BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJ.)

First Semester 2023-24 CS F425 Deep Learning

COMPREHENSIVE EXAMINATION [Closed Book]

Date: 20th December 2023 Weightage: 35 % Max. Marks: 70 Duration: 180 minutes

Part A [CLOSED BOOK] **[28 Marks]**

Important Instructions:

1. The exam has **THREE** parts – A, B and C. Part A and B are closed book and are provided to you in the beginning of the exam. The **recommended** time for them is total two hours. You can collect Part C (which is Open Book) whenever you submit Parts A and B.
2. This is Part A containing short-answer type questions.
3. Any overwritten answers would not be considered for a recheck request.
4. For Q1 & Q2: write your answers in the grid provided on the right.

Q1 answers (MCQ):		Q2 answers (True/False):		MARKS	
i)		i)		Marks of Q1: Marks of Q2: Marks of Q3:	
ii)		ii)			
iii)		iii)			
iv)		iv)			
v)		v)		TOTAL MARKS:	
vi)		vi)			
vii)		vii)		Recheck request:	
viii)		viii)			
ix)		ix)			
x)		x)			

1. Multiple Choice Questions **(+1 for right answer, -0.5 for wrong)**. Only one correct answer. [1*10 = 10]

i). Which of the following activation function can lead to the vanishing gradient problem?
A). ReLU, B). tanh, C). Leaky ReLU, D). None of these.

ii). Which of the following techniques can NOT help prevent a model from overfitting?
A). Data augmentation, B). Dropout, C). Early stopping, D). None of these

iii). After training a neural network, you observe a large gap between the training accuracy (95%) and the test accuracy (35%). Which of the following methods can be used to reduce this gap?
A). Generative adversarial network., B). Sigmoid activation, C). RMSprop optimizer, D). Dropout.

iv). Which of the following regularization methods leads to weight sparsity?
A). L1 regularization, B). L2 regularization, C). Early stopping, D). None of these.

v) Which of the following layers is generally NOT a part of a CNN?
A) Convolutional Layer B) Pooling Layer C) Code Layer D) Fully connected Layer

vi). Which of the below can you use to solve the exploding gradient problem?
A) Use SGD optimization, B) Oversample minority classes, C). Increase the batch size, D) Impose gradient clipping.

vii). If the input to the CNN of size 24x24 is convolved with a kernel of size 7x7 and same padding is used, what will be the size of the output matrix? Consider stride of 1.
A) 18x18 B) 24x24 C) 17x17 D) Cannot be determined with the information provided

viii). The convolution operation doesn't fully use the pixels at the corners of an image. This is resolved by the use of:
A) Padding B) Striding C) Kernels D) Pooling

ix). Which of the following is true about dropout?

A) Dropout leads to sparsity in the trained weights B) At test time, dropout is applied with inverted keep probability

C) Larger the keep probability of a layer, stronger the regularization of the weights in that layer D) None of these

x). Which of the following is TRUE about Momentum?

A) It helps in accelerating SGD in a relevant direction B) It helps SGD in avoiding local minima

C) It helps in faster convergence D) All of these

2. Answer as TRUE or FALSE. No reasoning or justification required. **(+1 for right answer, -0.5 for wrong)** [1*10=10]

i) Convolutional networks generally have more parameters than their equivalent fully connected Networks

ii) Autoencoders are able to compress data and thus can be used as a generic data compression algorithm.

iii) The output of the autoencoder will not be exactly the same as the input, and thus they are “lossy”.

iv) Autoencoders are considered a supervised learning technique since they produce the reconstructed image using the original image as an input.

v) An autoencoder can be forced to learn useful features by adding random noise to its inputs and making it recover the original noise-free data.

vi) Apart from being an optimization technique, Batch normalization also acts as a regularizer and often eliminates the need for using Dropout.

vii) Regularization is intended to reduce the training error as well as the generalization error.

viii) Pooling layers involve many fixed computations *and hence* they slow down the computation in a neural network.

ix) the basic concept behind RNNs is that RNNs use recurrent features from dataset to find the best optimization.

x) In general, training a GAN involves alternating periods where the discriminator trains for one or more epochs followed by the generator being trained for one or more epochs

3. If the input is of size 256x256x6 and the neural network structure is as indicated in the first column below, calculate the output feature map dimensions for each layer. [1*8=8]

The notation follows the convention:

- CONV-K-N denotes a convolutional layer with N filters, each them of size KxK. Padding and stride parameters are always 0 and 1 respectively.
- POOL-K indicates a KxK pooling layer with stride K and padding 0.
- FC-N stands for a fully-connected layer with N neurons.

Write your answer in the space provided in the table below.

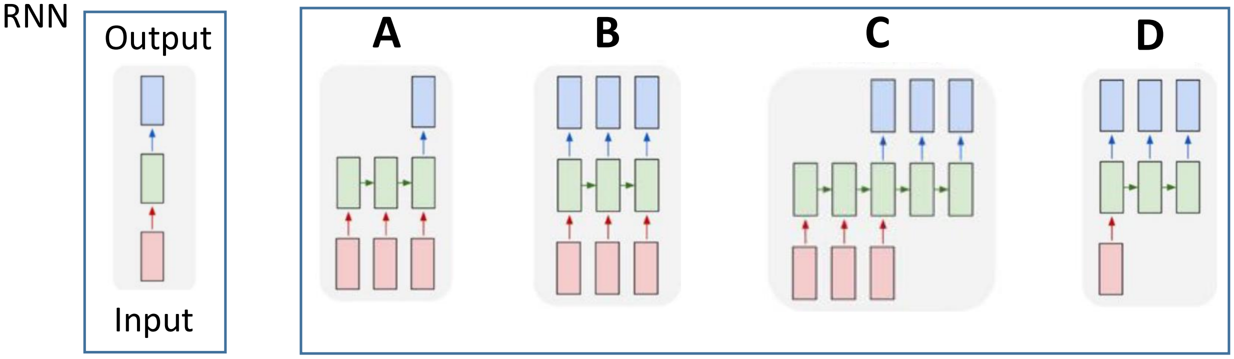
Layer	**Feature map dimensions**
INPUT	256x256x6
CONV-57-64	
POOL-2	
CONV-5-32	
POOL-2	
CONV-5-64	
POOL-2	
POOL-2	
FC-9	

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJ.)
First Semester 2023-24 CS F425 Deep Learning
COMPREHENSIVE EXAMINATION [Closed Book]
Date: 20th December 2023 Weightage: 35 % Max. Marks: 70 Duration: 180 minutes

Part B [CLOSED BOOK] **[22 Marks]**

This is Part B, Closed Book. Together with Part A, you are recommended to finish this in two hours. Once you submit Parts A and B, you can collect Part C [Open Book].

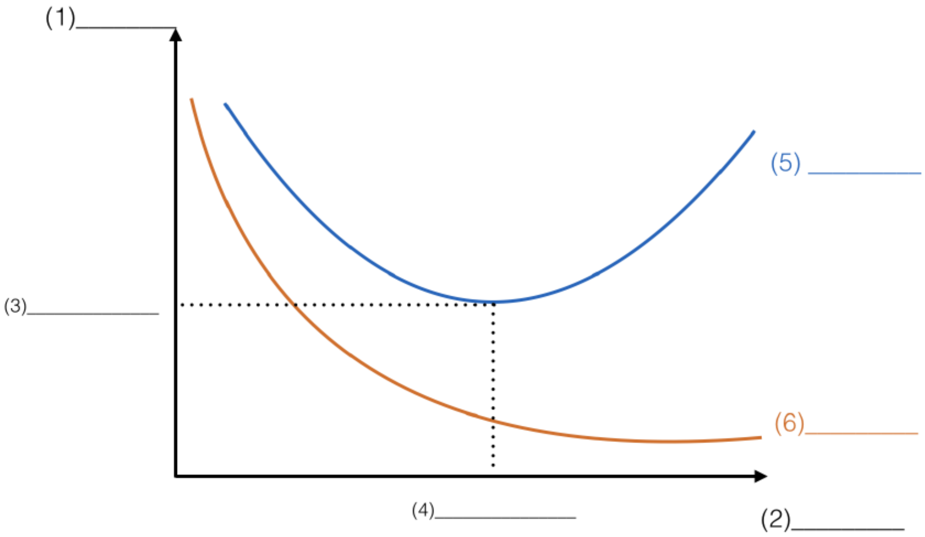
Q.1. Consider the following types of sequence modelling scenarios represented using an unfolded recurrent neural network (RNN) over time-steps: [1+1+1+1=4]



Categorize each of the following applications into any <u>one</u> of the above types i.e. A, B, C, or D. No reasoning or explanation required. (Note: No marks will be awarded if more than one answer is written for an application):

i). Image captioning, ii). Sentiment prediction, iii). Machine translation, iv). Video frame classification.

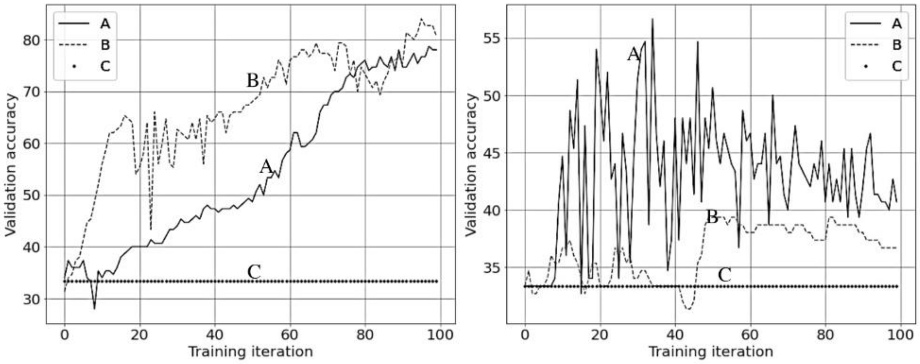
Q.2. Fill in the blanks in the following graph with regard to the regularization method "Early stopping". [3]



Q.3. You use vanilla (batch) gradient descent to optimize your loss function, but realise you are getting poor training loss. You notice that you're not shuffling the training data and feel that it might be a cause. Would shuffling the training data help in this regard? Give a clear YES or NO as an answer and then give a 1-2 lines justification. [2]

Q.4. Suppose we train two different deep CNNs to classify images:

Using ReLU as the activation function, and (ii) using sigmoid as the activation function. For each of them, we try initializing weights with the three different initialization methods, while the biases are always initialized to all zeros. We plot the validation accuracies with different training iterations below: [3]



(i) CNN with ReLU as the activation

(ii) CNN with sigmoid as the activation

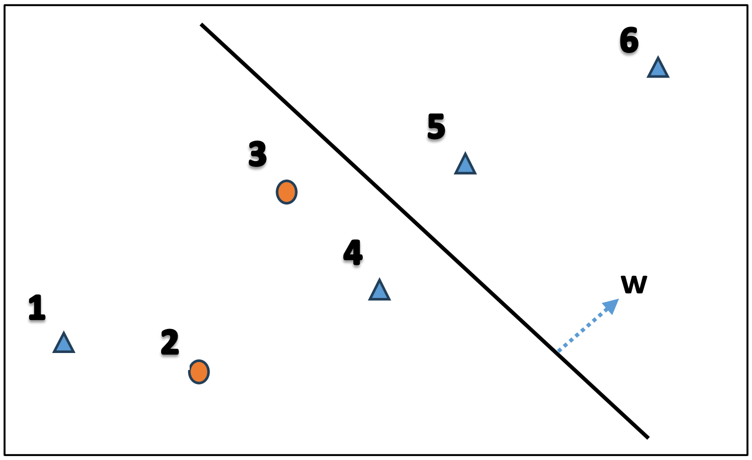
What is the weight initialization method for A, B, and C in the above plots from zero initialization, Xavier initialization, and Kaiming He initialization? (Answer with only one initialization method for A, B and C.)

Q.5 You are solving the binary classification task of classifying images as "car vs. no car". You design a CNN with a single output neuron. The final output of your network, $\hat{y}$ is given by:

$\hat{y} = \sigma(ReLU(z))$ (where z, as usual, is w.x + b)

You classify all inputs with a final value $\hat{y} \geq 0.5$ as car images. What problem are you going to encounter? Justify. [2]

Q.6. Given N training data points $\{x_i, y_i,\ i = 1{:}N,\}$, $x_i \in R^d$, labels $y_i \in \{1, -1\}$. We need a linear classifier $f(x) = sign(w.x)$ (read as w dot x) optimizing the loss function $L(z) = e^{-z}$, for $z = y.(w.x)$. Here, ▲ represents data point of class 1 and ● represents data point of class -1: [6+2=8]



a). Explain the penalties given by this loss functions for the different data points (1 to 6) shown above in the plot.

b). Derive the stochastic gradient descent update Δw for $L(z)$.