GEOMETRIC MODELLING AND UNDERSTANDING OF DEEP NEURAL NETWORKS

PIYUSH KAUL

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

GEOMETRIC MODELLING AND UNDERSTANDIBG OF DEEP NEURAL NETWORKS

A THESIS

submitted by

Piyush Kaul

2015EEZ7544

under the guidance of

Prof. Brejesh Lall

for the award of the degree

of

Doctor of Philosophy

Department Of Electrical Engineering

INDIAN INSTITUTE OF TECHNOLOGY DELHI

April 2023

CERTIFICATE

This is to certify that the thesis titled **Geometric Modelling and Understanding of Deep Neural Networks**, submitted by **Piyush Kaul**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. Brejesh Lall
Professor
Deptartment of Electrical Engineering
IIT-Delhi, 600 036

ACKNOWLEDGEMENTS

This research work would be impossible without the support and guidance of some amazing people. I would like to take the opportunity to thank all who guided, supported, and collaborated with me on this memorable journey. First and foremost, I would like to thank my respectable advisor Prof. Brejesh Lall for his wise guidance, unwavering support, and endless encouragement throughout this work. He helped in numerous ways on all aspects of this work, both showing the way throughout this research work and more generally on numerous personal and professional issues. Without his constant supervision and able guidance, this work would not have been a reality. I would also like to thank my Student Research Committee (SRC) members Prof. Shiv Dutt Joshi, Prof. Alok Kumar, and Prof. Sumantara Dutta Roy for their valuable comments, suggestions, and constant encouragement during my bi-yearly seminars. I would also like to convey many thanks to the peer reviewers of my work for their precious time, support and suggestions. I would also like to thank all the lab staff and my fellow Ph.D. and MSR students at IIT Delhi who were always there with their help and guidance. I would also acknowledge the excellent facilities and resources provided by IIT Delhi. Finally, I would also like to thank my mother and wife who had to sacrifice a lot to enable me to pursue this research. I would also like to remember my late father, who has been a constant source of inspiration for me.

Piyush Kaul

ABSTRACT

KEYWORDS: Deep Learning; Riemannian Geometry; Curvature; Information Geometry; Information Bottleneck; Natural Gradient Descent;

Deep neural networks are currently the best-performing algorithms for a variety of pattern recognition tasks, including object detection, object classification, speech recognition, etc. However, theoretically, these are still poorly understood due to a lack of mathematical models for the entire network, which hinders predictably improving and customizing these for new tasks. We utilize concepts from differential geometry, signal processing theory, and information theory. to develop an understanding of deep neural networks to better analyze, understand, and visualize their effectiveness in varied pattern recognition tasks. We also utilize these concepts for introducing improvements for better and faster training of neural networks. Lastly, we also show the interrelationship between existing theoretical paradigms of neural networks to get a broader and deeper understanding of how and why neural networks work in all aspects as described above.

We give a background on learning theory, curvature and linearization, information bottleneck theory, and information geometry which is required to understand our work on neural networks. We then explain our contribution to understanding the curvature view of neural networks by utilizing the theory of Riemannian geometry and direct measurement of Riemannian curvature. We propose a method for the calculation of Riemannian curvature of neural networks using pre-defined transformations and analyzing the behavior from derived Ricci scalers. These methods are empirically tested with various datasets and transformations, and the observations are interpreted. We also explain how this curvature view can be further utilized to define geodesics on Riemannian manifolds to see the impact of training on learning. We next explain the theory of information geometry and the concept of Fisher information as applied to neural networks and explain how it can be utilized and optimized. We explain the Kronecker factorization-based method of optimizing the calculation of natural gradient descent. We share a new algorithm, namely projective Kronecker factor approximate curvature, and show how it can help in faster convergence using a low-complexity implementation of natural gradient descent. We also estimate the complexity of

the new algorithm and compare it with state of art both analytically and experimentally.

We next explain how information theory can be utilized in form of the information bottleneck principle to understand neural network architecture and the impact of training on generalization. We further extend the interpretation specifically to explain the effectiveness of skip connections with both MLP and CNN-type architectures. We also define new metrics based on information bottleneck theory and do measurements for multiple architectures to show the impact of skip connections on learning. We finally look at the inter-relationship between the mutual information-based information plane interpretation of neural networks and the fisher information-based interpretation. We look at how the fisher information and mutual information are related and analyze the relationships in the context of autoencoders to show how the autoencoder behaves differently with no generalization phase for the latter part of the network.

Contents

List of Tables

List of Figures

ABBREVIATIONS

AMGD	Average Minimum Geodesic Distance
AGD	Average Geodesic Distance
CNN	Convolutional Neural Network
DPI	Data Processing Inequality
DNN	Deep Neural Network
ERM	Empirical Risk Minimization
EDGE	Edge Dependency Graph Estimator
ESPCN	Efficient sub-pixel convolutional neural network
FIM	Fisher Information Matrix
HR	High Resolution
IB	Information Botttleneck
IL	Information Lagrangian
KFAC	Kronecker Factored Approximate Curvature
KLD	Kullback–Leibler divergence
KFC	Kronecker Factored Curvature for Convolutional Networks.
LR	Low Resolution
LSTM	Long Short Term Memory
MLP	Multi-Layer Perceptron
MSE	Mean Square error
NGD	Natural Gradient Descent
PCA	Principal Component Analysis
PKFAC	Projective Kronecker Factored Approximate Curvature
PSNR	Peak Signal to Noise Ratio
RBF	Restricted Boltzmann Machine.
ResCNN	Resnet-like-skip CNN
ResMLP	Resnet-like-skip MLP
SGD	Stochastic Gradient Descent

SIFT Scale Invariant Feature Transform

SVD Singular Value Decomposition

NOTATION

$\mathbf{h}$	Prediction function
$\mathbf{X}$	Input to network/layer
$\mathbf{Y}$	Output of network/layer
$\mathscr{L}_{\mathbf{d}}(\mathbf{h})$	Loss function over dataset d and prediction function h
$\varepsilon_{\mathbf{app}}$	Approximation Error
$\varepsilon_{\mathbf{est}}$	Estimation Error
σ	A curve on manifold
$\beta_n(X)$	n'th Betti numbers on members of set X
$\mathbf{g}$	Riemannian Metric
$\mathbf{R_N}$	Complexity.
$\mathbf{D_N}$	Distortion.
$\phi_{,\alpha}$	Derivative of ϕ over co-ordinate α.
$\vec{\mathbf{A}}$	Vector Field.
$\tilde{\mathbf{A}}$	Dual Vector Field.
$\blacksquare_{\alpha}^{\beta}$	A matrix which transforms vectors from co-ordinates system β to coordinate system α
$\blacksquare_{\alpha\beta}^{\mu}$	The Christoffel symbol for μ^{th} component of derivative of $\frac{\partial \vec{e}_\gamma}{\partial x^\kappa}$.
$\mathbf{R}_{\beta\mu\nu}^{\alpha}$	Riemannian curvature as a tensor of the form $\binom{1}{3}$.
$\mathbf{R}_{\alpha\beta\mu\nu}$	Riemannian curvature as a tensor of the form $\binom{0}{4}$.
$\mathbf{R}_{\alpha\beta}$	Ricci tensor
$\mathbf{R}$	Ricci scaler
$\mathbf{k_{Gauss}}$	Gaussian curvature.
$\mathbf{T_x M}$	Tangent space at point x, on a manifold M.
$\mathbf{R^m}$	Euclidean space with dimension m.
$\boldsymbol{W}_{\mathbf{l}}$	Weights for layer l in neural network.
$\boldsymbol{a}_{\mathbf{l}}$	Output of activation function of layer l.
$\bar{\boldsymbol{a}}_{\mathbf{l}}$	Output of activation function of layer l with homogeneous co-ordinate.
$\mathscr{D}\boldsymbol{\theta}$	Gradient of log-likelihood function with parameters θ.
$\boldsymbol{F}$	Fisher Matrix.
$\mathbb{E}$	Expectation Operator.
$\otimes$	Kronecker product.
$\boldsymbol{F}^{-1}$	Inverse Fisher Matrix.
$\blacksquare\mathbf{h}$	The gradient of objective function
$\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}$	Kronecker Factors of Fisher Matrix
$\boldsymbol{Q}$	Eigenvector matrix.
$\mathbf{O}()$	Order of complexity.
$\mathbf{H}(\mathbf{X})$	Entropy of X
$\mathbf{I}(\mathbf{X};\mathbf{Y})$	Mutual information between X and Y
$\mathbf{C_l}$	Capture Ratio for layer l

tr(.) Trace of a matrix.