Vanishing Gradient Descent

→ Backpropagation

Gradient Descent

80's & 90's

$$w_{new} = w_{old} - \eta \frac{\delta L}{\delta w_{old}}$$

- w_{new}, w_{old} ✓
- η: Learning Rate ≈ 0.001
- $\frac{\delta L}{\delta w_{old}}$: Gradient
- $\frac{\delta L}{\delta w_{old}}$: Really small

Learning gets stopped

$w_{new} \approx w_{old}$ → Vanishing gradient Descent

$$\text{Gradient} = \frac{\delta L}{\delta w_{old}} = \frac{w_{old} - w_{new}}{\eta}$$

Multiple hidden layers

Activation function → Sigmoid

derivative → 0 to 0.5

tanh

chain Rule

$$\frac{\delta L}{\delta w_{11}'} = \frac{\delta L}{\delta \hat{y}} * \frac{\delta \hat{y}}{\delta z} * \frac{\delta z}{\delta o_{11}} * \frac{\delta o_{11}}{\delta w_{11}'}$$

$0 < \frac{\delta L}{\delta w} < 1$

$$= 0.1 * 0.1 * 0.1 * 0.1$$

$$= 0.00001$$

→ Really small value

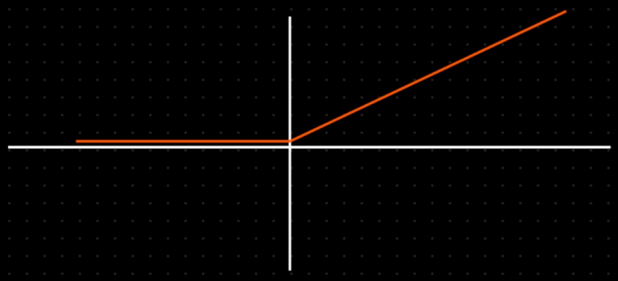
The vanishing gradient problem is a challenge that occurs during the training of artificial neural networks when gradients become very small and diminish as they are propagated back through the network.

In machine learning, the vanishing gradient problem is encountered when training neural networks with gradient-based learning methods and backpropagation. In such methods, during each training iteration, each neural network weight receives an update proportional to the partial derivative of the loss function with respect to the current weight.[1]

The problem is that as the network depth or sequence length increases, the gradient magnitude typically is expected to decrease (or grow uncontrollably), slowing the training process.[1] In the worst case, this may completely stop the neural network from further learning

* ReLU ⟶ avoid issue of vanishing gradient issue

$$\max(0, x_i)$$



Exploding Gradient Descent

$$\frac{\delta L}{\delta w_{11}'} = \frac{\delta L}{\delta \hat{y}} * \frac{\delta \hat{y}}{\delta z} * \frac{\delta z}{\delta o_{11}} * \frac{\delta o_{11}}{\delta w_{11}'}$$

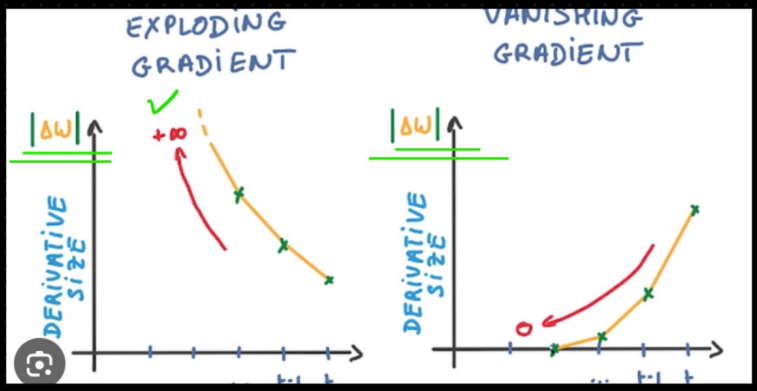
$$\approx 10 * 10 * 10 * 10$$

$$= 10000$$

drastic change

big

$$w_{new} = w_{old} - \eta \frac{\delta L}{\delta w_{old}}$$



Epoche = 20

Loss = 0.8

Loss = 0.7

Loss = 0.6

Loss = 0.5 ✓

Loss = 0.5 ✓

Vanishing gradient descent