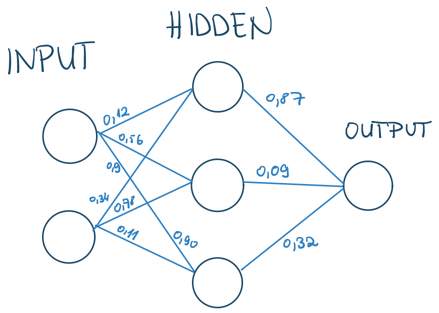
Homework 9 (Bonus sheet)

Group members:
Dhvaniben Jasoliya
Leutrim Uka
Nicola Horst
Taqueer Rumaney
Yuvraj Dhepe

1



(a) $x_1 = [1.0 \quad 1.0]$; $\quad$ Output $= \psi\left(\psi\left(x_1 \cdot W^{(1)} + b^{(1)}\right) \cdot W^{(2)} + b^{(2)}\right)$

$$x_1 \cdot W^{(1)} = \underset{1\times 2}{\begin{bmatrix} 1.0 & 1.0 \end{bmatrix}} \times \underset{3\times 2}{\begin{bmatrix} 0.12 & 0.34 \\ 0.56 & 0.78 \\ 0.90 & 0.11 \end{bmatrix}^T} = \underset{1\times 2}{\begin{bmatrix} 1.0 & 1.0 \end{bmatrix}} \times \begin{bmatrix} 0.12 & 0.56 & 0.90 \\ 0.34 & 0.78 & 0.11 \end{bmatrix}_{2\times 3}$$

$$= \begin{bmatrix} 0.12+0.34 & 0.56+0.78 & 0.9+0.11 \end{bmatrix} = \begin{bmatrix} 0.46 & 1.34 & 1.01 \end{bmatrix} \leftarrow \text{add } b^{(1)T}$$

$$x_1 \cdot W^{(1)} + b^{(1)} = \begin{bmatrix} 0.46 & 1.34 & 1.01 \end{bmatrix} + \begin{bmatrix} 0.21 & 0.43 & 0.65 \end{bmatrix}$$

$$= \begin{bmatrix} 0.67 & 1.77 & 1.66 \end{bmatrix} \leftarrow \text{apply ReLU}$$

$$\underbrace{\psi(x_1 \cdot W^{(1)} + b^{(1)})}_{H^{(1)}} = \begin{bmatrix} \max(0, 0.67) & \max(0, 1.77) & \max(0, 1.66) \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 0.67 & 1.77 & 1.66 \end{bmatrix}}_{\text{output of hidden layer}}$$

$$H^{(1)} \cdot W^{(2)} = \begin{bmatrix} 0.67 & 1.77 & 1.66 \end{bmatrix} \cdot \begin{bmatrix} 0.87 \\ 0.09 \\ 0.32 \end{bmatrix} = (0.67 \cdot 0.87) + (1.77 \cdot 0.09) + (1.66 \cdot 0.32) = 0.5829 + 0.1593 + 0.5312$$

$$= 1.2734 \quad \leftarrow \text{add } b^{(2)}$$

$$H^{(1)} \cdot W^{(2)} + b^{(2)} = 1.2734 + 0.54 = 1.8134 \quad \leftarrow \text{apply ReLU}$$

$$\psi(H^{(1)} W^{(2)} + b^{(2)}) = 1.8134$$

$$\boxed{\rightarrow \text{The output of the network with } x_1 = [1.0 \quad 1.0] \text{ is } 1.8134}$$

b) $x_2 = [0.0 \quad 1.0]$

$$x_2 \cdot W^{(1)} + b^{(1)} = \begin{bmatrix} 0.0 & 1.0 \end{bmatrix} \cdot \begin{bmatrix} 0.12 & 0.56 & 0.90 \\ 0.34 & 0.78 & 0.11 \end{bmatrix} + b^{(1)} = \begin{bmatrix} 0.34 & 0.78 & 0.11 \end{bmatrix} + \begin{bmatrix} 0.21 & 0.43 & 0.65 \end{bmatrix}$$

$$= \begin{bmatrix} 0.55 & 1.21 & 0.76 \end{bmatrix} \leftarrow \text{ReLU}$$

$$\underbrace{\psi(x_2 \cdot W^{(1)} + b^{(1)})}_{H^{(1)}} = \begin{bmatrix} 0.55 & 1.21 & 0.76 \end{bmatrix} \quad \text{(output of hidden layer)}$$

$$H^{(1)} \cdot W^{(2)} + b^{(2)} = 0.8306 + 0.54 = 1.3706 \leftarrow \text{ReLU}$$

$$\psi(H^{(1)} W^{(2)} + b^{(2)}) = 1.3706$$

$$\boxed{\rightarrow \text{The output of the network with } x_2 = [0.0 \quad 1.0] \text{ is } 1.3706.}$$

(b) If we have a fixed output $y_{fix} = 2.$ and a fixed input $x_{fix} = [1., 1.]$, we can modify the formula we used in (a):

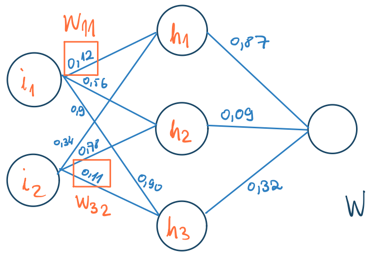
$$y_{fix} = \psi(x_{fix} \cdot W^{(1)} + b^{(1)}) \cdot W^{(2)} + b^{(2)}$$

$$\Rightarrow b^{(2)} = y_{fix} - \psi(x_{fix} \cdot W^{(1)} + b^{(1)}) \cdot W^{(2)}$$

$$b^{(2)} = 2 - [0.67 \;\; 1.77 \;\; 1.66] \cdot \begin{bmatrix} 0.87 \\ 0.09 \\ 0.32 \end{bmatrix} = 2 - 1.2734 = 0.7266$$

2

$$a_0^L[x] = 1,813 \qquad y = 1$$



$$W^{(1)} = \begin{array}{cc} \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \end{bmatrix} & \begin{matrix} \} h_1 \\ \} h_2 \\ \} h_3 \end{matrix} \\ \begin{matrix} \underbrace{\quad}_{i_1} & \underbrace{\quad}_{i_2} \end{matrix} & \end{array}$$

(i) $$\frac{\partial C}{\partial w_{11}^1} = a_1^0[x]\, \delta_1^1$$

$$\delta^2 = \psi'(z_j^L) \cdot (a_1^L[x] - y_1) = \mathrm{ReLU}(1{,}813)' \cdot (1{,}813 - 1) = 1 \cdot (0{,}813) = 0{,}813$$

$$\delta_1^1 = ((W^{(2)})^T \cdot \delta^2)_1 \;\; \psi'(z_1^1)$$

$$\rightarrow \delta_1^1 = \left(\begin{bmatrix} 0{,}87 \\ 0{,}09 \\ 0{,}32 \end{bmatrix}^T \cdot 0{,}813 \right) \cdot \psi'(0{,}67) = \begin{bmatrix} 0{,}87 \\ 0{,}09 \\ 0{,}32 \end{bmatrix} \cdot 0{,}813 =$$

$$a_1^0[x] = \sigma(z_1^0) = \mathrm{ReLU}(1) = 1$$

$$\frac{\partial C}{\partial w_1^1} = a_1^0[x]\, \delta_1^1 = 1 \cdot \begin{bmatrix} 0{,}707 \\ 0{,}073 \\ 0{,}260 \end{bmatrix} = \begin{bmatrix} \boxed{0{,}707} \\ 0{,}073 \\ 0{,}260 \end{bmatrix} \rightarrow \frac{\partial C}{\partial w_{11}^1}$$

(ii) $$\frac{\partial C}{\partial w_{32}^1} = a_2^0[x]\, \delta_2^1 \qquad j = 3 \quad k = 2 \quad l = 1$$

$$\delta_j^l = \delta_3^1 = (W^{(2)T} \cdot \delta^2)_3 \; \psi'(z_3^1) = \underbrace{\begin{bmatrix} 0{,}87 \\ 0{,}09 \\ 0{,}32 \end{bmatrix}}_{W^{(2)}} \cdot \underbrace{0{,}813}_{\delta^2} \cdot \psi'(\underbrace{1{,}66}_{z_3^1}) = \begin{bmatrix} 0{,}87 \\ 0{,}09 \\ 0{,}32 \end{bmatrix} \cdot 0{,}813 \cdot 1 = \begin{bmatrix} 0{,}707 \\ 0{,}073 \\ 0{,}260 \end{bmatrix}$$

$$\frac{\partial C}{\partial w_2} = a_3^0[x]\, \delta_3^1 = \sigma(z_2^0) \cdot \begin{bmatrix} 0{,}707 \\ 0{,}073 \\ 0{,}260 \end{bmatrix} = \max\{0, 1\} \begin{bmatrix} 0{,}707 \\ 0{,}073 \\ 0{,}260 \end{bmatrix} = \begin{bmatrix} 0{,}707 \\ 0{,}073 \\ \boxed{0{,}260} \end{bmatrix} \rightarrow \frac{\partial C}{\partial w_{32}^1}$$

(iii) $$\frac{\partial C}{\partial b_2^1} = \delta_2^1 = 0{,}07$$

③ The architecture of the neural network:

Layers (depth) & neurons (width):
- 1 input layer with 4 neurons
- 2 hidden layers with 100 neurons each
- 1 output layer with 3 neurons

Activation function:
- ReLU is used after both hidden layers $(\max\{0, x\})$
- No activation function used in the output layer*

Cost function:
- Cross Entropy: $$C(y, f_\theta(x)) = -\log\left(\frac{e^{f_\theta(x)_y}}{\sum_i e^{f(x)_i}}\right)$$

* Although no activation function is explicitly applied to the output layer, PyTorch's CrossEntropyLoss() implicitly applies softmax to the raw outputs.