Explainable AI Planning in Urban Traffic

Marlon Pereira da Silva
Pontifical Catholic University of Rio Grande do Sul (PUCRS)
marlon.p@edu.pucrs.br

Abstract

The decision-making process in Artificial Intelligence is often opaque, generating an idea of a 'black box' to the public in general - and even to specialists. This causes a lack of transparency in the results of the models, and general mistrust in AI-based solutions. Explainability - the capacity to make processes and decisions understandable for a human - can facilitate understanding an AI solution. This paper will explore the idea of explainability on Automated Planning within the context of urban traffic . To accomplish this, we will use Pddl+, which allows us to model entities such as processes and events in communion with a model-based reconciliation approach.

Introduction

Artificial Intelligence is occupying ever-growing importance in multiple domains like healthcare, finance, and law. Nowadays, many systems will consider outputs generated by AI frameworks at some step of the decision-making process. However, modern complex AI techniques such as deep learning are naturally opaque, making it very hard for a human - even a domain expert - to understand how a given method has resulted. This problem becomes even more critical when we consider the widespread use of AI and systems that use a mixed approach that an AI-generated output is the input of a human-based decision process. This opaqueness generates hardships for integrating human and AI resources on hybrid methods and building trust with civil society.

Although not a recent development, the concept of explainability comes into the limelight. An explainable AI is one such that its inner workings can be understandable by humans and which factors are taken into account to justify the choice of a solution (Chakraborti et al. 2017). A significant consequence is that it becomes possible for a human to audit and correct a solution if it has adopted undesirable criteria - for example, considering race or gender to decide if a person can take a loan with a finance company.

Initially focused on Machine Learning techniques, explainability permeates other subsets of AI, such as planning. Planning concerns itself with the course of actions needed to achieve a specific goal (Ghallab, Nau, and Traverso 2004). More specifically, classical planning aims to identify a sequence of steps, also known as a plan, that drives the initial state of this particular problem towards the desired state (Fox, Long, and Magazzeni 2017). The Planning Domain Definition Language (PDDL) is a standard encoding language for automated planning tasks, with many planning systems offering support (Borgo, Cashmore, and Magazzeni 2018). To better represent mixed discrete-continuous domains, an extension of PDDL named PDDL+ was created, which incorporates fully-featured autonomous processes and trigger-based events, allowing a more precise tool for modeling (Batusov and Soutchanski 2019).

Recent works exploring the intersection of explainability and planning treats concepts such as explicability and predictability of plans, focusing on the human interpretation of such plans (Zhang et al., 2017) and model reconciliation, examining the differences and correlations between different human and AI model representations (Chakraborti et al. 2017). Besides that, we have some guiding questions to determine if a given system is explainable:

- Why did it do that? And why did it not do something else?
- Why is what it is proposing better than the alternatives?
- Why cannot this be done?
- Why is there a reason to replan at this point (or not)?

In this work, we propose the use of techniques of explainability to generate an human-understandable model of planning. We aim to implement an PDDL+ model of urban traffic, and to translate such model into a human-friendly specification.

Technical Approach

The proposed project will be built through the use of PDDL+ and the OPTIC planner (Benton, Coles and Coles 2012) to formalise the domain and generate adequate plans for it. The main objective of this model will be to offer the optimal solution for resource allocation of public transportation, given a set of agents with distinct coordinates and objectives. After that, we will use a number of operations to convert the AI-based model to an human-friendly model (Cashmore et al. 2019). The planner will be executed against two scenarios, one considering a single start point and another one considering multiple origins.

Project Management

- Week 1 (5/17): Planning techniques (PDDL+)
- Week 2 (5/24): Planning techniques (PDDL+), model building
- Week 3 (5/31): Model building, generate explanations
- Week 4 (6/7): Model building, generate explanations, testing/tunning
- Week 5 (6/14): Testing/tunning, evaluate results
- Week 6 (6/21): Write report and presentation

Conclusion

In this work, we will use the PDDL+ specification to build a model of urban traffic and use that model to generate a tool that can give intelligible answers about the steps of a plan to an observer. To do that, we will model the scenario, create a plan for it, and then create a model-based explanation of the domain. In doing that, we aim to allow researchers and decision-makers to understand the processes and outputs of an automated planning solution. The results can contribute to building transparency and trust between AI systems and society as a whole.

References

Batusov, V.; and Soutchanski, M. 2019. A logical semantics for pddl+. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, 40–48.

Benton, J.; Coles, A.; Coles, A. 2012. Temporal Planning with Preferences and Time-Dependent Continuous Costs. Proceedings of the International Conference on Automated Planning and Scheduling, 22(1).: AAAI Press. Borgo, R.;

Cashmore, M.; and Magazzeni; D. 2018. Towards Providing Explanations for AI Planner Decisions. *IJCAI-18 Workshop on Explainable AI.*

Cashmore, M. et al. 2019. Towards explainable AI planning as a service. *arXiv preprint arXiv:1908.05059*

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*

Fox, M.; Derek L.; and Daniele M. 2017. Explainable planning. *IJCAI-17 workshop on Explainable AI* abs/1709.10256.

Ghallab, M.; Nau, D.; and Traverso, P. 2004. *Automated Planning: theory and practice*. Elsevier.

Nau, D.; Ghallab, M.; and Traverso, P. 2004. *Automated Planning: Theory amp; Practice*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2018. Handling Model Uncertainty and Multiplicity in Explanations via Model Reconciliation. In *ICAPS*

Zhang et al., 2017. Plan explicability and predictability for robot task planning. In IEEE International Conference on Robotics and Automation (ICRA)