Activate or Not: Learning Customized Activation

presentation by Mikhail Liz

November 2020

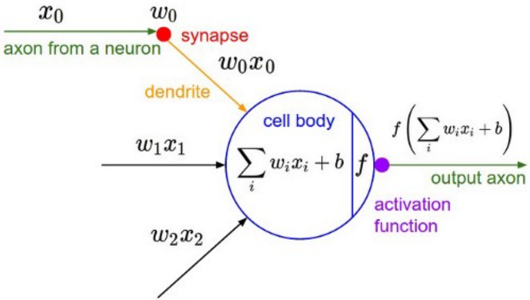
Content

- Introduction
- Sigmoid Activation Function
- Rectified Linear Unit Activation Function
- ACON Activation Function
- Results

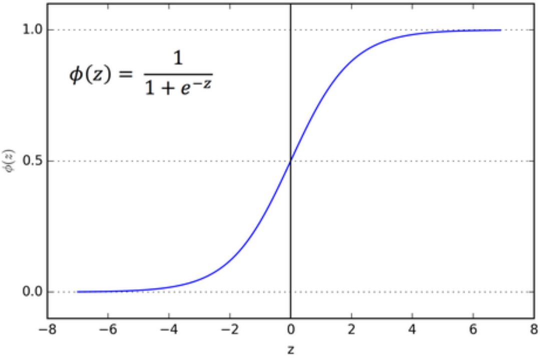
In artificial neural networks, the activation function of a node defines the output of that node given an input or set of inputs. Two types of activation functions:

- Linear Activation Function
- Non-linear Activation Functions

We need the activation function to introduce nonlinear real-world properties to artificial neural networks.



The sigmoid function curve looks like a S-shape.



Sigmoid Activation Function

Advantages and disadvantages

Advantages

- Exists between zero to one
- The function is differentiable
- The function is monotonic but function's derivative is not

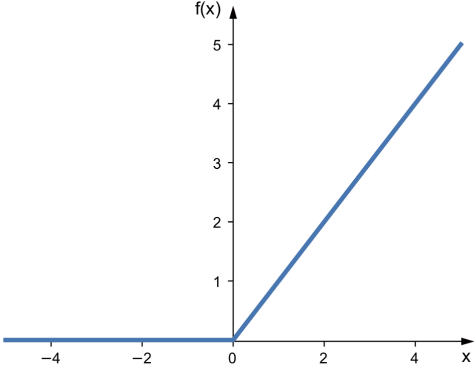
Disadvantages

- The vanishing gradient problem

ReLU Activation Function

Function type

Equation for ReLU function: $f(x) = max(0, x)$



ReLU Activation Function

Advantages and disadvantages

Advantages

- It is easy and fast to calculate the derivative
- Sparsity of activation

Disadvantages

- Dying ReLU problem

ACON Activation Function

Approximation of the maximum function

Equation for smooth maximum function:

$$S_\beta(x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} x_i e^{\beta x_i}}{\sum_{i=1}^{n} e^{\beta x_i}}$$

Approximation of the ReLU function:

$$
\begin{aligned}
S_\beta\left(\eta_a(x), \eta_b(x)\right) &= \eta_a(x) \cdot \frac{e^{\beta\eta_a(x)}}{e^{\beta\eta_a(x) + e^{\beta\eta_b(x)}} + \eta_b(x) \cdot \frac{e^{\beta\eta_b(x)}}{e^{\beta\eta_a(x) + e^{\beta\eta_b(x)}}}} \\
&= \eta_a(x) \cdot \frac{1}{1 + e^{-\beta(\eta_a(x) - \eta_b(x))} + \eta_b(x) \cdot \frac{1}{1 + e^{-\beta\left(\eta_b(x) - \eta_a(x)\right)}}} \\
&= \eta_a(x) \cdot \sigma\left[\beta\left(\eta_a(x) - \eta_b(x)\right)\right] + \eta_b(x) \cdot \sigma\left[\beta\left(\eta_b(x) - \eta_a(x)\right)\right] \\
&= \left(\eta_a(x) - \eta_b(x)\right) \cdot \sigma\left[\beta\left(\eta_a(x) - \eta_b(x)\right)\right] + \eta_b(x)
\end{aligned}
$$

$$\text{ACON-A: } S_\beta\left(\eta_a(x), \eta_b(x)\right) \text{ with } \eta_a(x) = x,\ \eta_b(x) = 0$$
$$\text{ACON-B: } S_\beta\left(\eta_a(x), \eta_b(x)\right) \text{ with } \eta_a(x) = x,\ \eta_b(x) = px$$
$$\text{ACON-C: } S_\beta\left(\eta_a(x), \eta_b(x)\right) \text{ with } \eta_a(x) = p_1 x,\ \eta_b(x) = p_2 x$$
$$\text{Meta-ACON: ACON-C with } \beta \text{ trainable parameter}$$

ACON Activation Function

Property of the ACON

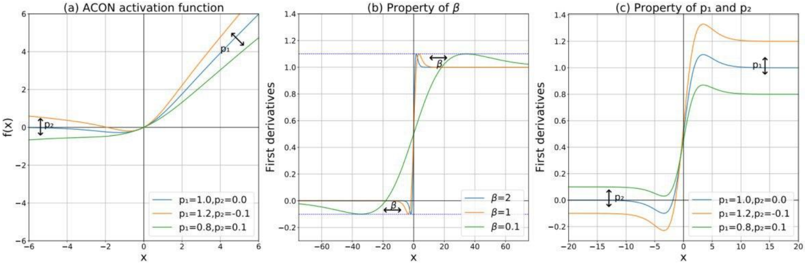


Figure 2: The ACON activation function and its first derivatives. (a) The ACON-C activation function with fixed β (see Fig. 3 for the influence of β); (b-c) The first derivatives with fixed $p_1 \& p_2$ (b) and fixed β (c). β controls how fast the first derivative asymptotes to the upper/lower bounds, which are determined by p_1 and p_2.

Results

Comparison of the meta-ACON

	ReLU			meta-ACON		
	FLOPs	# Params.	Top-1 err.	FLOPs	# Params.	Top-1 err.
MobileNetV1 0.25	41M	0.5M	47.6	41M	0.6M	**40.9** $_{(+6.7)}$
MobileNetV2 0.17	42M	1.4M	52.6	42M	1.9M	**46.2** $_{(+6.4)}$
ShuffleNetV2 0.5x	41M	1.4M	39.4	41M	1.7M	**34.8** $_{(+4.6)}$
MobileNetV1 0.75	325M	2.6M	30.2	326M	3.1M	**26.4** $_{(+3.8)}$
MobileNetV2 1.0	299M	3.5M	27.9	299M	3.9M	**25.0** $_{(+2.9)}$
ShuffleNetV2 1.5x	301M	3.4M	27.4	304M	6.0M	**24.7** $_{(+2.7)}$
ResNet-18	1.8G	11.7M	30.3	1.8G	11.9M	**28.4** $_{(+1.9)}$
ResNet-50	3.9G	25.5M	24.0	3.9G	25.7M	**22.0** $_{(+2.0)}$
ResNet-101	7.3G	44.1M	22.8	7.3G	44.1M	**21.1** $_{(+1.7)}$
ResNet-152	11.3G	60.0M	22.3	11.3G	60.1M	**20.5** $_{(+1.8)}$

Thank you for your attention!