國立臺灣大學電機資訊學院電信工程學研究所

碩士論文

Graduate Institute of Communication Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

影像中使用者感興趣區域偵測之資料集
Vanishing Node

張文于
Wen-Yu Chang

指導教授：林宗男
Advisor: Tsung-Nan Lin

中華民國 108 年 7 月
July, 2019

國立臺灣大學碩士學位論文

口試委員會審定書

影像中使用者感興趣區域偵測之資料集

Vanishing Node

本論文係張文于君 (R06942064) 在國立臺灣大學電信工程學研究所完成之碩士學位論文，於民國 108 年 7 月 8 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

所　長：

誌謝

感謝…

Acknowledgements

I’m glad to thank…

摘要

關鍵字： 深度學習

Abstract

It is well known that the problem of vanishing/exploding gradients creates a challenge when training deep networks. In this paper, we show another phenomenon, called *vanishing nodes*, that also increases the difficulty of training deep neural networks. As the depth of a neural network increases, the network's hidden nodes show more highly correlated behavior. This correlated behavior results in great similarity between these nodes. The redundancy of hidden nodes thus increases as the network becomes deeper. We call this problem "*Vanishing Nodes*." This behavior of vanishing nodes can be characterized quantitatively by the network parameters, which is shown analytically to be proportional to the network depth and inversely proportional to the network width. The numerical results suggest that the degree of vanishing nodes will become more evident during back-propagation training. Finally, we show that vanishing/exploding gradients and vanishing nodes are two different challenges that increase the difficulty of training deep neural networks.

Keywords: Deep Learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep neural networks (DNN) have succeeded in various fields, including computer vision [6], speech recognition [5], machine translation [11], medical analysis [10] and human games [9]. Some results are comparable to or even better than those of human experts.

State-of-the-art methods in many tasks have recently used increasingly *deep* neural network architectures. The performance has improved as networks have been made *deeper*. For example, some of the best-performing models [3, 4] in computer vision have included hundreds of layers.

Moreover, recent studies have found that as the depth of a neural network increases, problems such as vanishing or exploding gradients make the training process more challenging. [1, 2] investigated this problem deeply and suggested that initializing weights in appropriate scales can prevent gradients from vanishing or exploding exponentially. [7, 8] also studied how vanishing/exploding gradients arise via *mean field theory* and provided a solid theoretical discriminant to determine whether the propagation of gradients is vanishing/exploding.

Inspired by previous studies, we investigated the correlation between hidden nodes and discovered that a phenomenon that we call *vanishing nodes* can also affect the capability of a neural network. In general, the hidden nodes of a neural network become highly correlated as the network becomes deeper. The correlation between nodes implies the similarity between them, and high degree of similarity between nodes produces redundancy. Because a sufficient number of effective nodes is needed to approximate an arbitrary function, the redundancy of nodes in hidden layers may debilitate the representation capability of the entire network. Thus, as the depth of the network increases, the redundancy of hidden nodes may increase and hence affect the network's trainability. We name this phenomena as "*Vanishing Nodes*."

We propose a *Vanishing Node Indicator (VNI)*, which is the weighted average of squared correlation coefficients, as the quantitative metric for vanishing nodes. VNI can be theoretically approximated via the results on the spectral density of the end-to-end Jacobian. The approximation of VNI depends on the network parameters, including the width, the depth, the distribution of weights, and the activation functions, and it is shown to be simply proportional to the network depth and inversely proportional to the network width.

In addition, the numerical results show that back-propagation training also intensifies the correlations of hidden nodes when we consider a deep network. We find that although we use a relatively large network width, the correlations of hidden nodes may still increase during the training process.

Finally, we show that vanishing/exploding gradients and vanishing nodes are two different problems, so that the two problems may arise from specific conditions. The experimental results show that the likelihood of failed training increases as the depth of the network increases. The training will become much more difficult due to lack of network representation capability.

This paper is organized as follows: some related works are discussed in Section **??**. The vanishing nodes phenomenon is introduced in Section **??**. Theoretical analysis and a quantitative metric are reported in Section **??**. Section **??** compares the vanishing nodes with vanishing/exploding gradients. Section **??** reports the experimental results and Section **??** gives our conclusions.

Bibliography

[1] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 9:249–256, 2010.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *IEEE International Conference on Computer Vision*, 2015.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision*, pages 630–645, 2016.

[5] G. Hinton, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, B. Kingsbury, and T. Sainath. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29:82–97, 2012.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1106–1114, 2012.

[7] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Neural Information Processing Systems*, 2016.

[8] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. *International Conference on Learning Representations (ICLR)*, 2017.

[9] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, I. S. Nal Kalchbrenner, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 529(7587):484–489, 2016.

[10] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[11] Y. Wu, M. Schuster, Z. Chen, M. N. Quoc V. Le, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap. *arXiv:1609.08144*, 2016.