ISSN 1870-4069

Causal Based Q-Learning

Arquímides Méndez Molina, Ivan Feliciano Avelino,
Eduardo F.Morales, L. Enrique Sucar

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Coordinación de Ciencias Computacionales,
Mexico

arquimides.mendez@gmail.com, ivan.felavel@gmail.com

Abstract. Reinforcement learning and Causal Inference are indispensable part of machine learning. However, they are usually treated separately, although that both are directly relevant to problem solving methods. One of the challenges that emerge in Reinforcement Learning, is the trade-off between exploration and exploitation. In this work we propose to use causal models to attend the learning process of an agent. The causal models helps to restrict the search space by reducing the actions that an agent can take through interventional queries like: *Would I have achieved my goal if I had drop the passenger off here?*. This simulates common sense that lightens the time it takes the trial and error approach. We attack the classic taxi problem and we show that using causal models in the Q-learning action selection step leads to higher and faster jump-start reward and convergence, respectively.

Keywords: reinforcement learning, causal models, taxi domain.

1 Introduction

Reinforcement learning (RL) is the study of how an agent can learn to choose actions that maximize its future rewards through interactions with an environment [18]. RL is a technique to solve complex sequential decision making problems in several domains as healthcare, economics, robotics, among others. Existing studies apply RL algorithms in discovering optimal policies for a targeted problem, but ignores the abundant causal relationships present in the target domain.

Causal inference (CI) is another learning paradigm concerned at uncovering the cause-effect relationships between different variables [16,15]. CI addresses questions like: If I desire this outcome, what action do I need to take? So it may provide the information for an intelligent system to predict what may happen next so that it can better plan for the future. Given a causal structure of a system it is possible to predict what would happen if some variables are intervened, estimate the effect of confounding factors that affect both an intervention and its outcome, but also, predict the outcomes of cases that are never observed before.

Both reinforcement learning (RL) and causal inference have evolved independently and practically with no interaction between them, despite the fact that

both are directly relevant to problem solving processes. Nonetheless, recent work has focused on connecting these fields [8,9,20,5]. The goal of these works is to show how RL can be made more robust and general through causal mechanisms or vice versa. Also, a growth in what some are beginning to call (CausalRL) [12] is expected to become an indispensable part of General Artificial Intelligence. What CausalRL does seems to mimic human behaviors, learning causal effects from an agent communicating with the environment and then optimizing its policy based on the learned causal relations.

One of the challenges that emerge in Reinforcement Learning, is the trade-off between try new actions (exploration) and select the best action based on previous experience (exploitation) in a given state. Traditional exploration and exploitation strategies are undirected and do not explicitly chase interesting transitions. Using predictive models is a promising way to cope with this problem. In particular, these models may hold causal knowledge, that is, causal relationships.

In the present investigation we propose a method to guide the action selection in an RL algorithm using one or more causal models as oracles. The agent can consult those oracles to not perform actions that lead to unwanted states or choose the best option. This helps the agent learn faster since it will not move blindly. Through interventions in the causal model, we can make queries of the type *What if I do ...?*, e.g., If I drop the passenger off here, will my goal be achieved? This type of interventions can help to reduce the search space. An important distinction is that, in order to use a causal model as in favor of a reinforcement learning algorithm, we do not need it to be complete. In other words, we can think of one or several partial models that express relationships between variables of one or several subtasks of the general task we are trying to solve.

The remainder of this paper is organized as follows. Section 2 reviews related works. Section 3 describes in a very general way some concepts used in the proposal. Section 4 describes the proposed method. In Section 5 the experimental set-up is described and the main results presented. Finally, in Section 6, conclusions and future research directions are given.

2 Related Work

RL and CI have been widely explored separately [16,18]. Nevertheless, there are recent studies that are looking to connect the concepts of these two areas to set something they call *Causal Reinforcement Learning*, a paradigm that unites both approaches to solve problems that cannot be solved individually in each discipline [1,11]. The authors in [8], from a psychological approach, establish that the model used in model-based reinforcement learning algorithms it is causal. Taking an action in a state causes both a reward and a transition to a new state. However, the manipulationist mechanism is not addressed or explained.

Some other works have focus on handling confounders (those variables that affect action and output) in classic RL problems [2,12,7]. Besides that, it has been show that causal reasoning can arise from RL [4,13].

The idea of using knowledge from causal models to avoid or reduce trial-and-error learning in RL has not been explored, as far as we know. Authors in [14] propose a new method to speed up RL training through the use of a property that they define as *state-action permissibility*.

The main idea is to have a predictor that guides the action selection step. The predictor classifies whether an action leads to an optimal solution given the action and the current state. What distinguishes our work from this one is the use of causal model composed of state variables, actions and goals. Instead of consulting the model for predictions we propose to make intervention type queries so the agent is in the second rung of the ladder of causation.

3 Background

The definition of causality is that X causes Y, $X \rightarrow Y$, if and only if an intervention or manipulation in X has an effect on Y, keeping everything else constant [17].

A *graphical causal model* is a pair $M = \langle D, \Theta_D \rangle$ consisting of a *causal structure* D and a set of parameters compatible with D. A causal structure of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a different variable, and each arc represents a direct relationship among the corresponding variables [16]. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $P(u_i)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$, independently of all other u.

To better illustrate the above, consider the following example. Travis is a taxi driver whose main goal is to pick up a passenger at a certain point (passenger position) and take him to his destination (destination position) and drop him off there. For Travis, meeting his goal is based on his common sense. He doesn't try to pick a passenger when there is no passenger, drop him off there when he doesn't has arrived to the goal position, etc. We can create a causal model from the rules that guide Travis.

The parameters of our causal model can be defined as Boolean variables like in the set of equations 1, where $u_1, u_2 \in \{True, False\}$, u_3, c_4, c_5 can take some value that characterizes some position in the environment, e.g., coordinates in a map (c_4 and c_5 can be constant values). The rest of $u_i, u_i' \in \{True, False\}$ variables can be seen as unusual behaviors.

Let's suppose the case when $onDestinationLocation = False$, even when the taxi is on the same position as the passenger, maybe the passenger position has been updated without notifying the taxi driver, in this scenario $u_6' = True$.

The counterpart happens when $u_6 = True$, then the taxi is on the passenger position, see eq. 1 (the corresponding causal structure is shown in Figure 1):

$$\begin{aligned}
pickup &= u_1, \\
dropoff &= u_2, \\
cabPosition &= u_3, \\
destinationPosition &= c_4, \\
passengerPosition &= c_5, \\
onDestinationPosition &= [(destinationPosition = cabPosition) \vee u_6] \wedge \neg u'_6, \\
onPassengerPosition &= [(passengerPosition = cabPosition) \vee u_7] \wedge \neg u'_7, \\
inTheCab &= [(pickup = True \wedge onPassengerLocation = True) \\
&\quad \vee u_8)] \wedge \neg u'_8, \\
goal &= [(dropoff = True \wedge inTheCab = True \wedge \\
&\quad onDestinationLocation = True) \vee u_9] \wedge \neg u'_9.
\end{aligned} \tag{1}$$

Causal models, unlike probabilistic models, can serve to predict the effect of *interventions*. Interventions allow us to make queries of the type: Would the passenger be inside the taxicab if we make sure that the passenger is picked up here?. An *intervention*, which we denote by $do(X_i = x_i)$, means removing the equation $x_i = f_i(pa_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations [16]. The new model represents the system's behavior under the intervention $do(X_i = x_i)$ and, when solved for the distribution of X_j, produces the *causal effect* of X_i on X_j, which is denoted $P(x_j|do(X_i = x_i))$.

For example, to intervene on the variable $inTheCab$ in our example would be to set to one despite of whether the passenger was picked up. We would represent this by replacing the equation $inTheCab = pickup \times onPassengerLocation$ with $inTheCab = True$. Graphically, we can think of the intervention as "breaking the arrows" pointing into $inTheCab$.

4 Proposed Method

Our hypothesis is that causal inference can assist RL in learning value functions or policies more efficiently through the use of causal relations between state variables or between actions and state variables and therefore reducing the state or action space significantly.

To that end we proposed a method which consists of applying Algorithm 3 as a modification of the exploitation stage of Q-learning [19]. In general the method operates as follows. The agent observes a state, and through queries to one or more causal models, selects the action likely to allow the agent to meet a goal. The parameters of each causal model are given by a probabilistic SEM. The variables of the model are divided in three sets: state variables X, actions A and targets Z. The variables are defined as follows: $x = f_x(Pa_x), x \in X$,

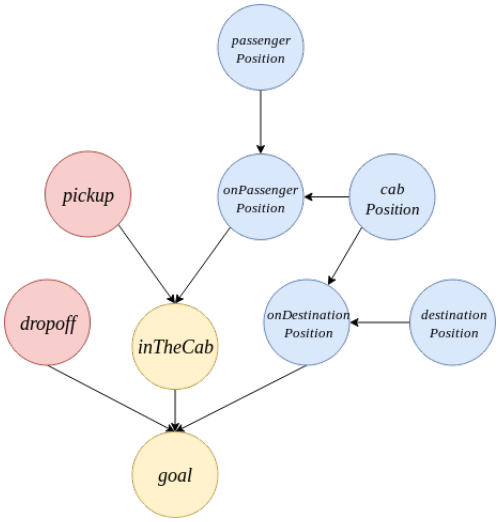

Fig. 1. Causal structure D for set of equations 1. The color of the nodes indicates to which set of variables corresponds. Red for actions (A), Yellow for target variables (Z) and blue for state variables (X). (Best seen in color).

$z = f_z(Pa_z), z \in Z$ where $Pa_x \subseteq X \cup A$ and $Pa_z \subseteq X \cup Z \cup A$. From the taxi example, the corresponding variables from Equation 1 for X, A, Z can be set as follows:

$$\begin{aligned} X = \{&passengerPosition, onPassengerPosition, cabPosition, \\ &onDestinationPosition, destinationPosition\}, \\ A = \{&pickUp, dropOff\}, \\ Z = \{&inTheCab, goal\}. \end{aligned}$$

In Algorithm 3, B is a set of observable instantiated variables, i.e., given the agent's observation we assign values to state variables from X. We assume that interventionist and observation distributions are already given so simply ask for $P(z|do(a), B)$ to obtain the causal effect in Algorithm 3 step 4. For our proposed method to work, the following assumptions must be meet:

- Non-empty set Z of target variables, can be ordered by a priority function.
- Non-empty set A of actions variables, containts only boolean variables.
- The agent can select only one action in a given state.
- All parameters of each Causal Model are defined.

5 Experimental Set-Up and Results

To show that our approach promises to be a way to improve RL we integrate it into the classical Q-learning algorithm. We replace the exploration step in

Algorithm 1: Q-Learning

input : $< S, A, R >$
output: Table Q

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q(e.g., ϵ - greedy)
        Take action a, observe r, s'
        $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma max_{a'} Q(s', a')) - Q(s, a))]$
        $s \leftarrow s'$
    until s is terminal or invalid
return Q

Algorithm 2: Causal Q-Learning

input : $< S, A, R >, G$
output: Table Q

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        $a \leftarrow$ interventional based selection using (s, G)
        If $(a = None)$:
            Choose a from s using policy derived from Q(e.g., ϵ - greedy)
        Take action a, observe r, s'
        $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma max_{a'} Q(s', a')) - Q(s, a))]$
        $s \leftarrow s'$
    until s is terminal or invalid
return Q

ϵ-greedy method to choose the actions by our method that queries the model. The problem to solve is the classical taxi task [6]. Figure 11 graphically shows the problem. A 5×5 grid world dwelled by a taxi agent. There are four locations in this world, marked as R, B, G, and Y.

The taxi problem is episodic. In each episode, the taxi starts in a randomly-chosen square. There is a passenger at one of the four locations (chosen randomly), and that passenger wishes to be transported to one of the four locations (also chosen randomly). The taxi must go to the passenger's location, pick up the passenger, go to the destination location , and drop the passenger off there. The episode ends when the passenger is deposited at the destination location.

There are six primitive actions in this domain: (a) four navigation actions that move the taxi one square North, South, East, or West; (b) a Pickup action; and (c) a Drop off action. The six actions are deterministic. There is a reward of -1 for each action and an additional reward of +20 for successfully delivering

Algorithm 3: Action selection based on interventional queries.

Input : A state s sense by the agent, A set of causal models G, A set Z of target variables of every $g \in G$ ordered by a priority function
Output: An action a.

$B \leftarrow get\_state\_observable\_values(s)$
foreach $z \in Z$ **do**
  foreach $a \in parents(z)$ *where* a *is an action variable* **do**
    $p \leftarrow P(z = True|do(a = True), B)$
      ▷ Here we get the causal effect on the target variable z through an intervention in the action variable a using the causal model g containing z.
    if $p > 0.5$ **then**
      return a
    end
  end
end
return *None*

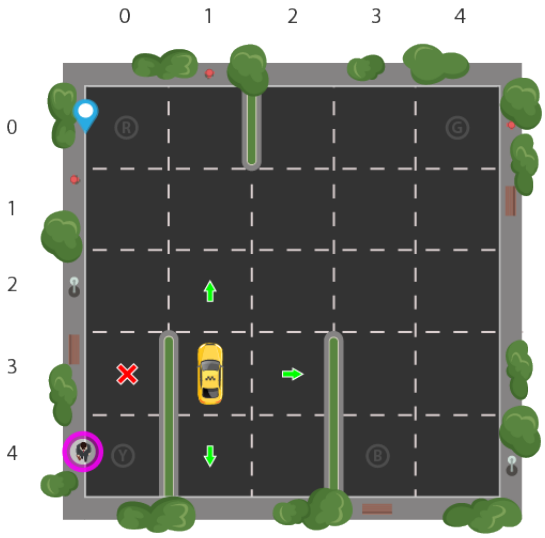

Fig. 2. Sketch of the taxi enviroment [10].

the passenger. There is also a 10 point penalty for illegal pick-up and drop-off actions [6]. There are 500 possible states: 25 squares, 5 locations for the passenger (including when he's inside the cab), and 4 destinations.

The causal model that is consulted to choose the actions is the presented in Section 3, extending it to queries on movement actions, so that the agent does not try move to positions where there are obstacles. For ease, we got rid of the u_i variables.

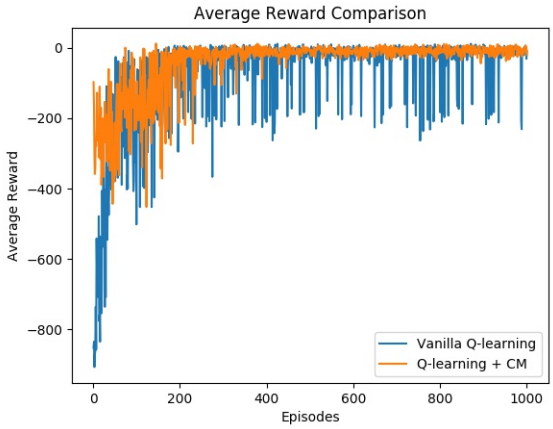

Fig. 3. Average reward of Vanilla Q-learning and Q-learning guided by a causal model.

From the model in Figure 1 the color of the nodes indicates to which set of variables corresponds. Red for actions (A), yellow for target variables (Z) and blue for state variables (X). Since the environment is deterministic, there is no need to compute a probability for the value of a target variable. Instead, we evaluate whether the value of the target variable is $True$ given the action and B.

As our baseline we implement a vanilla version of the Q-learning algorithm and we compare it with our version to which we denominate Q-learning + Causal Model (CM). We run 50 times each version of the algorithm and in each execution we compute the average reward per episode. Also, we set a qualifying mark based on the one established by Open AI Gym [1]. For this, we consider that the algorithm had reached an optimal reward once the average reward is equal to 9. So we assume that the algorithm that achieve it a smaller number of episodes is faster. On average, vanilla Q-learning reaches that reward in 95 episodes and Q-learning + CM in 65 episodes. In order to validate the results that the guided Q-learning version of the algorithm performs better than the vanilla version, we use the Wilcoxon Mann-Whitney rank sum test[3] with $p < 0.001$ to find statistical significant differences.

Figure 3 show the average reward per episode in both version of the algorithm for (average over 10 experiments). From the plot we can observe that our guided version starts with a higher reward. This is to be expected because, the agent doesn't start blindly. For a range of episodes there is no difference between the methods. However after a couple of hundred episodes, the Q-learnig guided by a causal model seems to converge and keeps more stable.

[1] `https://gym.openai.com/envs/Taxi-v1/`

6 Conclusions

Reinforcement Learning has proved to be successful in decision making problems. On the other hand, causal inference is clearly a novel but relevant and related area with untapped potential for any learning task. The use of causal models to provide auxiliary knowledge to an RL algorithm is a barely explored area. However, from the results obtained, we can see that this type of knowledge has the potential to accelerate RL. Although the problem attacked is simple because all the causes we have are direct and observable, the experimental results show that using causal models in the Q-learning action selection step leads to higher and faster jump-start reward and convergence, respectively. As future work we would like to try this action selection framework in Deep RL algorithms to solve more complex problems. Coping with more complex problems involves tasks not covered in this work, for example, undefined model parameters, incomplete causal structure or an unreliable causal model. In addition, we would like to explore the possibility that the causal model could also be learned during the training of the RL algorithm.

References

1. Bareinboim, E.: Causal reinforcement learning. `https://crl.causalai.net/` (2019)
2. Bareinboim, E., Forney, A., Pearl, J.: Bandits with unobserved confounders: A causal approach. In: Advances in Neural Information Processing Systems. pp. 1342–1350 (2015)
3. Colas, C., Sigaud, O., Oudeyer, P.Y.: A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms (2019)
4. Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., Kurth-Nelson, Z.: Causal reasoning from meta-reinforcement learning. arXiv preprint arXiv:1901.08162 (2019)
5. Dasgupta, I., Wang, J.X., Chiappa, S., Mitrovic, J., Ortega, P.A., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., Kurth-Nelson, Z.: Causal reasoning from meta-reinforcement learning. CoRR abs/1901.08162 (2019), `http://arxiv.org/abs/1901.08162`
6. Dietterich, T.G.: Hierarchical reinforcement learning with the maxq value function decomposition. J. Artif. Int. Res. 13(1), 227–303 (Nov 2000), `http://dl.acm.org/citation.cfm?id=1622262.1622268`
7. Forney, A., Pearl, J., Bareinboim, E.: Counterfactual data-fusion for online reinforcement learners. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1156–1164. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017), `http://proceedings.mlr.press/v70/forney17a.html`
8. Gershman, S.J.: Reinforcement learning and causal models. The Oxford handbook of causal reasoning p. 295 (2017)
9. Ho, S.: Causal learning versus reinforcement learning for knowledge learning and problem solving. In: The Workshops of the The Thirty-First AAAI Conference on Artificial Intelligence, Saturday, February 4-9, 2017, San Francisco, California,

USA. AAAI Workshops, vol. WS-17. AAAI Press (2017), http://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15182
10. Kansal, S.: Reinforcement q-learning from scratch in python with openai gym (2018), https://www.learndatasci.com/tutorials/reinforcement-q-learning-scratch-python-openai-gym/
11. Lu, C.: Introduction to causalrl (Jan 2019), https://causallu.com/2018/12/31/introduction-to-causalrl/
12. Lu, C., Schölkopf, B., Hernández-Lobato, J.M.: Deconfounding reinforcement learning in observational settings. CoRR abs/1812.10576 (2018), http://arxiv.org/abs/1812.10576
13. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. CoRR abs/1905.10958 (2019), http://arxiv.org/abs/1905.10958
14. Mazumder, S., Liu, B., Wang, S., Zhu, Y., Yin, X., Liu, L., Li, J., Huang, Y.: Guided exploration in deep reinforcement learning (2019), https://openreview.net/forum?id=SJMeTo09YQ
15. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Penguin Books Limited (2018)
16. Pearl, J.: Causality: models, reasoning, and interference. Cambridge University Press (2009)
17. Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science 28(7), 075310 (2018), https://doi.org/10.1063/1.5025050
18. Sutton, R.S., Barto, A.G.: Reinforcement learning: an introduction. The MIT Press. (2018)
19. Watkins, C.J., Dayan, P.: Q-learning. Machine learning 8(3-4), 279–292 (1992)
20. Yu, C., Dong, Y., Liu, J., Ren, G.: Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV. BMC Med. Inf. & Decision Making 19-S(2), 19–29 (2019), https://doi.org/10.1186/s12911-019-0755-6