Towards Interpretable Explanations for Transfer Learning in Sequential Tasks

Towards Interpretable Explanations for Transfer Learning in Sequential Tasks

Ramya Ramakrishnan and **Julie Shah**
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Massachusetts 02139

Abstract

People increasingly rely on machine learning (ML) to make intelligent decisions. However, the ML results are often difficult to interpret and the algorithms do not support interaction to solicit clarification or explanation. In this paper, we highlight an emerging research area of interpretable explanations for transfer learning in sequential tasks, in which an agent must explain how it learns a new task given prior, common knowledge. The goal is to enhance a user's ability to trust and use the system output and to enable iterative feedback for improving the system. We review prior work in probabilistic systems, sequential decision-making, interpretable explanations, transfer learning, and interactive machine learning, and identify an intersection that deserves further research focus. We believe that developing adaptive, transparent learning models will build the foundation for better human-machine systems in applications for elder care, education, and health care.

Introduction

People increasingly rely on machine learning (ML) to make intelligent decisions. However, many high-performing ML algorithms are black boxes to people in that they are difficult to understand and cannot easily support human interaction. By providing transparent representations and explaining the reasoning behind decisions, machines can better support human interaction, and people can better provide guidance to help the machine learn more quickly. Sequential decision-making problems pose an additional challenge in that an agent must consider sequences of actions to maximize expected utility in the future, which can be difficult to describe in an interpretable way. Recent works (Khan, Poupart, and Black 2011; Elizalde et al. 2009; Dodson et al. 2013) have developed approaches to provide explanations for sequential problems represented as Markov decision processes (MDPs). However, many of these works assume that an agent is learning a task with no prior knowledge. When both the ML agent and person have already acquired a knowledge base that can be adapted for new tasks, the reasoning involves explaining the decision using similar, previously learned tasks.

Position Statement

In this paper, we discuss related work and identify an emerging area of interpretable explanations for transfer learning in sequential tasks. Recently, there has been a growing emphasis on interactive systems that involve human-machine teamwork (Amir et al. 2015; Nikolaidis and Shah 2013). To facilitate effective collaboration, machines must be able to explain their decisions and interact fluidly with people. In this paper, we define "interpretable explanation" as communication of an agent's reasoning for decision-making to improve human understanding of the learning process. Recent works have developed approaches for interpretable ML models, interactive ML, explanations for sequential-decision making problems, and transfer learning in which an agent learns a new task given prior knowledge. However, there is little work in the intersection of these areas, in which a machine must progressively learn tasks using transfer learning and provide interpretable explanations as it learns so people can accordingly provide feedback. Explaining an agent's learning process as it learns new tasks is potentially useful because prior work (Lombrozo 2012) has shown that explanations have a strong positive impact on learning. Developing a transparent, interactive learning system can be useful in many applications, such as activity planning for elder care and course sequence advising for education.

Related Work

We highlight five areas of related work, shown in Figure 1, and discuss their relevance to our position.

Rule-Based vs. Probabilistic Systems

Many prior works in rule-based and expert systems have developed methods for generating explanations. Some early works (Clancey 1983; Swartout 1977) provided explanations using execution traces, showing how a rule lead to a final conclusion. While execution traces can help users debug a system, showing the entire trace can be overly complex without providing important reasons for a decision. In (Wick and Thompson 1992), an alternate approach was used to generate a justification for a *final* decision, which can be more interpretable than describing the complex reasoning process that lead to a decision.

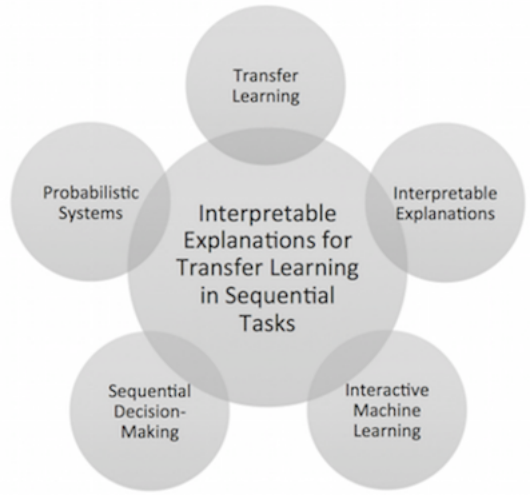

Figure 1: We identify five relevant areas of prior work and highlight the new emerging area of interpretable explanations for transfer learning in sequential tasks.

Explanations for rule-based systems, both those that describe the line of reasoning and those that justify a final decision, may not be as informative for stochastic settings. While all possible options can be enumerated and considered independently, rule-based systems cannot efficiently reason over these alternatives in relation to their uncertainties. For probabilistic tasks, it is important to develop a model and algorithm that can represent and reason over these relationships to provide a basis for producing informative explanations.

Representations vs. Explanations

There has been increasing interest in developing probabilistic ML models that use more interpretable representations, but many of them do not provide a framework to also generate explanations. For example, (Huysmans et al. 2011) evaluated the interpretability of various representations for classification tasks. They compared decision tables, decision trees, and rule-based representations and found that users preferred decision tables. In (Kim et al. 2015), an interpretable representation was developed for high-dimensional data using feature compression techniques. Several other works (Ustun and Rudin 2014; Chang et al. 2009) have focused on developing interpretable representations but they do not automatically provide the framework to generate explanations for these decisions.

For clustering problems, (Kim, Rudin, and Shah 2014) developed the Bayesian Case Model (BCM) to explain data through interpretable representations of clusters. Each cluster was represented using a prototype, an example data point in that cluster, and a subspace, a set of important features from that example that best characterized the cluster. Human subject experiments showed that people performed better on classification tasks using BCM than traditional ML algorithms. While the prototypes and subspaces can be used as explanations for the clusters in the data, they do not explain the reasoning process of the system, which is useful when trying to understand and make changes to the model. The approach also cannot be directly used for sequential problems where explanations must justify sequences of actions.

Single-Shot vs. Sequential Decisions

Many ML tasks often involve a single-shot decision, such as classifying or clustering a set of data points. There has been increasingly more work on making these single-shot ML systems more interpretable. In (Rüping 2006), classification models were modified to be more interpretable in three aspects: understandability, accuracy, and efficiency. To balance these, the approach used both global models that have interpretable structure as well as local models that give more details about specific parts of the model. One work (Letham et al. 2013) developed the Bayesian List Machine to generate a list of interpretable decision statements for high-dimensional data. In another work by (Bien and Tibshirani 2011), instead of generating decision rules, prototypes or examples from the data set were selected to represent the distinct categories. All of these works focus on making more interpretable models for a single-shot decision.

In contrast, sequential decision-making tasks consider sequences of related actions that maximize a measure of expected future utility. Explanations for sequential problems thus must explain this complex decision process so that people can understand and alter the model. Stochastic sequential decision-making problems are often represented as Markov decision processes (MDPs), and agents learn how to act robustly in these tasks through reinforcement learning (RL) (Sutton and Barto 1998). In the RL framework, an agent learns a task through repeated trial-and-error and finally learns a policy that specifies the action the agent should take at each state in the task.

Many works have developed methods to represent these tasks in an interpretable way using for example, Dynamic Bayesian Networks (DBNs) (Jonsson and Barto 2007) and decision trees (Tan 2014). Recently, there has been growing interest not only in developing interpretable representations, but also in generating explanations for these problems. In (Elizalde et al. 2009), the most relevant variable in an MDP was automatically determined by choosing the variable with the highest impact on expected utility given the current state and action. The relevant variable and optimal action were then combined with additional domain knowledge, represented as a frame hierarchy, to generate a verbal explanation.

In (Khan, Poupart, and Black 2011), explanations were generated for MDP policies by learning a Minimum Sufficient Explanation (MSE). This method computed occupancy frequencies, which represented the expected number of times the agent reached each state using a particular policy. It then used pre-defined templates for the explanations (e.g. "$\underline{ActionName}$ is likely to take you to $\underline{State_1}$ about $\underline{\lambda}$ times, which is as high as any other action"), where the states, actions, and occupancy frequencies were populated during task execution. While (Khan, Poupart, and Black 2011) provided a mathematically grounded explanation, (Dodson et al. 2013) provided a framework to generate more interpretable explanations by combining an MDP-based explainer and a case-based explainer. The MDP-based

explanation provided information about predicted future outcomes, while the case-based explanation provided previous cases that were similar to the current one. We believe that developing similar systems to explain complex sequential decisions using previous experiences remains a challenging problem that requires more attention.

Single Task vs. Transfer Learning

Many systems provide explanations for a given task, but few systems generate explanations based on similar previously learned situations. When a lifelong-learning agent faces a new task, it will use prior knowledge to more quickly learn each new task. It is important in these cases, for a machine to explain how it adapted prior knowledge so that people can understand the model and accordingly give feedback.

There is a large body of work in this area, known as transfer learning, in which an agent learns new tasks more quickly by using previously learned knowledge. Transfer learning can be applied to a variety of problems, such as classification, regression, and sequential decision-making. We focus on transfer learning for sequential tasks represented as MDPs (Taylor and Stone 2009) in which there are multiple previously learned tasks that the agent must use to learn. In (Fernández, García, and Veloso 2010), the Policy Reuse in Q-learning (PRQL) algorithm was developed to intelligently explore actions in a new task given previously learned policies. In a recent work, (Ramakrishnan 2015), the PRQL algorithm was modified to learn more quickly by taking better advantage of prior knowledge. Another work, (Ammar et al. 2014), developed an automated measure of similarity between MDPs, which compared state-action-state tuples from the transition functions of different tasks. However, the approach used deep learning techniques, which are sensitive to parameters and are used often as black box methods that are difficult to understand.

While these transfer learning approaches allow an agent to learn more quickly in new tasks, they have not been designed with the aim of explaining the transfer process to a person. Including the human in the agent's learning process requires representing tasks in an interpretable way, identifying high-level relationships between similar tasks, adapting to a new task using prior knowledge, explaining this adaptation process, and updating the model using human feedback. This is a challenging problem and requires much attention, as it can lead to powerful interactive systems that learn and generalize to new situations.

Offline vs. Interactive ML

Finally, we highlight the difference between ML algorithms that learn offline without human interaction and interactive ML algorithms in which people modify the model in short increments (Fails and Olsen Jr 2003). In (Bekkerman et al. 2007), documents were clustered interactively according to user preferences. Users provided examples of task-specific features and iteratively corrected the system when it made errors. (Kim 2015) also developed an interactive tool for clustering, in which users indicated representative examples or important features, and the clusters were updated accordingly. In (Amershi, Fogarty, and Weld 2012), an interactive ML system, called ReGroup, was developed to allow people to interactively create social groups. The system learned a personalized model of group membership and suggested members and group characteristics based on this model.

In the context of interactive reinforcement learning, the TAMER framework, developed by (Knox and Stone 2009), modeled a human's reinforcement function, which a robot then used to select actions likely to result in the most positive human reinforcement. (Griffith et al. 2013) developed another approach in which interactive feedback was directly converted into a policy that guided the robot's learning. In (Alexandrova et al. 2014), robot learning was improved through an interactive action visualization tool in which users edited actions to guide the robot.

Interactive systems, as presented in these works, must be able to learn quickly so that humans can interact and provide feedback in real-time. Thus, it is important to use an appropriate representation and algorithm when developing human-in-the-loop learning systems. For sequential decision-making problems in which an agent is explaining a complex learning process to a person, the agent should represent tasks and similarities between tasks in an interpretable way, provide explanations of its reasoning, and learn quickly so that people can interact and provide feedback to the system in real-time. We believe developing such a system would be an important step to making more fluid and intelligent human-machine systems.

Impact

Interactive transfer learning using interpretable explanations is an important area for research because it would allow an agent to progressively learn new tasks by interacting with and receiving feedback from a person. It is a challeging and non-trivial problem to learn which prior task is most similar to a new task, adapt the solution, provide a high-level interpretable explanation, and accurately incorporate human feedback into the model. We believe this is an important area to further explore because it can be especially useful for a variety of applications that require an adaptive learning system, such as assistive robotics and intelligent tutoring systems. Prior work in cognitive psychology (Lombrozo 2012) shows that explanations, to oneself or to others, can significantly improve learning. Thus, having an interactive system that generates interpretable explanations can be an important step in making more intelligent adaptive systems.

Conclusion

In this position paper, we have identified an important and emerging area of interactive transfer learning for sequential-decision making problems using interpretable explanations. There has been great emphasis on making ML models achieve higher performance and more accuracy on a variety of tasks, but there has been relatively little focus on making these models more interpretable. We identify a gap in research on generating explanations for sequential decision-making problems in which an agent adapts and learns from previously learned knowledge. We see this as a growing area in ML because it provides a link between people and ma-

chines and can provide a communication channel for both to improve based on the other's strengths. More work in this area can provide a foundation for more effective, interactive adaptive learning systems for applications such as assistive robotics, medical assistance, and disaster response.

References

Alexandrova, S.; Cakmak, M.; Hsiao, K.; and Takayama, L. 2014. Robot programming by demonstration with interactive action visualizations. In *Robotics: science and systems*, 48–56.

Amershi, S.; Fogarty, J.; and Weld, D. 2012. ReGroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 21–30. ACM.

Amir, O.; Grosz, B. J.; Gajos, K. Z.; Swenson, S. M.; and Sanders, L. M. 2015. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare.

Ammar, H. B.; Eaton, E.; Taylor, M. E.; Mocanu, D. C.; Driessens, K.; Weiss, G.; and Tuyls, K. 2014. An automated measure of MDP similarity for transfer in reinforcement learning. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Bekkerman, R.; Raghavan, H.; Allan, J.; and Eguchi, K. 2007. Interactive clustering of text collections according to a user-specified criterion. In *IJCAI*, 684–689.

Bien, J., and Tibshirani, R. 2011. Prototype selection for interpretable classification. *The Annals of Applied Statistics* 2403–2424.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J. L.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *NIPS*, 288–296.

Clancey, W. J. 1983. The epistemology of a rule-based expert system – a framework for explanation. *Artificial intelligence* 20(3):215–251.

Dodson, T.; Mattei, N.; Guerin, J. T.; and Goldsmith, J. 2013. An english-language argumentation interface for explanation generation with Markov decision processes in the domain of academic advising. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3(3):18.

Elizalde, F.; Sucar, E.; Noguez, J.; and Reyes, A. 2009. Generating explanations based on Markov decision processes. In *MICAI: Advances in Artificial Intelligence*. Springer. 51–62.

Fails, J. A., and Olsen Jr, D. R. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, 39–45. ACM.

Fernández, F.; García, J.; and Veloso, M. 2010. Probabilistic policy reuse for inter-task transfer learning. *Robotics and Autonomous Systems* 58(7):866–871.

Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C. L.; and Thomaz, A. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *NIPS*.

Huysmans, J.; Dejaeger, K.; Mues, C.; Vanthienen, J.; and Baesens, B. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51(1):141–154.

Jonsson, A., and Barto, A. 2007. Active learning of dynamic Bayesian networks in Markov decision processes. In *Abstraction, Reformulation, and Approximation*. Springer. 273–284.

Khan, O.; Poupart, P.; and Black, J. 2011. Automatically generated explanations for Markov decision processes. *Decision Theory Models for Applications in Artificial Intelligence: Concepts and Solutions* 144–163.

Kim, B.; Patel, K.; Rostamizadeh, A.; and Shah, J. 2015. Scalable and interpretable data representation for high-dimensional, complex data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Kim, B.; Rudin, C.; and Shah, J. A. 2014. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In *NIPS*, 1952–1960.

Kim, B. 2015. *Interactive and Interpretable Machine Learning Models for Human Machine Collaboration*. Ph.D. Dissertation, MIT.

Knox, W. B., and Stone, P. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*, 9–16. ACM.

Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2013. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model.

Lombrozo, T. 2012. Explanation and abductive inference. *Oxford handbook of thinking and reasoning* 260–276.

Nikolaidis, S., and Shah, J. 2013. Human-robot cross-training: Computational formulation, modeling and evaluation of a human team training strategy. In *HRI*, 33–40. IEEE Press.

Ramakrishnan, R. 2015. Perturbation training for human-robot teams. Master's thesis, MIT.

Rüping, S. 2006. *Learning interpretable models*. Ph.D. Dissertation.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Swartout, W. R. 1977. A digitalis therapy advisor with explanations. In *Proceedings of the 5th international joint conference on Artificial intelligence-Volume 2*, 819–825. Morgan Kaufmann Publishers Inc.

Tan, M. 2014. Learning a cost-sensitive internal representation for reinforcement learning. In *Proceedings of the Eighth International Workshop on Machine Learning*, 358–362.

Taylor, M. E., and Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. *JMLR* 10:1633–1685.

Ustun, B., and Rudin, C. 2014. Methods and models for interpretable linear classification. *arXiv preprint arXiv:1405.4047*.

Wick, M. R., and Thompson, W. B. 1992. Reconstructive expert system explanation. *Artificial Intelligence* 54(1):33–70.