CHAPTER – 10

Rectified Linear Unit

1. Rectified linear unit introduction
2. Relu functionality
3. Leaky relu.

- The main aim of rectifying linear unit is removing the vanishing GD problem.
- Normally the varnishing GD problem coming when a greater number of derivative terms get multiplied.
- If we take a sigmoid activation function, the values of derivative of sigmoid lies between 0 to 0.25. these are very small values. After multiply all these small values in chain rule will cause varnishing GD problem.
- Not only in sigmoid function, in threshold activation function also creates same problem.

Note: The vanishing gradient descent occurs due to two reasons:

1. Low values of derivative terms.
2. Chain rule in loss function.

Relu Functionality:

- This problem overcome by using *Relu* activation function. now let's understand how *Relu* works.

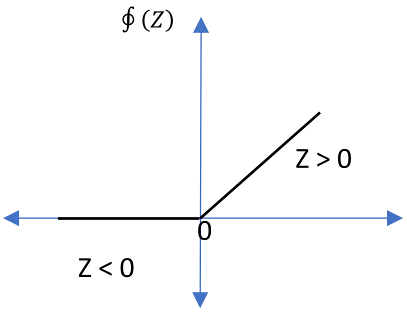

Figure 2 - Relu graph

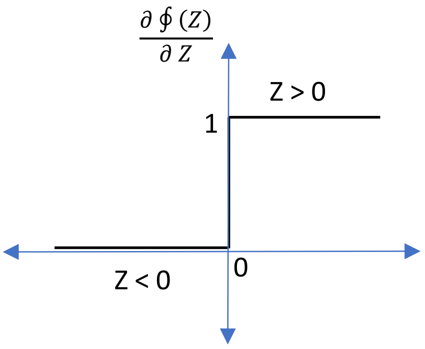

Figure 1 - Derivative Relu graph

- The formula of *Relu* activation function is:

$$\oint (Z) = \; max\;(Z, 0)$$

$$At\;normal\;condition\;:\;\begin{cases} When\;z\;is\;negitive\;\oint(Z) = 0 \\ When\;z\;is\;positive\;\oint(Z) = z \end{cases}$$

- The derivative of *Relu* activation function values are:

$$\frac{\partial \oint (Z)}{\partial\, Z} = \begin{cases} 1 \;\; if \;\; z = \; +ve\;or\;z\; > 0 \\ 0 \;\; if \;\; z = \; -ve\;or\;z < 0 \end{cases}$$

- Let's check above values in two cases with chain rule for updating weights.

Case – 1:

Consider the values of all derivative terms as 1 in chain rule of $\frac{\partial L}{\partial \omega_{old}}$.

$$\frac{\partial L}{\partial \omega_{old}} = 1\; \times 1\; \times 1$$

And consider $\alpha\; = 1$, then ω_{new} is very different from ω_{old}. ω_{new} will update normally and there is no problem in this case.

Case – 2:

In this case take one of the derivative terms as 0 in the child rule.

$$\frac{\partial L}{\partial \omega_{old}} = 1\; \times 0\; \times 1$$

$$\text{Finally,} \quad \omega_{new} \;=\; \omega_{old} \;-\; \alpha\;[0]$$

$$\omega_{new} \;=\; \omega_{old}$$

But actually, it is not correct because Relu function value zero even the value of z is not -ve in other derivative terms.

If we observe in this case no weight changes happening. And this case exactly creates the dead neuron means this particular neuron will be deactivated due to a single zero of derivative term in child rule.

Note: This particular problem we can overcome by using leaky Relu concept.

Leaky Relu:

- The *leaky relu* takes a small functional value instead of derivative term value when $z < 0$.

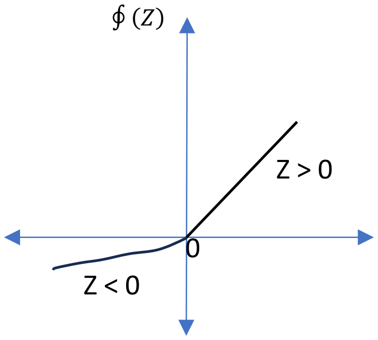

Figure 3 – Leaky Relu graph

- In this case the *Leaky Relu* graph will change as shown in the figure.

$$\frac{\partial \oint (Z)}{\partial Z} = \begin{cases} 1 & if \quad z > 0 \\ 0.01(z) & if \quad z < 0 \end{cases}$$

- Now the derivative terms will not be zero, but those are considered as nearby zero. And it will consider the dead neurons which are created by zero values of derivative terms.