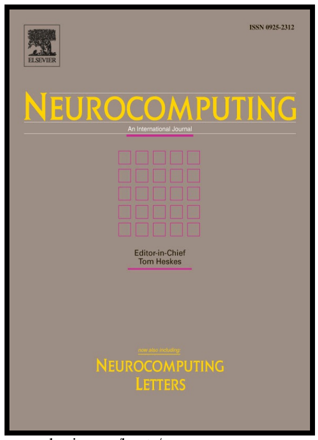
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

Bounded Activation Functions for Enhanced Training Stability of Deep Neural Networks on Visual Pattern Recognition Problems

Shan Sung Liew, Mohamed Khalil-Hani, Rabia Bakhteri

PII: S0925-2312(16)30879-7
DOI: http://dx.doi.org/10.1016/j.neucom.2016.08.037
Reference: NEUCOM17449

To appear in: *Neurocomputing*

Received date: 13 January 2016
Revised date: 9 August 2016
Accepted date: 10 August 2016

Cite this article as: Shan Sung Liew, Mohamed Khalil-Hani and Rabia Bakhteri Bounded Activation Functions for Enhanced Training Stability of Deep Neura Networks on Visual Pattern Recognition Problems, *Neurocomputing* http://dx.doi.org/10.1016/j.neucom.2016.08.037

Bounded Activation Functions for Enhanced Training Stability of Deep Neural Networks on Visual Pattern Recognition Problems

Shan Sung Liew[a,*], Mohamed Khalil-Hani[a], Rabia Bakhteri[b]

[a] VeCAD Research Laboratory, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Machine Learning Developer Group, Sightline Innovation, #202, 435 Ellice Ave, Winnipeg, MB R3B 1Y6, Canada

Abstract

This paper focuses on the enhancement of the generalization ability and training stability of deep neural networks (DNNs). New activation functions that we call bounded rectified linear unit (ReLU), bounded leaky ReLU, and bounded bi-firing are proposed. These activation functions are defined based on the desired properties of the universal approximation theorem (UAT). An additional work on providing a new set of coefficient values for the scaled hyperbolic tangent function is also presented. These works result in improved classification performances and training stability in DNNs. Experimental works using the multilayer perceptron (MLP) and convolutional neural network (CNN) models have shown that the proposed activation functions outperforms their respective original forms in regards to the classification accuracies and numerical stability. Tests on MNIST, *mnist-rot-bg-img* handwritten digit, and AR Purdue face databases show that significant improvements of 17.31%, 9.19%, and 74.99% can be achieved in terms of the testing misclassification error rates (MCRs), applying both mean squared error (MSE) and cross-entropy (CE) loss functions This is done without sacrificing the computational efficiency. With the MNIST dataset, bounding the output of an activation function results in a 78.58% reduction in numerical instability, and with the mnist-rot-bg-img and AR Purdue databases the problem is completely eliminated. Thus, this work has demonstrated the significance of bounding an activation function in helping to alleviate the training instability problem when training a DNN model (particularly CNN).

Keywords: Activation function, output boundary, generalization performance, training stability, deep neural network, convolutional neural network.

1. Introduction

Since the introduction of artificial neural networks (ANNs) in the last few decades, many research works have been conducted to explore the capability of ANNs in various reasoning and decision based applications. Inspired by how a biological brain works, an ANN possesses the ability to learn from experience, enabling it to solve sophisticated problems that are deemed too difficult for conventional methods.

The expressiveness and approximation properties of a neural network (NN) often rely heavily on its structure. It is believed that deeper models are more powerful in approximating functions, and can capture more information due to their large learning capacities. This is motivated from the biological point of view, that in general our brains and visual systems are composed of multiple stages of information processing. This leads to the development of the deep neural network (DNN) model, which can provide better and deeper abstraction.

In general, there are three main features that characterize a DNN: its model or architecture (connection pattern between neurons), the corresponding learning algorithm (method to search for the optimum weights) and its activation function (transfer function that provides the nonlinearity behavior [1, 2]). Different connection patterns can create different types of neuron layers that exhibit various characteristics. Examples of neuron layers include convolutional and pooling layers. A suitable learning algorithm is usually decided upon the DNN model based on how it can lead to a better convergence within the shortest possible time period.

However, the impact of an activation function on the generalization performance and training stability of a DNN is often ignored. There is a lack of consensus on how to select a good activation function for an NN model, and a specific one may not be suitable for all applications. This is especially true for problem domains where the numerical boundaries of the inputs and outputs are the main considerations.

In addition, the training process is heavily dependent on the choice of the activation function. As most supervised learning algorithms are based on the backward propagation of the error gradients, the tendency at which an activation function saturates during the back-propagation is one of the main concerns. Output saturation can result in poorer convergence, which is undesirable. Also, since an activation function is applied to the outputs of all neurons in most cases, its computational complexity will contribute heavily to the overall execution time.

Training stability of an NN model defines the magnitude of diversity detected during the training process [3]. A good train-

*Corresponding author. Tel.: +60 168588653
Email addresses: `ssliew2@live.utm.my` (Shan Sung Liew), `khalil@fke.utm.my` (Mohamed Khalil-Hani), `rbakhteri@sightlineinnovation.com` (Rabia Bakhteri)