School of Computer Science and Engineering

TSIS-8: Transformer for Text Translation (EN→ES)

Conducted by:

Mustafin Bexultan

Date	*Class time*
Friday	14:00 – 15:00

November 21, 2025

Table of Contents

1 Executive Summary

This report details the manual calculation of a multi-layer perceptron training process, replicating the specific handwritten logic found in the seminar notes. We perform Forward Propagation and Backpropagation for a network with topology 2-2-1 over three iterations.

The system is trained to map inputs $X_1 = 0.35, X_2 = 0.9$ to a target value of 0.5 using Sigmoid activation and Gradient Descent.

2 Network Architecture & Setup

The following diagram illustrates the network state during the first iteration, including initial weights and calculated hidden layer outputs.

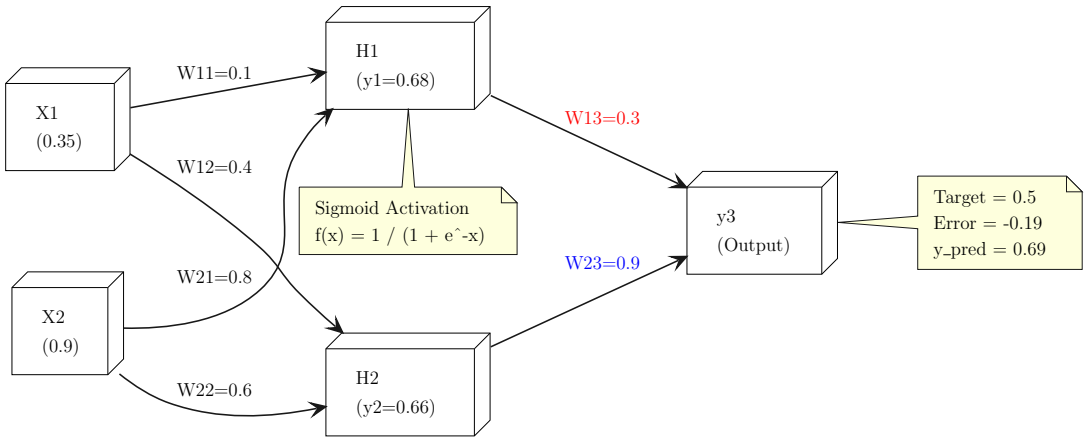

Figure 1: Visual representation of the network structure and initial parameters (generated via PlantUML).

2.1 Initial Parameters

- **Target (y_{target}):** 0.5.
- **Learning Rate (η):** 0.1.
- **Activation Function:** $f(x) = \frac{1}{1+e^{-x}}$.
- **Weights to Update:** $W_{13} = 0.3, W_{23} = 0.9$.

3 Calculation Process

3.1 Iteration 1

1. Forward Propagation
Hidden layer outputs (pre-calculated): $y_1 \approx 0.68, y_2 \approx 0.66$.
Output summation:

$$S_{out} = (y_1 \cdot W_{13}) + (y_2 \cdot W_{23}) = (0.68 \cdot 0.3) + (0.66 \cdot 0.9) = 0.204 + 0.594 = 0.801$$

Prediction (y_{pred}):

$$y_{pred} = \frac{1}{1+e^{-0.801}} \approx \mathbf{0.69}$$

2. Error & Gradient

$$E = y_{target} - y_{pred} = 0.5 - 0.69 = \mathbf{-0.19}$$

Gradient (δ_3) for output layer:

$$\delta_3 = y_{pred}(1-y_{pred}) \cdot E = 0.69 \cdot 0.31 \cdot (-0.19) \approx \mathbf{-0.0406}$$

3. Weight Update

$$\Delta W = \eta \cdot y_{in} \cdot \delta$$

$$W_{13}^{new} = 0.3 + (0.1 \cdot 0.68 \cdot -0.0406) = 0.3 - 0.0027 = \mathbf{0.2973}$$

$$W_{23}^{new} = 0.9 + (0.1 \cdot 0.66 \cdot -0.0406) = 0.9 - 0.0027 = \mathbf{0.8973}$$

3.2 Iteration 2

Using updated weights $W_{13} = 0.2973, W_{23} = 0.8973$.

1. Forward: Sum $= 0.68(0.2973) + 0.66(0.8973) \approx 0.794$.
$y_{pred} = \sigma(0.794) \approx \mathbf{0.688}$.

2. Error: $E = 0.5 - 0.688 = -0.188$.
3. Gradient: $\delta_3 = 0.688(1-0.688)(-0.188) \approx -0.0403$.

4. Update:

$$W_{13}^{new} = 0.2973 - 0.0027 = \mathbf{0.2946}$$

$$W_{23}^{new} = 0.8973 - 0.0027 = \mathbf{0.8946}$$

3.3 Iteration 3

Using updated weights $W_{13} = 0.2946, W_{23} = 0.8946$.

1. Forward: $y_{pred} \approx \mathbf{0.687}$.
2. Error: $E = -0.187$.
3. Gradient: $\delta_3 \approx -0.0402$.

4. Update:

$$W_{13}^{new} = 0.2946 - 0.0027 = \mathbf{0.2919}$$

$$W_{23}^{new} = 0.8946 - 0.0026 = \mathbf{0.8920}$$

4 Key Results (Training Log)

Table 1: Evolution of Output Layer weights over 3 iterations.

i (Iter)	**j (Weight)**	W_{old}	W_{new}	δ **(Gradient)**	η
1	W_{13}	0.3000	0.2973	-0.0406	0.1
1	W_{23}	0.9000	0.8973	-0.0406	0.1
2	W_{13}	0.2973	0.2946	-0.0403	0.1
2	W_{23}	0.8973	0.8946	-0.0403	0.1
3	W_{13}	0.2946	0.2919	-0.0402	0.1
3	W_{23}	0.8946	0.8920	-0.0402	0.1

5 Highlights

- **Error Reduction:** The error decreased monotonically from $|0.19|$ to $|0.187|$.
- **Weight Adjustment:** Both weights W_{13} and W_{23} decreased, as the network attempts to lower the total sum fed into the sigmoid function to push the output 0.69 down towards 0.5.
- **Visuals:** The included diagram (Fig. 1) accurately reflects the topology derived from the source notes.