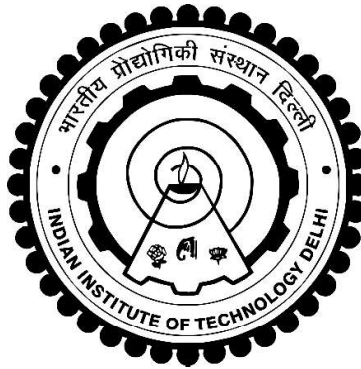


GEOMETRIC MODELLING AND UNDERSTANDING OF DEEP NEURAL NETWORKS

PIYUSH KAUL



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

GEOMETRIC MODELLING AND UNDERSTANDING OF DEEP NEURAL NETWORKS

A THESIS

submitted by

Piyush Kaul

2015EEZ7544

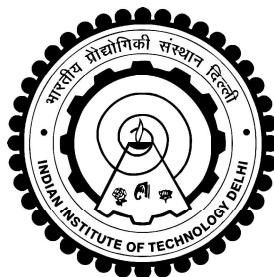
under the guidance of

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

Department 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 inter-relationship 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 scalars. 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

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
LIST OF TABLES	vii
LIST OF FIGURES	ix
ABBREVIATIONS	x
NOTATION	xii
1 INTRODUCTION	1
1.1 Objective and Scope	2
1.1.1 Motivation	2
1.1.2 Objective	2
1.1.3 Scope	2
1.2 Background	3
1.2.1 Neural Net Architecture	3
1.2.2 Learning Theory	8
1.2.3 Topology and Manifolds	9
1.2.4 Topology vs Geometry	13
1.2.5 Characterization by Homological Complexity	13
1.2.6 Characterization by Scattering Transform	19
1.2.7 Characterization by Curvature	24
1.2.8 Characterization by Information Theory	30
1.3 Literature Survey Summary	32
1.3.1 Riemannian Geometry	32
1.3.2 Information Geometry	34

1.3.3	Information Bottleneck and Skip Connections	36
1.3.4	Combining Shannon and Fisher Information	37
1.4	Organization of the Thesis	37
2	Riemmanian Curvature of Neural Networks	39
2.1	Introduction	39
2.1.1	Riemmanian Curvature	40
2.2	Curvature in Neural Nets	45
2.3	Experimental Measurements	47
2.3.1	One dimensional manifold classifier	47
2.3.2	CNN classifier with MNIST	48
2.3.3	Linearly separable case	51
2.3.4	CNN Classification with CIFAR-10	51
2.3.5	Results	52
2.4	Discussion	57
2.5	Distance Measures on Riemannian Manifolds	57
2.5.1	Method for Calculating Geodesics Distance	62
2.5.2	Relationship between Accuracy Metrics and Geometry	62
2.5.3	Results	64
2.6	Summary	64
3	Information Geometry and PKFAC	68
3.1	Introduction	68
3.2	Kronecker Factored Approximate Gradient	69
3.3	Projective KFAC	71
3.3.1	Estimating Rank	71
3.3.2	Projection Algorithm	72
3.3.3	Computing Inverse	73
3.3.4	Damping	74
3.3.5	Random Projections	75
3.3.6	Complexity	75
3.4	Experimental Measurement	77

3.4.1	Autoencoder	77
3.4.2	Multi Layer Perceptron with Fashion MNIST	77
3.4.3	LeNet-5 with Fashion MNIST	78
3.4.4	Resnet and with CIFAR10	79
3.4.5	Super-Resolution with ESPCN	79
3.5	Summary	82
4	Understanding Skip Connections through Information Bottleneck	89
4.1	Introduction	89
4.2	Method of Analysis	90
4.2.1	Capture Ratio	90
4.2.2	MI measurement - EDGE	91
4.2.3	Datasets	92
4.2.4	Experimental Setup	92
4.3	Experimental Results	93
4.3.1	Effect of Skip Connections: MLP vs ResMLP	93
4.3.2	Effect of Skip Connections: CNN vs RCNN	95
4.3.3	More on the Compression Phase	97
4.4	Summary	98
5	Combining Shannon and Fisher Information	104
5.1	Introduction	104
5.2	Fisher Information in Neural Networks	104
5.3	Shannon Information in Neural Networks	105
5.4	Using both Fisher and Shannon Information	107
5.5	Experimental Measurements	108
5.6	Summary	112
6	Conclusions and Future Work	113
6.1	Conclusions	113
6.2	Future Work	114
A	LIST OF PUBLICATIONS	122

List of Tables

1.1	Upper and lower bounds on the growth Betti numbers	15
2.1	Mean Curvature with Depth at various rotations	60
3.1	Complexity Estimates	76
3.2	Accuracy results with MLP and Fashion MNIST	80
3.3	Accuracy results with Lenet-5 and Fashion MNIST	80
3.4	Mean accuracy results with Resnet18 and CIFAR10	81
3.5	PSNR results using ESPCN and BSD300 datase	81

List of Figures

1.1	Multi Layer Perceptrons	4
1.2	CNN Architecture	5
1.4	Multi Layer Perceptrons	7
1.5	CNN Architecture	7
1.6	A Manifold	11
1.7	Tangent Space	12
1.8	Sphere and Torus	14
1.9	D1 and D2	16
1.10	Barcode from persistent homology	17
1.11	Homology of Neural Networks	18
1.12	Wavelet Filters. [56]	22
1.13	Scattering Transform	23
1.14	Scattering Coefficients	24
1.15	Curvature Propagation	26
1.16	Curvature Propagation	28
1.17	Information plane	33
2.1	A rectangular closed loop to derive Riemannian Curvature	43
2.2	Osculating Circle	44
2.3	Algorithm for calculating curvature	47
2.4	Training Set for Spiral Classification	48
2.5	Classifier regions for trained neural net	49
2.6	Mnist digit 9 used for Measurement	50
2.7	Softmax value with MNIST classification	50
2.8	Test Images for CIFAR-10	52
2.9	Ricci scaler for spiral data	52
2.10	Ricci Scaler: one hidden layer	54

2.11 Ricci Scaler: four hidden layers	55
2.12 Ricci Scaler Averaged	56
2.13 Ricci Scaler for 9 vs 1 digit classification	57
2.14 Ricci Curvature with Depth	58
2.15 Ricci Scaler CIFAR10	58
2.16 Ricci scaler vs α in CIFAR-10	59
2.18 Geodesic distance from single point	66
2.19 Geodesic distance between two points	66
3.1 Block Diagonal Approximation	70
3.2 Algorithm for PKFAC	83
3.3 Autoencoder with MNIST: SGD vs NGD	84
3.4 Loss and accuracy with MLP and Fashion MNIST	85
3.5 Loss and accuracy for LeNet-5 with Fashion MNIST	86
3.6 Loss and accuracy with ResNet18 and CIFAR10	87
4.1 RMLP. FC stands for fully connected layer	93
4.2 MI paths with and without skip connections	99
4.3 Relevant hidden information	100
4.4 MI paths for CNN	101
4.5 Relevant information for CNN	102
4.6 Train test accuracies	103
5.1 MLP: Log Determinant of inverse Fisher	108
5.2 MLP: Trace of inverse Fisher Information Matrix	109
5.3 Autoencoder Block Diagram	110
5.4 Autoencoder Log Determinent FIM	110
5.5 Autoencoder Trace FIM	111
5.6 Pseudo-Autoencoder Log Determinent	111

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 Bottleneck
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
$\mathcal{L}_d(\mathbf{h})$	Loss function over dataset d and prediction function $\mathbf{h}$
ε_{app}	Approximation Error
ε_{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.
ϕ, α	Derivative of ϕ over co-ordinate α .
$\mathbf{A}$	Vector Field.
$\tilde{\mathbf{A}}$	Dual Vector Field.
$\mathbf{M}_{\alpha}^{\beta}$	A matrix which transforms vectors from co-ordinates system β to coordinate system α
$\mathbf{M}_{\alpha\beta}^{\mu}$	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
$\mathbf{k}_{\text{Gauss}}$	Gaussian curvature.
$\mathbf{T}_x\mathbf{M}$	Tangent space at point x , on a manifold $\mathbf{M}$.
$\mathbf{R}^m$	Euclidean space with dimension m .
$\mathbf{W}_l$	Weights for layer l in neural network.
$\mathbf{a}_l$	Output of activation function of layer l .
$\tilde{\mathbf{a}}_l$	Output of activation function of layer l with homogeneous co-ordinate.
$\mathcal{D}\theta$	Gradient of log-likelihood function with parameters θ .
$\mathbf{F}$	Fisher Matrix.
$\mathbb{E}$	Expectation Operator.
$\otimes$	Kronecker product.
$\mathbf{F}^{-1}$	Inverse Fisher Matrix.
$\mathbf{h}$	The gradient of objective function
Ψ and Γ	Kronecker Factors of Fisher Matrix
$\mathbf{Q}$	Eigenvector matrix.
$\mathbf{O}()$	Order of complexity.
$\mathbf{H}(\mathbf{X})$	Entropy of $\mathbf{X}$
$\mathbf{I}(\mathbf{X}; \mathbf{Y})$	Mutual information between $\mathbf{X}$ and $\mathbf{Y}$
$\mathbf{C}_l$	Capture Ratio for layer l

tr(.) Trace of a matrix.