NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2023-2024

SC4001/CE4042/CZ4042 – NEURAL NETWORKS AND DEEP LEARNING

Nov/Dec 2023 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) Give brief answers to the following. Each part carries 2 marks.

   (i) What is the total number of learnable parameters in a 2-layer network with 5 neurons in each layer and receiving 2-dimensional inputs.

   (ii) Write a tensor with the shape [2, 1, 3] and integer elements, as in Python.

   (iii) One wants to initialize all the weights in a 3-layer deep neural networks to 0.4. Is this a good idea? And why?

   (iv) You want to solve a classification problem with a training sample of 50 patterns. But the training loss was high. Then you decide to train with 1000 samples. Is this good approach? And why?

   (v) You started training a neural network but the loss was neither decreasing or increasing. State two possible reasons.

   (vi) State two ways to handle local minima problem in gradient descent learning.

Note: Question No. 1 continues on Page 2

(vii) State one advantage and one disadvantage of using a small batch size for training.

(viii) State how you can fix underfitting in your neural network.

(16 marks)

(b) The output $\boldsymbol{y} \in \boldsymbol{R}^K$ of a linear neuron layer is given by $\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{x} + \boldsymbol{b}$ where $\boldsymbol{x} \in \boldsymbol{R}^n$is input, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are the weight matrix and bias vector of the layer, respectively. Given a target vector $\boldsymbol{d} \in \boldsymbol{R}^K$, the square error loss J of the layer is given by $J = (\boldsymbol{d} - \boldsymbol{y})^T(\boldsymbol{d} - \boldsymbol{y})$.

(i) Write the expressions for $\nabla_{\boldsymbol{y}}J$, $\nabla_{\boldsymbol{W}}J$ and $\nabla_{\boldsymbol{b}}J$.

(5 marks)

(ii) If L_2-norm weight regularization penalty $\beta\|\boldsymbol{W}\|^2$ is added to the loss J, write the expressions for $\nabla_{\boldsymbol{W}}J_1$ and $\nabla_{\boldsymbol{b}}J_1$ where $J_1$is the regularized loss. β is the penalty parameter.

(4 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 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 $Tanh$ activations functions and the output neuron is a logistic regression neuron. The weights of the network are initialized as indicated in the figure and all the biases are initialized to 0.5 (not shown).

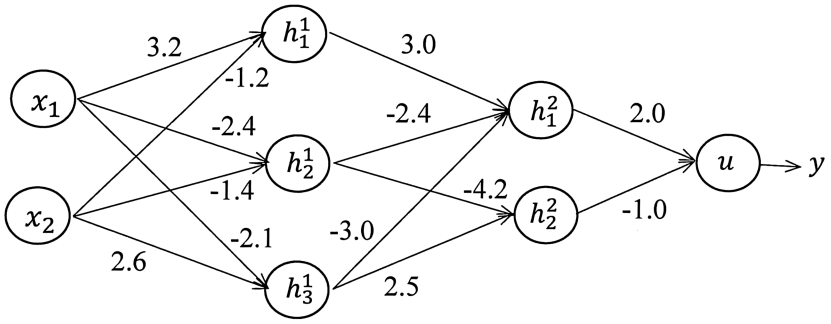

Figure Q2

Note: Question No. 2 continues on Page 3

The network is trained to produce a desired output $d = 0$ for an input $\boldsymbol{x} = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}$. You are to perform one iteration of stochastic gradient descent learning with the example $(\boldsymbol{x}, d)$. Give your answers rounded up to two decimal places.

(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)

(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 $Tanh$ 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 and a fully connected layer. 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.

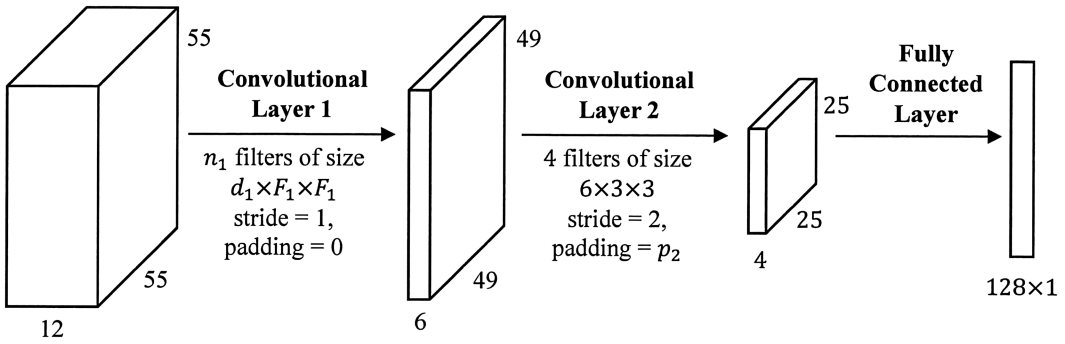

Figure Q3

(a) (i) Give the values of n_1, d_1, F_1 and p_2.

(4 marks)

(ii) Calculate the total number of parameters in each layer, namely the Convolutional Layer 1, Convolutional Layer 2 and Fully Connected Layer. Be reminded to account for the bias terms.

(4 marks)

(b) (i) Assume the same input volume $12 \times 55 \times 55$ and output volume $6 \times 49 \times 49$, replace the Convolution Layer 1 with a depthwise separable convolution, i.e., "depthwise + pointwise" convolutions. Assume stride = 1 and padding = 0. State the number of filters and the size of filter for each of the depthwise and pointwise convolution layers.

(4 marks)

(ii) Calculate the FLOPs of the original Convolution Layer 1, and the FLOPs of the new depthwise separable convolution designed by you in Q3(b)(i). Finally, compute the ratio of the FLOPs between the new and original layers.

(4 marks)

(iii) You want to apply batch normalization in your network. Explain why you should not choose a very small mini-batch size during your training.

(3 marks)

Note: Question No. 3 continues on Page 5

(c) The statements below are all related to autoencoders. Answer "TRUE" or "FALSE" to the following statements. Each part carries 1 mark.

(i) The primary goal of an autoencoder is to achieve high classification accuracy on labeled data.

(ii) Autoencoders can be used as a pre-training step for deep networks in situations where labeled data is limited.

(iii) Stacked autoencoders are created by training multiple autoencoders in parallel, with each one independently processing the input data.

(iv) Overcomplete autoencoders have a latent space dimension that is larger than the input dimension.

(v) Sparse autoencoders utilize a loss function that encourages the network to have many activations close to one in the bottleneck layer.

(vi) Autoencoders with a symmetrical architecture have different layers and neuron counts in the encoder and decoder parts, up to the bottleneck.

(6 marks)

4. (a) Consider a Jordan-type recurrent neural network (RNN) that receives 2-dimensional input patterns $\boldsymbol{x} \in \boldsymbol{R}^2$ and has one hidden layer. The RNN has two neurons in the hidden layer (which are initialized to zeros) and one neuron in the output layer. The hidden layer neurons have *Tanh* activation functions and the output layer neurons use *Sigmoid* activation functions.

The weight matrices $\boldsymbol{U}$ connecting the input to the hidden layer, the top-down recurrence weight matrix $\boldsymbol{W}$, and $\boldsymbol{V}$ connecting the hidden output to the output layer are given by

$$\boldsymbol{U} = \begin{pmatrix} 0.2 & 0.3 \\ 0.8 & 0.9 \end{pmatrix}, \boldsymbol{W} = \begin{pmatrix} 2.0 & 1.0 \end{pmatrix} \text{ and } \boldsymbol{V} = \begin{pmatrix} 0.2 \\ -0.2 \end{pmatrix}$$

All bias connections to neurons are set to 0.2. The output layer is initialized to an output of 1.0.

Note: Question No. 4 continues on Page 6

Find the output of the network for an input sequence $(\boldsymbol{x}(1), \boldsymbol{x}(2))$. Provide your answers rounded to four decimal places.

$$\boldsymbol{x}(1) = \begin{pmatrix} 1.5 \\ 1.0 \end{pmatrix} \text{ and } \boldsymbol{x}(2) = \begin{pmatrix} -3.0 \\ 2.0 \end{pmatrix}.$$

(6 marks)

(b) Select the correct option (A, B, C or D) for each question.

A. Both statements are TRUE.
B. Statement I is TRUE, but statement II is FALSE.
C. Statement I is FALSE, but statement II is TRUE.
D. Both statements are FALSE.

(i) Statement I: The self-attention mechanism in Transformers allows each token in the input sequence to focus on different parts of the sequence when producing the output.
Statement II: In Transformers, the number of self-attention heads is always fixed at one for all models and applications.

(2 marks)

(ii) Statement I: The Transformer model uses convolutional layers to capture local patterns within the sequence data.
Statement II: Layer normalization is a critical component in the Transformer's architecture, helping stabilize the activations throughout the network.

(2 marks)

(iii) Statement I: The "query", "key", and "value" matrices in the self-attention mechanism of Transformers are all derived from the same initial input embeddings.
Statement II: In the multi-head self-attention mechanism, different heads can potentially learn to attend to different parts or aspects of the input sequence.

(2 marks)

Note: Question No. 4 continues on Page 7

(c) The Transformer architecture employs a mechanism known as "positional encoding" to account for the order of words or tokens in a sequence.

(i) Explain how sinusoidal positional encoding is computed and justify why it might be beneficial over other potential positional encoding methods.

(5 marks)

(ii) Based on the sinusoidal positional encoding, calculate the positional encoding values for dimension [0, 10] when the position of a word in the sequence is 5 (assuming the initial position is 0) and the dimension of the embeddings is 512. Provide your answers rounded to four decimal places.

(4 marks)

(d) Explain the concept of "mode collapse" in the context of GANs and discuss its implications on the quality of the generated data.

(4 marks)

END OF PAPER

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

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.