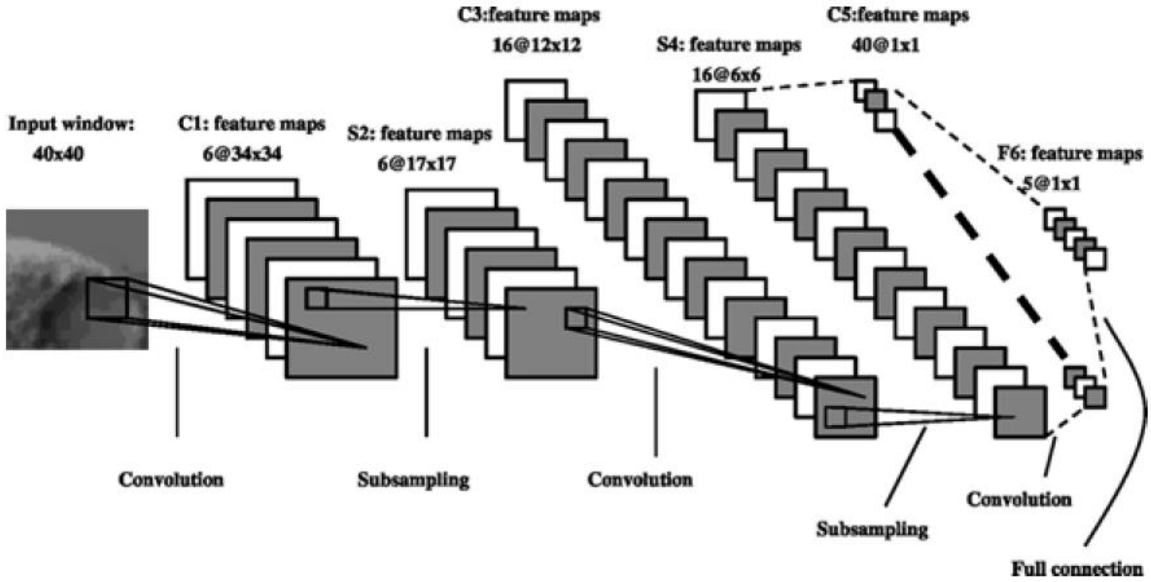
$\frac{\partial L}{\partial a_2^{(3)}}$ = - (target-out) = - (0.5-0.7) = 0.2

$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \cdot \frac{\partial z_1^{(3)}}{\partial w_5}$ = 0.65 * 0.75 * (1-0.75) * 0.75 = …

c) $\frac{\partial L}{\partial w_1} = \left(\frac{\partial L}{\partial a_1^{(3)}} \cdot \frac{\partial a_1^{(3)}}{\partial z_1^{(3)}} \cdot \frac{\partial z_1^{(3)}}{\partial a_1^{(2)}} + \frac{\partial L}{\partial a_2^{(3)}} \cdot \frac{\partial a_2^{(3)}}{\partial z_2^{(3)}} \cdot \frac{\partial z_2^{(3)}}{\partial a_1^{(2)}}\right) \cdot \frac{\partial a_1^{(2)}}{\partial z_1^{(2)}} \cdot \frac{\partial z_1^{(2)}}{\partial w_1} =$

= (0.65 * 0.75 * (1-0.75) * 0.4 + 0.2 * 0.7 * (1-0.7) * 0.2)* 0.75 * (1-0.75) * 0.5

Q4.



In the figure, you see a CNN architecture, where input is a 40x40 grayscale image, C1 is a conv. layer with 6 feature maps of 34x34, S2 is a max-pooling layer with 6 feature maps of 34x34, C3 is a conv. with 16 feature maps of 12x12, S4 is a max-pooling layer etc. Please fill in the missing information for the layers given below. Please note that there is no single correct answer, i.e. your choice of filter-size affects stride, padding etc. and vice-versa.

Answers:

C1)

Filter-size: 7x7x1 Stride: 1 Padding: 0

Number of parameters to learn while optimization: (7x7+1)x6=300

S2)

Filter-size: 2x2 Stride: 2 Padding: 0

Number of parameters to learn while optimization: 0

C3)

Filter-size: 6x6x6 Stride: 1 Padding: 0

Number of parameters to learn while optimization: (6x6x6+1)x16=217x16=3472

S4)

Filter-size: 2x2 Stride: 2 Padding: 0

Number of parameters to learn while optimization: 0

Q5. Which one of the following is true about max-pooling?

a) It allows a neuron in a network to have information about features in a larger part of the image, compared to a neuron at the same depth in a network without max pooling.
b) It increases the number of parameters when compared to a similar network without max pooling.
c) It increases the sensitivity of the network towards the position of features within an image.

Answer: a

Q6. Explain why dropout in a neural network acts as a regularizer.

Q7. Apply max-pooling on the following image patch (use 3x3 filters and a stride of 2):

2	5	7	1	6
7	4	7	9	1
0	3	6	6	2
8	4	2	6	9
5	0	8	1	4

Q8. When the input is 2-dimensional, you can plot the decision boundary of your neural network and clearly see if there is overfitting. How do you check overfitting if the input is 10-dimensional?

Answer: Compute cost/loss and accuracies in the validation and training set. If there is a significant difference between validation and training set, then you have overfitting problem.

Q9. What's the risk with tuning hyperparameters using the test dataset?

Answer: The model will not generalize well to unseen data because it overfits the test set. Tuning model hyperparameters to a test set means that the hyperparameters may overfit to that test set. If the same test set is used to estimate performance, it will produce an overestimate. Using a separate validation set for tuning and test set for measuring performance provides unbiased, realistic measurement of performance.

Q10. You are building a binary classifier for recognizing cucumbers (y=1) vs. watermelons (y=0). Which one of these activation functions would you recommend using for the output layer?

a) ReLU b) Leaky ReLU c) Sigmoid d) tanh

Answer: c

Q11. T or F?

____ If we are training a neural network for a multi-class classification, number of units in the output layer is equal to the number of classes.

____ Number of units in the input layer of a neural network is equal to the number of features

+ 1(bias unit).
____ Sigmoid outputs are not zero-centered, tanh function outputs are zero-centered
____ While initializing weights, centering around zero is not enough, we should adjust the magnitude of the weights close to zero.

Answers: T T T T

Q12. You have an input volume that is 32x32x16, and apply max pooling with a stride of 2 and a filter size of 2. What is the output volume?

Answer: 16x16x16

Q13. Suppose you have an input volume of dimension 64x64x16. How many parameters would a single 1x1 convolutional filter have (including the bias)?

Answer: 17

Q14. Which of the following do you typically see as you move to deeper layers in a ConvNet?

a) H and W increases, while number of channels C decreases
b) H and W decreases, while number of channels C also decreases
c) H and W increases, while number of channels C also increases
d) H and W decrease, while number of channels C increases

Answer: d