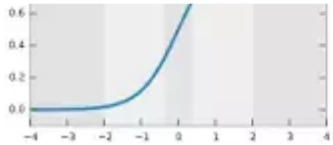
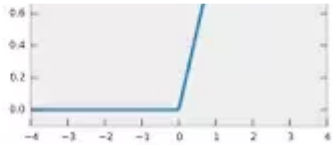
How does ELU activation function help convergence, and what's its advantages over ReLU or sigmoid or tanh function?

1 Answer

Saiprasad Koturwar, studied at Indian Institute of Technology, Bombay
Answered Sep 15, 2016

ELU(Exponential linear unit) function takes care of the Vanishing gradient problem. The other mentioned activation functions are prone to reaching a point from where the gradient of the functions does not change.

Lets start with the advantages that relu gives over sigmoid and tanh (tanh and sigmoid are simailar) so if you look at the following plot sigmoid/tanh gets saturated for large values of x, so as your activation value increases the corresponding gradient approaches zero and the corrosponding neurons effectively learn nothing. But with relu you do not have that problem as you will get finite gradient no matter what the value of x.

mean value of activation is zero you get on (which would have mean activation zero) your overall network becomes linear and which will effectively be equal to a single layer network and with linear networks there is very little that you can learn from the data, that's why we use non linear activation functions. Now what ELU does is that it tries to make the mean activation close to zero and as it is an exponential function it does not saturate(I have not used this activation function yet), you can conclude this from ELU graph.