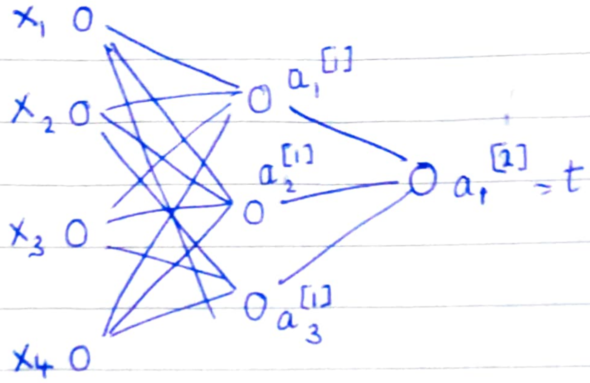
Given network :



Loss function : $\mathcal{L}^{(i)}(t^{(i)}, y^{(i)}) = -y^{(i)} \log t^{(i)} - (1-y^{(i)}) \log(1-t^{(i)})$

Cost : $J(W, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}^{(i)}(t^{(i)}, y^{(i)})$

Now, to generate predictions, apply forward propogation :
For instance i, we have :

$$z^{[1](i)} = W^{[1]} x^{(i)} + b^{[1]} \quad -①$$

$$a^{[1](i)} = g^{[1]}(z^{[1](i)}), \text{ where } g^{[1]} = ReLU \quad -②$$

$$= \begin{cases} z_j^{[1](i)}, & z_j^{[1]} > 0 \quad \forall j \\ 0, & \text{otherwise} \end{cases}$$

$$z^{[2](i)} = W^{[2]} a^{[1](i)} + b^{[2]} \quad -③$$

$$a^{[2](i)} = t^{(i)} = g^{[2]}(z^{[2](i)}), \text{ where } g^{[2]} = \sigma \quad -④$$

$$= \frac{1}{1 + e^{-z^{[2](i)}}}$$

$$\mathcal{L}^{(i)}(t^{(i)}, y^{(i)}) = -y^{(i)} \log t^{(i)} - (1-y^{(i)}) \log(1-t^{(i)})$$

Vectorizing across all instances, we have :

$$Z^{[1]} = W^{[1]} X + b^{[1]} \quad -⑤$$

$$A^{[1]} = ReLU(Z^{[1]}) \quad -⑥$$

$$Z^{[2]} = W^{[2]} A^{[1]} + b^{[2]} \quad -⑦$$

$$A^{[2]} = \sigma(Z^{[2]}) = T \quad -⑧$$

$$J(W, b) = (-Y \log^T(T) - (1-Y) \log^T(1-T))/m.$$

Now, parameters are updated via gradient descent:

$$W^{[l]} = W^{[l]} - \frac{\partial J}{\partial W^{[l]}} \cdot \alpha$$

$$b^{[l]} = b^{[l]} - \frac{\partial J}{\partial b^{[l]}} \cdot \alpha$$

To compute derivatives, use backpropogation:

For layer 2, we have; for each instance i:

$$W_{jk}^{[2]} = W_{jk}^{[2]} - \frac{\partial \mathcal{L}^{(i)}}{\partial W_{jk}^{[2]}} \cdot \alpha \Rightarrow W_k^{[2]} = W_k^{[2]} - \frac{\partial \mathcal{L}^{(i)}}{\partial W_k^{[2]}} \alpha$$

$$b_j^{[2]} = b_j^{[2]} - \frac{\partial \mathcal{L}^{(i)}}{\partial b_j^{[2]}} \cdot \alpha \Rightarrow b^{[2]} = b^{[2]} - \frac{\partial \mathcal{L}^{(i)}}{\partial b_j} \alpha$$

Now, $\frac{\partial \mathcal{L}^{(i)}}{\partial W_k^{[2]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial t^{(i)}} \frac{\partial t^{(i)}}{\partial z^{[2](i)}} \frac{\partial z^{[2](i)}}{\partial W_k^{[2]}}$

$$= \frac{\cancel{+}t^{(i)} - y^{(i)}}{\cancel{t^{(i)}(1-t^{(i)})}} \cancel{t^{(i)}(1-t^{(i)})} \, a_k^{[1](i)}$$

$$= (t^{(i)} - y^{(i)}) a_k^{[1](i)}$$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial b^{[2]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial t^{(i)}} \frac{\partial t^{(i)}}{\partial z^{[1](i)}} \frac{\partial z^{[2](i)}}{\partial b^{[2]}}$$

$$= (t^{(i)} - y^{(i)}) \cdot 1$$

Vectorizing across instances & k:

$$\frac{\partial J}{\partial W_k^{[2]}} = \frac{1}{m}\begin{bmatrix} \frac{\partial \mathcal{L}^{(1)}}{\partial W_k^{[2]}} & \frac{\partial \mathcal{L}^{(2)}}{\partial W_k^{[2]}} & \cdots & \frac{\partial \mathcal{L}^{(3)}}{\partial W_k^{[2]}} \end{bmatrix}$$

$$= \frac{1}{m}\begin{bmatrix} t^{(1)} - y^{(1)} & t^{(2)} - y^{(2)} & \cdots & t^{(m)} - y^{(m)} \end{bmatrix} \begin{bmatrix} a_k^{[1](1)} \\ a_k^{[1](2)} \\ a_k^{[1](3)} \\ \vdots \\ a_k^{[1](m)} \end{bmatrix}$$

$$= \frac{1}{m}(T - Y) a_k^{[1]T}$$

$$\frac{\partial J}{\partial W^{[2]}} = \frac{1}{m}(T - Y) A^{[1]T}$$

$$\frac{\partial J}{\partial b^{[2]}} = \frac{1}{m}\sum_{i=1}^{m} t^{(i)} - y^{(i)} = \text{np.sum}(T-Y,\ \text{axis}=1,\ \text{keepdims}=\text{True})$$

Now, for layer 1, for each instance i we have :

$$\frac{\partial \mathcal{L}^{(i)}}{\partial W_{jk}^{[1]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial t^{(i)}} \frac{\partial t^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial a_j^{[1](i)}} \frac{\partial a_j^{[1](i)}}{\partial z_j^{[1](i)}} \frac{\partial z^{[1](i)}}{\partial W_{jk}^{[1]}}$$

$$= \frac{t^{(i)} - y^{(i)}}{\cancel{t^{(i)}(1-t^{(i)})}} \cancel{t^{(i)}(1-t^{(i)})}\, W_j^{[2]}\, ReLU'(z_j^{[1](i)})\, x_k^{(i)}$$

$$= (t^{(i)} - y^{(i)})\, W_j^{[2]}\, ReLU'(z_j^{[1](i)})\, x_k^{(i)}$$

where $ReLU'(x) = \begin{cases} 1 & x > 0 \\ 0/0 & x = 0 \\ 0 & x < 0 \end{cases}$

$$\frac{\partial \mathcal{L}^{(i)}}{\partial b_j^{[1]}} = \frac{\partial \mathcal{L}^{(i)}}{\partial t^{(i)}} \frac{\partial t^{(i)}}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial a_j^{[1](i)}} \frac{\partial a_j^{[1](i)}}{\partial z_j^{[1](i)}} \frac{\partial z_j^{[1](i)}}{\partial b_j^{[1]}}$$

$$= (t^{(i)} - y^{(i)})\, W_j^{[2]}\, ReLU'(z_j^{[1](i)}) \cdot \cancel{x}$$

Vectorizing across instances :

$$\frac{\partial J}{\partial W_{jk}^{[1]}} = \frac{1}{m}\begin{bmatrix} \frac{\partial \mathcal{L}^{(1)}}{\partial W_{jk}^{[1]}} & \frac{\partial \mathcal{L}^{(2)}}{\partial W_{jk}^{[1]}} & \cdots & \frac{\partial \mathcal{L}^{(m)}}{\partial W_{jk}^{[1]}} \end{bmatrix}$$

$$= \frac{1}{m}\begin{bmatrix} t^{(1)} - y^{(1)}\, W_j^{[2]}\, ReLU'(z_j^{[1](1)}) \\ t^{(2)} - y^{(2)}\, W_j^{[2]}\, ReLU'(z_j^{[1](2)}) \\ \vdots \\ \end{bmatrix}^T \begin{bmatrix} x_k^{(1)} \\ x_k^{(2)} \\ \vdots \\ x_k^{(m)} \end{bmatrix}$$

$$= \cancel{\frac{1}{m}(T-Y)(W_j^{[2]} \odot ReLU'(z_j^{[1]}))\, X_k^T}$$

$$\cancel{\frac{\partial J}{\partial W^{[1]}} = \frac{1}{m}(T-Y)(W^{[2]} \odot ReLU'(z^{[1]}))\, X^T}$$

$$= \frac{W_j^{[2]T}}{m}(T-Y) \odot ReLU'(z_j^{[1]})\, x_k^T$$

$$\frac{\partial J}{\partial W^{[1]}} = W^{[2]^T}(T-Y) \odot ReLU'(Z^{[1]})\, X^T$$

$$\frac{\partial J}{\partial b_j^{[1]}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial \mathcal{L}^{(i)}}{\partial b_j^{(i)}}$$

$$= np.sum(W_j^{[2]}(T-Y) \odot ReLU'(z_j^{[1]}), axis=1, keepdims=T$$

$$\frac{\partial J}{\partial b_j^{[1]}} = \frac{1}{m} np.sum(W^{[2]^T}(T-Y) \odot ReLU'(z^{[1]}), axis=1, keepdims=T)$$

Final update equations for gradient descent:

$$W^{[2]} = W^{[2]} - \alpha (T-Y) A^{[1]^T}/m$$

$$b^{[2]} = b^{[2]} - \alpha \sum_{axis=1}(T-Y) / m$$

$$W^{[1]} = W^{[1]} - \frac{\alpha}{m} W^{[2]^T}(T-Y) \odot ReLU'(Z^{[1]}) X^T$$

$$b^{[1]} = b^{[1]} - \frac{\alpha}{m} \sum_{axis=1} (W^{[2]^T}(T-Y) \odot ReLU'(Z^{[1]}))$$