NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2024-2025

SC4001/CE4042/CZ4042 – NEURAL NETWORKS AND DEEP LEARNING

Nov/Dec 2024 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 7 pages.
2. Answer **ALL** questions.
3. This is an open-book examination.
4. All questions carry equal marks.

---

1. (a) State whether each of the following statements is "TRUE" or "FALSE" with explanation. Each subquestion carries 1 mark.

   (i) The reason for using L1 or L2 regularization is for the model to converge toward the minimum training error more quickly.

   (ii) If two classes of samples are separable by a straight line, we can train a network without any non-linear activation in its hidden layers.

   (iii) Using Dropout always results in a lower evaluation error.

   (iv) Backpropagation is used to compute the optimal structure of a neural network, including the number of layers and neurons.

   (v) The *ReLU* (Rectified Linear Unit) activation function can introduce negative values in the output.

   (vi) Adding more layers in a neural network does not always result in a lower evaluation error.

Note: Question No. 1 continues on Page 2

(vii) Using K-fold cross-validation is more often conducted when the dataset is small than when the dataset is large.

(7 marks)

(b) Give brief answers to the following. Each part carries 2 marks.

(i) State the shape of tensor [[[1], [0]], [[3], [4]], [[-2], [-1]]] .

(ii) For Xavier/Glorat uniform weight initialization schemes, why are the gain values different for different activation functions?

(iii) How would one decide to increase the number of layers (i.e. depth) versus the number of neurons per layer (i.e. width) in a neural network?

(6 marks)

(c) Figure Q1 shows a dataset with each example belonging to one of the two classes, displayed in the space of its two features x_1 and x_2. You are to design a discrete perceptron network (i.e. with unit step activation functions) with a single hidden layer to classify the examples.

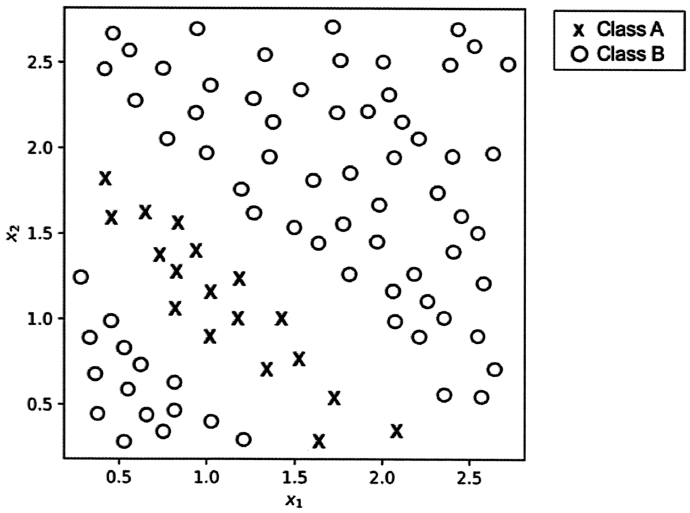

Figure Q1

(i) Draw the decision boundaries separating the two classes to design the network.

(2 marks)

Note: Question No. 1 continues on Page 3

(ii) State the number of neurons in the hidden layer.

(2 marks)

(iii) Draw the network indicating the values of all weights and biases.

(8 marks)

2. The 3-layer feedforward neural network shown in Figure Q2 receives two-dimensional inputs $(x_1, x_2) \in \boldsymbol{R^2}$ and produces a one-dimensional output y. The first hidden layer consists of three neurons and the second hidden layer consists of two neurons. All hidden neurons have *ReLU* activation functions and the output neuron is a logistic regression neuron. The weights of the network are initialized as indicated in Figure Q2 and all the biases are initialized to 0 (not shown).

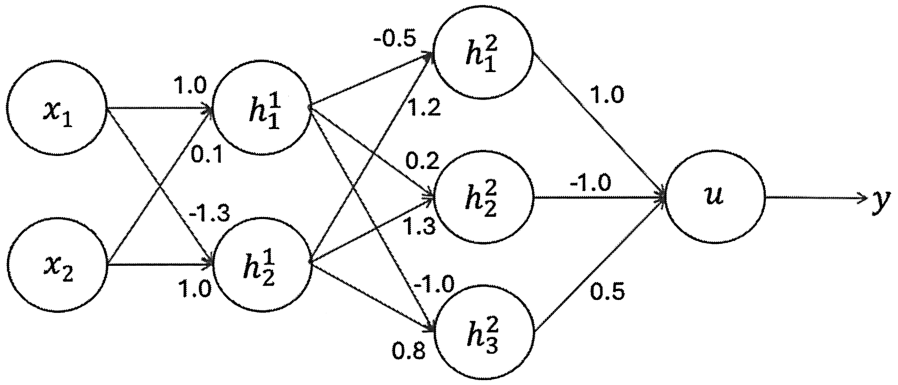

Figure Q2

The network is trained to produce a desired output $d = 0$ for an input $\boldsymbol{x} = \begin{pmatrix} -1.0 \\ 1.5 \end{pmatrix}$. You are to perform one iteration of stochastic gradient descent learning with the example $(\boldsymbol{x}, d)$. Give your answers rounded up to three significant figures.

(a) Write initial weight matrices $\boldsymbol{W}$ and bias vectors $\boldsymbol{b}$, connected to the three layers.

(3 marks)

(b) Find the synaptic inputs $\boldsymbol{u}$ and outputs $\boldsymbol{h}$ of the two hidden layers.

(4 marks)

Note: Question No. 2 continues on Page 4

(c) Find the output y and the cross-entropy at the output layer.

(4 marks)

(d) Find the derivatives $f'(\boldsymbol{u})$ at the two hidden layers with respect to synaptic input $\boldsymbol{u}$, where f is the $ReLU$ activation function.

(2 marks)

(e) Find gradients $\nabla_{\boldsymbol{u}} J$ of the cost J with respect to activations $\boldsymbol{u}$, at the three layers.

(6 marks)

(f) Find gradients $\nabla_{\boldsymbol{W}} J$, and $\nabla_{\boldsymbol{b}} J$ of the cost J with respect to weights $\boldsymbol{W}$, and biases $\boldsymbol{b}$, respectively, at the three layers.

(6 marks)

3. Figure Q3 depicts a network that consists of two convolutional layers. The size of input or output volume is represented as $D \times H \times W$, where D is the number of channels, and $H \times W$ is the spatial size.

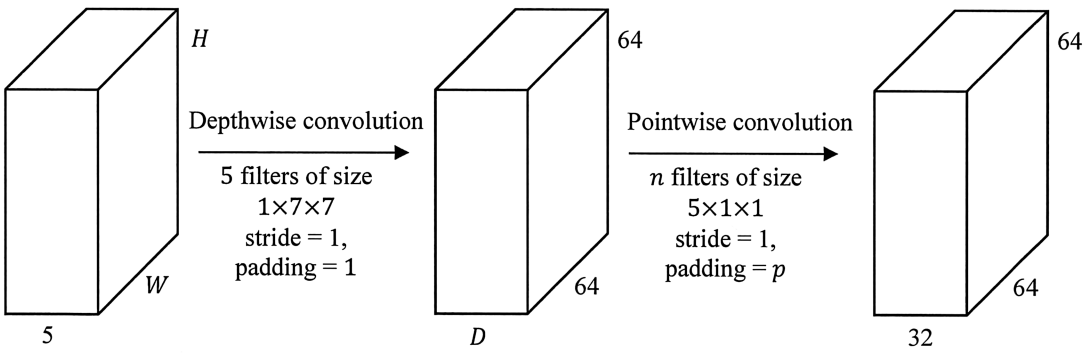

Figure Q3

(a) (i) Give the values of H, W, D , n and p.

(4 marks)

Note: Question No. 3 continues on Page 5

(ii) Calculate the FLOPs of the depthwise convolution and pointwise convolution layers, respectively.

(4 marks)

(b) You are training a neural network and need to implement batch normalization on a given layer. For a mini-batch of size $N = 5$, you have the following input activations for a particular feature/dimension:

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \\ x^{(5)} \end{bmatrix} = \begin{bmatrix} 10 \\ 12 \\ 14 \\ 16 \\ 18 \end{bmatrix}$$

The batch normalization parameters are:

- Scale parameter (gamma): $\gamma = 1.5$
- Shift parameter (beta): $\beta = -2$
- Small constant for numerical stability: $\epsilon = 1 \times 10^{-5}$

Get the final output $\mathbf{y}$ of the batch normalization layer. For all calculations, round your answers to four decimal places where necessary. Show all your working steps for full credit.

(8 marks)

(c) Answer "TRUE" or "FALSE" to the following statements. Each sub-question carries 1 mark.

(i) Backpropagation Through Time (BPTT) is an algorithm used to train RNNs by unrolling the network over time.

(ii) Long Short-Term Memory (LSTM) networks are immune to the vanishing gradient problem due to their gating mechanisms.

(iii) In a Recurrent Neural Network (RNN), the hidden state at time t is calculated by applying a nonlinear activation function to the weighted sum of the input at time t and the hidden state at time $t + 1$.

(iv) In an RNN, using a *Tanh* activation function in the hidden layer helps mitigate the vanishing gradient problem.

Note: Question No. 3 continues on Page 6

(v) The "cell state" in an LSTM can be considered as a conveyor belt that runs straight down the entire chain with only minor linear interactions.

(5 marks)

(d) Sparse autoencoders are designed to learn features that are robust and useful by encouraging sparsity in the hidden units. Given a dataset where each input **x** is a one-hot encoded vector, justify whether a sparse autoencoder would be appropriate for learning meaningful representations of such data. Support your answer with theoretical reasoning.

(4 marks)

4. (a) Complete the following statements by filling in the blanks. These statements pertain to the self-attention mechanism in Transformers. Answers to be indicated in the answer sheet.

(i) The self-attention mechanism allows the model to weigh the significance of different tokens in the input sequence when encoding a particular token, effectively capturing _____ dependencies.

(1 mark)

(ii) In multi-head self-attention, the outputs from each attention head are _____ and then passed through a final _____ layer to produce the combined output of the multi-head attention mechanism.

(2 marks)

(iii) The multi-head cross-attention in decoder takes the _____ and _____ from the encoder, and takes _____ from the previous layer of the decoder.

(3 marks)

(iv) The scaling factor $1/\sqrt{D_k}$ is essential for the effective computation of attention weights in Transformers. It prevents _____ from causing the Softmax function to produce near-binary outputs, which would impede learning. By ensuring the attention scores are appropriately scaled, the model maintains stable _____ and learns more effectively across different layers and attention heads.

(2 marks)

Note: Question No. 4 continues on Page 7

(b) Consider a scaled dot-product self-attention mechanism where you have a single query vector and multiple key and value vectors

$$\text{Query, } \mathbf{q} = [2 \quad 3], \quad \text{Key, } \mathbf{K} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{Value, } \mathbf{V} = \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix}$$

Determine the final output $y = \text{Attention}(\mathbf{q}, \mathbf{K}, \mathbf{V})$. Assume $D_k = 2$. For all calculations, round your answers to four decimal places where necessary. Show all your working steps for full credit.

(14 marks)

(c) Explain the potential consequences when the discriminator in a Generative Adversarial Network (GAN) significantly outperforms the generator. How does this imbalance affect the training dynamics and the overall performance of the GAN?

(3 marks)

END OF PAPER

CE4042 NEURAL NETWORK & DEEP LEARNING
CZ4042 NEURAL NETWORK & DEEP LEARNING
SC4001 NEURAL NETWORK & DEEP LEARNING

CONFIDENTIAL

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.