Neural Network: Diagram and Dry Run Calculations

1 Neural Network Diagram

The neural network consists of an input layer with 4 nodes, a hidden layer with 3 nodes (using ReLU activation), and an output layer with 1 node (summing the hidden layer outputs). The weights $W \in R^{3\times 4}$ connect the input to the hidden layer, and biases $b \in R^3$ are added to the hidden layer pre-activations. The output y is the sum of the hidden layer activations, and the loss is $L = y^2$.

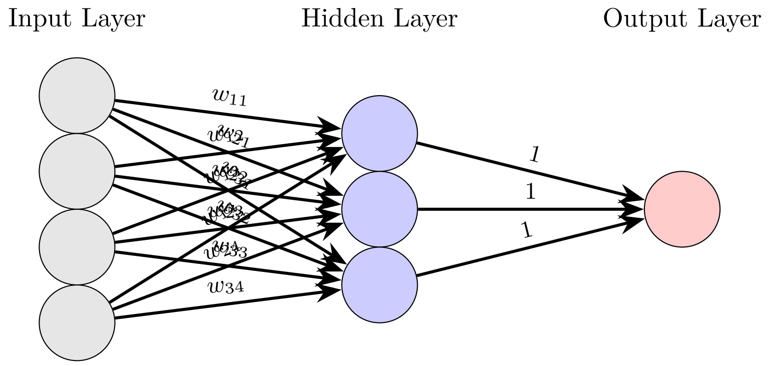

Figure 1: Neural network architecture: 4 input nodes, 3 hidden nodes with ReLU activation, and 1 output node.

2 Equations

The following equations govern the neural network's forward and backward passes.

2.1 Forward Pass

- Linear transformation:
$$z = Wx + b$$
where $W \in R^{3\times 4}$, $x \in R^4$, $b \in R^3$.
- Activation function (ReLU):
$$a = \text{ReLU}(z) = \max(0, z)$$
- Output (sum of activations):
$$y = \sum_{i=1}^{3} a_i$$
- Loss function:
$$L = y^2$$

2.2 Backward Pass

- Gradient of loss w.r.t. output:
$$\frac{dL}{dy} = 2y$$
- Gradient of output w.r.t. activations:
$$\frac{dy}{da} = \begin{bmatrix}1\\1\\1\end{bmatrix} \in R^3$$
- Gradient of loss w.r.t. activations:
$$\frac{dL}{da} = \frac{dL}{dy} \cdot \frac{dy}{da}$$
- Gradient of activations w.r.t. pre-activations (ReLU derivative):
$$\frac{da}{dz} = \begin{cases}1 & \text{if } z_i > 0\\ 0 & \text{otherwise}\end{cases}$$
- Gradient of loss w.r.t. pre-activations:
$$\frac{dL}{dz} = \frac{dL}{da} \cdot \frac{da}{dz}$$
- Gradient of loss w.r.t. weights:
$$\frac{dL}{dW} = \frac{dL}{dz} \cdot x^T$$
- Gradient of loss w.r.t. biases:
$$\frac{dL}{db} = \frac{dL}{dz}$$

2.3 Parameter Updates

- Update weights:
$$W \leftarrow W - \eta \cdot \frac{dL}{dW}$$
- Update biases:
$$b \leftarrow b - \eta \cdot \frac{dL}{db}$$
where $\eta = 0.001$ is the learning rate.

3 Dry Run Calculations

We perform a dry run for two iterations of gradient descent, using the following initial parameters:

- Inputs: $x = \begin{bmatrix}1 & 2 & 3 & 4\end{bmatrix}^T$
- Weights:
$$W = \begin{bmatrix}0.1 & 0.2 & 0.3 & 0.4\\ 0.5 & 0.6 & 0.7 & 0.8\\ 0.9 & 1.0 & 1.1 & 1.2\end{bmatrix}$$
- Biases: $b = \begin{bmatrix}0.1 & 0.2 & 0.3\end{bmatrix}^T$
- Learning rate: $\eta = 0.001$

3.1 Iteration 1

3.1.1 Forward Pass

Step 1: Compute $z = Wx + b$:

$$Wx = \begin{bmatrix} 0.1 \cdot 1 + 0.2 \cdot 2 + 0.3 \cdot 3 + 0.4 \cdot 4 \\ 0.5 \cdot 1 + 0.6 \cdot 2 + 0.7 \cdot 3 + 0.8 \cdot 4 \\ 0.9 \cdot 1 + 1.0 \cdot 2 + 1.1 \cdot 3 + 1.2 \cdot 4 \end{bmatrix} = \begin{bmatrix} 0.1 + 0.4 + 0.9 + 1.6 \\ 0.5 + 1.2 + 2.1 + 3.2 \\ 0.9 + 2.0 + 3.3 + 4.8 \end{bmatrix} = \begin{bmatrix} 3.0 \\ 7.0 \\ 11.0 \end{bmatrix}$$

$$z = Wx + b = \begin{bmatrix} 3.0 \\ 7.0 \\ 11.0 \end{bmatrix} + \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix} = \begin{bmatrix} 3.1 \\ 7.2 \\ 11.3 \end{bmatrix}$$

Step 2: Compute $a = \textbf{ReLU}(z)$: Since $z_i > 0$ for all i, $\text{ReLU}(z) = z$:

$$a = \begin{bmatrix} 3.1 \\ 7.2 \\ 11.3 \end{bmatrix}$$

Step 3: Compute $y = \sum a_i$:

$$y = 3.1 + 7.2 + 11.3 = 21.6$$

Step 4: Compute loss $L = y^2$:

$$L = 21.6^2 = 466.56$$

3.1.2 Backward Pass

Step 5: Gradient of loss w.r.t. y:

$$\frac{dL}{dy} = 2y = 2 \cdot 21.6 = 43.2$$

Step 6: Gradient of y w.r.t. a:

$$\frac{dy}{da} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Step 7: Gradient of loss w.r.t. a:

$$\frac{dL}{da} = \frac{dL}{dy} \cdot \frac{dy}{da} = 43.2 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 43.2 \\ 43.2 \\ 43.2 \end{bmatrix}$$

Step 8: Gradient of a w.r.t. z: Since $z = \begin{bmatrix} 3.1 & 7.2 & 11.3 \end{bmatrix}^T$, all $z_i > 0$:

$$\frac{da}{dz} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Step 9: Gradient of loss w.r.t. z:

$$\frac{dL}{dz} = \frac{dL}{da} \cdot \frac{da}{dz} = \begin{bmatrix} 43.2 \\ 43.2 \\ 43.2 \end{bmatrix}$$

Step 10: Gradient of loss w.r.t. weights:

$$\frac{dL}{dW} = \frac{dL}{dz} \cdot x^T = \begin{bmatrix} 43.2 \\ 43.2 \\ 43.2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 43.2 & 86.4 & 129.6 & 172.8 \\ 43.2 & 86.4 & 129.6 & 172.8 \\ 43.2 & 86.4 & 129.6 & 172.8 \end{bmatrix}$$

Step 11: Gradient of loss w.r.t. biases:

$$\frac{dL}{db} = \frac{dL}{dz} = \begin{bmatrix} 43.2 \\ 43.2 \\ 43.2 \end{bmatrix}$$

3.1.3 Parameter Updates

Step 12: Update weights:

$$\eta \cdot \frac{dL}{dW} = 0.001 \cdot \begin{bmatrix} 43.2 & 86.4 & 129.6 & 172.8 \\ 43.2 & 86.4 & 129.6 & 172.8 \\ 43.2 & 86.4 & 129.6 & 172.8 \end{bmatrix} = \begin{bmatrix} 0.0432 & 0.0864 & 0.1296 & 0.1728 \\ 0.0432 & 0.0864 & 0.1296 & 0.1728 \\ 0.0432 & 0.0864 & 0.1296 & 0.1728 \end{bmatrix}$$

$$W = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.5 & 0.6 & 0.7 & 0.8 \\ 0.9 & 1.0 & 1.1 & 1.2 \end{bmatrix} - \begin{bmatrix} 0.0432 & 0.0864 & 0.1296 & 0.1728 \\ 0.0432 & 0.0864 & 0.1296 & 0.1728 \\ 0.0432 & 0.0864 & 0.1296 & 0.1728 \end{bmatrix} = \begin{bmatrix} 0.0568 & 0.1136 & 0.1704 & 0.2272 \\ 0.4568 & 0.5136 & 0.5704 & 0.6272 \\ 0.8568 & 0.9136 & 0.9704 & 1.0272 \end{bmatrix}$$

Step 13: Update biases:

$$\eta \cdot \frac{dL}{db} = 0.001 \cdot \begin{bmatrix} 43.2 \\ 43.2 \\ 43.2 \end{bmatrix} = \begin{bmatrix} 0.0432 \\ 0.0432 \\ 0.0432 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix} - \begin{bmatrix} 0.0432 \\ 0.0432 \\ 0.0432 \end{bmatrix} = \begin{bmatrix} 0.0568 \\ 0.1568 \\ 0.2568 \end{bmatrix}$$

Step 14: Loss:

$$\text{Loss} = 466.56$$

3.2 Iteration 2

3.2.1 Forward Pass

Step 1: Compute $z = Wx + b$:

$$Wx = \begin{bmatrix} 0.0568 \cdot 1 + 0.1136 \cdot 2 + 0.1704 \cdot 3 + 0.2272 \cdot 4 \\ 0.4568 \cdot 1 + 0.5136 \cdot 2 + 0.5704 \cdot 3 + 0.6272 \cdot 4 \\ 0.8568 \cdot 1 + 0.9136 \cdot 2 + 0.9704 \cdot 3 + 1.0272 \cdot 4 \end{bmatrix} = \begin{bmatrix} 0.0568 + 0.2272 + 0.5112 + 0.9088 \\ 0.4568 + 1.0272 + 1.7112 + 2.5088 \\ 0.8568 + 1.8272 + 2.9112 + 4.1088 \end{bmatrix} = \begin{bmatrix} 1.704 \\ 5.704 \\ 9.704 \end{bmatrix}$$

$$z = Wx + b = \begin{bmatrix} 1.704 \\ 5.704 \\ 9.704 \end{bmatrix} + \begin{bmatrix} 0.0568 \\ 0.1568 \\ 0.2568 \end{bmatrix} = \begin{bmatrix} 1.7608 \\ 5.8608 \\ 9.9608 \end{bmatrix}$$

Step 2: Compute $a =$ ReLU(z): Since $z_i > 0$ for all i, $\text{ReLU}(z) = z$:

$$a = \begin{bmatrix} 1.7608 \\ 5.8608 \\ 9.9608 \end{bmatrix}$$

Step 3: Compute $y = \sum a_i$:

$$y = 1.7608 + 5.8608 + 9.9608 = 17.5824$$

Step 4: Compute loss $L = y^2$:

$$L = 17.5824^2 = 309.139$$

3.2.2 Backward Pass

Step 5: Gradient of loss w.r.t. y:

$$\frac{dL}{dy} = 2 \cdot 17.5824 = 35.1648$$

Step 6: Gradient of y w.r.t. a:

$$\frac{dy}{da} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Step 7: Gradient of loss w.r.t. a:

$$\frac{dL}{da} = 35.1648 \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 35.1648 \\ 35.1648 \\ 35.1648 \end{bmatrix}$$

Step 8: Gradient of a **w.r.t.** z: Since $z = \begin{bmatrix} 1.7608 & 5.8608 & 9.9608 \end{bmatrix}^T$, all $z_i > 0$:

$$\frac{da}{dz} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Step 9: Gradient of loss w.r.t. z:

$$\frac{dL}{dz} = \begin{bmatrix} 35.1648 \\ 35.1648 \\ 35.1648 \end{bmatrix}$$

Step 10: Gradient of loss w.r.t. weights:

$$\frac{dL}{dW} = \begin{bmatrix} 35.1648 \\ 35.1648 \\ 35.1648 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 35.1648 & 70.3296 & 105.4944 & 140.6592 \\ 35.1648 & 70.3296 & 105.4944 & 140.6592 \\ 35.1648 & 70.3296 & 105.4944 & 140.6592 \end{bmatrix}$$

Step 11: Gradient of loss w.r.t. biases:

$$\frac{dL}{db} = \begin{bmatrix} 35.1648 \\ 35.1648 \\ 35.1648 \end{bmatrix}$$

3.2.3 Parameter Updates

Step 12: Update weights:

$$\eta \cdot \frac{dL}{dW} = 0.001 \cdot \begin{bmatrix} 35.1648 & 70.3296 & 105.4944 & 140.6592 \\ 35.1648 & 70.3296 & 105.4944 & 140.6592 \\ 35.1648 & 70.3296 & 105.4944 & 140.6592 \end{bmatrix} = \begin{bmatrix} 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \\ 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \\ 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \end{bmatrix}$$

$$W = \begin{bmatrix} 0.0568 & 0.1136 & 0.1704 & 0.2272 \\ 0.4568 & 0.5136 & 0.5704 & 0.6272 \\ 0.8568 & 0.9136 & 0.9704 & 1.0272 \end{bmatrix} - \begin{bmatrix} 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \\ 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \\ 0.0351648 & 0.0703296 & 0.1054944 & 0.1406592 \end{bmatrix} = \begin{bmatrix} 0.0216352 & 0.0432704 \\ 0.4216352 & 0.4432704 \\ 0.8216352 & 0.8432704 \end{bmatrix}$$

Step 13: Update biases:

$$\eta \cdot \frac{dL}{db} = 0.001 \cdot \begin{bmatrix} 35.1648 \\ 35.1648 \\ 35.1648 \end{bmatrix} = \begin{bmatrix} 0.0351648 \\ 0.0351648 \\ 0.0351648 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.0568 \\ 0.1568 \\ 0.2568 \end{bmatrix} - \begin{bmatrix} 0.0351648 \\ 0.0351648 \\ 0.0351648 \end{bmatrix} = \begin{bmatrix} 0.0216352 \\ 0.1216352 \\ 0.2216352 \end{bmatrix}$$

Step 14: Loss:

$$\text{Loss} = 309.139$$

4 Final Parameters

After two iterations, the final parameters are:

$$W = \begin{bmatrix} 0.0216352 & 0.0432704 & 0.0649056 & 0.0865408 \\ 0.4216352 & 0.4432704 & 0.4649056 & 0.4865408 \\ 0.8216352 & 0.8432704 & 0.8649056 & 0.8865408 \end{bmatrix}$$

$$b = \begin{bmatrix} 0.0216352 \\ 0.1216352 \\ 0.2216352 \end{bmatrix}$$