MNIST Activation Functions Comparison

Comparing Sigmoid, Tanh, and ReLU

**Obje

* How different activations affect training speed
* The vanishing gradient problem
* Gradient flow analysis
* Practical recommendations

Problem Statement

The Problem

Classi

Activation	Range	Key Property
Sigmoid	(0,1)	Probabilistic
Tanh	(-1,1)	Zero-centered
ReLU	[0,inf)	No vanishing

MNIST Dataset

Input Data

Property	Value
Training samples	60,000
Test samples	10,000
Image size	28 x 28
Classes	10 (digits 0-9)

Preprocessing Steps

* Flatten: 28x28 -> 784 vector
* Normalize: 0-255 -> 0-1 range

Neural Network Architecture

Model Structure

Input L

* Optimizer: Adam (learning_rate=0.001)
* Loss: Sparse Categorical Cross-Entropy
* Epochs: 20
* Batch size: 128

All moc

Activation Functions Explained

Sigmoid: f(x) = 1/(1+e^-x)

* Range: (0, 1)
* Max derivative: 0.25 (causes vanishing gradients!)

Tanh: f

* Range: (-1, 1)
* Zero-centered (better than sigmoid)

ReLU: f

* Range: [0, infinity)
* Derivative = 1 for positive (no vanishing!)

The Vanishing Gradient Problem

What Is It?

Gradie

* Sigmoid max derivative = 0.25
* Each layer multiplies gradient by <= 0.25
* After 4 layers: 0.25^4 = 0.004 (nearly zero!)

Why Re

* ReLU derivative = 1 for positive inputs
* Gradients pass through unchanged
* Early layers can still learn effectively

Experiment Results

Performance Comparison

Model	Accuracy	Avg Time	Gradient Mag
Sigmoid	97.5%	2.5s	0.0001
Tanh	97.8%	2.3s	0.0005
ReLU	98.2%	2.0s	0.0050

Key Finding

ReLU gr

Key Observations

1. Convergence Speed

ReLU

Sigmoid: 0.0001	Tanh: 0.0005	ReLU: 0.0050

4. All Models Work for MNIST

MNIST is

When to Use Each Activation

Recommendations

Layer Type	Recommended Activation
Hidden layers (default)	ReLU
Binary classification output	Sigmoid
Multi-class output	Softmax
RNN hidden states	Tanh
LSTM/GRU gates	Sigmoid

Special Cases

* If dead neurons occur: Use Leaky ReLU
* For transformers: GELU is common

Interview Key Points

Top Questions

Q: Wh

A: ReLU derivative = 1 for positive, no gradient shrinking

Q: When

A: Binary output layer, LSTM gates

Q: What

A: Neurons stuck at 0, fix with Leaky ReLU

Key Nu

Conclusion

What We Learned

* Activation functions critically affect training
* ReLU is the default for hidden layers
* Vanishing gradients are a real problem
* Choose activation based on layer purpose

Action

* Start with ReLU for hidden layers
* Use Sigmoid for binary output
* Use Softmax for multi-class output
* Monitor gradient magnitudes in deep networks

Remem