Neural Network Backpropagation

1-4-1 Architecture: Complete Worked Example

Handwritten Notes Conversion

1 Network Architecture

Network Structure: 1 input $\rightarrow$ 4 hidden (ReLU) $\rightarrow$ 1 output (Linear)

- **Input**: $x = 2.0$
- **Target**: $y = 5.0$

2 Network Parameters

2.1 Input to Hidden Layer Weights

Weight matrix $\mathbf{W}_1 \in \mathbb{R}^{4\times 1}$ (4 hidden neurons, 1 input):

$$\mathbf{W}_1 = \begin{pmatrix} 1.0 \\ -3.0 \\ 0.5 \\ 4.0 \end{pmatrix} \tag{1}$$

Hidden layer bias $\mathbf{b}_1 \in \mathbb{R}^{4\times 1}$:

$$\mathbf{b}_1 = \begin{pmatrix} 5.0 \\ -2.0 \\ 0.0 \\ 2.0 \end{pmatrix} \tag{2}$$

2.2 Hidden to Output Layer Weights

Weight matrix $\mathbf{W}_2 \in \mathbb{R}^{1\times 4}$ (1 output, 4 hidden neurons):

$$\mathbf{W}_2 = \begin{pmatrix} -2.4 & 1.8 & -1.0 & -1.5 \end{pmatrix} \tag{3}$$

Output layer bias $b_2 \in \mathbb{R}$:

$$b_2 = -1.0 \tag{4}$$

3 Forward Pass

3.1 Computing Hidden Layer Activations

For each hidden neuron $i = 1, 2, 3, 4$:

$$a_i = \text{ReLU}(W_{1i} \cdot x + b_{1i}) = \text{ReLU}(W_{1i} \cdot 2.0 + b_{1i}) \tag{5}$$

Hidden neuron 1:

$$a_1 = \text{ReLU}(1.0 \times 2.0 + 5.0) = \text{ReLU}(7.0) = 7.0 \tag{6}$$

Hidden neuron 2:

$$a_2 = \text{ReLU}(-3.0 \times 2.0 + (-2.0)) = \text{ReLU}(-8.0) = 0.0 \tag{7}$$

Hidden neuron 3:

$$a_3 = \text{ReLU}(0.5 \times 2.0 + 0.0) = \text{ReLU}(1.0) = 1.0 \tag{8}$$

Hidden neuron 4:

$$a_4 = \text{ReLU}(4.0 \times 2.0 + 2.0) = \text{ReLU}(10.0) = 10.0 \tag{9}$$

Hidden layer output:

$$\mathbf{a} = \begin{pmatrix} 7.0 \\ 0.0 \\ 1.0 \\ 10.0 \end{pmatrix} \tag{10}$$

3.2 Computing Output

Pre-activation:

$$\begin{aligned} z &= \mathbf{W}_2\mathbf{a} + b_2 && (11)\\ &= -2.4 \times 7.0 + 1.8 \times 0.0 + (-1.0) \times 1.0 + (-1.5) \times 10.0 + (-1.0) && (12)\\ &= -16.8 + 0.0 - 1.0 - 15.0 - 1.0 && (13)\\ &= -33.8 && (14) \end{aligned}$$

Final output (linear activation):

$$\hat{y} = z = -33.8 \tag{15}$$

4 Loss Computation

Using Mean Squared Error:

$$\begin{aligned} L &= \frac{1}{2}(\hat{y} - y)^2 && (16)\\ &= \frac{1}{2}(-33.8 - 5.0)^2 && (17)\\ &= \frac{1}{2}(-38.8)^2 && (18)\\ &= \frac{1}{2}(1505.44) && (19)\\ &= 752.72 && (20) \end{aligned}$$

5 Backward Pass

5.1 Output Layer Gradients

Error at output:

$$\delta_{\text{out}} = \frac{\partial L}{\partial z} = \hat{y} - y = -33.8 - 5.0 = -38.8 \tag{21}$$

Gradients for output weights:

$$\frac{\partial L}{\partial W_{21}} = \delta_{\text{out}} \times a_1 = -38.8 \times 7.0 = -271.6 \tag{22}$$

$$\frac{\partial L}{\partial W_{22}} = \delta_{\text{out}} \times a_2 = -38.8 \times 0.0 = 0.0 \tag{23}$$

$$\frac{\partial L}{\partial W_{23}} = \delta_{\text{out}} \times a_3 = -38.8 \times 1.0 = -38.8 \tag{24}$$

$$\frac{\partial L}{\partial W_{24}} = \delta_{\text{out}} \times a_4 = -38.8 \times 10.0 = -388.0 \tag{25}$$

Gradient for output bias:

$$\frac{\partial L}{\partial b_2} = \delta_{\text{out}} = -38.8 \tag{26}$$

5.2 Hidden Layer Gradients

Computing hidden layer deltas:

For $i = 1, 2, 3, 4$:

$$\delta_i = \frac{\partial L}{\partial a_i} \times \frac{\partial a_i}{\partial z_i} = (W_{2i} \times \delta_{\text{out}}) \times \text{ReLU}'(z_i) \tag{27}$$

Where $\text{ReLU}'(z) = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{if } z \leq 0 \end{cases}$

Hidden neuron 1 ($z_1 = 7.0 > 0$, so ReLU' = 1):

$$\delta_1 = W_{21} \times \delta_{\text{out}} \times 1 = -2.4 \times (-38.8) \times 1 = 93.12 \tag{28}$$

Hidden neuron 2 ($z_2 = -8.0 \leq 0$, so ReLU' = 0):

$$\delta_2 = W_{22} \times \delta_{\text{out}} \times 0 = 1.8 \times (-38.8) \times 0 = 0.0 \tag{29}$$

Hidden neuron 3 ($z_3 = 1.0 > 0$, so ReLU' = 1):

$$\delta_3 = W_{23} \times \delta_{\text{out}} \times 1 = -1.0 \times (-38.8) \times 1 = 38.8 \tag{30}$$

Hidden neuron 4 ($z_4 = 10.0 > 0$, so ReLU' = 1):

$$\delta_4 = W_{24} \times \delta_{\text{out}} \times 1 = -1.5 \times (-38.8) \times 1 = 58.2 \tag{31}$$

5.3 Input Layer Weight Gradients

Gradients for input-to-hidden weights:

$$\frac{\partial L}{\partial W_{11}} = \delta_1 \times x = 93.12 \times 2.0 = 186.24 \tag{32}$$

$$\frac{\partial L}{\partial W_{12}} = \delta_2 \times x = 0.0 \times 2.0 = 0.0 \tag{33}$$

$$\frac{\partial L}{\partial W_{13}} = \delta_3 \times x = 38.8 \times 2.0 = 77.6 \tag{34}$$

$$\frac{\partial L}{\partial W_{14}} = \delta_4 \times x = 58.2 \times 2.0 = 116.4 \tag{35}$$

Gradients for hidden layer biases:

$$\frac{\partial L}{\partial b_{11}} = \delta_1 = 93.12 \tag{36}$$

$$\frac{\partial L}{\partial b_{12}} = \delta_2 = 0.0 \tag{37}$$

$$\frac{\partial L}{\partial b_{13}} = \delta_3 = 38.8 \tag{38}$$

$$\frac{\partial L}{\partial b_{14}} = \delta_4 = 58.2 \tag{39}$$

6 Summary of All Gradients

6.1 Output Layer Gradients

$$\frac{\partial L}{\partial \mathbf{W}_2} = \begin{pmatrix} -271.6 & 0.0 & -38.8 & -388.0 \end{pmatrix} \tag{40}$$

$$\frac{\partial L}{\partial b_2} = -38.8 \tag{41}$$

6.2 Hidden Layer Gradients

$$\frac{\partial L}{\partial \mathbf{W}_1} = \begin{pmatrix} 186.24 \\ 0.0 \\ 77.6 \\ 116.4 \end{pmatrix} \tag{42}$$

$$\frac{\partial L}{\partial \mathbf{b}_1} = \begin{pmatrix} 93.12 \\ 0.0 \\ 38.8 \\ 58.2 \end{pmatrix} \tag{43}$$

7 Parameter Updates

Using gradient descent with learning rate α:

7.1 Output Layer Updates

$$W_{21}^{\text{new}} = W_{21}^{\text{old}} - \alpha \times (-271.6) = -2.4 + 271.6\alpha \tag{44}$$

$$W_{22}^{\text{new}} = W_{22}^{\text{old}} - \alpha \times 0.0 = 1.8 \tag{45}$$

$$W_{23}^{\text{new}} = W_{23}^{\text{old}} - \alpha \times (-38.8) = -1.0 + 38.8\alpha \tag{46}$$

$$W_{24}^{\text{new}} = W_{24}^{\text{old}} - \alpha \times (-388.0) = -1.5 + 388.0\alpha \tag{47}$$

$$b_2^{\text{new}} = b_2^{\text{old}} - \alpha \times (-38.8) = -1.0 + 38.8\alpha \tag{48}$$

7.2 Hidden Layer Updates

$$W_{11}^{\text{new}} = W_{11}^{\text{old}} - \alpha \times 186.24 = 1.0 - 186.24\alpha \tag{49}$$

$$W_{12}^{\text{new}} = W_{12}^{\text{old}} - \alpha \times 0.0 = -3.0 \tag{50}$$

$$W_{13}^{\text{new}} = W_{13}^{\text{old}} - \alpha \times 77.6 = 0.5 - 77.6\alpha \tag{51}$$

$$W_{14}^{\text{new}} = W_{14}^{\text{old}} - \alpha \times 116.4 = 4.0 - 116.4\alpha \tag{52}$$

$$b_{11}^{\text{new}} = b_{11}^{\text{old}} - \alpha \times 93.12 = 5.0 - 93.12\alpha \tag{53}$$

$$b_{12}^{\text{new}} = b_{12}^{\text{old}} - \alpha \times 0.0 = -2.0 \tag{54}$$

$$b_{13}^{\text{new}} = b_{13}^{\text{old}} - \alpha \times 38.8 = 0.0 - 38.8\alpha \tag{55}$$

$$b_{14}^{\text{new}} = b_{14}^{\text{old}} - \alpha \times 58.2 = 2.0 - 58.2\alpha \tag{56}$$

8 Key Observations

- **ReLU Effect**: Hidden neuron 2 had negative pre-activation (-8.0), so its output was 0 and it contributes no gradient updates.
- **Large Error**: The network prediction (-33.8) is very far from target (5.0), resulting in large gradients.
- **Gradient Flow**: Only neurons with positive pre-activations (neurons 1, 3, 4) receive gradient updates due to ReLU.
- **Learning Direction**: Most gradients are large, indicating significant parameter changes needed in the next iteration.