Lahore Campus

Deep Learning (CS5102)

Sessional-I Exam

Date: September 21st 2024

Course Instructor(s)

Ms. Mamoona Akbar

Total Time (Hrs):	**1**
Total Marks:	**50**
Total Questions:	**3**

Roll No ________ Section ________ Student Signature ________

Do not write below this line

Attempt all the questions.

CLO #2: Understand and design the structure of deep neural networks

CLO #3: Understand the different layers and their operation

Q1: Part A [10 marks]

Consider image data with size 400*400*3 where 3 is the depth of the data. The data is passed through a network with the following layers. Calculate the output size and the number of weights/tuneable parameters from each layer.

- Convolution layer with 50 filters of size 5*5 with stride 2.
- Convolution layer with 60 filters of size 7*7 with stride 3.
- Fully connected layer with 50 neurons.
- Fully connected layer with 40 neurons.

Q1: Part B [10 marks]

For the following building blocks in a convolution neural network (CNN), write down list of parameters and hyper-parameters:

	Parameter	**Hyper Parameters**
Convolution Layer		
Polling Layer		
Fully connected layer		
Other Components		

Q1: Part C [5 marks]

Choose the correct Options

1. Using "neural style transfer" (as seen in class), you want to generate an RGB image of the Great Wall of China that looks like it was painted by Picasso. The size of your image is 100x100x3 and you are using a pre-trained network with 1,000,000 parameters. At every iteration of gradient descent, how many updates do you perform?
   30,000

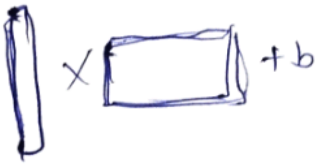
   a. 10,000
   b. 30,000 100x100*3 = 30000
   c. 1,000,000
   d. 1,030,000 ✓

2. Mini-batch gradient descent is a better optimizer than full-batch gradient descent to avoid getting stuck in saddle points.
   a. True
   b. False

SGD > MBD > B

3. Cross entropy loss can be applied to classification problem with ________ number of classes
   a. 1
   b. 2
   c. any
   d. none of the above

4. The pooling layer adds ______________ parameters/weights for every feature in a data point.
   a. 0
   b. 1
   c. 2
   d. none of the above



5. Which technique prevents overfitting in a neural network by randomly dropping out neurons during training?
   a. Dropout
   b. Batch Normalization
   c. L1 Regularization
   d. L2 Regularization

CLO #1: Understand the basic concepts of Deep Learning

Q2: Answer the following question [5 marks]

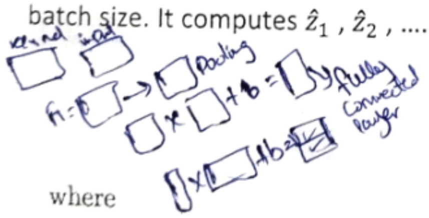
a) Difference between RMSprop and Nesterov Accelerated Gradient (NAG)

b) Describe different weight initialization techniques and when we used that.

CLO #3: Understand the different layers and their operation

Q3: [20 marks]

Using batch normalization in neural networks requires computing the mean and variance of a tensor. Suppose a batch normalization layer takes vectors $z_1, z_2, \cdots, z_m$ as input, where m is the mini-batch size. It computes $\hat{z}_1, \hat{z}_2, \ldots\ldots, \hat{z}_m$ according to



$$\hat{z}_i = \frac{z_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

where

$$\mu = \frac{1}{m}\sum_{i=1}^{m} z_i, \quad \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(z_i - \mu)^2.$$

It then applies a second transformation to obtain $\tilde{z}_1, \tilde{z}_2, \cdots, \tilde{z}_m$ using learned parameters γ and β as

$$\tilde{z}_i = \gamma\hat{z}_i + \beta.$$

In this question, you can assume that $\epsilon = 0$.

0059

National University of Computer and Emerging Sciences

Lahore Campus

a) You forward-propagate a mini-batch of *m*=4 examples in your network. Suppose you are at a batch normalization layer, where the immediately previous layer is a fully connected layer with 3 units. Therefore, the input to this batch normalization layer can be represented as the below matrix:

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \end{array}$$
$$\begin{bmatrix} 12 & 14 & 14 & 12 \\ 0 & 10 & 10 & 0 \\ -5 & 5 & 5 & -5 \end{bmatrix}$$

What are $\hat{z}_i$? Please express your answer in a 3×4 matrix.

b) Continue with the above setting. Suppose γ=(1,1,1), and β=(0,−10,10). What are $\widehat{z}_i$. ? Please express your answer in a 3×4 matrix.

1×3 3×4

c) Describe the differences of computations required for batch normalization during training and testing.

d) Describe how the batch size during testing affect testing results.

[1 1 1] [

1×3

Deep learning for Perception (CS4045)

Date: Nov. 2st 2024

Course Instructor(s)

Ms. Mamoona Akbar

Total Time (Hrs):	1
Total Marks:	35
Total Questions:	2

Roll No ____________ Section ____________ Student Signature ____________

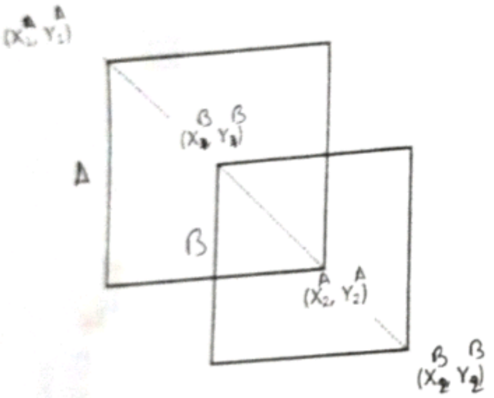
Do not write below this line

Attempt all the questions.

CLO #4: Be able to read and understand state of the art architectures of computer vision

Q1: [10 + 3 + 2 + 5 + 3+ 2 marks]

a) Consider the following bounding boxes:


$(\overset{A}{X_1}, \overset{A}{Y_1})$
$(\overset{B}{X_1}, \overset{B}{Y_1})$
A
B
$(\overset{A}{X_2}, \overset{A}{Y_2})$
$(\overset{B}{X_2}, \overset{B}{Y_2})$

Upper bounding box is the ground truth and has coordinates: $(\overset{A}{X1}, \overset{A}{Y1}), (\overset{A}{X2}, \overset{A}{Y2})$ = (100,150), (250,300)

Lower bounding box is the predicted one and has coordinates: $(\overset{B}{X3}, \overset{B}{Y3}), (\overset{B}{X4}, \overset{B}{Y4})$ = (130,170), (270,360)

To Do: Please calculate the intersection over union.

b) What is the purpose of the non-max suppression function in Yolo? Give specific reasoning

c) Suppose you have two anchor boxes, write Y(predicted) equations for different scenarios.

d) Suppose you want to apply 128 filters of 3x3 on a volume of 256x256x64 to output a volume of 256x256x128. How many operations will be required to apply this convolution layer? If we add a bottleneck layer of 8 1x1 convolutions, how many total number of operations will required? Show complete working.

e) Difference between ROI polling and ROI align

f) What are some limitations of R-CNN, and how does Faster R-CNN address them?

CLO #5: Be able to read and understand the research paper in the field

Q2: [1+2+7 marks]

a) Suppose you have an image of width 200 and height 450. The ROI region starts at 40*170 and ends at 170*250.

1. Extract the feature map by the deep convolution layer of scaling factor 32.
2. The resulting polling layer is 2*3. What is the size of each block of the polling layer
3. Each grid is further divided into 2*2. Find the linear interpolation of the highlighted grid

	▨

The subset of Image intensity after applying the feature map is given below

	2	3	4	5	6	7
3	0.5	0.1	0.3	0.7	0.2	0.5
4	0.2	0.8	0.2	0.8	0.2	0.8
5	0.3	0.3	0.3	0.3	0.3	0.5
6	0.5	0.5	0.5	0.5	0.5	0.5
7	0.6	0.6	0.6	0.6	0.6	0.6
8	0.5	0.3	0.5	0.3	0.5	0.3
9	0.8	0.5	0.8	0.5	0.8	0.5

National University of Computer and Emerging Sciences 0051

Lahore Campus

Deep Learning for Perception (CS4045)

Date: December 16st 2024

Course Instructor(s)

Ms. Mamoona Akbar

Final Exam

Total Time (Hrs):	**3**
Total Marks:	**95**
Total Questions:	**4**

Roll No ______ Section ______ Student Signature ______

Do not write below this line

Attempt all the questions.

CLO # 2: Implement deep learning models using popular frameworks

Q1: [10 + 10 + 5 + 5 + 5 = 35 marks]

a) You are designing a convolutional neural network with the following layers. For each layer calculate the number of weights, biases and volume sizes of the output for that layer. If not specified assume stride size 1 and pooling zero.

Layer	Number of weights	Number of biases	Volume dimensions
Input	0	0	32x32x3
Conv Layer with filter size 5x5 and number of filters 8	600	8	28x28x8
Maxpooling with filter size 2x2 and stride size=2	256	80	14x14x8
Conv Layer with filter size 3x3 and number of filters 16	2304	16	12x12x16
Maxpooling with filter size 2x2 and stride size=2	1024	160	6x6x16
Conv Layer (same) with filter size 3x3 and number of filters 32	9216	32	4x4x32
Maxpooling with filter size 2x2 and stride size=2	4096	320	2x2x32
FC layer with 64 units	8192	64	2x2x32
Output Layer to classify 5 classes	640	5x32	5x2x2x32

b) Now, we estimate CNNs' computation overhead by counting the FLOPs (floating point operations). For simplicity, we only consider the forward pass.

Given x1; x2; · · · ; xn all scalars, we assume:

- A scalar multiplication xi · xj accounts for one FLOP;
- A scalar addition xi + xj accounts for one FLOP;
- A max operation max(x1; x2; · · · ; xn) accounts for n - 1 FLOPs.

All other operations do not account for FLOPs.

How many FLOPs will be conducted by the forward pass of the network given in part a? You can assume that a linear activation function is applied in all the layers.

Layer	Number of FLOPs
Input	0
Conv Layer with filter size 5x5 and number of filters 8	470400 470400
Maxpooling with filter size 2x2 and stride size=2	50175 50175
Conv Layer with filter size 3x3 and number of filters 16	16588 16588
Maxpooling with filter size 2x2 and stride size=2	36863 36863
Conv Layer (same) with filter size 3x3 and number of filters 32	73728 73728
Maxpooling with filter size 2x2 and stride size=2	16383 16383
FC layer with 64 units	4096 4096
Output Layer to classify 5 classes	20480 20480

c) You are benchmarking runtimes for layers commonly encountered in CNNs. Which of the following layers: Convolution, pooling, fully connected would you expect to be the fastest (in terms of floating point operations)?

d) You are benchmarking runtimes for layers commonly encountered in CNNs. Which of the following layers: Convolution, pooling, fully connected would you expect to be the fastest (in terms of floating point operations)? if asked for slower

e) Suppose you want to apply 128 filters of 3x3 on a volume of 256x256x64 to output a volume of 256x256x128. How many operations will be required to apply this convolution layer? If we add a bottleneck layer of 8 1x1 convolutions, how many total number of operations will required? Show complete working.

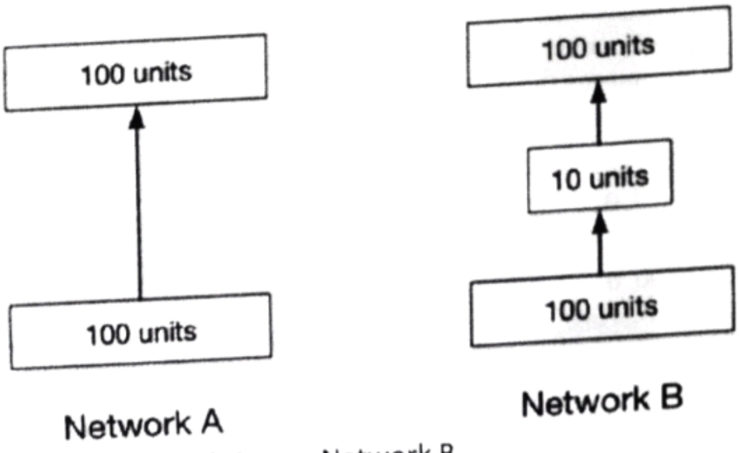
256 × 256 × 3 × 3 × 64 × 128 = 4831838208   Total: 4865392640

256 × 256 × 1 × 1 × 64 × 8 = 33554432

CLO # 1: Understand the theoretical foundations of deep learning, including neural networks and optimization techniques.

Q2: [3 + 5 + 5 + 3 + 4 = 20 marks]

a) Consider the following two multilayer perceptrons, where all of the layers use linear activation functions.



Network A

Network B

Give one advantage of Network A over Network B
Give one advantage of Network B over Network A

b) Give a method to fight vanishing gradient in fully-connected neural networks. Assume we are using a network with Sigmoid activations trained using SGD.

c) You are trying to build a model that predicts whether a person will click on a given advertisement. Your friend told you that creating an ensemble of models (averaging the results of multiple models), can lead to better predictions. You try 2 approaches:
• You train a very large model, which takes many hours to train.
• You train 10 slightly smaller models and average the results of each. The entire process takes a smaller amount of time compared to the first model.

Both these models give you an accuracy of 90% on your training set. You need to pick an approach to deploy on a single-GPU machine that needs to provide predictions in real-time. Less inference time is of the utmost importance here. Which approach is more suitable and why?

d) You come across a nonlinear function that passes 1 if its input is nonnegative, else evaluates to 0, A friend recommends you use this non-linearity in your convolutional neural network with the Adam optimizer. Would you follow their advice? Why or why not?

e) Rearragnge the steps of gradient desent algorithm
I. Calculate error between predicted and actual value
II. Riterate untill you find best weights for the network
III. Pass an input through the networkand get values from the output layer
IV. Initialize random weights and bias
V. Go to each neuron which contributes to the error and change its weights to reduce the error

IV
II
III
I
V

CLO #3 : Apply deep learning algorithms to solve real-world problems such as image classification, object detection, and natural language processing.

Q3: [5 + 5 + 10 + 15 = 35 marks]

a) A very useful embedding for a small collection of English words is created via a Word2Vec Skip-Gram Network with 6 hidden nodes. Given 4 words in this collection that are all quite different from one another semantically, which of the following is the most likely embedding for these 4 words? Explain your answer.

I. 111110, 101111, 111011, 110111
II. 100001, 111000, 011110, 000111
III. 101000, 001100, 011000, 010100
IV. 111000, 011100, 011000, 110000
V. All 4 options (A-D) are equally likely

b) Recall the LSTM architecture. Suppose you want the memory cell to sum its inputs over time. What values should the update gate and forget gate take? Justify your answer.

c) Compute y predicted

$$\mathbf{h}_t = g(\mathbf{V}\mathbf{x}_t + \mathbf{U}\mathbf{h}_{t-1} + \mathbf{c})$$
$$\hat{\mathbf{y}}_t = \mathbf{W}\mathbf{h}_t + \mathbf{b}$$

Above are the equations of a simple RNN. You can assume that **g is RELU function and biases are all zero. Below are** given the transpose of one hot encodings of the words: **Hello, RNN, the, to, welcome. Use these encodings and weight** matrices: U, V, W compute the output of the RNN for the sentence: **"welcome to the RNN"**

Hello = [1 0 0 0 0], RNN= [0 1 0 0 0], the=[0 0 1 0 0], to=[0 0 0 1 0], welcome=[0 0 0 0 1]

V is

1	2	-2	1	2
0	3	0	-1	0
-1	1	1	0	0

U is

1	0.5	1
0	1	0.5
-1	-1	2

4×3 3×1
4×1

W is

1	0	1
0	1	1
-1	1	0
1	-1	-1

d)

طالبعلم کتاب پڑھ رہا ہے۔ *(The student is reading a book.)*

A, a [0.1 , 0.2 , 0.1, 0.1]

Word (Urdu)	
طالبعلم (Student)	[0.6, 0.4, 0.2, 0.7]
کتاب (Book)	[0.8, 0.1, -0.4, 0.5]
پڑھ (Read)	[0.7, 0.3, 0.1, 0.6]
رہا (Is)	[0.5, 0.2, 0.0, 0.4]
ہے	[0.4, 0.1, 0.0, 0.3]

Assume the following weight matrix

$$W_Q = \begin{bmatrix} 0.3 & 0.3 & 0.3 & 0.3 \\ 0.4 & 0.4 & 0.4 & 0.4 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix}, \quad W_K = \begin{bmatrix} 0.4 & 0.4 & 0.4 & 0.4 \\ 0.5 & 0.5 & 0.5 & 0.5 \\ 0.3 & 0.3 & 0.3 & 0.3 \\ 0.2 & 0.2 & 0.2 & 0.2 \end{bmatrix}, \quad W_V = \begin{bmatrix} 0.6 & 0.6 & 0.6 & 0.6 \\ 0.2 & 0.2 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.3 & 0.3 \\ 0.4 & 0.4 & 0.4 & 0.4 \end{bmatrix}$$

1. **Computer query of کتاب**
2. **Compute key and value of "Read"**
3. **Compute cross self-attention of کتاب**

CLO #4: Evaluate the performance of deep learning models using appropriate metrics and techniques.

Q4: [5 marks]

List three uses of autoencoders other than dimensionality reduction and how autoencoders can be used for your mentioned applications.