Different Activation Function characteristics explanation

Activation Functions:

1. **Step**
2. **Sigmoid**
3. **Tanh**
4. **Relu**
5. **Elu**
6. **Selu**

Submitted by:

Name: Md. Mintajur Rahman Emon
Id: 20-43870-2

Department of Computer Science

Faculty of Science & Technology

American International University - Bangladesh

Activation Function: In general, an activation function is a function used in artificial neural networks which outputs a small value for small inputs, and a larger value if its inputs exceed a threshold. If the inputs are large enough, the activation function "fires" to the next neuron, otherwise it does nothing. In CNN, the purpose of introducing an activation function is to give neural network nonlinear expression ability, so that it can better fit the results, as well as improve the accuracy. There are numbers of activation functions. Six of them are discussed below:

1. Step Function:

A step function changes an input value into a binary 0 or 1 output. Its calculating efficiency is great.

Equation:

A step function f: R → R can be written in the form:

$$f(x) = \sum_{i=0}^{n} \alpha_i \chi_{A_i}(x)$$

for all real numbers x.

If n ≥ 0, α_i are real numbers and A_i are intervals, then the indicator function of A is χ_A, and it can be written as below:

$$\chi_A(x) = \begin{cases} 1; if\ x \in A, \\ 0; if\ x \notin A \end{cases}$$

Advantages:

1. It can create a binary classifier
2. It is a simple function as it's mechanism indicates the output will be either 0 or, 1.

Disadvantages:

1. This function is not useful when there are multiple classes in the target variable.
2. The gradient of the step function is zero which causes a hindrance in the back propagation process.

2. Sigmoid Function:

A sigmoid function is a bounded, differentiable, real function that is defined for all real input values and has a non-negative derivative at each point and exactly one inflection point. It's graph is sigmoidal (s - shaped).

Equation:

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

Advantages:

1. It gives smooth gradient, thereby, preventing jumps in output values.
2. It's one of the normalized functions.
3. With 1 or 0, it makes a clear prediction.

Disadvantages:

1. It vanishes gradient.
2. This is not a zero centric function.
3. It is computationally expensive function.

3. Tanh Function:

Tanh is also like logistic sigmoid but better. The range of the tanh function is from (-1 to 1). Tanh is also sigmoidal (s - shaped). It squashes numbers to range [-1,1]

Equation:

$$\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \begin{cases} \xrightarrow[x\to-\infty]{} -1 \\ \xrightarrow[0>x\gg-1]{} e^{2x}-1 \\ \xrightarrow[x\to 0]{} 0 \\ \xrightarrow[0<x\ll+1]{} e^{2x}-1 \\ \xrightarrow[x\to+\infty]{} +1 \end{cases}.$$

Advantages:

1. The negative inputs will be mapped strongly negative and the zero inputs will be mapped near zero in the tanh graph.

2. This is zero centric.
3. The function is differentiable.

Disadvantage:

1. It kills gradient when saturated.
2. It's computationally expensive.

4. ReLu Function:

The rectified linear activation function or ReLU is a non-linear function or piecewise linear function that will output the input directly if it is positive, otherwise, it will output zero. It is the most commonly used activation function in neural networks, especially in Convolutional Neural Networks (CNNs) & Multilayer perceptrons.

Equation:

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

Advantages:

1. Does not saturate (in +region)
2. Very computationally efficient
3. Converges much faster than
4. Sigmoid/Tanh in practice (e.g. 6x)
5. Actually more biologically plausible than sigmoid

Disadvantages:

1. Don't give zero centered output.
2. Large weight updates can mean that the summed input to the activation function is always negative, regardless of the input to the network.

5. Elu Function:

Exponential Linear Unit or its widely known name ELU is a function that tend to converge cost to zero faster and produce more accurate results. Different to other activation functions, ELU has a extra alpha constant which should be positive number. ELU is very similar to RELU except negative inputs.

Equation:

$$f(x) = \begin{cases} 0 & if\ x > 0 \\ \alpha(\exp(x) - 1) & if\ x \leq 0 \end{cases}$$

Advantages:

1. Tend to converge faster than ReLU (because mean ELU activations are closer to zero)
2. Better generalization performance than ReLU.
3. Fully continuous.
4. Fully differentiable.
5. Does not have a vanishing gradients problem.
6. Does not have an exploding gradients problem.
7. Does not have a dead relu problem.

Disadvantages:

1. During the test time ELU perform slower than ReLU and its variants.
2. It is computationally slower due to the use of computationally expensive exponential function.

6. Selu Function:

Scaled Exponential Linear Unit (SELU) activation function is scaled variant of ELU activation function. It uses two fixed parameters α and λ, and the value of these is derived from the inputs. However, for standardized inputs (mean of 0 and standard deviation of 1) the suggested values are α=1.6733, λ=1.0507.

Equation:

$$\text{selu}(x) = \lambda \begin{cases} x & \text{if } x > 0 \\ \alpha e^x - \alpha & \text{if } x \leqslant 0 \end{cases}$$

Advantages:

1. It provides self-normalization.
2. It solves the vanishing or exploding gradients problem.

Disadvantages:

1. It is not yet used widely in practice.
2. More research on architectures such as CNNs and RNNs using Selus is needed for wide sread industry use.