Calculation of Gradients

Charitha
AI24BTECH11014

March 15, 2025

Backpropagation, an algorithm, used to train neural networks by adjusting the weights based on the gradient of the loss function. In this document, the gradients of the loss function with respect to the weights in a neural network are calculated. Gradient descent, an optimization technique, updates the weights by moving in the direction of the negative gradient of the loss function, ensuring it points towards the steepest descent and hence minimizing the error.

1 Gradient with respect to $W^{(2)}$

The mean squared loss function L is given by:

$$L = \frac{1}{2N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

Taking the derivative of L with respect to $W^{(2)}$:

$$\frac{\partial L}{\partial W^{(2)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial O_i} \cdot \frac{\partial O_i}{\partial W^{(2)}}.$$

$$\frac{\partial L}{\partial \hat{y}_i} = -(y_i - \hat{y}_i)$$

Since $\hat{y}_i = \sigma(O)$, we use the sigmoid derivative:

$$\sigma'(O) = \hat{y}_i(1 - \hat{y}_i) = \frac{\partial \hat{y}_i}{\partial O_i}$$

Thus, the error term for the output layer is:

$$\delta_O = -(y_i - \hat{y}_i) \cdot \sigma'(O).$$

Since, $O = ZW^{(2)}$, $\frac{\partial O_i}{\partial W^{(2)}} = Z$ The gradient is given by:

$$\nabla W^{(2)} = \frac{1}{N} Z^T \delta_O. \tag{1}$$

2 Gradient with respect to $W^{(1)}$

To compute the gradient for $W^{(1)}$, we propagate the error from the output layer to the hidden layer:

2.1 Breaking Down Each Derivative

Using the chain rule, we express the gradient as:

$$\frac{\partial L}{\partial W^{(1)}} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial O_i} \cdot \frac{\partial O_i}{\partial Z_i} \cdot \frac{\partial Z_i}{\partial H_i} \cdot \frac{\partial H_i}{\partial W^{(1)}}.$$

- The first two terms $\frac{\partial L}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial O_i}$ are already computed in δ_O.
- The term $\frac{\partial O_i}{\partial Z_i}$ is simply $W^{(2)T}$, since $O = ZW^{(2)}$.
- The term $\frac{\partial Z_i}{\partial H_i}$ is the derivative of the activation function, $\sigma'(H)$.
- Finally, since $H = XW^{(1)}$, the derivative $\frac{\partial H_i}{\partial W^{(1)}}$ simplifies to X.

Thus, we can write:

$$\delta_H = \left(\delta_O W^{(2)T}\right) \odot \sigma'(H).$$

And the gradient for the hidden layer weights is:

$$\nabla W^{(1)} = \frac{1}{N} X^T \delta_H. \tag{2}$$