UNIVERSITY OF
CAMBRIDGE
Department of Computer
Science and Technology

Ain't Nobody Got Time For That: Budget-aware Concept Intervention Policies

Thomas Yuan

Downing College

May 2024

Submitted in partial fulfillment of the requirements for the
Computer Science Tripos, Part III

Total page count: 15

Main chapters (excluding front-matter, references and appendix): 6 pages (pp 8–13)

Main chapters word count: 467

Methodology used to generate that word count:

```
$ make wordcount
gs -q -dSAFER -sDEVICE=txtwrite -o - \
   -dFirstPage=6 -dLastPage=11 report-submission.pdf | \
egrep '[A-Za-z]{3}' | wc -w
467
```

Declaration

I, Thomas Yuan of Downing College, being a candidate for the Computer Science Tripos, Part III, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

Signed: Thomas Yuan

Date: April 10, 2024 -

Abstract

Regular supervised learning Machine Learning models learn to predict the labels of inputs. Concept BottleNeck Models (CBMs) are ML models designed to increase the interpretability of model predictions by decomposing a model into two submodels, splitting the original process into predicting a set of human-interpretable concepts / features present in the input, then predicting the label using these concepts. Since these concepts are human-interpretable, it is much more easier to understand the reasoning behind the predicted labels, thus mitigating some of the potential dangerous downsides associated with using ML models as "black-box" models, especially in fields where these predictions can have significant consequences to human life, such as medicine, criminal justice, autonomous vehicles, etc.

During inference time, professionals can intervene on CBMs by correcting the predicted concepts leading to more accurate predicted labels. Due to the costs associated with performing such an intervention, the question of what concepts to intervene on in order to maximize the accuracy of the model becomes an important research question. This project focuses on answering this research question. This project attempts to model the costs associated with using experts to perform interventions as a budget, and thus the main research question of this project is "How can we determine the concepts to intervene on for a given budget for a set of inputs and the corresponding model predictions?"

This project focuses on answering the above research questions. It first investigates the differences between greedy and non-greedy models, showing that non-greedy algorithms can outperform its greedy counterparts. It then investigates the performance of greedy models, building on top of existing methods by incorporating surrogate models to model the distribution of concepts. The output of these surrogate models are then used by an ML model that learns to predict the next concept to intervene on in each step. The project then investigates using Reinforcement Learning and these surrogate models to train a non-greedy model that learns to predict an entire sequence of interventions for given inputs and corresponding CBM outputs. Lastly, the project then investigates if it is necessary to train the prediction model simultaneously with the CBMs.

Acknowledgements

This project would not have been possible without the wonderful support of my lovely supervisors Mateo Espinosa Zarlenga, Dr Mateja Jamnik and Dr. Zohreh Shams. I would also like to thank my friends and family for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine Learning models are universal approximation solutions to problems and have been viewed and used traditionally as "black-box" solutions, where users simply query the model and receive an answer to the problem without necessarily knowing the exact reasoning process behind it. Over the past decade, the increased application of ML models along with this "black-box" property has raised many concerns, especially in areas where decisions made are critical to human safety, such as in medicine or automated driving. To increase interpretability, researchers developed concept bottleneck models that map input to a set of human-interpretable concepts, and then concepts to the actual predictions. This increases the interpretability of the model as humans can understand the basis of the machine learning model predictions via the high-level concepts that it trains to predict which are then used to produce output. Another additional benefit of these types of machine-learning models is that when used in practice, experts can intervene in the predicted concepts to the correct concepts to generate more accurate model output. Given that experts have limited time, determining the order of concepts to ask experts to intervene on to maximize the accuracy of the model given a limited budget, becomes an interesting problem. This project focuses specifically on trying to solve this question using Reinforcement Learning and Surrogate models to model conditional probabilities. The end goal of this project is to be able to train RL models that will be able to produce non-greedy (or simply greedy, but more optimal) policies that can outperform existing greedy policy approaches.

Chapter 2

Background

Chapter 3

Related Work

Chapter 4

Design and implementation

Chapter 5

Evaluation

Chapter 6

Summary and conclusions

Bibliography

Appendix A

Technical Details