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

神經元消失：使深層神經網路難以訓練的另一種現象

Vanishing Nodes: Another Phenomena That Makes Training Deep Neural Networks Difficult

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

神經元消失：使深層神經網路難以訓練的另一種現象

Vanishing Nodes: Another Phenomena That Makes Training Deep Neural Networks Difficult

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, 梯度消失, 機器學習理論

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, Vanishing gradient, Learning theory

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Deep neural networks (DNN) have succeeded in various fields, including computer vision [?], speech recognition [?], machine translation [?], medical analysis [?] and human games [?]. 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 [?, ?] 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. [?, ?] investigated this problem deeply and suggested that initializing weights in appropriate scales can prevent gradients from vanishing or exploding exponentially. [?, ?] 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.

Chapter 2

Related Work

Problems in the training of deep neural networks have been encountered in several studies. For example, [?, ?] investigated vanishing/exploding gradient propagation and gave weight initialization methods as the solution. [?] suggested that vanishing/exploding gradients might relate to the sum of the reciprocals of the hidden layer widths. [?, ?] stated that saddle points are more likely than local minima to be a problem for training deep neural networks. [?, ?, ?] exposed the *degradation* problem: the performance of a deep neural network degrades as the depth increases.

The correlation between the nodes of hidden layers within a deep neural network is the main focus of this paper, and several kinds of correlations have been discussed in the literature. [?] surveyed the propagation of the correlation between two different inputs after several layers. [?, ?] suggested that the input features must be whitened (i.e., zero-mean, unit variances and uncorrelated) to achieve a faster training speed.

Dynamical isometry is one of the conditions that make ultra-deep network training more feasible. [?] reported dynamical isometry to theoretically ensure depth-independent learning speed. [?, ?] suggested several ways to achieve dynamical isometry for various settings of network architecture, and [?, ?] practically trained ultra-deep networks in various tasks.

Chapter 3

Vanishing Nodes: correlation between hidden nodes

In this section, the correlation of hidden-layer neurons is investigated. If a pair of neurons is highly correlated (for example, the correlation coefficient is equal to $+1$ or -1), one of the neurons becomes redundant. Great similarity between nodes may reduce the effective number of neurons within a network. In some cases, the correlation of hidden nodes may disable the entire network. This phenomenon is called *Vanishing Nodes*.

First, consider a deep feed-forward neural network with depth L. For simplicity of analysis, we assume all layers have the same width N. The weight matrix of layer l is $\mathbf{W}_l \in \mathbb{R}^{N\times N}$, the bias of layer l is $\mathbf{b}_l \in \mathbb{R}^N$ (a column vector), and the common activation function of all layers is $\phi(\cdot) : \mathbb{R} \to \mathbb{R}$. The input of the network is $\mathbf{x}_0$, and the nodes at output layer L denote $\mathbf{x}_L$. The pre-activation of layer l is $\mathbf{h}_l \in \mathbb{R}^N$ (a column vector), and the post-activation of layer l is $\mathbf{x}_l \in \mathbb{R}^N$ (a column vector). That is, $\forall l \in \{1, ..., L\}$,

$$\mathbf{h}_l = \mathbf{W}_l \mathbf{x}_{l-1} + \mathbf{b}_l, \quad \mathbf{x}_l = \phi(\mathbf{h}_l). \tag{3.1}$$

The variance of node i is defined as $\sigma_i^2 \triangleq \mathbb{E}_{\mathbf{x}_0}[(x_{l(i)} - \overline{x_{l(i)}})^2]$, and the squared correlation coefficient (ρ_{ij}^2) between nodes i and j can be computed as $\rho_{ij}^2 \triangleq \frac{\mathbb{E}_{\mathbf{x}_0}[(x_{l(i)}-\overline{x_{l(i)}})(x_{l(j)}-\overline{x_{l(j)}})]^2}{\mathbb{E}_{\mathbf{x}_0}[(x_{l(i)}-\overline{x_{l(i)}})^2]\mathbb{E}_{\mathbf{x}_0}[(x_{l(j)}-\overline{x_{l(j)}})^2]}$, where ρ_{ij}^2 ranges from 0 to 1. Nodes $x_{l(i)}$ and $x_{l(j)}$ are highly correlated only if the magnitude of the correlation coefficient between two nodes ρ_{ij} is nearly 1. ρ_{ij}^2 indicates the magnitude of similarity between node i and node j. If ρ_{ij} is close to $+1$ or -1, then node i can be approximated in a linear fashion by node j. Great similarity indicates redundancy. If nodes of hidden layers exhibit great similarity, the effective number of nodes will be much lower than the original network width. Therefore, we call this phenomena *Vanishing Node Problem*.

In the following section, we propose a metric to measure the quantitative property of vanishing nodes for a deep feed-forward neural network. Theoretical analysis of the metric indicates that the quantitative property of vanishing nodes is proportional to the network depth and inversely proportional to the network width. The quantity is shown analytically to depend on the statistical property of weights and the nonlinear activation function.

3.1 Vanishing Node Indicator

Consider the network architecture defined in eqn. (**??**). In addition, the following assumptions are made: (1) The input $\mathbf{x}_0$ is zero-mean, and the features in $\mathbf{x}_0$ are independent and identically distributed. (2) All weight matrices $\mathbf{W}_l$ in each layer are initialized from the same distribution with variance σ_w^2/N. (3) All the bias vectors $\mathbf{b}_l$ in each layer are initialized to zero.

The input-output Jacobian matrix $\mathbf{J} \in \mathbb{R}^{N \times N}$ is defined as the first-order partial derivative of the output layer with respect to the input layer, which can be rewritten as $\frac{\partial \mathbf{x}_L}{\partial \mathbf{x}_0} = \prod_{l=1}^{L} \mathbf{D}_l \mathbf{W}_l$, where $\mathbf{D}_l \triangleq diag(\phi'(\mathbf{h}_l))$ is the derivative of point-wise activation function ϕ at layer l. To conduct a similar analysis as [**?**], consider the first-order forward approximation: $\mathbf{x}_L - \overline{\mathbf{x}_L} \approx \mathbf{J}\mathbf{x}_0$. Therefore, the covariance matrix of the nodes ($\mathbf{C} \in \mathbb{R}^{N \times N}$) at the output layer can be computed as

$$\mathbf{C} \triangleq \mathbb{E}_{\mathbf{x}_0}[(\mathbf{x}_L - \overline{\mathbf{x}_L})(\mathbf{x}_L - \overline{\mathbf{x}_L})^T] \approx \mathbb{E}_{\mathbf{x}_0}[(\mathbf{J}\mathbf{x}_0)(\mathbf{J}\mathbf{x}_0)^T] = \mathbf{J}\mathbb{E}_{\mathbf{x}_0}[\mathbf{x}_0\mathbf{x}_0^T]\mathbf{J}^T = \sigma_x^2 \mathbf{J}\mathbf{J}^T, \quad (3.2)$$

where σ_x^2 is the common variance of features in $\mathbf{x}_0$, and the expected values are calculated with respect to the input $\mathbf{x}_0$. For notational simplicity, we omit the subscript $\mathbf{x}_0$