Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2024-2025

Comprehensive Examination
(EC-3 Regular)

Course No. : DSECLZG524
Course Title : Deep Learning
Nature of Exam : Open Book
Weightage : 40%
Duration : 2 Hours
Date of Exam : 07-09-2025

No. of Pages = 03
No. of Questions = 06

Note to Students:

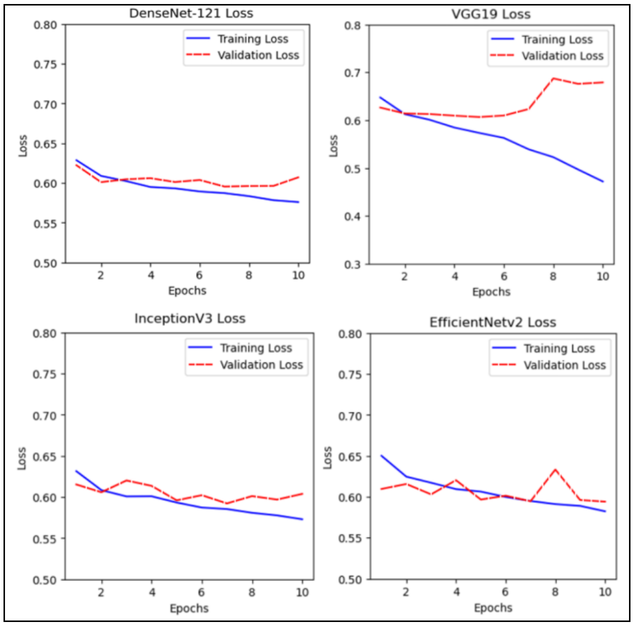
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1. Answer the following questions: (limit answers to 1-2 sentences) **[5 Marks]**

A. Suppose you are training neural networks on 100x100 images to predict 5 classes. Neural network A consists of a single linear layer followed by a softmax output activation. Neural network B consists of a sequence of layers with dimensions 128, 512, and 32, respectively, followed by a softmax output activation. Assuming that both neural networks are trained using an identical procedure (e.g. batch size, learning rate, epochs, etc), and neither contains hidden activations, what can you generally expect about the relative performance of A and B on the test data? (2M)

B. Suppose you set up and train a neural network on a classification task and converge to a final loss value. Keeping everything in the training process the same (e.g. learning rate, optimizer, epochs). It is possible to reach a lower loss value by ONLY changing the network initialization. (2M)

C. Four deep neural network models are trained on a classification task, and below are the plots of their losses. Based on these plots, which model is overfitting? (1M)



Q.2. Answer the following **[5 Marks]**

A. Assume a company asks you to develop an application that can predict which kind of bird is depicted in a given image. Which kind of task is this? List and explain the individual steps you would follow to solve this problem using deep learning. (end-to-end steps get full marks) (2M)

B. Draw the computational graph corresponding to the following equations:

$$o = a(\mathbf{w} \cdot \mathbf{x} + b)$$

$$o = f(x) = f\Big(g_1(h_1(x), h_2(x)),\ g_2(h_2(x), h_3(x))\Big)$$

Clearly show the flow of inputs, intermediate nodes (e.g., weighted sum, hidden functions), and final output. (3M)

Q.3. You are given a CNN architecture defined in PyTorch as follows: **[10 Marks]**

```
nn.Sequential(
    nn.Conv2d(1, 5, kernel_size=3, stride=1, padding=1),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=2, stride=2),
    nn.Conv2d(5, 10, kernel_size=5, stride=1, padding=0),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=2, stride=2),
    nn.Conv2d(10, 20, kernel_size=2, stride=2, padding=0),
    nn.ReLU(),
    nn.MaxPool2d(kernel_size=5, stride=2),
    nn.Flatten(),
    nn.Linear(2800, 6)
)
```

Assume that the input is a grayscale image of size 192 × 256 × 1. Now, answer the following questions:

A. Compute the output size (height × width × depth) after each (Conv and Pooling) layer (3.5M)

B. Calculate the number of trainable weights and biases in each convolutional layer (1.5M)

C. Verify that the flattened input to the fully connected layer matches the expected dimension (1M)

D. Compute the number of trainable parameters in the fully connected (Linear) layer. (1M)

E. Briefly explain in 1-2 sentences why a CNN is an adequate choice of architecture for image classification (1M)

F. Why is ReLU (and its variants) often preferred over sigmoid/tanh in modern CNNs? Provide two clear reasons (1-2 sentences). (1M)

G. If we add Batch Normalization after Conv1 and Conv3 and Dropout (p=0.5) before FC, explain the role of each and how they impact training/generalization. (1M)

Q.4. **I)** You are given a simple RNN encoder with the following configuration: **[10 Marks]**

- Vocabulary embeddings (each word mapped to a 2-dimensional vector):

$$x_{\text{Deep}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ x_{\text{learning}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ x_{\text{works}} = \begin{bmatrix} -1 \\ -1 \end{bmatrix},\ x_{\langle stop \rangle} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- Input-to-hidden weight matrix:

$$W_{hx} = \begin{bmatrix} 0 & -1 \\ 2 & 0 \\ 1 & 1 \end{bmatrix}$$

- Recurrent weight matrix: $W_{hh}=I_{(identity)}$
- Initial hidden state: $h_0=0$
- Biases: all zero.
- Activation: ReLU applied elementwise.

Compute the final hidden state h_4 of the encoder for the input sequence "Deep learning works <stop>"

Show your step-by-step calculations, including:

A. The weighted sum at each time step. (2M)
B. The hidden state after applying ReLU at each time step. (3M)
C. The final hidden state h_4. (2M)

II) Imagine you are building a model to predict the sentiment of a movie review. The review is long (200+ words), but the key sentiment is expressed only near the end.

A. Explain in 1–2 sentences why a vanilla RNN would struggle with this task and how it impacts training. (1.5M)
B. In 1–2 sentences, describe which gating mechanism in an LSTM helps solve this issue and how it preserves the critical information. (1.5M)

Q.5. Consider a sequence of three tokens with embeddings: **[5 Marks]**

$X_1=[1,0]$, $X_2=[0,1]$, $X_3=[1,1]$

Assume query (Q), key (K), and value (V) matrices are the identity matrix. Answer the following:

A. Compute the raw attention scores between X_1 and all tokens (including itself). (2M)

B. Apply the softmax function to normalize these scores. (2M)

C. In one sentence, explain what self-attention achieves in this example. (1M)

Q.6. A hospital wants to securely store and transmit medical images (X-rays, MRIs). Since storage is limited, they decide to use an autoencoder to compress the images before saving them. **[5 Marks]**

A. Which type of autoencoder would be most suitable in this scenario, and why? (2M)

B. Suppose noise is present in the transmission of images. Which variant of autoencoder should be used, and how does it improve robustness? (2M)

C. Briefly state one practical advantage of using deep autoencoders over linear methods like PCA in this case. (1M)