TOPOLOGICAL PRUNER

A NEURAL NETWORK PRUNER USING TOPOLOGICAL DATA ANALYSIS

W.M.M.J.U. Perera

199481D

Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science

Department of Computational Mathematics

University of Moratuwa,

Sri Lanka

March 2022

Declaration

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

W. M. M. J. U. Perera

Signature of Student

Date:03/03/2022

Supervised by

Dr. Subha Fernando

Dr. Ashwini Amarasinghe

Signature of Supervisor(s)

Date:27/06/2021

Dedication

I dedicate my dissertation to my family and many friends.
A special feeling of gratitude to my loving parents, Mr. Susantha Perera and Mrs. Manel Irangani whose words of encouragement and push for tenacity ring in my ears
and also to my wife Sherine Weerasinghe who was the best source of inspiration.

Acknowledgement

First and foremost, I am grateful to my supervisors, Dr. Subha Fernando and Dr. Ashwini Amarasinghe for their valuable advices, continuous support, and patience during my research. Their immense knowledge and ample experience encouraged me all the time during my academic research and daily life. I would also like to extend my thankfulness to Prof. Asoka Karunananda for his guidance to prepare thesis materials and directing in the correct path. My sincere gratitude goes to all other lecturers and non-academic staff members who helped me to make this project a success.

I would also like to thank all my teachers from the kindergarten to the Master's level. Specially, Prof. U.N.B. Dissanayeke and Dr. Shelton Perera for strengthening my foundation in Mathematics. I would not be able to complete this research without the passion and curiosity in Mathematics, which they planted within myself.

Finally, I would like to express my gratitude to my parents, my wife and my friends. Without their tremendous understanding and encouragement in the past few years, it would not be impossible for me to complete my study programme.

Abstract

Architectural damage due to neural network pruning has been a research problem. To recover the accuracy loss, after pruning, pruned neural network needed to be trained further for a certain time period to gain the accuracy back. If the damage done by the pruning process is severe, some layers can collapse and at worse, the entire model may become untrainable. Therefore, pruning process needs to be done carefully to prevent any significant damage to the neural network. Although some existing approaches have been used to overcome this issue by identifying the salience of a neuron with respect to the overall architecture, it is not computationally efficient. Further, the exiting solutions do not count the topological meaning of the neural network architecture during the pruning process. We believe that identifying the salience of neuron with respect to the layer is sufficient to avoid severe damages to the overall architecture.

Topology, the champion of mathematical shapes, has been introduced to solve the aforesaid problem. We introduce 'Topological Pruner', a novel pruner that uses a genetic algorithm powered by a topological fitness function to identify removable neurons of each layer of a pre trained neural network. After pruning is done, the model is retrained so that the parameters of the remaining neuron can be readjusted to recover the model. As per to our knowledge this is the first ever attempt to use persistence homology, a topological tool for pruning.

Number of parameters, FLOPs and recovery time of the new pruner is evaluated on CIFAR10 dataset on VGG-16 architecture against L1Filter Pruner, L2Filter Pruner and FPGM Pruner. Evaluation results show that the new pruner competes well with the existing pruners. We conclude that, topological data analysis can be used to explain the recoverability and mitigate damage cause by neural network pruning.

Table of Content

List of Figures

List of Tables

List of Abbreviations

CBIS : Component based software engineering

FLOPS: Floating-point Operations per Second

GA : Genetic Algorithms

NAS : Neural Architecture Search

NNP : Neural Network Pruning