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: ??

Main chapters (excluding front-matter, references and appendix): 1 pages (pp ??–??)

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 9, 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

Chapter 2

Background

Chapter 3

Related Work

Chapter 4

Design and implementation

Chapter 5

Evaluation

Chapter 6

Summary and conclusions

Appendix A

Technical Details