Birla Institute of Technology and Science, Pilani

Work Integrated Learning Programmes Division

M. Tech. in AIML

II Semester 2022-2023

Mid-Semester Test
(EC2 - Makeup)

Course Number	AIMLCZG511
Course Name	Deep Neural Networks
Nature of Exam	Open Book
Weight-age for grading	30
Duration	2 hrs
Date of Exam	

* Pages	3
* Questions	5

---

1. A neural network designed using Tensorflow Keras is given below. Students are instructed to type the answers in the textbox of the portal.

```
net = tf.keras.layers.Sequential()
net.add(
        tf.keras.layers. InputLayer(input_shape = ((256*256*3),)),
        tf.keras.layers.Dense(1048, activation='relu'),
        tf.keras.layers.Dense(512, activation='relu'),
        tf.keras.layers.Dropout(0.4),
        tf.keras.layers.Dense(256, activation='relu'),
        tf.keras.layers.Dense(128, activation='relu'),
        tf.keras.layers.Dense(64, activation='relu'),
        tf.keras.layers.Dropout(0.6),
        tf.keras.layers.Dense(32, activation='relu'),
        tf.keras.layers.Dense(16, activation='relu'),
        tf.keras.layers.Dense(8, activation='softmax'))
```

   (a) What is the objective of the neural network? What is the input given to the network? What is the expected output? How deep and wide is the network. [2]

   (b) Justify the choice activation function in output layer. Instead of Relu activation function, justify the choice of using Tanh activation function. [1]

   (c) Two dropout statements are added in the code. How many additional parameters are learned because of this? If drop out is added after the last statement in the given code, how will it affect the network? [1]

   (d) Write the code snippet for adding the optimizer of your choice. Justify the choice of the optimizer. Assume any other relevant information. [1]

   (e) What will the following code snippet do to the network. [1]

```
cb = tf.keras.callbacks.EarlyStopping(monitor='loss', patience=7)
history = net.fit(epochs=200, batch_size=32, callbacks=[cb])
```

Rubrics

(a) Objective: 8-class classification of 256 by 256 color images 0.5 marks
Input is 256 by 256 colour image and Output is 8 class classification 0.5 marks
Depth = 8 Width = 1048 1 marks

(b) Relu for hidden layers as input is image. Relu as computationally efficient. 0.5 marks
Tanh is computationally expensive when compared to Relu.

(c) 0 additional parameters. 0.5 marks
Adding dropout after last layer will add drop output neurons. Classification may be incorrect. 0.5 marks

(d) Correct code 0.5 marks
Justification of optimizer 0.5 marks

(e) Stops the training when last 7 historical loss values are converging.

2. (a) Figure 1 below plots the loss when batch gradient descent is used for training. Which optimizers plots the loss in figure 2 and 3. [1]

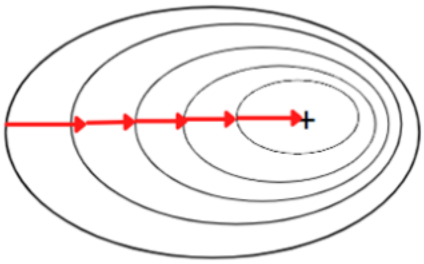


Figure 1

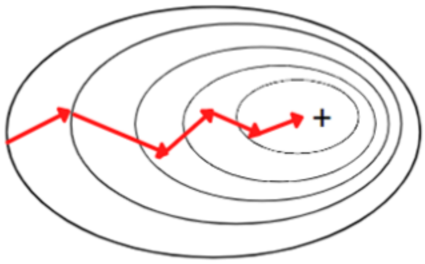


Figure 2

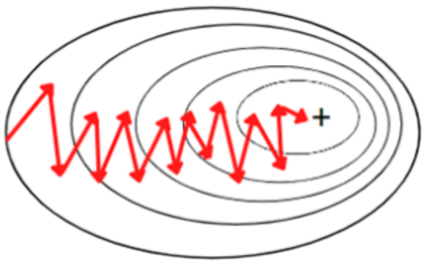


Figure 3

Figure 1

(b) In figure 2, assume that the learning rate used was 0.5. Redraw the plot to show the effect of increasing the learning rate to 2 and decreasing the learning rate to 0.01. [2]

(c) You are given a simple neural network with a single hidden layer containing two neurons, and an output layer containing one neuron. All neurons use the sigmoid activation function. The weights and biases of the network are as follows: [3]

- Weights from input to hidden layer: $w_1 = 0.5, w_2 = -0.6$
- Biases in hidden layer: b1 = 0.1, b2 = -0.2
- Weights from hidden layer to output: w3 = 0.7, w4 = -0.8
- Bias in output layer: b3 = 0.3

Given an input x = 0.75, calculate the output of the network and mean squared loss if the desired output is 1.25. Use the sigmoid activation function. What will be the effect in the loss if the loss function used is

$$L(w, b) = \frac{1}{2}(d - \hat{y})^2 + \lambda||w^2||$$

Rubrics

(a) Fig 2 mini batch sgd and fig 3 is sgd. 2*0.5 marks

(b) Increased learning rate: may not converge or diverge 0.5 marks
Decreased learning rate: slow learning, more oscillations 0.5 marks

(c)

$$
\begin{aligned}
h1 &= sigmoid(w1 * x + b1) = sigmoid(0.5 * 0.75 + 0.1) = 0.6168 \qquad \text{0.5 marks} \\
h2 &= sigmoid(w2 * x + b2) = sigmoid(-0.6 * 0.75 - 0.2) = 0.3427 \qquad \text{0.5 marks} \\
output &= sigmoid(w3 * h1 + w4 * h2 + b3) \\
&= sigmoid(0.7 * 0.6168 + (-0.8) * 0.3427 + 0.3) = 0.6125 \qquad \text{0.5 marks} \\
mse &= 0.5 * (1.25 - 0.6125)^2 = 0.2032 \qquad \text{0.5 marks} \\
L(w,b) &= 0.5 * (1.25 - 0.6125)^2 + \lambda[0.5^2 + (-0.6)^2 + 0.7^2 + (0.8)^2] \\
&= 0.2032 + \lambda\, 1.74 \qquad \text{0.5 marks}
\end{aligned}
$$

Assume Lambda value in the above equation. This introduces more loss, reducing overfitting. 0.5 marks

3. (a) Consider a two input XNOR gate and simulate a perceptron algorithm for it, where, learning rate=0.02 and threshold = 0.2. [2]

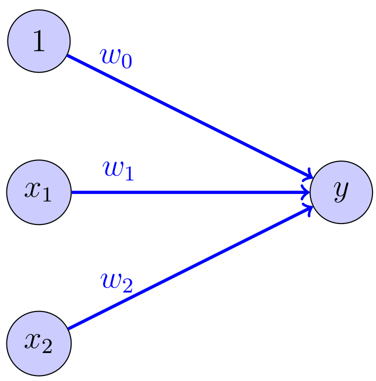
(b) Represent using a multilayer neural network $A \odot B \odot C \odot D \odot E$ where $\odot$ represents XNOR. What will be the optimal depth and width for this network. [3]

(c) You'd like to train a fully-connected neural network with 7 hidden layers, each with 8 hidden units. The input is 30-dimensional and the output is a binary. What is the total number of trainable parameters in your network? [1]

Rubrics

(a)

-1	-1	-1
-1	1	1
1	-1	1
1	1	-1

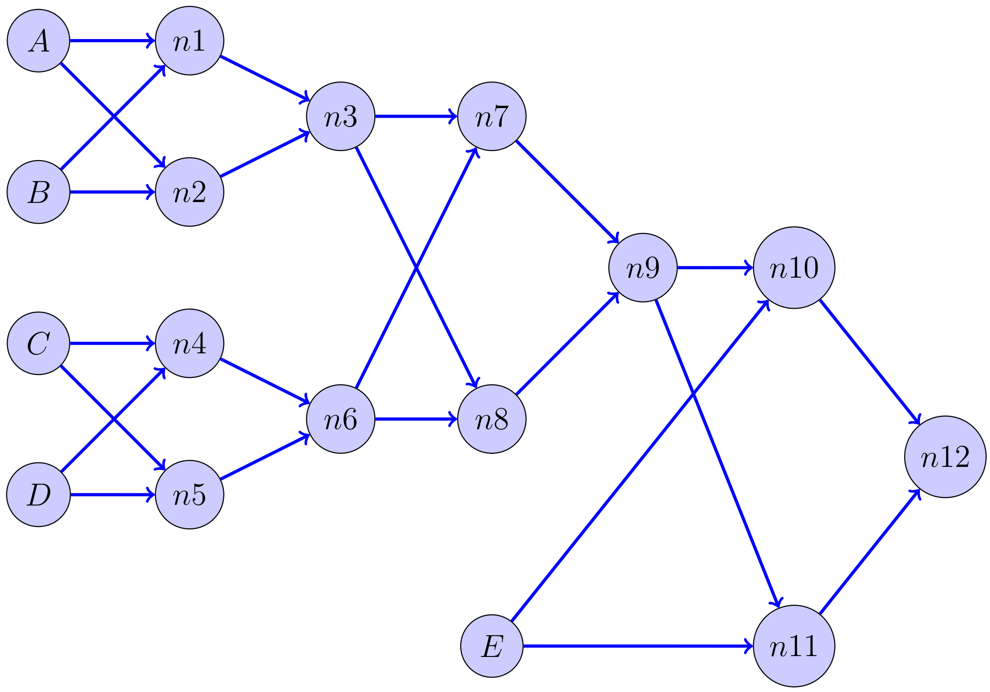


$$
\begin{aligned}
\text{Equations } y &= sign(w_1x_1 + w_2x_2 + w_0) \\
w_i &= w_i + \eta(d-y)x_i \\
\text{Assume } w_1 &= w_2 = 0 \\
\text{First input } y &= sign(0+0-0.2) = -1 \qquad \text{equals d(x). No parameter update} \\
\text{Second input } y &= sign(0+0-0.2) = -1 \qquad \text{not equals d(x). Parameter update} \\
w_1 &= 0 + 0.02*(1-(-1))*(-1) = -0.04 \\
w_2 &= 0 + 0.02*(1-(-1))*(+1) = 0.04 \\
w_0 &= -0.2 + 0.02*(1-(-1)) = -0.16 \\
\text{Third input } y &= sign(-0.04*1 + 0.04*(-1) - 0.16) \\
&= -1 \qquad \text{not equals d(x). Parameter update} \\
w_1 &= -0.04 + 0.02*(1-(-1))*(1) = 0.0 \\
w_2 &= 0.04 + 0.02*(1-(-1))*(-1) = 0.0 \\
w_0 &= -0.16 + 0.02*(1-(-1)) = -0.12 \\
\text{Fourth input } y &= sign(0*1 + 0*1 - 0.12) = -1 \qquad \text{equals d(x). No Parameter update}
\end{aligned}
$$

(b) XNOR can be represented by 3 neurons in triangular pattern similar to XOR. This is one solution.

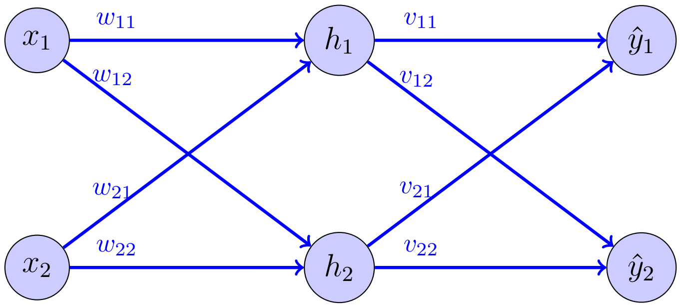
$$
\begin{aligned}
n &= 5 \\
width &= 3(n-1) = 12 \\
depth &= 2log_2n = 2*3 = 6
\end{aligned}
$$

Diagram 2 marks, width 0.5 and depth 0.5



(c) (30+1)*8 + (8+1)*8*7 + (8+1)*1 or (30+1)*8 + (8+1)*8*7 + (8+2)*2 = 761 or 772

4. Consider the following network structure. You can assume the initial weights. Assume bias to be zero for easier computations. Given that $< x_1, x_2, \hat{y}_1, \hat{y}_2 >=< 1, 1, 0, 1 >$ where $\hat{y}$ is the target. Assume $\beta = 0.9$ and $\eta = 0.01$.



(a) Compute the forward propagation and generate the output. Use Relu for hidden layers and Sigmoid activation function for output layer. [2]

(b) Compute the Softmax loss function for both outputs. [1]

(c) Let the initial weights that assumed be the weights [at time (t-1). Compute the weights v_{21}, w_{12} and w_{22} at time t using SGD. [1.5]

(d) Let the weight at time t be the ones computed in part (c). Compute the weights v_{21}, w_{12} and w_{22} at $(t + 1)$ when momentum is used. [1.5]

Rubrics

(a) Award 1 mark if only equations are written. Substitute assumed weights and compute values, then award 2 marks in total. If assumed weights are all same among multiple students, report to IC.

$$h1 = relu(w_{11} * 1 + w_{21} * 1 + 0) \tag{1}$$
$$h2 = relu(w_{12} * 1 + w_{22} * 1 + 0) \tag{2}$$
$$\hat{y_1} = sigmoid(v_{11}h_1 + v_{21}h_2) \quad \text{Substitute values of eqs: 1,2} \tag{3}$$
$$\hat{y_2} = sigmoid(v_{12}h_1 + v_{22}h_2) \quad \text{Substitute values of eqs: 1,2} \tag{4}$$

(b) 1 marks for loss computation. Writing equations alone will be awarded 0 marks.

$$loss \quad L = \frac{e^{\hat{y_1}}}{e^{\hat{y_1}} + e^{\hat{y_2}}}; \frac{e^{\hat{y_2}}}{e^{\hat{y_1}} + e^{\hat{y_2}}}$$

(c) Computation of weights using SGD. Award 1 mark if only equations are written. If all

3 weights of equations 13, 14, 15 computed, then award 3*0.5 marks.

$$\frac{\partial L}{\partial \hat{z}} = \frac{\partial L}{\partial \hat{y}} * \frac{\partial \hat{y}}{\partial \hat{z}} = \frac{\partial L}{\partial \hat{y}} * 1 \qquad \text{Substitute value of eq: 6} \tag{5}$$

$$\frac{\partial L}{\partial v_{21}} = \frac{\partial L}{\partial \hat{z}} * \frac{\partial \hat{z}}{\partial v_{21}} = \frac{\partial L}{\partial \hat{z}} * h_2 \qquad \text{Substitute value of eq: 7} \tag{6}$$

$$\frac{\partial L}{\partial h_2} = \frac{\partial L}{\partial \hat{z}} * \frac{\partial \hat{z}}{\partial h_2} = \frac{\partial L}{\partial \hat{z}} * v_{21} \qquad \text{Substitute value of eq: 6} \tag{7}$$

$$\frac{\partial L}{\partial z_2} = \frac{\partial L}{\partial h_2} * \frac{\partial h_2}{\partial z_2} = \frac{\partial L}{\partial h_2} * 1 \qquad \text{Substitute value of eq: 9} \tag{8}$$

$$\frac{\partial L}{\partial w_{12}} = \frac{\partial L}{\partial z_2} * \frac{\partial z_2}{\partial w_{12}} = \frac{\partial L}{\partial z_2} * x_1 \qquad \text{Substitute value of eq: 10} \tag{9}$$

$$\frac{\partial L}{\partial w_{22}} = \frac{\partial L}{\partial z_2} * \frac{\partial z_2}{\partial w_{22}} = \frac{\partial L}{\partial z_2} * x_2 \qquad \text{Substitute value of eq: 10} \tag{10}$$

$$v_{21} = assumed\, v_{21} - \eta \frac{\partial L}{\partial v_{21}} \qquad \text{Substitute value of eq: 8} \tag{11}$$

$$w_{12} = assumed\, w_{12} - \eta \frac{\partial L}{\partial w_{12}} \qquad \text{Substitute value of eq: 11} \tag{12}$$

$$w_{22} = assumed\, w_{22} - \eta \frac{\partial L}{\partial w_{22}} \qquad \text{Substitute value of eq: 12} \tag{13}$$

(d) Computation of weights using Momentum. Award 1 mark if only equations are written. If all 3 weights of equations 16, 17, 18 computed, then award 3*0.5 marks.

$$v_2 = v_2(eqn13) - \eta \frac{\partial L}{\partial v_2} + \beta(v_2(eqn13) - assumed\, v_2) \tag{14}$$

$$w_{12} = w_{12}(eqn14) - \eta \frac{\partial L}{\partial w_{12}} + \beta(w_{12}(eqn14) - assumed\, w_{12} \tag{15}$$

$$w_{22} = w_{22}(eqn15) - \eta \frac{\partial L}{\partial w_{22}} + \beta(w_{22}(eqn14) - assumed\, w_{22} \tag{16}$$

5. (a) Represent the function

$$f(x, y) = x^4 - 32x^2 + y^4 - 18y^2$$

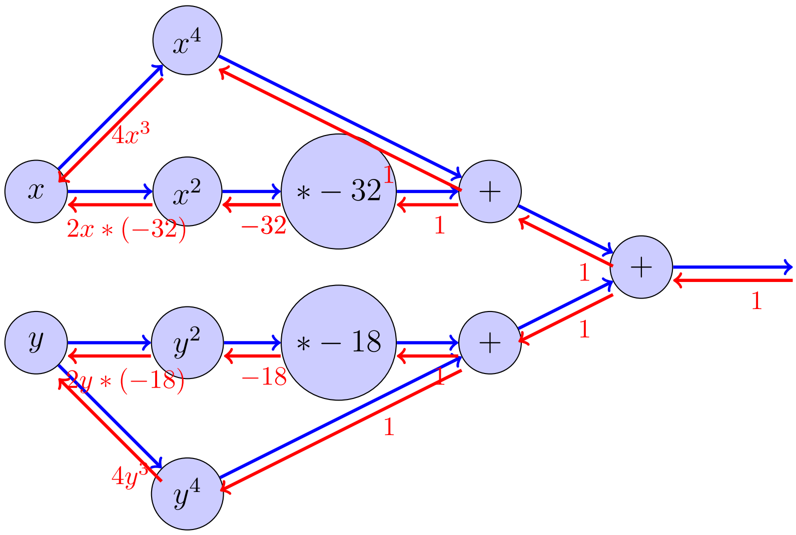
using a computation graph. Evaluating this function at x = 2, and y = 2. Find the first derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ using computation graph through backpropagation. [3]

(b) Compute the Hessian of the function

$$f(x, y) = x^4 - 32x^2 + y^4 - 18y^2$$

You are given two point (0,0) and (4,3). Among these two points, which points are the local minima or local maxima or both. [3]

(a) Graph 1.5 mark computing of partial derivatives using graph 1.5 marks



$$f(x,y) = x^4 - 32x^2 + y^4 - 18y^2$$
$$\frac{\partial f}{\partial x} = 4x^3 - 64x$$
$$\frac{\partial f}{\partial y} = 4y^3 - 36x$$

$$f(x,y) = x^4 - 32x^2 + y^4 - 18y^2$$
$$H = \begin{bmatrix} 12x^2 - 64 & 0 \\ 0 & 12y^2 - 36 \end{bmatrix}$$
$$\nabla^2 f(0,0) = \begin{bmatrix} -64 & 0 \\ 0 & -36 \end{bmatrix}$$
$$\nabla^2 f(4,3) = \begin{bmatrix} 128 & 0 \\ 0 & 72 \end{bmatrix}$$

(b) local minimum is (4,3) and point of local maximum is (0,0)
Local minima 1.5 marks and local maxima 1.5 marks