Name: _____

Computing ID: _____

Signature: _____

Q1 (10 points): Choose the best answer in each of the questions below.

1. A multi-layer neural network model trained using stochastic gradient descent on the same dataset with different initializations for its parameters is guaranteed to learn the same parameters.
   A. True B. False

2. The idea of backpropagation is essentially Chain Rule mathematically and Dynamic Programming computationally.
   A. True B. False

3. In a simple MLP model with 8 neurons in the input layer, 5 neurons in the hidden layer and 1 neuron in the output layer. What is the size of the weight matrices between the input and hidden layer and hidden and output layer?
   A. 5x8,1x5 B. 8x5, 1x5 C. 8x5, 5x1 D. 5x1, 8x5

4. Changing Sigmoid activation to ReLu will help to get over the vanishing gradient issue.
   A. TRUE, B. FALSE

5. Which of the following functions can be used as an activation function in the output layer if we wish to predict the probabilities of n classes (p1, p2..pn) such that sum of p over all n equals to 1?
   A. Softmax, B. ReLu, C. Sigmoid, D. Tanh

6. Which of the following statements about pooling in CNN models is INCORRECT?
   a. Pooling layers provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map.
   b. A pooling layer is a new layer added before applying a nonlinear layer (or activation layer).
   c. Two common pooling methods are average pooling and max pooling.
   d. One intuitive reasoning behind pooling is that, it can progressively reduce the spatial size of the representation to reduce the amount of parameters in the network, and hence to also control overfitting.

7. Which of the following statement is true regarding dropout?
   1: Dropout gives a way to approximate by combining many different architectures
   2: Dropout demands high learning rates
   3: Dropout can help preventing overfitting
   A. Both 1 and 2 B. Both 1 and 3 C. Both 2 and 3 D. All 1, 2 and 3

8. What steps can we take to prevent overfitting in a Neural Network?
   A. Data Augmentation B. Early Stopping C. Dropout D. All of the choices

9. Which of the following techniques perform similar operations as dropout in a neural network?
   A. Bagging B. Boosting C. Stacking D. None of these

10. Which of the following gives non-linearity to a neural network?
A. Stochastic Gradient Descent B. Rectified Linear Unit C. Convolution function D. None of these

Q2 (10 points): Answer each of the following questions:

A. What are the steps for using a gradient descent algorithm?
1. Calculate error between the actual value and the predicted value
2. Reiterate until you find the best weights of network
3. Pass an input through the network and get values from output layer
4. Initialize random weight and bias
5. Go to each neurons which contributes to the error and change its respective values to reduce the error

B. Suppose you have a dataset from where you have to predict three classes. Then which of the following configuration you should use in the output layer?
1. Activation function = softmax, loss function = cross entropy
2. Activation function = sigmoid, loss function = cross entropy
3. Activation function = softmax, loss function = mean squared error
4. Activation function = sigmoid, loss function = mean squared error

C. In a neural network, which of the following causes the loss not to decrease faster?
1. Stuck at a local minima
2. High regularization parameter
3. Slow learning rate
4. All of the above

D. Making your network deeper by adding more parametrized layers will always...
1. slow down training and inference speed.
2. reduce the training loss.
3. improve the performance on unseen data.
4. None of above

E. Which of the following is true about dropout?
1. Dropout leads to sparsity in the trained weights
2. At test time, dropout is applied with inverted keep probability
3. The larger the keep probability of a layer, the stronger the regularization of the weights in that layer
4. None of the above

F. Which of the following would you consider to be valid activation functions to train a neural net in practice?
1. $f(x) = - \min(2, x)$
2. $f(x) = 0.9x + 1$
3. Both are good choices.

G. Which of the following can be significant causes of the vanishing gradient problem in deep learning? Choose all correct answers.
1. The second derivative being equal to zero in an ReLU activation function.
2. Saturation of the sigmoid activation function.
3. Many hidden layers in the same network.
4. Using the CPU instead of a GPU.

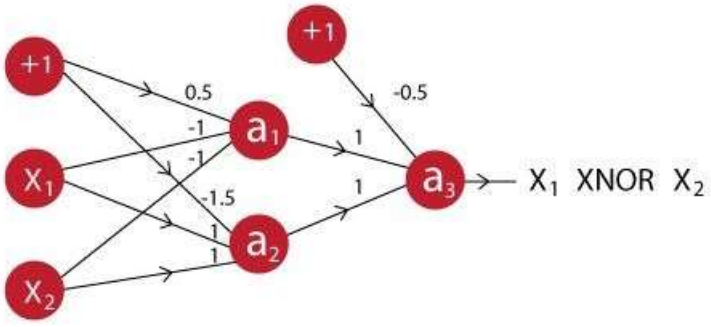
H. Autoencoding by a neural network is (choose all correct answers)
1. trivial to achieve when the hidden layer and input layer have the same size.
2. a relatively simple task, but a useful support for more challenging tasks.
3. useful for finding compression codes of the input patterns.
4. easily regularized to produce sparse embeddings.

I. In deep learning, regularization is: (choose all correct answers)
1. often achieved by using dropout
2. aided by dense hidden-layer representations
3. an attempt to reduce testing error while keeping training error low
4. an attempt to enhance generalization

J. Which of the following is true about the softmax function? (choose all correct answers)
1. It is generally more useful for classification problems than for regression problems.
2. It can scale a vector of activations.
3. It can be used to simulate competition among same-layer neurons by magnifying the strengths of the more active neurons and reducing those of the less active neurons.
4. None of above.

Q3 (5 pints):

1. (2 points) A network is created when we multiple neurons stack together. Let us take an example of a neural network simulating an XNOR function.



You can see that the last neuron takes input from two neurons before it. The activation function for all the neurons is given by:

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

Suppose X1 is 0 and X2 is 1, what will be the output for the above neural network?

2. (3 points) Can you represent the following Boolean function with a single unit from a neural network? If yes, show the weights. If not, explain why not in 1-2 sentences.

A	B	f(A,B)
1	1	0
0	0	0
1	0	1
0	1	0

Q4 (15 pints): Short Answer

A. (2 points) You're training a neural network and notice that the validation error is significantly lower than the training error. Name two possible reasons for this to happen.

B. (3 points) Why sigmoid activation function could cause vanishing gradient problem? How does ReLu activation function alleviate this problem?

C. (2 points) Explain why we need activation functions.

D. (2 points) Why are convolutional layers more commonly used than fully-connected layers for image processing?

E. (3 points) Deep networks typically require a lot of data to train from scratch, but can be "fine tuned" for another task quickly. Explain this in terms of the features computed in the layers.

F. (3 points) Assume we have a set of data from patients who have visited UVA hospital during the year 2021. A set of features (e.g., temperature, height) have been also extracted for each patient. Our goal is to decide whether a new visiting patient has any of diabetes, heart disease, or Alzheimer (a patient can have one or more of these diseases). We have decided to use a neural network to solve this problem. We have two choices: either to train a separate neural network for each of the diseases or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Which method do you prefer? Justify your answer.