Data-dependent Resource Allocation and Routing in Multi-path Neural Networks for Vision-based Learning

M H G Dumindu Tissera

(188013F)

Thesis submitted in partial fulfillment of the requirements for the degree Doctor of Philosophy

Department Electronic and Telecommunication Engineering

University of Moratuwa

Sri Lanka

September 2023

DECLARATION

I declare that this is my own work and this 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 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).

Signature: UOM Verified Signature Date: 2023/09/10

The above candidate has carried out research for the PhD thesis under my supervision.

Signature of the Supervisor(s): Date: 10 September 2023

UOM Verified Signature

Dr. Ranga Rodrigo

Abstract

As the depth of a neural network increases, the non-linearity and more parameters allow it to learn more complex functions. While network deepening has been proven effective, there is still an opportunity for efficient feature extraction within a layer that will improve the overall performance for the complexity of the network. Widening networks by adding more filters to each layer is the naive approach towards strengthening layer-wise feature extraction. It is an inefficient scaling option, considering the number of parameters being quadratic with the number of filters employed per layer. In contrast, parallel extractors in each layer provide an efficient scaling option. However, without context-dependent input allocation among these processes, such parallel computations tend to learn similar features, collapsing to a single computation.

Thus, it is vital to study the parallel stacking of computations layer-wise and design a routing method that allocates incoming feature maps to these computations. The expected outcome is to group homogeneous feature maps in parallel layers and employ exclusive filter sets to each of the groups (paths) so that the filter sets of each path can specialize in extracting features exclusive to each group.

To allow the network input to be routed end-to-end over such parallel paths, we propose data-dependent parallel resource allocation methods layer-wise. Given a layer of parallel tensors, we first employ sub-networks that produce gating coefficients to weigh cross-connections to the next layer of parallel tensors. Then, the next layer's parallel tensors are constructed by getting summations of the current layer's tensors, each weighted by the corresponding gating coefficient. We demonstrate that our multi-path networks outperform previous widening and adaptive feature extraction, ensembles, and deeper networks with comparable complexity using image recognition challenges.

To further regularize gating sub-networks, we think of a gating network's path allocation as a soft clustering of its input feature maps. Thus, we propose a neural mixture model-based clustering objective to use as a regularization loss for the gating networks, which We first study as a standalone neural network-based clustering approach. The proposed clustering framework uses a neural network to learn cluster distributions in mixture modeling instead of tuning human-defined distributions. We adopt the Expectation-Maximization (EM) algorithm to train the network and perform batch-wise EM iterations where the forward pass acts as the E-step and the backward pass as the M-step. For image clustering, we use the mixture-based EM objective as the clustering objective, along with consistency optimization. Our networks outperform traditional and single-stage deep clustering methods that still depend on k-means.

Finally, we propose using this clustering objective to regulate gating networks to get distributed gating activation patterns. We show that the skewed gating patterns can be improved with such regularization loss as a local regularization. We further present the need for a global regularization method that takes the end task performance into account. We also suggest extending research towards sparse resource allocation, along with gating networks to handle more diversity.

***Keywords*- Multi-Path Networks, Data-Dependent Routing, Deep Clustering, Neural Mixture Models**

DEDICATION

To the resilient people of Sri Lanka,

who have faced the challenges of the country's financial crisis with unwavering determination and faith,

who continue to endure and strive for a better future for the nation.

This dissertation is a tribute to your strength and commitment to a brighter tomorrow.

ACKNOWLEDGMENTS

I would like to express my deep appreciation and gratitude to the following individuals for their invaluable contributions to the completion of my Ph.D. thesis.

First and foremost, I would like to thank my adviser, Dr. Ranga Rodrigo, for his unwavering support, guidance, and encouragement throughout my Ph.D. studies. His expertise, insight, and mentorship have been invaluable to my academic and personal growth.

I also extend my gratitude to Dr. Subha Fernando for her technical support, which has contributed significantly to the success of this research. Additionally, I am grateful to Prof. Sanath Jayasena and Dr. Jayathu Samarawickrama for their constructive suggestions and insightful comments, which have helped refine and improve this thesis's technical content.

I would like to acknowledge the generous support of CodeGen International (PVT) Ltd and its CEO Dr. Harsha Subashinghe, who provided funding for this research. I am grateful for their support, which has allowed me to pursue my academic goals.

I would also like to thank Prof. Dileeka Dias, Prof. Sanath Jayasena, Dr. Ranga Rodrigo, Dr. Subha Fernando, Dr. Jayathu Samarawickrama, Dr. Harsha Subhasinsghe, and CodeGen International (PVT) Ltd for their role in the establishment of QBITS Lab at the University of Moratuwa and the Ph.D. program.

I extend my sincere appreciation to Prof. Ruwan Udayanga and Dr. Pratha-

pasinghe Dharmawansha, the Research Coordinators, for coordinating the program and providing feedback on technical content. I am also indebted to Dr. Charith Chiththraranjan, the chair of the progress review panel, for reviewing and providing valuable feedback on my work.

I thank my colleagues and collaborators, Rukshan Wijesinghe, Kasun Vithanage, Alex Xavier, and Pubudu Ekanayake, for their outstanding technical contributions and unwavering support throughout this research.

Finally, I thank the thesis examiners Prof Salman Khan, Dr. Sadeep Jayasumana and Dr. Shehan Perera, and PhD examination chair Prof. Gihan Dias for providing valuable suggestions.

Thank you for your invaluable contributions, guidance, and support in making this achievement possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES