Reykjavík University Project Report, Thesis, and Dissertation Template

by

Sigurður Helgason

Thesis of 60 ECTS credits submitted to the School of Computer Science
at Reykjavík University in partial fulfillment
of the requirements for the degree of
Master of Science (M.Sc.) in Computer Science

December 2019

Examining Committee:

Yngvi Björnsson, Supervisor
Professor, Reykjavík University, Iceland

Tough E. Questions, Examiner
Associate Professor, Massachusetts Institute of Technology, USA

Reykjavík University Project Report, Thesis, and Dissertation Template

Sigurður Helgason

December 2019

Abstract

this is my abstract written in the language of English I am testing thought

alas I knew

ok so now I'm trying something new

asdf

test

Herkænsku mat á djúptauganetum

Sigurður Helgason

desember 2019

Útdráttur

this is my abstract written in the language of íslensku þæöasdf

Important!!! Read the Instructions!!!

If you have not already done so, LaTeX the `instructions.tex` to learn how to setup your document and use some of the features. You can see a (somewhat recent) rendered PDF of the instructions included in this folder at `instructions-publish.pdf`. There is also more information on working with LaTeX at `http://samvinna.ru.is/project/htgaru/how-to-get-around-projects-publish.pdf`. This includes common problems and fixes.

This page will disappear in anything other than draft mode.

I dedicate this thesis to my family who while not understanding most of what I do always support me. The friends that still want to talk to me even though my schedule has rendered me unmeetupwithable. Lastly but certainly not least, Hulda Lilja who always easead my mind during my spurts of imposter syndrome, and fear of failure. Don't Panic.

Acknowledgements

So long, and thanks for all the fish.

Acknowledgements are optional; comment this chapter out if they are absent Note that it is important to acknowledge any funding that helped in the workThis work was funded by 2020 RANNIS grant "Survey of man-eating Minke whales" 1415550. Additional equipment was generously donated by the Icelandic Tourism Board.

Preface

This dissertation is original work by the author, Sigurður Helgason.

The preface is an optional element explaining a little who performed what work. See `https://www.grad.ubc.ca/sites/default/files/materials/thesis_sample_prefaces.pdf` for suggestions.

List of publications as part of the preface is optional unless elements of the work have already been published. It should be a comprehensive list of all publications in which material in the thesis has appeared, preferably with references to sections as appropriate. This is also a good place to state contribution of student and contribution of others to the work represented in the thesis.

Contents

List of Figures

List of Tables

List of Abbreviations

MSc	Masters of Science
ML	Machine Learning
AI	Artificial Intelligence

This part of the dissertation introduces the concepts and the field.

Chapter 1

Introduction

State the objectives of the exercise. Ask yourself: Why did I design/create the item? What did I aim to achieve? What is the problem I am trying to solve? How is my solution interesting or novel?

In the world of deep learning is exciting, we have models that are able to examine images and very reliably be able to identify it's content. Deep learning models have beaten Opthalmologists in identifing diabetic retinopathy, they've identified cancer cells where others have not. Deep learning models have even predicted the likelyhood a person will die in the following year. These models have done these things extremely accurately, cheap, and fast.

The field of examining deep learning models to gain a richer understanding in how they work, and what makes them so much better than humans at various tasks is still new. This is the field of Explainable Artificial Intelligence (XAI). If we wish to continue using artificial intelligence (AI) and machine learning (ML), to improve our lives we must examine how they work, this is not only to improve ourselves but these explanations are a legal requirement now in many areas, namely, legal, and medical. Recently, XAI has generally been focused on Model Interpretability, in particular focus on gaining insights on the concepts learned by Deep Neural Networks.

This thesis takes a look at how we can examine an artificial neural network (ANN) to understand which higher level concepts (HLC) it deems important for a given state within games. These games can be simple like Tic-Tac-Toe or Breakthrough and

extremely complicated like Chess or Go. If an ANN values a HLC greatly which we as humans don't find very important, and yet the ANN is able to defeat us there is reason to think that HLC should be of more importance to us.

The method of examining HLC's within games has generally been done by examining the current state of the game by evaluating them w.r.t. a heuristic. A heuristic within games are a evaluation of the end cost for a state given just the state, for example, the amount of pieces left within a game of chess. Intuitively, the amount of pieces left is a good estimate for a state in chess, this is a higher level concept we use to evaluate a chess position. The piece amount is much lower level than other concepts we use, grand master chess players evaluate a position w.r.t. states where the king is safe from attack, or the structure of the pawn positions. The idea is to examine a neural networks evaluation of a state regarding those higher level concepts, if human players evaluate the king safety of a state low but the neural network highly values it, and the neural network plays better than the player.

The chess engine DeepBlue was the first chess engine to win a chess match agains a chess grand master. It's power in the world of chess was a marvel. DeepBlue used tree- search to evaluate it's position and possible moves, doing an immense amount of calcluations in order to arrive at the best move it could given the time constraints. The tree search approach is dependent on knowing which board state is good and what board state is bad, This can be easy when the boardstate is such that you're guaranteed to have been checkmated in one move, and it's incredibly hard when you boardstate is such that both you and your opponent have 16 pieces and are only a handfull of moves in. Generally the boardstate is a function of the pieces you have, how safe they are, and how opportunistic you are to capture your opponents pieces. These values are not obvious and require a domain expert to describe the situations and give the values for these cases.

The process of evaluating chess positions did not change for many years. With the revival of deep neural networks this changed. A neural network is a network of connected nodes (neurons) that each how a specific weight. These networks are fed input which is passed through the network and it is updated by the weights of the

neurons, the input is morphed into a more comprehensible value and these values are then used to generate some output. The output can be a binary classification for instance input=Image output={cat, dog}. Processing the pixels and altering them to numerical values to find the output cat or dog is not intuitive and understanding of the values of the neurons in the network is not considered fruitful.

This approach of chess engines is still popular today but not the reigning champion anymore. In todays best chess engines deep neural networks are used to

Deep neural network machine learning models have an amazing impact on modern society today, we use them to know which items to purchase, to know which people we should love, and we use them to drive us to work. These tasks given to deep learning models are all niceties for us to enjoy. However, as our dependance on them increases and their astounding success in any and all tasks continues, we will want to utilize them for all things. This leads to a utopia where decision making is removed from the hands of humans. The question we must ask ourselves is when do we want to know why a decision was taken, is it only when the decision taken is not preferrable to us. For instance, if a deep neural network model used by your lending authority decides you are not worty of a loan. These questions should have answers but as it stands today no such answer is possible.

We can use methods like saliency maps to map a deep learning model emphasis on pixels to hopefully understand what portions the model is focusing on. This method however, tells us nothing about what

results that are nothing but amazing.

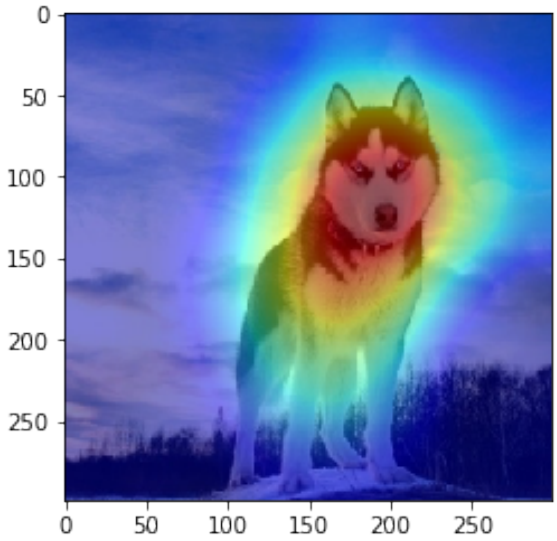

Figure 1.1: Example of a models saliency map for an image of a dog

1.1 Background

1.1.1 Explainable Artificial Intelligence

Within XAI many methods have been developed to try to evaluate ANN's, these methods are often referred as model interpretability methods. In the field image recognition there has been a lot of work examining which pixels of an image the model deems important. Arguably the most popular method for this is Saliency Maps (CITE ME). There the pixel values the model deems important are colored in s.t. a human can examine the image and get a sense of what portions of the image are important to the model, an example of a classification of a dog can be seen in Figure 1.1.

1.1.2 Reinforcement Learning

In 2017 a Google based artificial intelligence research company published a paper called 'Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm'. The paper described a Model based Reinforment learning agent and an accompanying Residual Deep Neural Network architecture that was able to master the game of chess without being given any idea of how to play the game of chess, and others. The architecture used in this paper is heavily inspired by that architecture.

1.1.3 Model Interpretability

In 2018 the paper 'Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)' was published where an approach to model interpretability was introduced which is able to distinguish

The deep neural network architecture that is used in this paper is based on the AlphaZero paper published by DeepMind,

This part of the dissertation describes the work done.

Chapter 2

Methods

2.1 Monte Carlo Tree Search

To train the models

This section discusses the various machine learning methods utilized throughout this project and discusses their applicability.

I'm a little teapot

I'm a smaller teapot

Chapter 3

Results

In this section you discuss any issues that came up while developing the system. If you found something particularly interesting, difficult, or an important learning experience, put it here. This is also a good place to put additional figures and data.ello world

This part of the dissertation talks about future work discussion concludes the work and

Chapter 4

Discussion

Here I will discuss the myriad of fields research like this can help for instance for health organizations, patients, taxpayers, and pharmaey

4.1 Summary

summarize the workey

4.2 Conclusion

conclude the work, discuss the significance of the workey