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: 16

Main chapters (excluding front-matter, references and appendix): 7 pages (pp 8–14)

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: May 7, 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

1.1 Concept BottleNeck Models

Machine Learning (ML) 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 (CBMs) [2] that predicts a set of human-interpretable concepts from input, and then uses the predicted concepts to predict labels. This increases the interpretability of the model as humans can understand the basis of the ML model predictions via the high-level intermediate concepts that the model is trained to predict.

Another additional benefit of these 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 query experts to intervene on, to maximize the accuracy of the model given a limited budget, becomes an important problem. Since the objective behind developing these models is to use them in real life, this project utilizes different datasets that reflect different situations in real life to measure the performance. This is further discussed in Section ??.

1.2 CEM and IntCEM

Current research has made significant progress on optimizing these models for interventions, including numerous studies on developing models to learn the optimal order of concepts to intervene on [1], most notably Intervention-aware Concept Embedding Models (IntCEMs) [7]. IntCEMs build upon Concept Embedding Models [6], a variant of

CBMs that utilize embeddings to represent the intermediate concepts such that models learn to encode information about unlabelled concepts while still preserving the valuable interpretability of CBMs. IntCEMs augment CEMs with an additional model that learns to predict the next concept to intervene on given the current state of the CBM, which is also used during training to increase the CBM's sensitivity to interventions. IntCEMs achieve state-of-the-art performance on the performance of interventions while still maintaining similar performance when no interventions are performed.

1.3 Reinforcement Learning

Despite the above-mentioned improvements, there is still a big gap between the performance of these intervention policies versus the best possible performance, i.e. the performance of intervention policies that have access to the ground truth output labels. Additionally, existing approaches such as IntCEM are greedy, which means that they learn to predict concepts that maximize the performance at each step. it has been speculated that non-greedy methods may outperform these existing greedy methods, with the objective being maximizing performance after intervening a certain number of concepts rather than maximizing performance at each step. One such approach that may be used to generate non-greedy intervention policies is Reinforcement Learning (RL) [5].

This project focuses specifically on trying to solve the question of finding a good intervention policy for a given budget using RL and Surrogate models to model conditional probabilities to guide the RL model, taking inspiration from an approach [3] in a similar setting, namely Active Feature Acquisition (AFA) [4]. This project successfully develops a novel RL-based method that when combined with existing methods from IntCEM to increase sensitivity to interventions, is able to learn an intervention policy and a Concept BottleNeck Model that outperforms existing non-greedy intervention policies and models for different budgets, while maintaining similar performance when no interventions are performed.

Chapter 2

Background

Chapter 3

Related Work

Chapter 4

Design and implementation

Chapter 5

Evaluation

Chapter 6

Summary and conclusions

Bibliography

[1] Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37:5948–5955, 06 2023.

[2] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[3] Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6450–6459. PMLR, 18–24 Jul 2021.

[4] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney. Active feature-value acquisition for classifier induction. In *Fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 483–486, 2004.

[5] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* A Bradford Book, Cambridge, MA, USA, 2018.

[6] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[7] Mateo Espinosa Zarlenga, Katherine M. Collins, Krishnamurthy Dj Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Appendix A

Technical Details