Indian Institute of Technology Patna
End Semester Examination
Course Name: Deep Learning
Course Code: MA4114/MA472
Date: 28 November 2024

Maximum Marks: 60 Time: 3 Hours

[Total Marks = 12]

1. **Fully Connected Neural Network**

(a) What is dropout in the context of FC layers? [Mark = 1]

(b) How does splitting a dataset into train, val and test sets help identify overfitting? [Marks = 2]

(c) You are designing a deep learning system to detect driver fatigue in cars. It is crucial that that your model detects fatigue, to prevent any accidents. Which of the following is the most appropriate evaluation metric: Accuracy, Precision, Recall, Loss Value. Explain your choice. [Marks = 2]

(d) You have a single hidden-layer neural network for a binary classification task. The input is $X \in \mathbb{R}^{n \times m}$, output $\hat{y} \in \mathbb{R}^{1 \times m}$, and true label $y \in \mathbb{R}^{1 \times m}$. The forward propagation equations are as follows:

$$z^{[1]} = W^{[1]}X + b^{[1]}$$
$$a^{[1]} = \sigma(z^{[1]})$$
$$\hat{y} = a^{[1]}$$

The cost function is:

$$J = -\sum_{i=1}^{m} \left[y^{(i)} \log(\hat{y}^{[i]}) + (1 - y^{(i)}) \log(1 - \hat{y}^{[i]}) \right]$$

Write the expression for $\frac{\partial J}{\partial W^{[1]}}$ as a matrix product of two terms. [Marks = 4]

(e) Consider a neural network encoder $z = \text{softmax}[f_\theta(X)]$. You can think of f_θ as an MLP for this example. z is the softmax output, and we want to discretize this output into a one-hot representation before passing it into the next layer.

Consider the operation one_hot, where one_hot(z) returns a one-hot vector with a 1 at the argmax location. For example:

$$\text{one\_hot}([0.1, 0.5, 0.4]) = [0, 1, 0].$$

Say we want to pass this output to another fully connected layer g_ϕ to get a final output y.

(i) Is there a problem with the neural network defined below? [Marks = 1]

$$y = g_\phi(\text{one\_hot}(\text{softmax}(f_\theta(X))))$$

(ii) Consider the following function:

$$z = S_\tau(f_\theta(X)) = \text{softmax}\left(\frac{f_\theta(X)}{\tau}\right)$$

Here dividing by τ means every element in the vector is divided by τ. Obviously, when $\tau = 1$, this is exactly the same as the regular softmax function. What happens when $\tau \to \infty$? What happens when $\tau \to 0$?
Hint: You don't need to prove these limits; just showing a trend and justifying is sufficient. [Marks = 2]

2. **Convolutional Neural Networks (CNN)** [Total Marks = 12]

(a) How does the size of the kernel in a convolutional layer affect the receptive field and the learning process of a CNN? [Marks = 1.5]

(b) Explain how skip connections in deep CNN architectures like ResNet help address the vanishing gradient problem during training and improve convergence. [Marks = 1.5]

(c) You are tasked with designing a CNN classifier. For each layer, calculate the following:

- The number of weights and biases.

- The dimensions of the associated feature maps.

The network is defined as follows:

- **CONV-K-N**: Represents a convolutional layer with N filters, each of size $K \times K$. Stride is set to 1, and padding is 0.
- **POOL-K**: Represents a $K \times K$ pooling layer with stride K and padding 0.
- **FC-N**: Represents a fully-connected layer with N neurons.

The architecture of the CNN is given below:

Layer	Activation Map Dimensions	Number of Weights	Number of Biases
INPUT	$128 \times 128 \times 3$	0	0
CONV-9-32			
POOL-2			
CONV-5-64			
POOL-2			
CONV-5-64			
POOL-2			
FC-3			

Table 1: CNN Architecture Overview

Fill in the table with the necessary calculations for the number of weights, biases, and feature map dimensions for each layer. [Marks = 5]

(d) Following the last FC-3 layer of your network, what activation must be applied? Given a vector a = [0.3, 0.3, 0.3], what is the result of using your activation on this vector? [Marks = 2]

(e) You start training your model and notice underfitting, so you decide to add data augmentation as part of your preprocessing pipeline. Given that you are working with images of handwritten digits, for each data augmentation technique, state whether or not the technique is appropriate for the task. If not, explain why not. [Marks = 2]

- Scaling slightly
- Flipping vertically or horizontally
- Rotating by 90 or 180 degrees
- Shearing slightly

3. RNNs, LSTMs and Transformer [Total Marks = 17]

(a) Why are vanishing or exploding gradients an issue for RNNs? [Marks = 3]

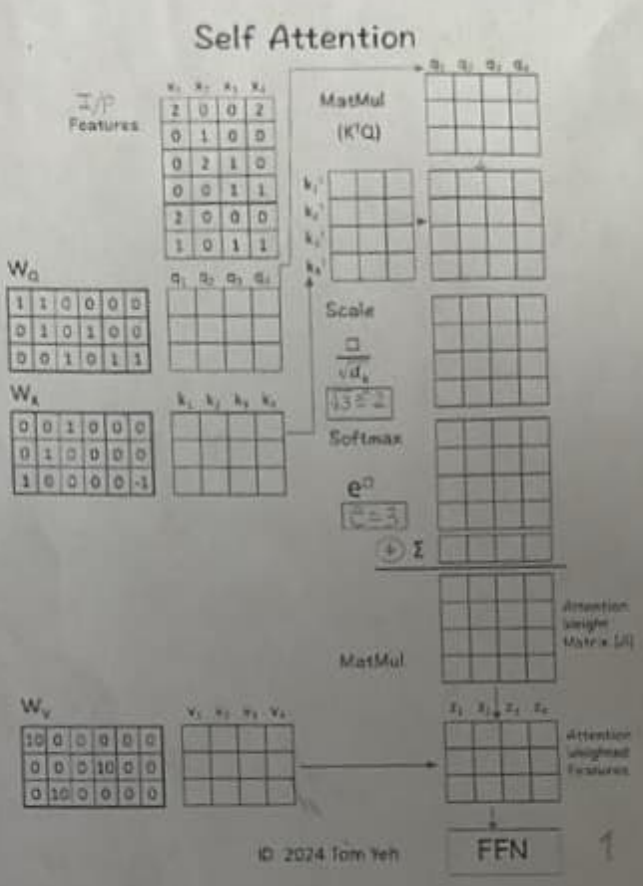
(b) Provide a detailed derivation of the equations for a standard Long Short-Term Memory (LSTM) cell. Discuss the role of each gate (input gate, forget gate, output gate) and the cell state. [Marks = 3]

(c) Given an RNN layer with:

- Input size: 50
- Hidden size: 200

What is the total number of parameters in the RNN layer, including weights for the input-to-hidden, hidden-to-hidden, and biases? [Marks = 3]

(d) Perform the right operations and fill in the blanks: [Marks = 6]



(e) Explain the concept of image patching in the Vision Transformer. How does the process of dividing an image into patches help in feeding data to the model? [Marks = 2]

4. **Autoencoder, Variational Autoencoder and Generative Adversarial Networks** [Total Marks = 19]

(a) What is the difference between an Autoencoder and an Encoder-Decoder architecture? [Marks: 2]

(b) Discuss the limitations of standard autoencoders in generative tasks. How do VAEs address these limitations? [Marks: 3]

(c) Direct sampling of an image from its distribution $P(X)$ is ill-posed because of the high dimensionality of the sampling space. Hence, we try to compute a distribution of a lower dimensional entity z conditioned on an image X, which is $P(z|X)$. Show that computing $P(z|X)$ is also intractable. How is this problem handled in a VAE? [Marks: 3]

(d) Derive the Evidence Lower Bound (ELBO) in the context of a Variational Autoencoder (VAE) by applying Kullback-Leibler (KL) divergence. [Marks: 4]

(e) Explain the role of the KL divergence term in the VAE loss. What happens when this term is weighted too heavily or too lightly? [Marks: 3]

(f) Is the KL divergence term in the VAE loss function symmetric? [Mark: 1]

(g) Explain the two main components of a Generative Adversarial Network (GAN) and describe the tasks performed by each component. [Marks: 3]