Gradient Decay: Why Deep Networks Failed

- Sigmoid (≤ 0.25 per layer)
- ReLU (≈ 0.5 per layer)
- Well-behaved (≈ 0.9 per layer)

Vanishing threshold

Gradient Magnitude

Layer Depth