Types of Neural Networks Activation Functions

Binary Step Function – Binary step function depends on a threshold value that decides whether a neuron should be activated or not.

Limitations:

- It cannot be used for multi-class classification problems.
- The gradient of the step function is zero, which causes a hindrance in the backpropagation process.

Linear Activation Function – The linear activation function, also known as "no activation," or "identity function" (multiplied x1.0), is where the activation is proportional to the input.

Limitations:

- It's not possible to use backpropagation as the derivative of the function is a constant and has no relation to the input x.
- All layers of the neural network will collapse into one if a linear activation function is used. No matter the number of layers in the neural network, the last layer will still be a linear function of the first layer. So, essentially, a linear activation function turns the neural network into just one layer.

Non-Linear Activation Functions –

The linear activation function shown above is simply a linear regression model.

Because of its limited power, this does not allow the model to create complex mappings between the network's inputs and outputs.

Limitations:

- They allow backpropagation because now the derivative function would be related to the input.
- They allow the stacking of multiple layers of neurons as the output would now be a non-linear combination of input passed through multiple layers.

Non-Linear Neural Networks Activation Functions

Sigmoid / Logistic Activation Function – This function takes any real value as input and outputs values in the range of 0 to 1.

Tanh Function (Hyperbolic Tangent) – Tanh function is very similar to the sigmoid/logistic activation function, and even has the same S-shape with the difference in output range of -1 to 1.

ReLU Function – ReLU stands for Rectified Linear Unit. Although it gives an impression of a linear function, ReLU has a derivative function and allows for backpropagation while simultaneously making it computationally efficient.

Leaky ReLU Function – Leaky ReLU is an improved version of ReLU function to solve the Dying ReLU problem as it has a small positive slope in the negative area.

Parametric ReLU Function – Parametric ReLU is another variant of ReLU that aims to solve the problem of gradient's becoming zero for the left half of the axis.

Exponential Linear Units (ELUs) Function – Exponential Linear Unit, or ELU for short, is also a variant of ReLU that modifies the slope of the negative part of the function

Softmax Function – It is described as a combination of multiple sigmoids. It calculates the relative probabilities. Similar to the sigmoid/logistic activation function, the SoftMax function returns the probability of each class.

It is most commonly used as an activation function for the last layer of the neural network in the case of multi-class classification.

Swish – It is a self-gated activation function developed by researchers at Google. Swish consistently matches or outperforms ReLU activation function on deep networks applied to various challenging domains such as image classification, machine translation etc.

Gaussian Error Linear Unit (GELU) – The Gaussian Error Linear Unit (GELU) activation function is compatible with BERT, ROBERTa, ALBERT, and other top NLP models. This activation function is motivated by combining properties from dropout, zoneout, and ReLUs.

Scaled Exponential Linear Unit (SELU) – SELU was defined in self-normalizing networks and takes care of internal normalization which means each layer preserves the mean and variance from the previous layers. SELU enables this normalization by adjusting the mean and variance**.**

How to choose the right Activation Function?

As a rule of thumb, you can begin with using the ReLU activation function and then move over to other activation functions if ReLU doesn't provide optimum results.

And here are a few other guidelines:

- ReLU activation function should only be used in the hidden layers.
- Sigmoid/Logistic and Tanh functions should not be used in hidden layers as they make the model more susceptible to problems during training (due to vanishing gradients).
- Swish function is used in neural networks having a depth greater than 40 layers.

Finally, a few rules for choosing the activation function for your output layer based on the type of prediction problem that you are solving:

- **Regression** - Linear Activation Function
- **Binary Classification**—Sigmoid/Logistic Activation Function
- **Multiclass Classification**—Softmax
- **Multilabel Classification**—Sigmoid

The activation function used in hidden layers is typically chosen based on the type of neural network architecture.

- **Convolutional Neural Network (CNN):** ReLU activation function.
- **Recurrent Neural Network:** Tanh and/or Sigmoid activation function.

Neural Network Activation Functions