REVISE SATURATED ACTIVATION FUNCTIONS

Arnaud WATUSADISI MAVAKALA

AMMI, AIMS-Senegal

August 20, 2021

Overview

Overview

INTRODUCTION

Motivation

The motivation was to contradict the conclusion of some previous work that the saturation property is the cause of slow convergence.

Overview

INTRODUCTION

Objectives

1. The objective was to review two commonly used saturated functions, the logistic sigmoid and the hyperbolic tangent (tanh).
2. We show that, with appropriate rescaling, the logistic sigmoid achieves comparable results to the tanh function.
3. Then, following the same argument, we improve tahn by penalizing the negative part.
4. We show that "penalized tanh" is comparable and even outperforms unsaturated functions

Overview

WHY TRAINING DEEP NEURAL NETWORKS IS HARD WITH THE LOGISTIC SIGMOID

- its slope in the linear regime is 1/4 rather than 1, so we need to initialize the weight is a 16 times smaller variance to keep the variance of the gradient of each layer the same.
- it has a non-zero mean, so the output variance increases linearly with the layer.

Rescale the logistic sigmoid to match the coefficient of the first two degrees with tanh and relu, In other words, we transform the logistic sigmoid by

$$scaled\ sigmoid(x) = 4 * sigmoid(x) - 2 = \frac{4}{1 + e^{x}} - 2$$

WHY TRAINING DEEP NEURAL NETWORKS IS HARD WITH THE LOGISTIC SIGMOID

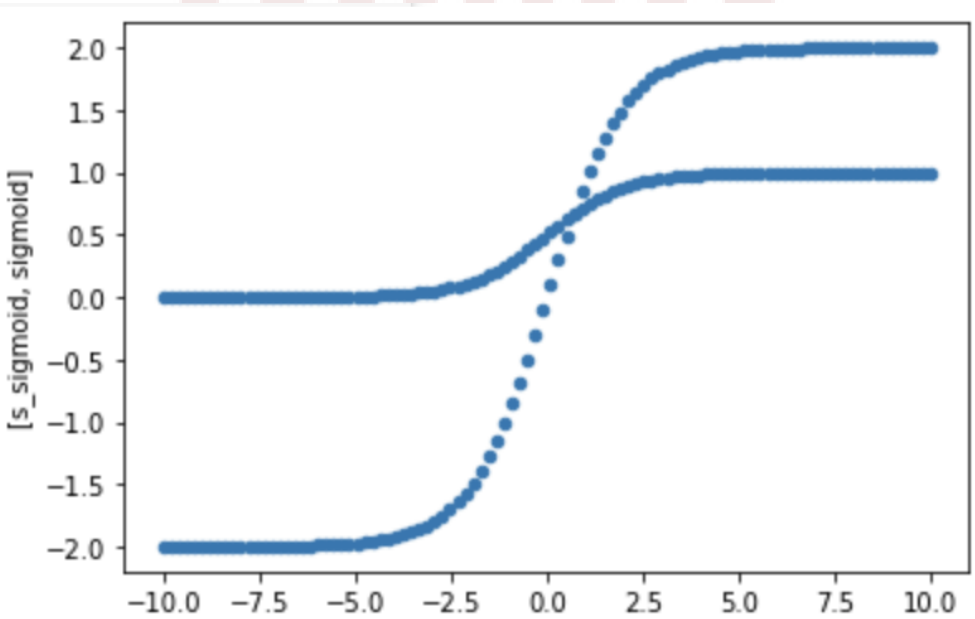

Figure: Scaled sigmoid and sigmoid

PENALIZED SATURATED ACTIVATION FUNCTIONS

the "leaky ReLU", which is presented as follows,

$$f(x) = \begin{cases} x & \text{if if } x > 0 \\ a * x & \text{otherwise} \end{cases}$$

the leaky ReLU can be seen as an improvement over the simple identical activation function f(x) = x that penalizes the gradient of the negative part. We propose the function "penalized tanh" which takes the form of

$$f(x) = \begin{cases} tanh(x) & \text{if if } x > 0 \\ a * tanh(x) & \text{otherwise} \end{cases}$$

Overview

Comparison

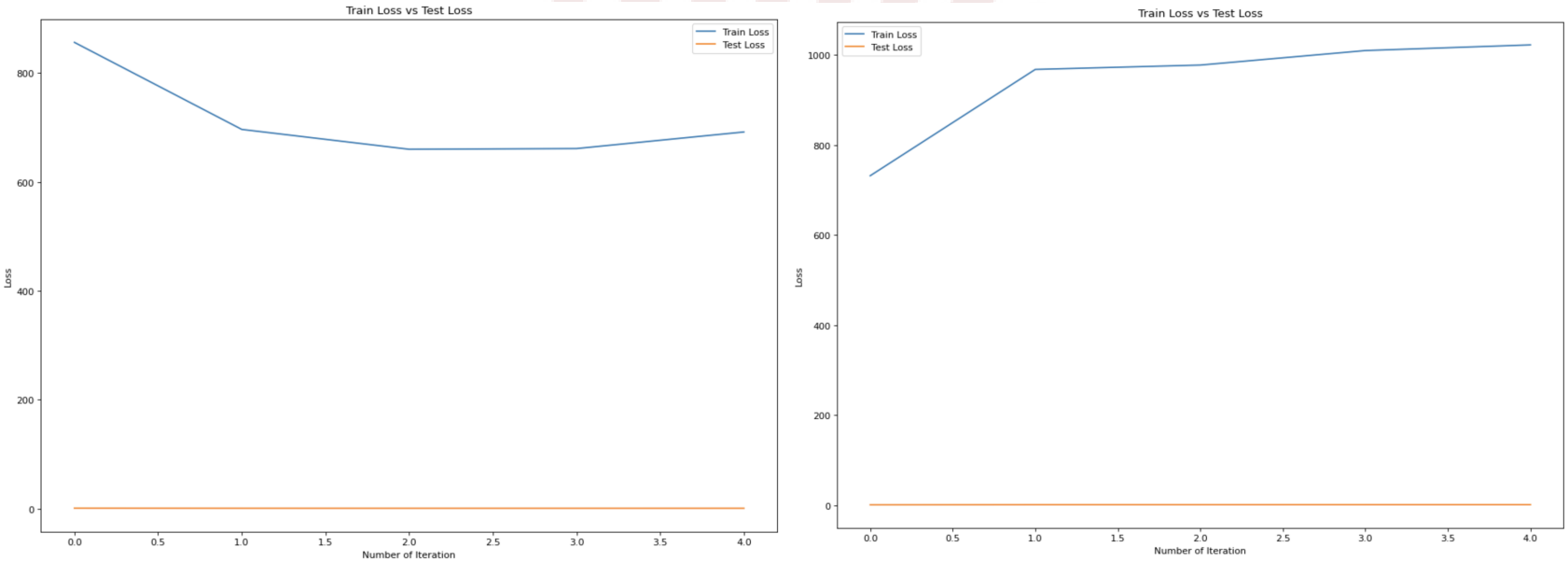

Figure: sigmoid and Scaled sigmoid

Comparison

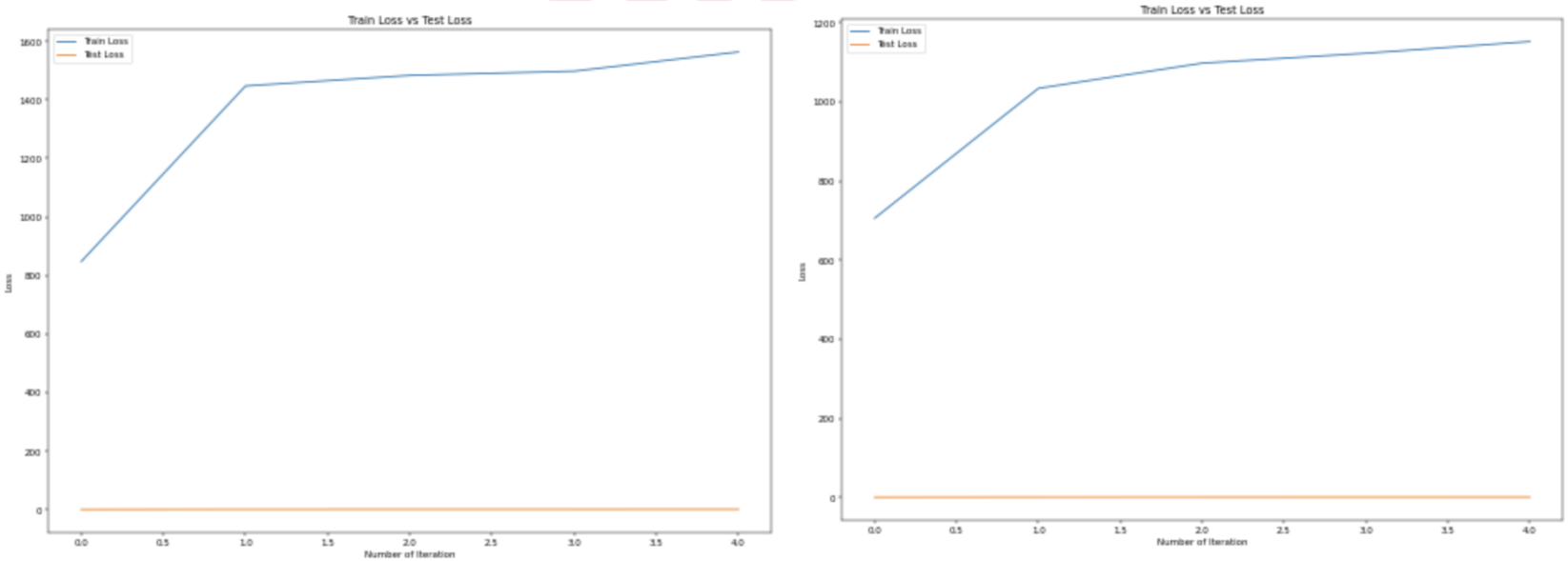

Figure: Tanh and penalized Tanh

Comparison

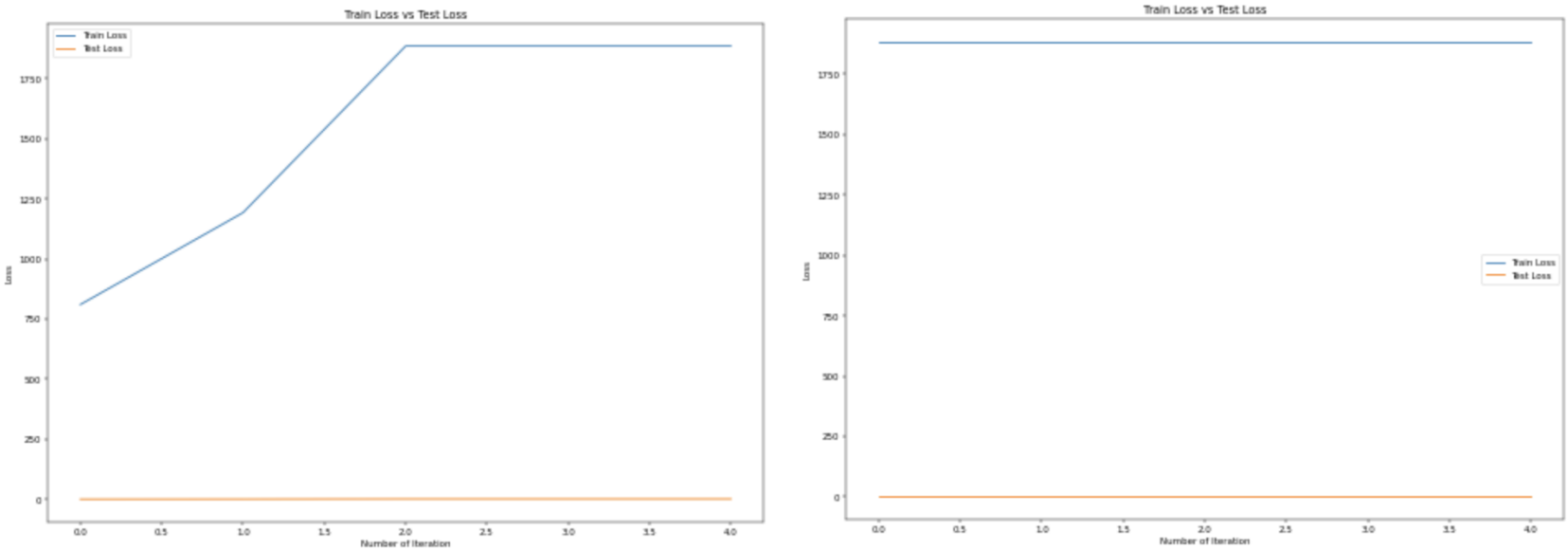

Figure: ReLu and Leaky ReLu

CONCLUSION

The outcome of the paper was to study the differences in network performance between using saturated activation functions and using unsaturated functions. The result found suggests that using the penalty on the negative part, the saturation of the activation function is comparable to ReLU and Leaky ReLU.

REFERENCES

- Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274, 2015.
- Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In International conference on artificial intelligence and statistics, pp. 249–256, 2010.
- Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853, 2015.

Thank you for your attention