What Is Saturating Gradient Problem

Asked 5 years, 3 months ago Modified 2 years ago Viewed 11k times

4

Can anyone explain what is `Saturating Gradient` problem? It would be nice if anyone can provide math details as well. Thank you in advance!

machine-learning neural-network deep-learning gradient-descent backpropagation

Share Improve this question Follow

edited Feb 10, 2018 at 8:42
Green Falcon
13.8k 9 54 96

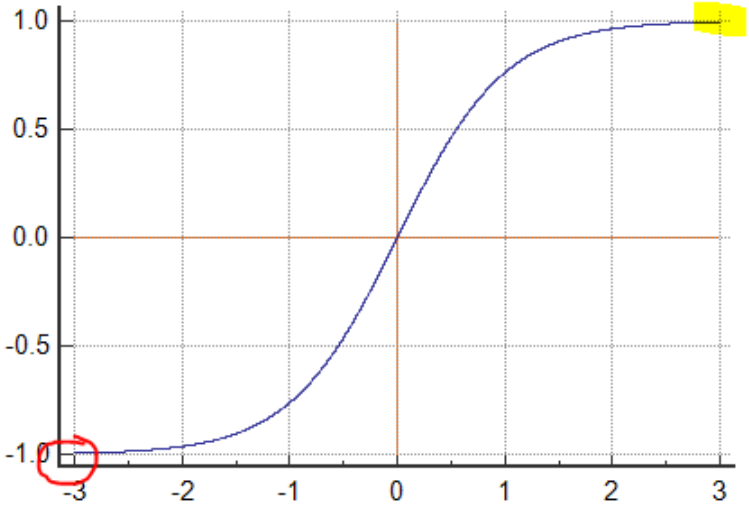
asked Feb 10, 2018 at 8:15
Stefan Radonjic
716 1 7 20

3 Answers

Sorted by: Highest score (default)

6

If you use sigmoid-like activation functions, like sigmoid and tanh, after some epochs of training, the linear part of each neuron will have values that are very big or very small. This means that the linear part will have a big output value regardless of its sign. Consequently, the input of sigmoid-like functions in each neuron which adds non-linearity will be far from the center of these functions.



In those locations, the gradient/derivative value is very small. Consequently, after numerous iterations, the weights get updated so slowly because the value of the gradient is very small. This is why we use the `ReLU activation function` for which its gradient doesn't have this problem. Saturating means that after some epochs that learning happens relatively fast, the value of the linear part will be far from the center of the sigmoid and it somehow saturates, and it takes too much time to update the weights because the value of gradient is small. You can take a look at here as a solution for this problem.

If I want to explain the math part, suppose that you are using sigmoid as the activation function. If σ represents sigmoid, its gradient is $\sigma(1-\sigma)$. Now suppose that your linear part, the input of sigmoid is a positive number which is too large, then sigmoid which is:

$$\frac{1}{1+e^{-x}}$$

will have a value near to one but smaller than that. On the other hand σ will be so close to zero, multiplying σ and $1-\sigma$ will result in a small value, which means the value of the gradient is very small. If the value of the linear part is very small, then σ will be close to zero and $1-\sigma$ will be close to 1 but smaller than that. Again, multiplying these will lead to a small value as the gradient.

Share Improve this answer Follow

edited Feb 21, 2021 at 13:25

cwallenwein
107 5

answered Feb 10, 2018 at 8:32

Green Falcon
13.8k 9 54 96

1 Thank you for clarifying that for me. I now understand perfectly what saturating means :) – Stefan Radonjic Feb 10, 2018 at 8:57

What if it happens with relu activation also? – Elbek Nov 13, 2019 at 3:48

@Elbek but its derivative is always one for positive inputs. – Green Falcon Nov 13, 2019 at 11:10

1

In neural networks, activation functions such as the logistic (sigmoid) and the hyperbolic tangent functions map any real values to a compact range of values. For example, the sigmoid function, S(x)= 1/(1+ e^(-x)) maps a set of real values x to between 0 and 1. To attain these boundaries of either 0 or 1, large magnitude negative or positive values of x are required. Therefore, a neuron is said to be saturated when extremely large weights cause the neuron to produce values (gradients) that are very close to the range boundary. If the gradient is constantly 0, no learning will take place in the neural network. Likewise, if the gradient is constantly 1, it most likely means that the neuron is over-fitting on training data and will likely perform poorly on test data.

Share Improve this answer Follow

edited May 18, 2021 at 19:28

answered May 18, 2021 at 19:15

rocksyne
111 3

0

According to the Cambridge Dictionary saturation means

> the act or result of filling a thing or place completely so that no more can be added

In this context, it refers to a function for which a bigger input will not lead to a relevant increase in output. So if the gradient is saturated (meaning it is extremely close to zero), a bigger upstream gradient doesn't lead to a bigger current gradient when applying the chain rule.

Share Improve this answer Follow

answered Feb 21, 2021 at 8:28

cwallenwein
107 5

Or it could mean that the gradient for $\sigma(1000)$ is not much different from the gradient for $\sigma(1001)$ – cwallenwein Feb 21, 2021 at 8:45