Adaptive Smooth Rectifier

MENGYING SUN, Michigan State University
XIAORAN TONG, Michigan State University

Through this project we explore an alternative form of the rectifier activation units of adaptive smoothness, and, with such alternative, the possibility of relatively lightweighted networks out perfroming the mainstream deep networds under problem domains other than image classification.

General Terms: AI, Algorithms, Performance

Additional Key Words and Phrases: ReLU

ACM Reference format:
Mengying Sun and Xiaoran Tong. 2017. Adaptive Smooth Rectifier. 1, 1, Article 1 (January 2017), 2 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The popular trend in the filed of Machine Learning has always been pursuing a "deeper" and "bigger" Artificial Neural Network (ANN), known as Deep Learning, ever since the revival of the classic Multiple Layered Perceptron (MLP) around 2006 [4]. Although the topology of an MLP and its capability of emulating functions of any complexity is well known for decades, going towards depth of tens and hundreds of layers was made possible by the greedy unsupervised pre-training [4], aided by the improvement in computation power and an explosion of data. Thansk to the introduction of rectified linear unit (ReLU) as a replacement of the classical sigmoid neurons [1], recently networks do not have to go through the pre-training procedure, and has been able to push their depth to a new high, at times even more than a thousand layers. As the current dominant trend, ReLU offers a number of advantages [1]. From the the perspective of neural computation, the asymmetric, half dominant activation curve can better mimics the actual biological activation of the neurons in the central neural system (CNS), plus, the network as whole respects the fact that only a small fraction the CNS is active at any moment[1]. From the function searching point of view, ReLU promotes model sparsity at neurons level that helps to reduce overfit, which means more room for deeper structures given the same tolerance of total complexity (e.g., the number of active nodes and connections at any time). ReLU also partially solved the issue of diminishing gradient when the large input signal is given to the traditional sigmoid or tangent activation units, which also elevates the potential of going "deeper".

Despite many victory scored by deep ANN [6, 9], going deeper does not come without a cost. Aside from the computation load and the memory constraint of the GPUs, the network still tends to overfit when the depth keeps growing, a recent trend is to randomly "dropout" a portion of neurons at any update cycle to promote each neuron extract different features, which improves generalization performce by a large margin [5]. Another noteworthy issue is the stuck of backpropagation signals since the neurons can be seen act as resistance, which could drastically chock the gradient from flowing further backwards if many such units in the frontal layers are in dorment state. Topological counter measures has been proposed along the way, such as residual learning based short circuiting [3] to help the flow of gradient, the companion trainers [7] to promote uniformed evolution, and the conglomeration of more than one the topological tweaks through implict ensembling [8], all of which aim to prevent part of of network, especially those of lower depth from stagnation. Asides from structure improvement, [2] also proposed a more general form of ReLU called PRelu, such that the negative part of the activation allows non-constant, thus avoiding 0 gradient. These augmentations, including the very adaptation of ReLU itself, all strictly follows the central doctrine that "deeper is better", which might be approximated or even out performed in domains other than imaging, acoustic, and texture based classification, by shallower networks of alternative structure components.

2 PROPOSAL

Motivated by [2], which grants the negtive part of the ReLU with non-zero gradient to prevent gradient from completely diminishing, we propose a adaptive form of softplus (ASP) to compare with ReLU. ASP is differentiable over $\mathcal{R}$, with trainable parameters to control its smoothness. In such way it allow the data to decide the proper shape of the activation unit, which can be as sharp as the mainstream ReLU, or as smooth as the softplus, while retaining advantages like asymmetry and high gradient upon large positive input that come with Relu.

In this study we propose to expriment networks that is not very deep under the question other than common benchmarks. One candidate domain is the disease prediction with genome that may favor smaller networks, since the data itself is extremely sparse thanks to the deep sequencing technology, yet highly correlated due to linkage disequilibrium. We would also explore the shallow network's performance with semi-supervised encoding tasks (i.e., non-linear PCA).

We would assess the property of the proposed ASP in comparison with Relu via the following criteria

- generalization error and convergence rate;
- The network sparsity under fixed or flexible smoothing parameter.

These criteria will be measured under the following scenarios

- Image classification (reference);
- Genome based prediction;
- Genome auto-endocing.

This work is supported by the National Science Foundation, under grant CNS-0435060, grant CCR-0325197 and grant EN-CS-0329609.
Author's addresses: G. Zhou, Computer Science Department, College of William and Mary; Y. Wu and J. A. Stankovic, Computer Science Department, University of Virginia; T. Yan, Eaton Innovation Center; T. He, Computer Science Department, University of Minnesota; C. Huang, Google; T. F. Abdelzaher, (Current address) NASA Ames Research Center, Moffett Field, California 94035.

For these project, we use fully connected networks.

REFERENCES

[1] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks.. In *Aistats*, Vol. 15. 275.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] G. E. Hinton and R. R. Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507. DOI: http://dx.doi.org/10.1126/science.1127647 arXiv:http://science.sciencemag.org/content/313/5786/504.full.pdf

[5] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580 (2012). http://arxiv.org/abs/1207.0580

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[7] Chen-Yu Lee, Saining Xie, Patrick W Gallagher, Zhengyou Zhang, and Zhuowen Tu. 2015. Deeply-Supervised Nets.

[8] Saurabh Singh, Derek Hoiem, and David Forsyth. 2016. Swapout: Learning an ensemble of deep architectures. In *Advances in Neural Information Processing Systems*. 28–36.

[9] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.