End Term Assignment

Mathematics for Deep Learning

Activation Function

Activation Function is the component of the Neural Network that adds non – linearity into the Dataset model. This enables the model to learn the complex data patterns (*Example – Images, Audio, etc*). Without this even a neural network would simply behave as a simple linear regression model.

Activation functions decide whether a neuron should be activated based on the weighted sum of inputs and a bias term.

Following are the important features of the Activation Function which makes the computation efficient and inexpensive: -

i. The activation function should satisfy the condition of non – linearity.
ii. The function should satisfy the condition of differentiability (Except certain points which are not of immense importance).
iii. The gradient of the function should have finite values and should be bounded.
iv. The function should not have the vanishing gradient issue.
v. The Computation should be inexpensive (unlike the sigmoidal function).
vi. The function should be numerically stable (unlike the swish function).

The Activation Function

The own developed definition of the Activation Function,

$$\boxed{f(x) = \begin{cases} \beta x - (1-\beta), & x < -1 \\ x, & -1 \le x \le 1 \\ \beta x + (1-\beta), & x > 1 \end{cases}}$$

Here, β is a constant having values in the range of 0 to 1.

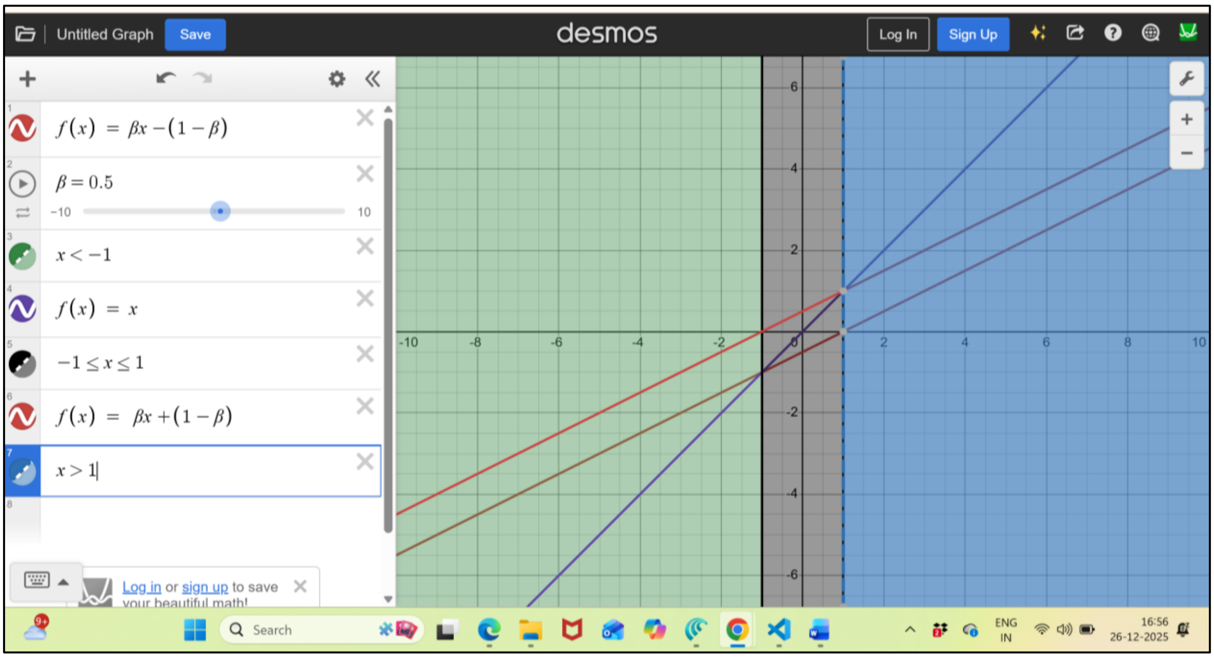
The Activation Function has the following properties: -

i. The function is linear in near the origin (*As shown in the graph below*) which ensures a smooth gradient flow.
ii. This Activation function has the Zero Centred Property (*like the Tanh(x) activation function*).

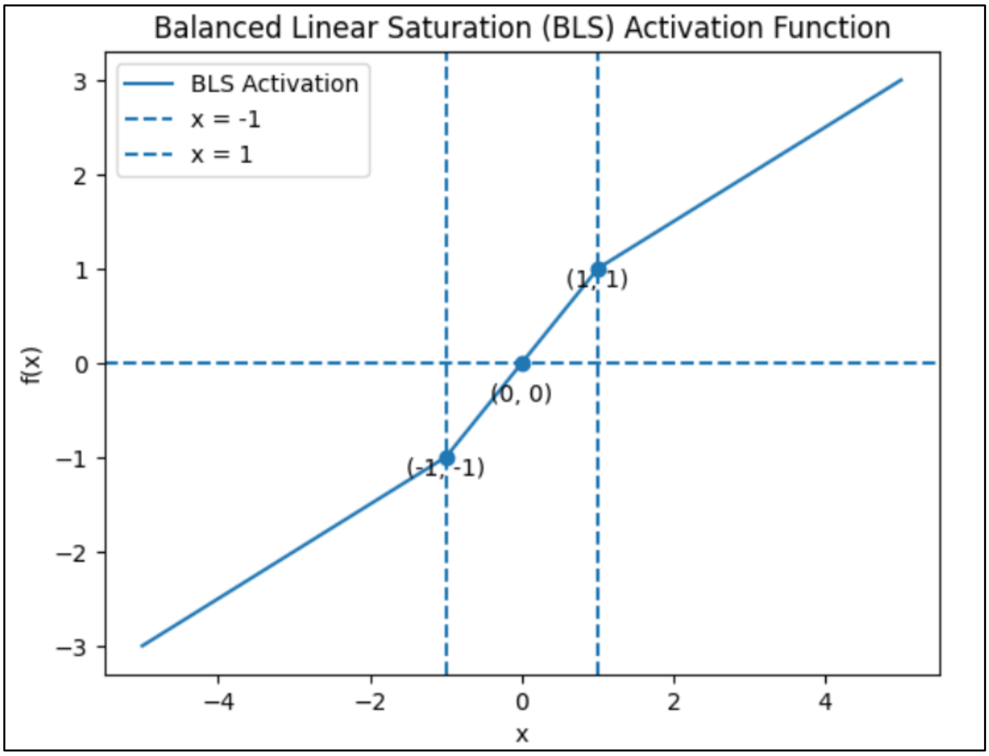
iii. The function ensures that the gradient of the function never becomes zero. Solves the Vanishing Gradient Issue.

iv. Is computationally inexpensive, as it uses only multiplication and addition. (*Unlike the case of the Sigmoidal Function in which the Computation becomes expensive because of the exponents and the divisions*)

The Graph of the Activation Function



The Graph for the Activation Function



The Mathematical Properties of the Activation Function

1. ***Non – Linearity into the function: -***
   Since, $f(ax+b) \neq af(x)+b$
   Hence, the function is Non – Linear in its nature.
   The Non – Linearity in the function is important for making the Neural Network to understand Complex Data Patterns.
2. **Differentiability: -**
   The derivative of the function is: -

$$\boxed{f(x) = \begin{cases} \beta x - (1-\beta), & x < -1 \\ x, & -1 \leq x \leq 1 \\ \beta x + (1-\beta), & x > 1 \end{cases}}$$

$$\boxed{f'(x) = \begin{cases} \beta, & |x| > 1 \\ 1, & |x| < 1 \end{cases}}$$

   The Activation function is differentiable at all the points except at +1 and -1 (*Similar to the case of the ReLU Function, which is not differentiable at 0*)
   Hence, the function has finite and bounded derivate, which is important for the Back Proportion.
3. **Vanishing Gradient Problem Resolved: -**
   The Gradient of the activation function becomes zero at the extreme points making the computation difficult to proceed ahead and stops the process in between.

$$\lim_{|x| \to \infty} f'(x) = 0$$

   But, in this Activation Function the gradient is either 1 or β,

$$\lim_{|x| \to \infty} f'(x) = \beta$$

   Hence, $0 \leq \beta \leq f'(x) \leq 1$
   Hence, the gradient is bounded and non – zero. This ensures that the vanishing gradient issue gets eliminated and the Neural Network performs a smooth Back Proportion.
4. **Gradient Stability: -**
   Let L be the Cost Function,
   Therefore,

$$\boxed{\left|\frac{\partial L}{\partial x}\right| = \left|\frac{\partial L}{\partial f(x)}\right| \cdot |f'(x)|}$$

   Since, $0 \leq \beta \leq f'(x) \leq 1$, and $\beta \in (0,1)$
   Therefore,

$$\boxed{\beta\left|\frac{\partial L}{\partial f(x)}\right| \leq \left|\frac{\partial L}{\partial x}\right| \leq \left|\frac{\partial L}{\partial f(x)}\right|}$$

Hence, in this case the gradient is neither becoming zero nor is going to ∞.

5. **Computationally Efficient: -**

   Since, the Activation function and its gradient involves only,

   - Addition
   - Multiplication
   - And, comparison

   Operations, this makes the computation overall efficient, as no exponential, powers and division operations are involved in the Computation.

   (*This function is as cheap as ReLU, as it only involves the comparison operation*)

6. **The Zero Centred Property: -**

   Since, $f(-x) = -f(x)$,

   This ensures that the function is symmetric about zero. This property improves the convergence during the time of gradient descent.

<u>The Summary of the Report</u>

The Activation function proposed here satisfies all the fundamental criteria of a well-behaved activation function: -

i. Nonlinearity
ii. Differentiability
iii. Computational efficiency
iv. Bounded gradient propagation
v. And numerical stability.

The Activation Function eliminates the problem of vanishing gradients while avoiding the dead neuron problem.

<u>Dead Neuron Problem: -</u>

i. The gradient in this situation becomes zero.
ii. The weights stop getting upgraded.
iii. The Activation Function contributes nothing to the Neural Network, as the gradient of the cost function also becomes zero.