

ReLU is extremely simple but it solved the biggest problems in deep learning.

1. ReLU prevents vanishing gradients

For positive inputs, derivative = 1.

Gradients stay strong and deep networks train properly.

2. ReLU adds non-linearity

The “kink” at 0 makes the network nonlinear.

Without this, the entire deep network collapses into a linear model.

3. ReLU creates sparse activations

Negative values become 0 → many neurons turn off.

This reduces overfitting and noise while making models faster and more robust.

4. ReLU is computationally cheap

Just $\max(0, x)$ — no exponentials like sigmoid/tanh.

Training becomes much faster.

5. ReLU gives clean feature interpretation

Positive = feature present

Zero = feature absent

This helps CNNs detect patterns reliably.

Summary:

ReLU works because it keeps gradients alive, adds nonlinearity, makes networks sparse, trains faster, and stabilizes CNNs. Its simplicity is exactly what makes it powerful.