1a) $4 \times 6 + 6 \times 2$

4 input nodes; 6 nodes in 1st hidden layer; 6 nodes; 2 nodes in 2nd hidden layer

$+ \; 2 \times 1 = 24 + 12 + 2 = 38$

2 nodes; 1 output node

If we add 1 more node to the 1st hidden layer, our calculation would be

$4 \cdot 7 + 7 \cdot 2 + 2 \cdot 1 = 28 + 14 + 2 = 44$

7 nodes

now 7 nodes in 1st hidden layer

Non-linear activation functions allow for neural networks to learn complex patterns in the data beyond simple linear regression.

If our function only contained linear activation functions, it would be the same as doing simple linear regression. By implementing non-linearity we can have the neural network model much more complex data and relationships.

b)

$$W^{(1)^T} x = \begin{bmatrix} 1 & 0 & 3 & -5 \\ -2 & 2 & -1 & 0 \\ 5 & 0 & -2 & 3 \\ 1 & -2 & -4 & 4 \\ -6 & 2 & 2 & -1 \\ 2 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 12 \\ 6 \\ 0.9 \end{bmatrix} = \begin{bmatrix} 14.5 \\ 16 \\ -4.3 \\ -43.4 \\ 29.1 \\ 32 \end{bmatrix} \quad \text{ReLU}\left(\begin{bmatrix} 14.5 \\ 16 \\ -4.3 \\ -43.4 \\ 29.1 \\ 32 \end{bmatrix} \right) = \begin{bmatrix} 14.5 \\ 16 \\ 0 \\ 0 \\ 29.1 \\ 32 \end{bmatrix}$$

c)

$$W^{(2)T} h_1 = \begin{bmatrix} -2 & 0 & 1 & -4 & 1 & 0 \\ 1 & 3 & -1 & 6 & 3 & -1 \end{bmatrix} \begin{bmatrix} 14.5 \\ 16 \\ 0 \\ 0 \\ 29.1 \\ 32 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 117.8 \end{bmatrix}$$

$$\sigma\left(\begin{bmatrix} 0.1 \\ 117.8 \end{bmatrix}\right) = \begin{bmatrix} \frac{1}{1+e^{-0.1}} \\ \frac{1}{1+e^{-117.8}} \end{bmatrix} \approx \begin{bmatrix} 0.525 \\ 1 \end{bmatrix}$$

d)

$$W^{(3)T} h_2 = \begin{bmatrix} -24 & 90 \end{bmatrix} \begin{bmatrix} 0.525 \\ 1 \end{bmatrix} = 77.4$$

$$\text{Loss} = (86 - 77.4)^2 = 73.96$$

e)

$$\frac{\partial L}{\partial W^{(3)}} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial W^{(3)}} = 2(W^{(3)T} h_2 - y_i) h^{(2)}$$

$$\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial W^{(2)}}$$

$$= 2(W^{(3)T} h^{(2)} - y_i) h^{(2)} \cdot h^{(1)} \sigma(W^{(2)T} h^{(1)})(1 - \sigma(W^{(2)T} h^{(1)}))$$