STAT 433 – Midterm Part I

1. During backpropagation, when the gradient passes backward through a sigmoid activation function, the gradient will always decrease in magnitude.
   A. True
   B. False
2. Suppose that you find that your model's training error looks so good (potential overfitting). What can you do to address this issue? (Check all that apply)
   A. Data augmentation
   B. Dropout
   C. Batch Normalization
   D. RMSprop Optimizer
3. Which of the following is true?
   A. Batch Normalization is an alternative method of dropout.
   B. Batch Normalization makes training faster.
   C. Batch Normalization is a non-linear transformation to give nonlinearity to the network.
   D. Batch Normalization is standardizing the data before training neural network.
4. You want to make the weights sparse and smaller. How can you do that? Why?

5. $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ has a similar performance as sigmoid function except that it is zero-centered. Write down tanh(x) in terms of $\sigma(x)$ where $\sigma(x) = 1/(1 + e^{-x})$. Show your work to get the full credit.

6. You have a single layer neural network for a binary classification with a sigmoid activation function as below. (X: nXm matrix, predicted y & true label y: 1 X m)

$$z = WX + b$$
$$h = \sigma(z)$$
$$\hat{y} = h$$
$$L = -\sum_{i}^{m} y_i \log \hat{y} + (1 - y_i) \log(1 - \hat{y_i})$$

What is $\frac{\partial L}{\partial W}$? Write your answer as a matrix-matrix multiplication.

7. (continued from the above question) suppose that you apply ReLU activation before sigmoid activation. i.e., $\hat{y} = \sigma(ReLU(z))$. Then you classify the object by checking if $\hat{y} \geq 0.5$ or $\hat{y} < 0.5$. What will happen? Why?

8. Suppose that your classmate finds an activation function that is similar to ReLU such that

$$f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases}$$

Will you use this? Why?

9. Provide two reasons why we are using convolutional layers instead of fully connected layers for image classification.

10. Consider to build a CNN for an image classification problem in which the layers are defined by the left column below. Fill the table below. Assume that width & height of the kernels (for Conv, Pool) are the same. Stride 1 Pad 1 for convolving layers. Stride 2 Pad 0 for Pooling layers. FC: a fully-connected layer.

	Output Size		Layer		
Layer	C	H/W	filters	kernel	Number of parameters
Input	3	32	-	-	0
Conv			16	3	
ReLU			-	-	
Pool			-	2	
BatchNorm			-	-	
Conv			16	3	
ReLU			-	-	
Pool			-	2	
Flatten		-	-	-	
FC	10	-	-	-	