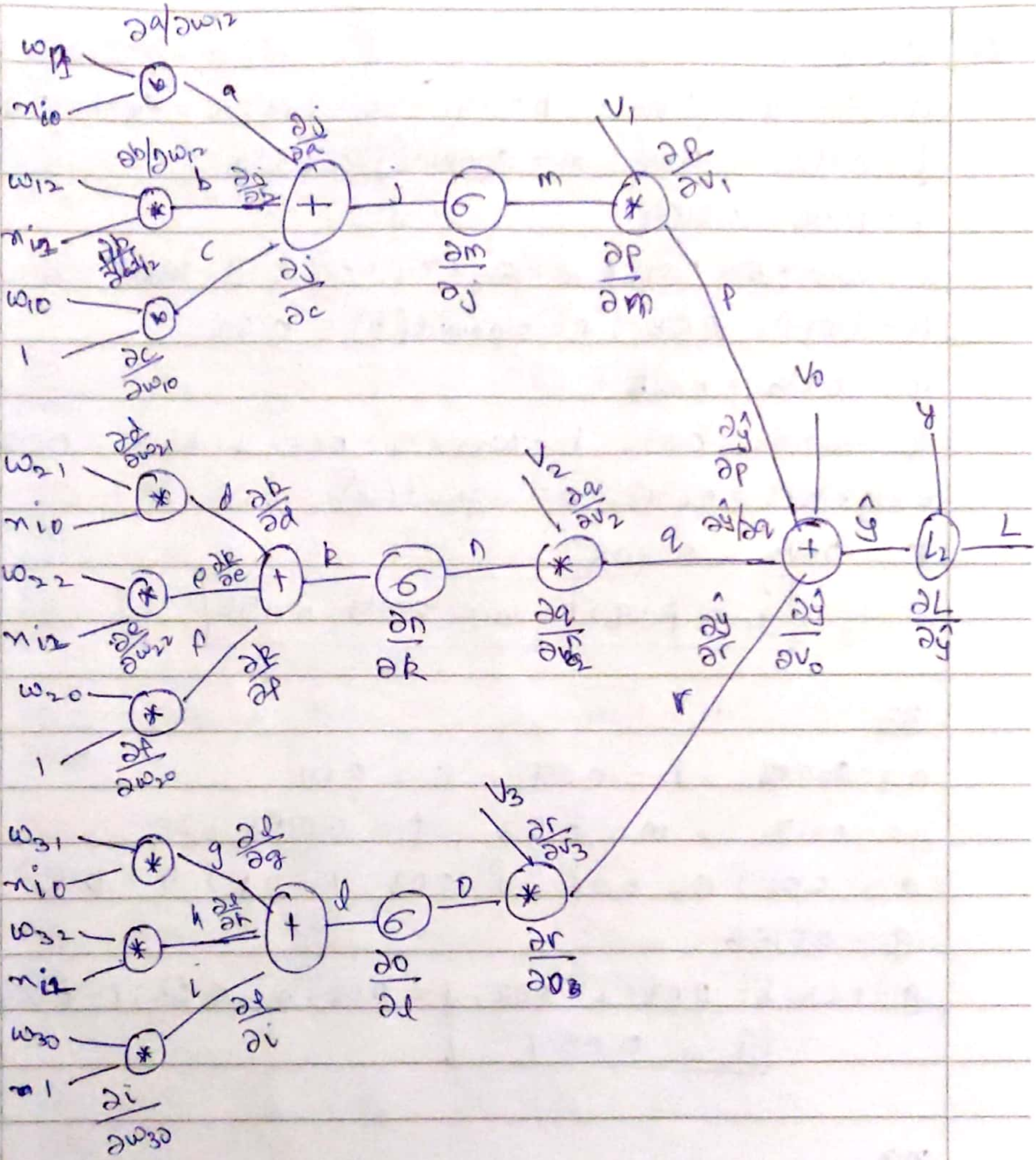
Pranav Behari Lal – A20417025

Assignment 2

Deep Learning – Spring 2019 – CS577

1a.



x10
w12
x11
w10
1
a
b
c
j
m
p
v1
w21
x10
w22
x11
w20
1
d
e
f
k
n
q
v2
v0
y
L2
L
w31
x10
w32
x11
w30
1
g
h
i
l
o
v3
r

b. $\underline{x_1}$

$a = w_{11} \times x_{10} = 0.02$; $b = w_{12} \times x_{11} = 0.06$; $c = w_{10} \times 1 = 0.01$

$j = a+b+c = 0.09$; $m = \text{sigmoid}(j) = 0.52$

$p = m \times v_1 = 0.0104$

$d = w_{21} \times x_{10} = 0.01$; $e = w_{22} \times x_{11} = 0.04$; $f = w_{20} \times 1 = 0.03$

$k = d+e+f = 0.08$; $n = \text{sigmoid}(k) = 0.52$

$q = n \times v_2 = 0.015$

$g = w_{31} \times x_{10} = 0.03$; $h = w_{32} \times x_{11} = 0.02$; $i = w_{30} \times 1 = 0.02$

$l = g+h+i = 0.07$; $o = \text{sigmoid}(l) = 0.52$

$r = o \times v_3 = 0.0208$

$$\boxed{\hat{y}_1 = p+q+r+v_0 = \cancel{0.24}\ 0.056}$$

$\underline{x_2}$

$a = 0.02$; $b = 0.09$; $c = 0.01$

$j = 0.12$; $m = 0.53$; $p = 0.0106$

$d = 0.01$; $e = 0.06$; $f = 0.03$; $k = 0.1$; $n = 0.52$

$q = 0.0156$

$g = 0.03$; $h = 0.03$; $i = 0.02$; $l = 0.08$; $o = 0.52$; $r = 0.0208$

$$\boxed{\hat{y}_2 = 0.057}$$

$\underline{x_3}$

$a = 0.04$; $b = 0.06$; $c = 0.01$; $j = 0.11$; $m = 0.53$; $p = 0.0106$

$d = 0.02$; $e = 0.04$; $f = 0.03$; $k = 0.09$; $n = 0.52$; $q = 0.0156$

$g = 0.06$; $h = 0.02$; $i = 0.02$; $l = 0.1$; $o = 0.52$; $r = 0.0208$

$$\boxed{\hat{y}_3 = 0.057}$$

c.) $\frac{\partial L}{\partial \hat{y}} = \cancel{-2(\hat{y}-y)} \; \cancel{-2(y-\hat{y})} ; \; -2(y-\hat{y})$ | $L = (y-\hat{y})^2$

$$\frac{\partial \hat{y}}{\partial v_0} = \frac{\partial \hat{y}}{\partial p} = \frac{\partial \hat{y}}{\partial q} = \frac{\partial \hat{y}}{\partial r} = 1$$

$$\Rightarrow \frac{\partial p}{\partial v_1} = m ; \; \frac{\partial q}{\partial v_2} = n ; \; \frac{\partial r}{\partial v_3} = 0 ; \cancel{o}.$$

$$\frac{\partial p}{\partial m} = v_1 ; \; \frac{\partial q}{\partial n} = v_2 ; \; \frac{\partial r}{\partial o} = v_3$$

$$\frac{\partial m}{\partial j} = m(1-m) ; \; \frac{\partial n}{\partial k} = n(1-n) ; \; \frac{\partial o}{\partial l} = o(1-o)$$

$$\frac{\partial j}{\partial a} = \frac{\partial j}{\partial b} = \frac{\partial j}{\partial c} = \frac{\partial k}{\partial d} = \frac{\partial k}{\partial e} = \frac{\partial k}{\partial f} = \frac{\partial l}{\partial g} = \frac{\partial l}{\partial h} = \frac{\partial l}{\partial i} = 1$$

$$\frac{\partial a}{\partial w_{11}} = x_{i0} ; \; \frac{\partial b}{\partial w_{12}} = x_{i1} ; \; \frac{\partial c}{\partial w_{10}} = 1$$

$$\frac{\partial d}{\partial w_{21}} = x_{i0} ; \; \frac{\partial e}{\partial w_{22}} = x_{i1} ; \; \frac{\partial f}{\partial w_{20}} = 1$$

$$\frac{\partial g}{\partial w_{31}} = x_{i0} ; \; \frac{\partial h}{\partial w_{32}} = x_{i1} ; \; \frac{\partial i}{\partial w_{30}} = 1$$

$$\Rightarrow \frac{\partial \hat{y}}{\partial v_0} = 1 ; \; \frac{\partial \hat{y}}{\partial v_1} = m ; \; \frac{\partial \hat{y}}{\partial v_2} = n ; \; \frac{\partial \hat{y}}{\partial v_3} = 0$$

$$\frac{\partial \hat{y}}{\partial w_{10}} = \frac{\partial \hat{y}}{\partial p} \cdot \frac{\partial p}{\partial m} \frac{\partial m}{\partial j} \cdot \frac{\partial j}{\partial c} \cdot \frac{\partial c}{\partial w_{10}} = 1 \cdot v_1 \cdot m(1-m) \cdot 1 \cdot 1$$

$$\frac{\partial \hat{y}}{\partial w_{11}} = \frac{\partial \hat{y}}{\partial p} \frac{\partial p}{\partial m} \frac{\partial m}{\partial j} \cdot \frac{\partial j}{\partial a} \frac{\partial a}{\partial w_{11}} = 1 \cdot v_1 \, m(1-m) \, 1 \cdot x_{i0}$$

$$\frac{\partial \hat{y}}{\partial w_{12}} = \frac{\partial \hat{y}}{\partial p} \cdot \frac{\partial m}{\partial m} \cdot \frac{\partial m}{\partial j} \cdot \frac{\partial j}{\partial b} \cdot \frac{\partial b}{\partial w_{12}} = 1 \cdot v_1 \, m(1-m) \cdot 1 \cdot x_{i1}$$

$|||^{ly}$, $\frac{\partial \hat{y}}{\partial w_{20}} = 1 \cdot v_2 \cdot n(1-n) \cdot 1 \cdot 1$; $\frac{\partial \hat{y}}{\partial w_{21}} = 1 \cdot v_2 \cdot n(1-n) \cdot 1 \cdot x_{i0}$

$$\frac{\partial \hat{y}}{\partial w_{22}} = 1 \cdot v_2 \cdot n(1-n) \cdot 1 \cdot x_{i1}$$

$\frac{\partial \hat{y}}{\partial w_{30}} = 1 \cdot v_3 \cdot o(1-o) \cdot 1 \cdot 1$; $\frac{\partial \hat{y}}{\partial w_{31}} = 1 \cdot v_3 \cdot o(1-o) \cdot 1 \cdot x_{i0}$

$$\frac{\partial \hat{y}}{\partial w_{32}} = 1 \cdot v_3 \cdot o(1-o) \cdot 1 \cdot x_{i1}$$

$\rightarrow$ $\underline{x_1}$

$$\frac{\partial l}{\partial \hat{y}} = -2(y-\hat{y}) = -2(8-0.056) = \cancel{-15.888} \; -15.888$$

$$\frac{\partial \hat{y}}{\partial v} = \begin{bmatrix} \frac{\partial \hat{y}}{\partial v_0} & \frac{\partial \hat{y}}{\partial v_1} & \frac{\partial \hat{y}}{\partial v_2} & \frac{\partial \hat{y}}{\partial v_3} \end{bmatrix} = \begin{bmatrix} 1 & 0.52 & 0.52 & 0.52 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial w} = \begin{bmatrix} \partial\hat{y}/\partial w_{10} & \partial\hat{y}/\partial w_{11} & \partial\hat{y}/\partial w_{12} \\ \partial\hat{y}/\partial w_{20} & \partial\hat{y}/\partial w_{21} & \partial\hat{y}/\partial w_{22} \\ \partial\hat{y}/\partial w_{30} & \partial\hat{y}/\partial w_{31} & \partial\hat{y}/\partial w_{32} \end{bmatrix}$$

$$= \begin{bmatrix} 0.00499 & 0.00499 & 0.00998 \\ 0.00748 & 0.00748 & 0.0149 \\ 0.00998 & 0.00998 & 0.0199 \end{bmatrix}$$

$$\Rightarrow \Delta v = \frac{\partial l}{\partial v} = \begin{bmatrix} -15.888 & -8.262 & -8.262 & -8.262 \end{bmatrix}$$

$$\Delta w = \frac{\partial l}{\partial w} = \begin{bmatrix} -0.0792 & -0.0792 & -0.1584 \\ -0.1187 & -0.1187 & -0.2375 \\ -0.1584 & -0.1584 & -0.3168 \end{bmatrix}$$

1d. $\Delta v_j = \sum_{i=1}^{m} (\hat{y}_j^{(i)} - y_j^{(i)}) \cdot z^{(i)}$

$\Delta w_j = \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \cdot v_j \; z_j^{(i)}(1-z_j^{(i)}) \, x^{(i)}$

In our example, $z^{(i)}$ is the vector $[m \; n \; o]$

→ For $x = [1,2]$ $y = [8]$

$\Delta v_j = (\hat{y}_j^{(i)} - y_j^{(i)}) \; \cancel{[m \; n \; o]} \; [1 \; m \; n \; o]$

(adding a bias term)

$\Rightarrow \Delta v_j = [-15.888 \quad -8.262 \quad -8.262 \quad -8.262]$

$\Delta w_j = (\hat{y}^{(i)} - y^{(i)}) \cdot v_j \cdot z_j^{(i)} (1 - z_j^{(i)}) \; x^{(i)}$

→ $\Delta w_1 = (-15.888) \cdot 0.02 \cdot 0.52 \, (1-0.52) \, [1 \;\; 1 \;\; 2]$

$\cancel{z_j} = z_1 = m$; adding a bias in x

$\Rightarrow \Delta w_1 = [-0.792 \quad -0.792 \quad -0.1584]$

Similarly, $\Delta w_2 = (\hat{y}^{(i)} - y^{(i)}) \, v_2 \cdot z_2^{(i)} (1 - z_2^{(i)} \; x^{(i)})$

$\rightarrow \Delta w_2 = [-0.1187 \quad -0.1187 \quad -0.2375]$

$\rightarrow \Delta w_3 = [\hat{y}^{(i)} - y^{(i)}] \, v_3 \, z_3^{(i)} \cdot (1 - z_3^{(i)}) \, x^{(i)}$

$= [-0.1584 \quad -0.1584 \quad -0.3168]$

These are the same results we got in part 1(c).

2.

a.) $f(x,y) = (2x+3y)^2$

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} = \begin{bmatrix} 4(2x+3y) \\ 6(2x+3y) \end{bmatrix}$$

b. $F(x,y) = \begin{bmatrix} x^2+2y \\ 3x+4y^2 \end{bmatrix}$

$$DF(x,y) = \begin{bmatrix} 2x & 2 \\ 3 & 8y \end{bmatrix}$$

$$DF(1,2) = \begin{bmatrix} 2 & 2 \\ 3 & 16 \end{bmatrix}$$

2c. $G(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$

$$F \circ G(x) = \begin{bmatrix} x^2 + 2x^2 \\ 3x + 4x^4 \end{bmatrix} = \begin{bmatrix} 3x^2 \\ 3x + 4x^4 \end{bmatrix}$$

$$DG(x) = \begin{bmatrix} 1 \\ 2x \end{bmatrix}$$

<u>CHAIN RULE:</u>

$$D(F \circ G)(x) = DF(G(x))\ G(x)$$

$$= \begin{bmatrix} 2x & 2 \\ 3 & 8x^2 \end{bmatrix} \begin{bmatrix} 1 \\ 2x \end{bmatrix}$$

$$= \begin{bmatrix} 6x \\ 3 + 16x^3 \end{bmatrix}$$

$$\Rightarrow D(F \circ G)(2) = \begin{bmatrix} 12 \\ 121 \end{bmatrix}$$

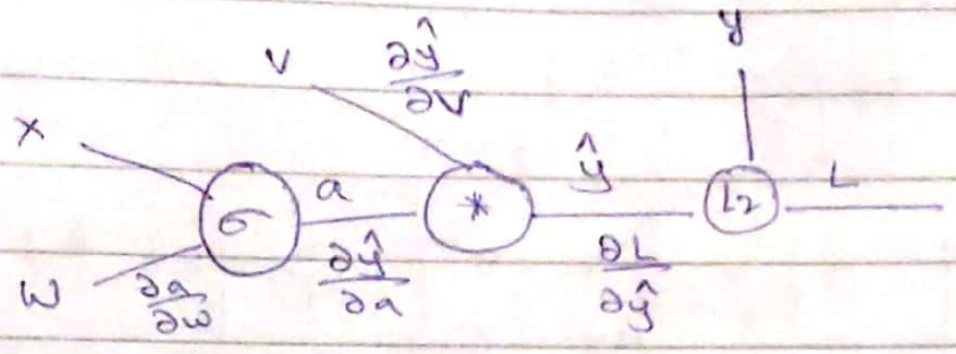
<u>DIRECTLY</u>

$$F \circ G(x) = \begin{bmatrix} 3x^2 \\ 3x + 4x^4 \end{bmatrix}$$

$$\Rightarrow D(F \circ G)(x) = \begin{bmatrix} 6x \\ 3 + 4x^3 \end{bmatrix}$$

$$\Rightarrow D(F \circ G)(2) = \begin{bmatrix} 12 \\ 121 \end{bmatrix}$$

2d.



$$x^T = \begin{bmatrix} 1 & 1 & 2 \end{bmatrix}$$

$$w = \begin{bmatrix} w_1^T \\ w_2^T \\ w_3^T \end{bmatrix} = \begin{bmatrix} 0.01 & 0.02 & 0.03 \\ 0.03 & 0.01 & 0.02 \\ 0.02 & 0.03 & 0.01 \end{bmatrix}$$

$$v = \begin{bmatrix} 0.01 & 0.02 & 0.03 & 0.04 \end{bmatrix}$$

FORWARD PASS

$$a = \text{sigmoid}(wx)$$

$$wx = \begin{bmatrix} 0.01 & 0.02 & 0.03 \\ 0.03 & 0.01 & 0.02 \\ 0.02 & 0.03 & 0.01 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0.09 \\ 0.08 \\ 0.07 \end{bmatrix}$$

$$a = \text{sigmoid}(wx) = \begin{bmatrix} 0.52 \\ 0.52 \\ 0.52 \end{bmatrix} = \begin{bmatrix} w_1 x \\ w_2 x \\ w_3 x \end{bmatrix}$$

$\hat{y} = va$ ($\because$ after adding a bias term in 'a')

$$\Rightarrow \hat{y} = \begin{bmatrix} 0.01 & 0.02 & 0.03 & 0.04 \end{bmatrix} \begin{bmatrix} 1 \\ 0.52 \\ 0.52 \\ 0.52 \end{bmatrix} = 0.056$$

BACKWARD PROPAGATION

$$L = (\hat{y} - y)^2$$

$$\frac{\partial L}{\partial \hat{y}} = 2(\hat{y} - y) = -15.888$$

$$\frac{\partial \hat{y}}{\partial v} = \begin{bmatrix} \frac{\partial \hat{y}}{\partial v_0} \\ \frac{\partial \hat{y}}{\partial v_1} \\ \frac{\partial \hat{y}}{\partial v_2} \\ \frac{\partial \hat{y}}{\partial v_3} \end{bmatrix} = \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.52 \\ 0.52 \\ 0.52 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial a} = \begin{bmatrix} \frac{\partial \hat{y}}{\partial a_1} \\ \frac{\partial \hat{y}}{\partial a_2} \\ \frac{\partial \hat{y}}{\partial a_3} \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$\frac{\partial a}{\partial w} = \begin{bmatrix} \frac{\partial a_1}{\partial w_1} \\ \frac{\partial a_2}{\partial w_2} \\ \frac{\partial a_3}{\partial w_3} \end{bmatrix} = \begin{bmatrix} \frac{\partial \, \text{sigmoid}(w_1 x)}{\partial w_1} \\ \frac{\partial \, \text{sigmoid}(w_2 x)}{\partial w_2} \\ \frac{\partial \, \text{sigmoid}(w_3 x)}{\partial w_3} \end{bmatrix} = \begin{bmatrix} a_1(1-a_1) \\ a_2(1-a_2) \\ a_3(1-a_3) \end{bmatrix} \begin{bmatrix} x^T \\ x^T \\ x^T \end{bmatrix}$$

$$= \begin{bmatrix} 0.52(1-0.52) \\ 0.52(1-0.52) \\ 0.52(1-0.52) \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.249 & 0.249 & 0.499 \\ 0.249 & 0.249 & 0.499 \\ 0.249 & 0.249 & 0.499 \end{bmatrix}$$

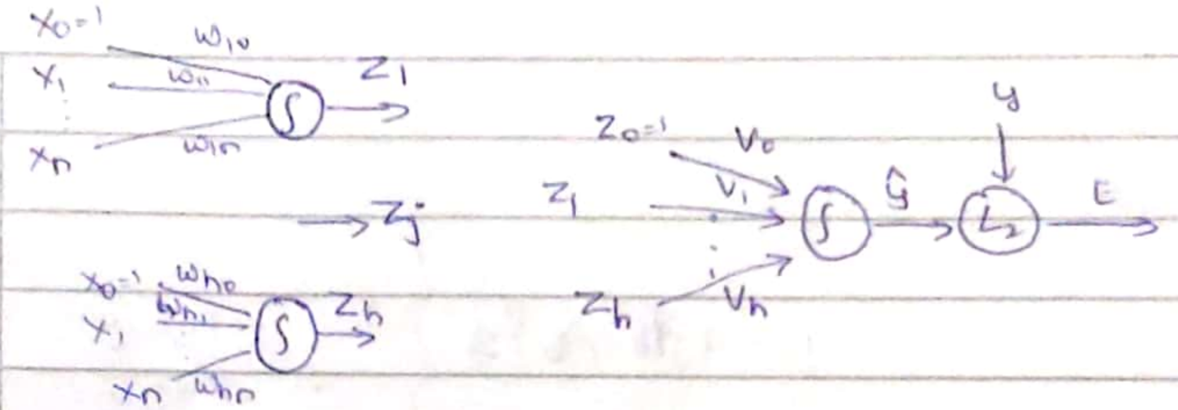
$$\Rightarrow \frac{\partial L}{\partial v} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial v} = \begin{bmatrix} -15.888 \\ -8.262 \\ -8.262 \\ -8.262 \end{bmatrix}$$

$$\cancel{\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a} \cdot \frac{\partial a}{\partial w}} = \cancel{\begin{bmatrix} -0.0792 & -0.0792 & -0.158 \end{bmatrix}}$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a} \frac{\partial a}{\partial w} = \begin{bmatrix} -0.0792 & -0.0792 & -0.1584 \\ -0.1187 & -0.1187 & -0.2375 \\ -0.1584 & -0.1584 & -0.3168 \end{bmatrix}$$

Scanned by CamScanner

3.



$$E = \frac{1}{2} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

$$\hat{y} = \text{sigmoid}(z^T v) = \text{sigmoid}(v_0 + z_1 v_1 + \cdots + z_h v_h)$$

$$z_j = \text{sigmoid}(w_j^T x)$$

$$\frac{\partial E}{\partial v} = \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v}$$

$$\frac{\partial E}{\partial \hat{y}} = \sum_{i=1}^{m} (\hat{y}_i - y_i)$$

$$\frac{\partial \hat{y}}{\partial v} = \frac{\partial\, \text{sigmoid}(z^T v)}{\partial v} = \hat{y}^{(i)}(1 - \hat{y}^{(i)}) \cdot z^{(i)}$$

$$\Rightarrow \boxed{\frac{\partial E}{\partial v} = \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot \hat{y}^{(i)}\left(1 - \hat{y}^{(i)}\right) \cdot z^{(i)}}$$

Implementation Question

Files:

1. HiddenLayer.py:
- Class for a Hidden Layer. Contains functions for forward and backward passes on a Hidden Layer.
- Uses Sigmoid Activation.
- Size of Hidden Layers is passed as a command line argument in the Main.py class.

2. OutputLayer.py
- Class for a Output Layer. Contains functions for forward and backward passes on the Output Layer.
- Uses Softmax Activation
- Size of Output Layer depends on the dataset fed to the model. It takes the value of the number of classes in the target variable.

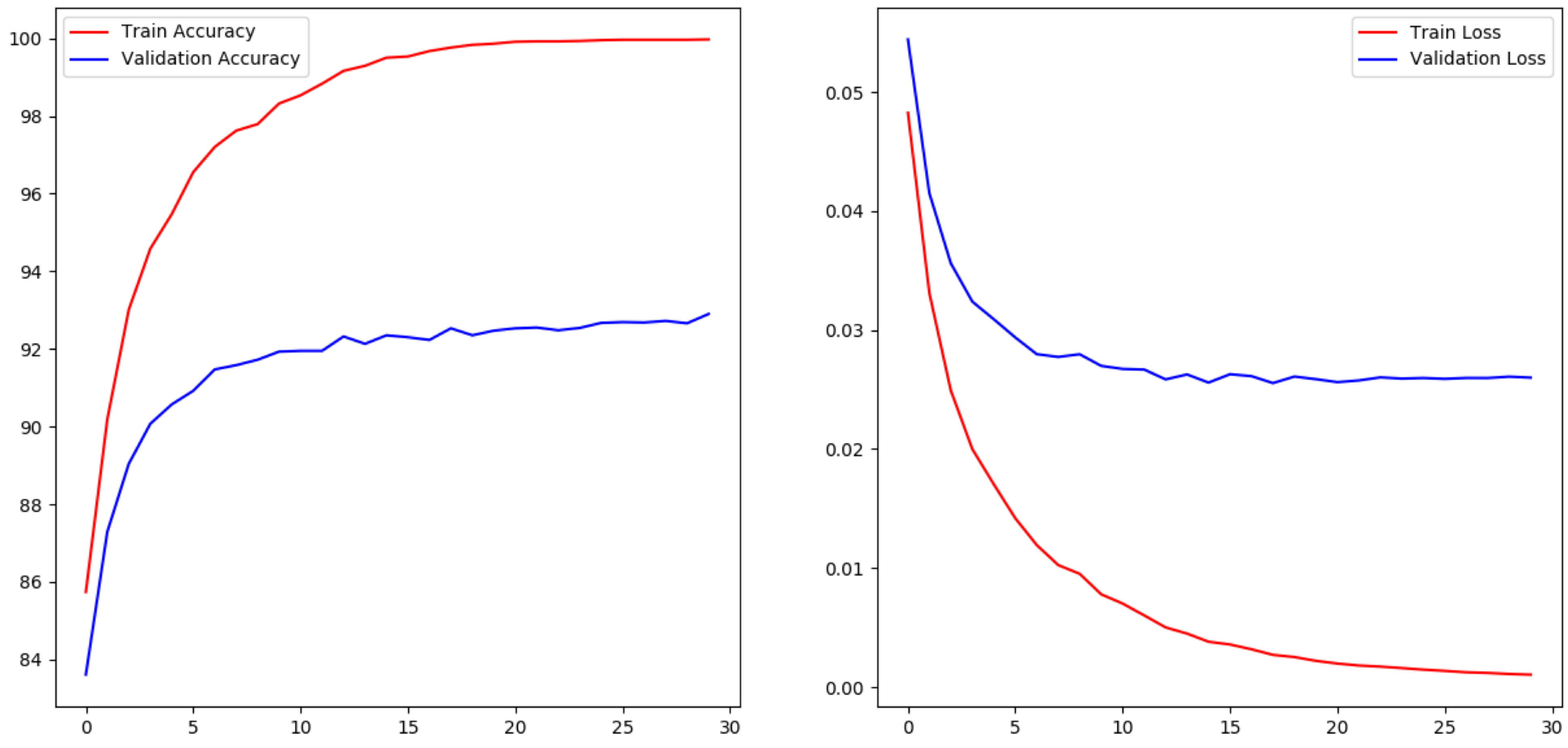
3. Model.py
- This class is used to define the Model.
- It instantiates the Hidden and Output Layers in accordance to the parameters fed to it. Number and size of Hidden Layers can be fed as command line parameters in the Main.py class.
- Contains functions for Forward and Backward pass through the entire model, and also evaluation and plotting functions.
- Uses Categorical Cross-Entropy as the loss function.

4. Main.py
- The Main driver file. Trains models on 2 different datasets and saves the metrics plots.
- Datasets used are – MNIST dataset from keras and IRIS dataset from UCI ML repository.

To execute, place the 4 files in same directory and run the following command:
Python Main.py –learning_rate *lr* –epochs *epochs* –hidden_layers *[array_of_hidden_layers]*

E.g: Python Main.py –learning_rate *0.001* –epochs 30 –hidden_layers [256, 64]

MNIST Dataset Results



IRIS Dataset Results