Activation Function Selection

Sachchida Nand Chaurasia

Assistant Professor

Department of Computer Science
Banaras Hindu University
Varanasi
Email id: snchaurasia@bhu.ac.in, sachchidanand.mca07@gmail.com

April 19, 2023

How to Choose the Right Activation Function for Neural Networks I

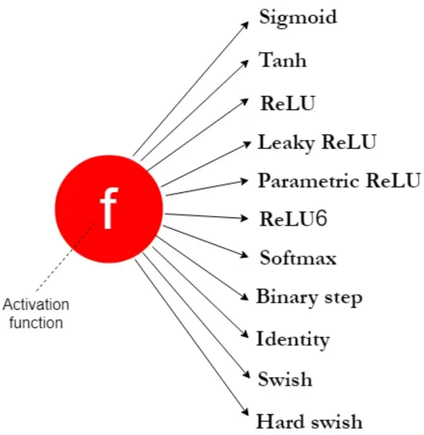


Figure 1: Different types of activation functions

How to Choose the Right Activation Function for Neural Networks II

Activations functions in different layers in a neural network

✓ A neural network typically consists of three types of layers: Input Layer, Hidden Layer(s) and Output Layer.

✓ The input layer just holds the input data and no calculation is performed. Therefore, no activation function is used there.

✓ We must use a non-linear activation function inside hidden layers in a neural network. This is because we need to introduce non-linearity to the network to learn complex patterns.

✓ Without non-linear activation functions, a neural network with many hidden layers would become a giant linear regression model that is useless for learning complex patterns from real-world data.

How to Choose the Right Activation Function for Neural Networks III

✓ The performance of a neural network model will vary significantly depending on the type of activation function we use inside the hidden layers.

✓ We must also use an activation function inside the output layer in a neural network. The choice of the activation function depends on the type of problem that we want to solve.

How to Choose the Right Activation Function for Neural Networks IV

Linear vs non-linear functions

✓ Most of the activation functions are non-linear. However, we also use linear activation functions in neural networks. For example, we use a linear activation function in the output layer of a neural network model that solves a regression problem.

✓ Some activation functions are made up of two or three linear components. Those functions are also classified as non-linear functions.

How to Choose the Right Activation Function for Neural Networks V

Different types of activation functions

1. Sigmoid activation function:

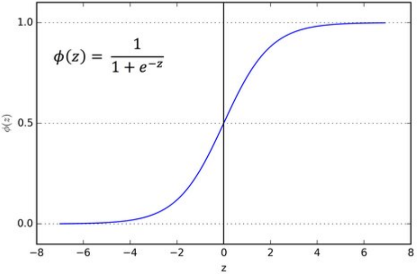


Figure 2: Sigmoid activation function

How to Choose the Right Activation Function for Neural Networks VI

Key features:

- This is also called the logistic function used in logistic regression models.
- The sigmoid function has an s-shaped graph.
- Clearly, this is a non-linear function.
- The sigmoid function converts its input into a probability value between 0 and 1.
- It converts large negative values towards 0 and large positive values towards 1.
- It returns 0.5 for the input 0. The value 0.5 is known as the threshold value which can decide that a given input belongs to what type of two classes.

How to Choose the Right Activation Function for Neural Networks VII

Usage:

- In the early days, the sigmoid function was used as an activation function for the hidden layers in MLPs, CNNs and RNNs.
- However, the sigmoid function is still used in RNNs.
- Currently, we do not usually use the sigmoid function for the hidden layers in MLPs and CNNs. Instead, we use ReLU or Leaky ReLU there.
- The sigmoid function must be used in the output layer when we build a binary classifier in which the output is interpreted as a class label depending on the probability value of input returned by the function.

How to Choose the Right Activation Function for Neural Networks VIII

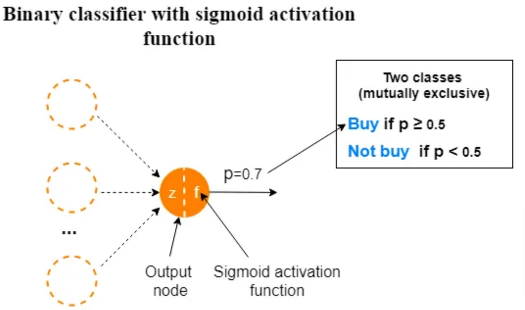
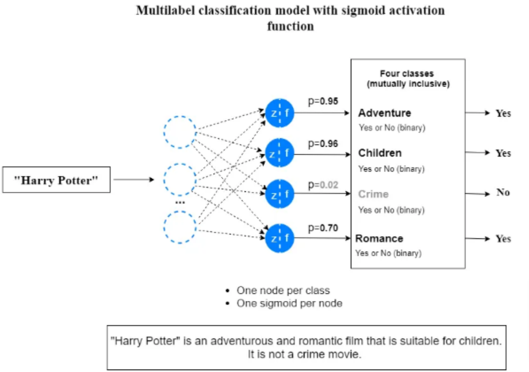

Binary classifier with sigmoid activation function
Two classes (mutually exclusive)
Buy if p ≥ 0.5
Not buy if p < 0.5
p=0.7
z
f
...
Output node
Sigmoid activation function

Figure 3: Binary classification with sigmoid function

- The sigmoid function is used when we build a multilabel classification model in which each mutually inclusive class has two outcomes. Do not confuse this with a multiclass classification model.

How to Choose the Right Activation Function for Neural Networks IX



- One node per class
- One sigmoid per node

"Harry Potter" is an adventurous and romantic film that is suitable for children. It is not a crime movie.

Figure 4: Multilabel classification with sigmoid function

Drawbacks: We do not usually use the sigmoid function in the hidden layers because of the following drawbacks.

How to Choose the Right Activation Function for Neural Networks X

- The sigmoid function has the vanishing gradient problem. This is also known as saturation of the gradients.
- The sigmoid function has slow convergence. Its outputs are not zero-centered. Therefore, it makes the optimization process harder.
- This function is computationally expensive as an e^z term is included.

How to Choose the Right Activation Function for Neural Networks XI

2. Tanh activation function

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{0.1}$$

Key features:

- The output of the tanh (tangent hyperbolic) function always ranges between -1 and +1.
- Like the sigmoid function, it has an s-shaped graph. This is also a non-linear function.
- One advantage of using the tanh function over the sigmoid function is that the tanh function is zero centered. This makes the optimization process much easier.

How to Choose the Right Activation Function for Neural Networks XII

- The tanh function has a steeper gradient than the sigmoid function has.

Usage:

- Until recently, the tanh function was used as an activation function for the hidden layers in MLPs, CNNs and RNNs.
- However, the tanh function is still used in RNNs.
- Currently, we do not usually use the tanh function for the hidden layers in MLPs and CNNs. Instead, we use ReLU or Leaky ReLU there.
- We never use the tanh function in the output layer.

How to Choose the Right Activation Function for Neural Networks XIII

Drawbacks: We do not usually use the tanh function in the hidden layers because of the following drawback.

- The tanh function has the vanishing gradient problem.
- This function is computationally expensive as an e^z term is included.

How to Choose the Right Activation Function for Neural Networks XIV

3. ReLU activation function

$$f(x) = max(0, x) \tag{0.2}$$

Key features:

- The ReLU (Rectified Linear Unit) activation function is a great alternative to both sigmoid and tanh activation functions.
- Inventing ReLU is one of the most important breakthroughs made in deep learning.
- This function does not have the vanishing gradient problem.
- This function is computationally inexpensive. It is considered that the convergence of ReLU is 6 times faster than sigmoid and tanh functions.

How to Choose the Right Activation Function for Neural Networks XV

- If the input value is 0 or greater than 0, the ReLU function outputs the input as it is. If the input is less than 0, the ReLU function outputs the value 0.
- The ReLU function is made up of two linear components. Because of that, the ReLU function is a piecewise linear function. In fact, the ReLU function is a non-linear function.
- The output of the ReLU function can range from 0 to positive infinity.
- The convergence is faster than sigmoid and tanh functions. This is because the ReLU function has a fixed derivate (slope) for one linear component and a zero derivative for the other linear component. Therefore, the learning process is much faster with the ReLU function.

How to Choose the Right Activation Function for Neural Networks XVI

- Calculations can be performed much faster with ReLU because no exponential terms are included in the function.

Usage:

- The ReLU function is the default activation function for hidden layers in modern MLP and CNN neural network models.
- We do not usually use the ReLU function in the hidden layers of RNN models. Instead, we use the sigmoid or tanh function there.
- We never use the ReLU function in the output layer.

How to Choose the Right Activation Function for Neural Networks XVII

Drawbacks:

- The main drawback of using the ReLU function is that it has a dying ReLU problem.
- The value of the positive side can go very high. That may lead to a computational issue during the training.

How to Choose the Right Activation Function for Neural Networks XVIII

4. Leaky ReLU activation function

$$f(x) = max(\alpha x, x) \tag{0.3}$$

Key features:

- The leaky ReLU activation function is a modified version of the default ReLU function.
- Like the ReLU activation function, this function does not have the vanishing gradient problem.

How to Choose the Right Activation Function for Neural Networks XIX

- If the input value is 0 greater than 0, the leaky ReLU function outputs the input as it is like the default ReLU function does. However, if the input is less than 0, the leaky ReLU function outputs a small negative value defined by αz (where α is a small constant value, usually 0.01 and z is the input value).
- It does not have any linear component with zero derivatives (slopes). Therefore, it can avoid the dying ReLU problem.
- The learning process with leaky ReLU is faster than the default ReLU.

Usage:

- The same usage of the ReLU function is also valid for the leaky ReLU function.

How to Choose the Right Activation Function for Neural Networks XX

5. ReLU6 activation function

   Key features:

   - The main difference between ReLU and ReLU6 is that ReLU allows very high values on the positive side while ReLU6 restricts to the value 6 on the positive side. Any input value which is 6 or greater than 6 will be restricted to the value 6 (hence the name).
   - The ReLU6 function is made up of three linear components. It is a non-linear function.

How to Choose the Right Activation Function for Neural Networks XXI

6. Softmax activation function:

$$softmax(z_i) = \frac{\exp{(z_i)}}{\sum_{j=1} \exp{(z_i)}} \tag{0.4}$$

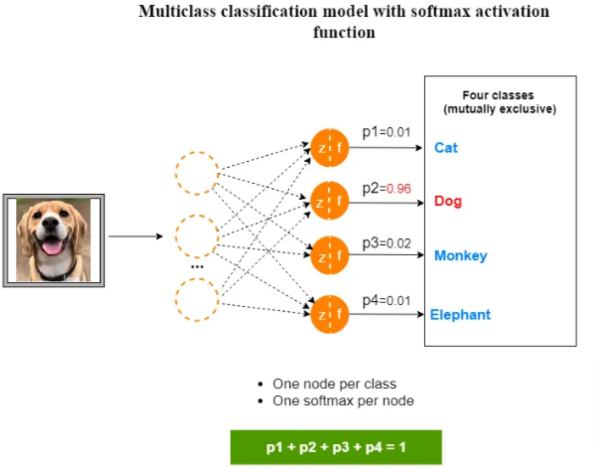
Key features:

- This is also a non-linear activation function.
- The softmax function calculates the probability value of an event (class) over K different events (classes). It calculates the probability values for each class. The sum of all probabilities is 1 meaning that all events (classes) are mutually exclusive.

Usage:

- We must use the softmax function in the output layer of a multiclass classification problem.

How to Choose the Right Activation Function for Neural Networks XXII



Multiclass classification model with softmax activation function

- One node per class
- One softmax per node

p1 + p2 + p3 + p4 = 1

Figure 5: Multiclass classification with softmax function

How to Choose the Right Activation Function for Neural Networks XXIII

- We never use the softmax function in the hidden layers.

How to Choose the Right Activation Function for Neural Networks XXIV

Summary

✓ Activation functions are just mathematical functions. The main feature that an activation function should have is that the function should be differentiable as this is a requirement for backpropagation in model training.

✓ Choosing the right activation function is the main challenge and it can be considered as a type of hyperparameter tuning in which the programmer manually chooses the activation function by understanding the problem definition and considering the performance of the model and the convergence of the loss function.

How to Choose the Right Activation Function for Neural Networks XXV

General guidelines to choose the right activation function:

- No activation function is required in the input layer nodes of a neural network. So, you don't need to worry about activation functions when you define the input layer.
- The output layer activation function depends on the type of problem that we want to solve. In a regression problem, we use the linear (identity) activation function with one node. In a binary classifier, we use the sigmoid activation function with one node. In a multiclass classification problem, we use the softmax activation function with one node per class. In a multilabel classification problem, we use the sigmoid activation function with one node per class.

How to Choose the Right Activation Function for Neural Networks XXVI

- We should use a non-linear activation function in hidden layers. The choice is made by considering the performance of the model or convergence of the loss function. Start with the ReLU activation function and if you have a dying ReLU problem, try leaky ReLU.
- In MLP and CNN neural network models, ReLU is the default activation function for hidden layers.
- In RNN neural network models, we use the sigmoid or tanh function for hidden layers. The tanh function has better performance.
- Only the identity activation function is considered linear. All other activation functions are non-linear.
- We never use softmax and identity functions in the hidden layers.

How to Choose the Right Activation Function for Neural Networks XXVII

- We use tanh, ReLU, variants of ReLU, swish and hard swish functions only in the hidden layers.