Week 1: Introduction to Deep Learning

Outline

This week introduces the fundamental concepts of **deep learning**, focusing on **feedforward neural networks (FNNs)**, **backpropagation**, and **activation functions**. Participants will explore how neural networks function as universal function approximators and understand why **multi-layer perceptrons (MLPs)** are essential for solving non-linearly separable problems like XOR.

1️⃣ Feedforward Neural Networks

A **feedforward neural network (FNN)** maps an input vector **X** to an output **Y** using layers of neurons:

$$h = \sigma(WX + b)$$

Multi-Layer Perceptron (MLP)

An MLP with a single hidden layer computes:

$$h = \sigma(W_1X + b_1)$$

$$y = \sigma(W_2h + b_2)$$

2️⃣ Activation Functions

Why Are They Needed?

Without non-linearity, an MLP collapses into a single linear function. Activation functions **enable complex decision boundaries**.

Common Activation Functions

Sigmoid:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

ReLU (Rectified Linear Unit):

$$ReLU(x) = \max(0, x)$$

Tanh:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

3️⃣ Gradient Descent & Backpropagation

Gradient Descent

Used to minimize the loss function **L** by updating weights **W**:

$$W_{\text{new}} = W - \eta \frac{\partial L}{\partial W}$$

Backpropagation

Backpropagation **computes gradients efficiently** using the chain rule:

Output Layer Gradient:

$$\frac{\partial L}{\partial W_2} = \delta_2 h^T$$

Hidden Layer Gradient:

$$\delta_1 = (\delta_2 W_2) \circ \sigma'(h)$$

4️⃣ Loss Functions

Mean Squared Error (MSE)

$$L = \frac{1}{N} \sum (y_{\text{true}} - y_{\text{pred}})^2$$

Cross-Entropy Loss

$$L = -\sum y \log \hat{y}$$

5️⃣ Weight Initialization

Xavier (Glorot) Initialization

$$W \sim \mathcal{N}\left(0, \frac{1}{\text{fan-in}}\right)$$

He Initialization (for ReLU)

$$W \sim \mathcal{N}\left(0, \frac{2}{\text{fan-in}}\right)$$

6️⃣ Summary of Key Equations

Concept	Equation
Feedforward Layer	$h = \sigma(WX + b)$
Backpropagation	$\frac{\partial L}{\partial W_2} = \delta_2 h^T$
Sigmoid Activation	$\sigma(x) = \frac{1}{1 + e^{-x}}$
ReLU Activation	$ReLU(x) = \max(0, x)$
MSE Loss	$L = \frac{1}{N} \sum (y_{\text{true}} - y_{\text{pred}})^2$
Cross-Entropy Loss	$L = -\sum y \log \hat{y}$

Recommended Reading:

- **Goodfellow:** Chapters 6 & 8
- **D2L:** Chapter 3