Explaining a Reinforcement Learning Agent via Prototyping

Anonymous Author(s)
Affiliation
Address
email

Abstract

While deep reinforcement learning has proven to be successful in solving control tasks, the "black-box" nature of an agent has received increasing concerns. We propose a prototype-based post-hoc *policy explainer*, ProtoX, that explains a black-box agent by prototyping the agent's behaviors into scenarios, each represented by a prototypical state. When learning prototypes, ProtoX considers both visual similarity and scenario similarity. The latter is unique to the reinforcement learning context, since it explains why the same action is taken in visually different states. To teach ProtoX about visual similarity, we pre-train an encoder using contrastive learning via self-supervised learning to recognize states as similar if they occur close together in time and receive the same action from the black-box agent. We then add an isometry layer to allow ProtoX to adapt scenario similarity to the downstream task. ProtoX is trained via imitation learning using behavior cloning, and thus requires no access to the environment or agent. In addition to explanation fidelity, we design different prototype shaping terms in the objective function to encourage better interpretability. We conduct various experiments to test ProtoX. Results show that ProtoX achieved high fidelity to the original black-box agent while providing meaningful and understandable explanations.

1 Introduction

Deep reinforcement learning (DRL) has been widely used to solve various control problems such as autonomous driving [1], gaming [2], etc. Despite their superior performance, DRL agents are often criticized for being "black-boxes" in nature and unable to provide human-understandable explanations for their actions. Popular solutions to interpreting a black-box agent include attention-based methods [3] and saliency maps [4, 5, 6]. These explanations, however, only inform users of where the agent "looks at", but do not provide a more specific explanation for *why an agent takes this action in a given state*. Some recent methods answer this question by formulating a classification problem and build a decision tree from state information to predict an action [7]. However, tree-based models face the limitation of feature representation, that they can only work with meaningful features in order to make sense to humans. They do not naturally apply to unstructured data such as images and texts unless human-understandable information is manually extracted.

This paper aims to *explain* a black-box agent that works with unstructured data. To make the explanations understandable, we make use of the concept of *prototypes* from recent DNN models for image classification [8, 9]. A prototype is an instance (e.g. an image) from data that is representative of a prominent feature for a certain class. For examples, the red-bellied woodpecker class could have a prototype for its red head and a prototype for its wings with a zebra-like pattern, etc. Then, when there is a new input, the input can be compared against all prototypes trained from data. If the input also has a red head and zebra-like pattern on the wings, the weighted sum of similarities in the red-bellied woodpecker class will be high, and thus the input will be predicted to be in this class.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this paper, we propose a policy explainer called ProtoX, which explains an action by relating an input state to a set of prototypical behaviors learned from the demonstrations from the agent. Figure 1 shows an example of ProtoX explanation for an agent playing the SuperMario game. ProtoX learns a set of prototypical states during training, which are most representative of the agent's behavior. Then, for a given input state where Mario stands in front of a pipe and the agent plays the action RIGHT + JUMP, ProtoX explains this action by collecting *evidence* from the similarities to the prototypical behaviors. The evidence is the sum of similarities between the input state and the prototypes, weighted by the action connections. An action connection represents how strongly a prototype supports the prediction of an action and in which direction. Thus, naturally, the action with the largest evidence (RIGHT + JUMP in Figure 1 has the largest evidence of 27.8) is predicted to be the action of the agent for the given state.

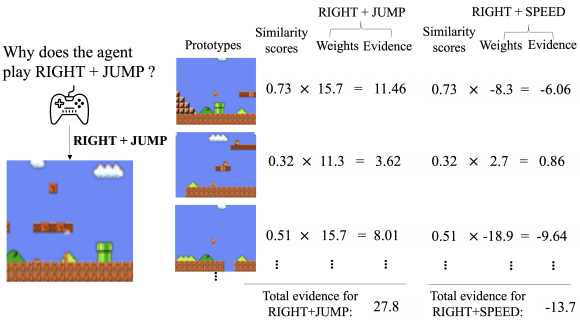


Figure 1: Mario jumps over the pipe because his situation is similar to general scenarios of jumping over an obstacle, a hole, or jumping to hit a question mark.

In summary, ProtoX prototypes an agent's behavior into a small set of representative behavior patterns, each represented by a prototype, and explains the agent's action based on these basic patterns. Therefore, a key determinant of the success of ProtoX is the prototypes. An important consideration when computing the similarity between an input and a prototype is the **visual** similarity. If two states look entirely different, the high similarity between them will confuse human users and discredit the explanation. In ProtoX, we design a self-supervised training step for the encoder to learn from demonstrations of experts, where images that are closer to each other in time and receive the same actions are considered more similar than those far away or receiving different actions. Thus, we create a siamese net [10, 11] that utilizes a modified form of the time-contrastive loss. Our pre-training routine yields encoders whose feature spaces cluster together states which occur closely in time and have the same corresponding actions. Consequently, visually similar states are closer together in the latent space of the encoder.

In the supervised learning setting, visual similarity would be the only consideration, given the nature of image classification tasks. However, in a reinforcement learning setting, two visually similar states could receive different actions (such as the flip points we define later), and two visually dissimilar states could represent a similar "scenario" that requires the same action. Therefore, the similarities need also need to take into account **scenario** similarity. Here, a scenario represents a collection of states in which an agent performed the same action for similar reasons. For example, in Figure 1, an action RIGHT+JUMP has a high similarity score to prototypes that represent different scenarios for jumping, which could be that Mario encounters a pipe (the 1st prototype), a hole (the 3rd prototype), or preparing to hit a question mark brick (the 2nd prototype). In fact, the scenario similarity brings more value to the explanation, since it helps users to think about what is common between two visually dissimilar states that leads to the same action. To allow the feature representations to incorporate scenario similarity, we add a near-isometry layer after the encoder, which applies a linear transformation of the output of encoder to allow some states to move closer to each other.

ProtoX is trained end-to-end via imitation learning, by learning from the demonstrations of the black-box agent to be explained. We choose behavioral cloning [12] in this paper because it requires minimal information from both the agent and the environment: only demonstrations from the agent and no access to the environment or the agent. Therefore, it is most applicable in practice. Many other imitation methods, such as GAIL-based methods and inverse reinforcement learning methods, require more information from the environment or/and the agent, such as knowing the transition functions, executing a policy during training, etc.

We evaluate ProtoX on OpenAI Gym classic control benchmark tasks and compare it with three baseline models. Since our model provides a type of post-hoc explanation, we assess the *fidelity* of the explanation in terms of accuracy in predicting a black-box agent's actions. We find that ProtoX achieves high fidelity while providing meaningful and human-understandable explanations via learned prototypes.

2 Related Work

RL interpretation can be categorized into two main groups: inherently interpretable RL models and post-hoc explanation approaches. Inherently interpretable RL models generate policies that are intrinsically understandable, such as Neurally Directed Program Search [13], fuzzy RL policies [14], hierarchical policies [15, 16, 17], or extract time step importance within episodes as strategy-level explanations [18]. Post-hoc methods, like the proposed ProtoX, aim to explain a given black-box agent to help users understand the agent's decision-making process towards an action. The goal is not good task performance but high fidelity in explaining a given black-box agent, regardless of the own performance of the agent.

Based on the type of data they work with, post-hoc methods can also be split into two areas. The first class of methods only work with structured data, such as genetic programming [19], VIPER[7] and other rule-based methods [20, 21, 22], expected consequences [23], causal lens [24], complementary RL [25], etc. These methods often adopt existing forms of interpretable models from supervised learning, such as rules [7, 25]. However, since the explanations need to directly use state features, they do not naturally work with unstructured data. To learn video games from pixels, they require meaningful features to be extracted *manually*. The other type of post-hoc methods work with structured data. Examples include saliency-based methods [26, 27, 28, 25, 4], reward decomposition [29], and interestingness elements [30]. The explanations, however, are either only at the global level or too abstract, such as saliency maps that only inform users of where the agent "looks at", but do not directly explain the policy.

Another stream of work that aims to provide some interpretability stem from the family of imitation learning. For example, inverse reinforcement learning (IRL) methods [31, 32, 33] aim at learning a reward function under which the observed demonstrations of an agent is near-optimal. The interpretability, however, is only achieved through the learned reward function with a pre-defined structure (e.g., linear combination of state features) rather than understanding the policy itself. Methods based on generative adversarial imitation learning (GAIL) [34] such as xGAIL [35] and infoGAIL [36] provide indirect explanations to the learned policy in terms of global and local feature importance or latent factors that govern the policy generation. However, none of them provide a direct explanation of the expert's actions under each state that are intuitive to human. More importantly, all of the above methods require intensive interactions with the environment or the agent during training (such as executing a policy), which are often not accessible in most practical situations. On the other hand, our proposed ProtoX model is completely offline and requires no environment or agent interaction. It also needs no additional information from the environment such as the transition function, reward function, etc.

3 Prototype-based Policy Explainer

We consider a discrete-time partially observable Markov decision processes (POMDP), which include video games such as the OpenAI Gym classic control games[37]. We denote the observation (e.g., frame of video game) at time t by o_t and the C latest observations at time t by $s_t = (o_t, o_{t-1}, \ldots, o_{t-C+1})$. We call s_t the state at time t and denote the set of all states across all times as a state space $\mathcal{S}$.[1] Let $\mathcal{A}$ be the action space of the POMDP and we consider a policy that chooses an action $a \in \mathcal{A}$ at time t based on the C latest observations, namely, a mapping $\pi : \mathcal{S} \to \mathcal{A}$.

Given a policy by an expert $\pi^* : \mathcal{S} \to \mathcal{A}$, our goal is to build a prototype-based policy explainer that is parameterized by ϕ, denoted as $\pi_\phi : \mathcal{S} \to \mathcal{A}$, called ProtoX, such that $\pi_\phi(s) = \pi^*(s), \forall s$. We train π_ϕ via behavioral cloning, which proceeds with two steps. First, we apply π^* to the POMDP to generate a set of state-action trajectories and decompose all trajectories into a dataset of the expert's actions at different states, denoted by $\mathcal{D} = \{(s_i, \pi^*(s_i))_i\}$. Next, we train a classifier to predict actions from states.

We present the architecture of ProtoX, which consists of four main components, an encoder, a near-isometry layer, a prototype layer, and a fully connected layer. See Figure 2 for an illustration of the architecture. Below, we describe its main components, and then formulate the learning objective which considers both the fidelity and the interpretability.

[1] We have a little abuse of terminology here as $\mathcal{S}$ is different from the traditional state space of a POMDP.

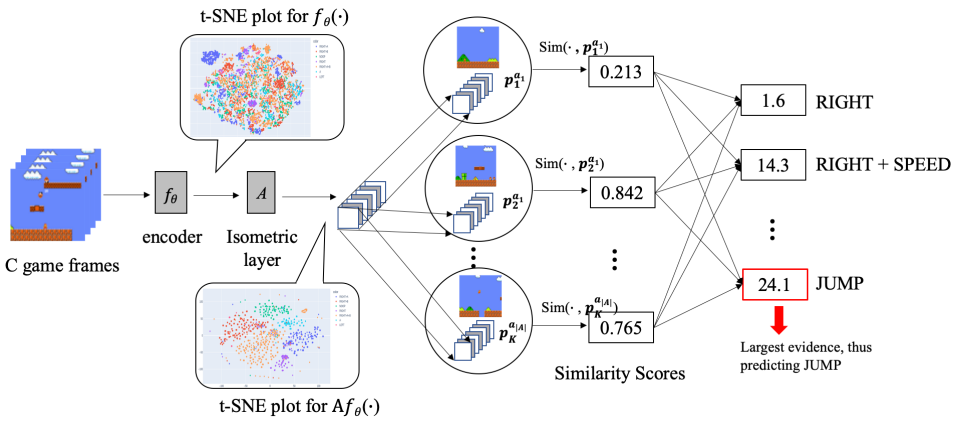


Figure 2: The model architecture of ProtoX.

3.1 Feature Representations Preserving Visual and Scenario Similarity

A key determinant of the success of ProtoX is that the feature representation of the states that can respect the visual similarity *and* reflect scenario similarity. Therefore, ProtoX needs to recognize task-relevant features from the images to benefit the downstream task, as well as understanding what is considered "visually similar". However, we do not have labeled data that supervises an encoder to determine which images are similar and which are not. Therefore, we use a self-supervised learning approach. We construct an objective via contrastive learning and train an encoder f_θ parameterized by θ using a siamese variational auto-encoder (VAE) [38]. The contrastive portion of our objective is similar to the time-contrastive triplet loss from [10], based on the idea that images that are closer to each other in time are more likely to be visually similar to each other. Our loss function differs slightly in that it uses a time-contrastive *quadruplet* loss. We use a quadruplet loss so that the encoder learns that states are the most similar when they occur close in time *and* are associated with the same action.

For a given state s_t, which we call an anchor state, a positive (visually similar) image s_t^+ is one that is at most δ away in time in an episode and receives the same action from π^*, i.e., $\pi^*(s_t) = \pi^*(s_t^+)$. We construct two types of negative (dissimilar) states. We find one negative state s_t^- among states that are within the same temporal neighborhood but receive a different action than the anchor. We find a second negative state s_t^{--} among states that are outside the temporal window. Since s_t^{--} is far away from s_t in time, it is highly likely they look dissimilar. Thus, a quadruplet $(s_t, s_t^+, s_t^-, s_t^{--}) \in \mathcal{S}^4$ is formulated. See Figures 3 for an illustration of the idea in Super Mario.

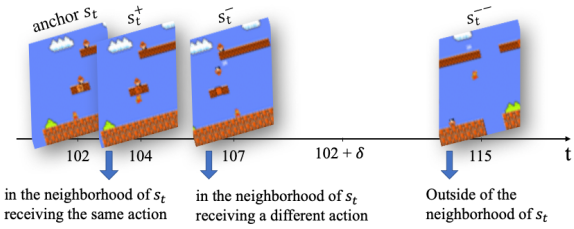


Figure 3: The model trains itself by trying to answer the following questions simultaneously: What is common between the s_t and s_t^+ frames? What is different between the similar-looking s_t and s_t^-? What is considered visually dissimilar between s_t and s_t^{--}?

The quadruplet loss is then defined on $(s_t, s_t^+, s_t^-, s_t^{--})$ as

$$\mathcal{L}_{\text{quadruplet}}(s_t, s_t^+, s_t^-, s_t^{--}) = \max(\|f_\theta(s_t) - f_\theta(s_t^+)\|_2^2 - \|f_\theta(s_t) - f_\theta(s_t^-)\|_2^2 + m_1, 0) \quad (1)$$
$$+ \max(\|f_\theta(s_t) - f_\theta(s_t^+)\|_2^2 - \|f_\theta(s_t^-) - f_\theta(s_t^{--})\|_2^2 + m_2, 0)$$

m_1 and m_2 are referred to as margins. m_1 is used so that, by minimizing $\mathcal{L}_{\text{quadruplet}}$, f_θ views anchors as at least m_1 closer to positives than to the first type of negatives. The use of m_2 is to ensure that the two types of negatives are pushed at least m_2 away from each other. The full loss function for the siamese VAE is the quadruplet loss plus the conventional VAE loss function. Pseudocode and further details for our pre-training algorithm can be found in the supplementary material.

Once f_θ is trained, it will be frozen for the down-stream training, such that the property discussed above is preserved.

Output of the Encoder The encoder is able to map visually similar states closer in their encodings. In general, we find that the most similar states to a given state are those that almost look identical. The less similar states are generally those that could occur in a small window of time before or after the given state. We show the efficacy of the encoder by aggregating the most similar 30 states together, i.e., the 30 states with the smallest $||f_\theta(x) - f_\theta(x')||_2$. See Figure 4 for an example that shows what states the pre-trained encoder views as similar in Super Mario Bros. level 1-1. This example shows that the states similar to the input are those where Mario is in the process of jumping over a hole in the ground.

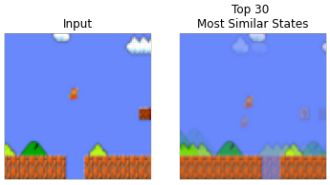


Figure 4: An input state x and top 30 states most similar to $f_\theta(x)$ overlaid.

Learning Scenario Similarity Note that the training of f_θ does not involve the downstream tasks, thus the feature representations, while respecting the visual similarity, may not consider the task relevance, i.e., the scenario similarity, which is task-specific and agent-specific. Similar failure modes in contrastive learning are explored in [39]. Therefore, to overcome the possibility of the f_θ failing to be useful to the downstream task, we add a linear layer $A : \mathbb{R}^{C' \times H' \times W'} \rightarrow \mathbb{R}^{C'H'W'}$ after the encoder. Since A is learned during training, it will further adapt the representations of the input states to the downstream RL task. Howsever, we also want to make sure this linear transformation of A does not undo ProtoX's ability to understand visual similarity. Thus, A needs to be nearly isometric, such that, via[40], if $\left\|A^\top A - I\right\|_2^2 \leq \delta$ (where $\delta > 0$), then for all data x, x'

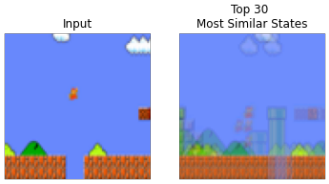


Figure 5: An input state x and top 30 states most similar to $Af_\theta(x)$ overlaid. ProtoX identifies the green pipe being similar to the hole since they both activate the action of JUMP.

$$\|Af_\theta(x) - Af_\theta(x')\|_2^2 \leq (1+\delta)\|f_\theta(x) - f_\theta(x')\|_2^2$$

Therefore, we will include a penalty term in the learning objective such that the isometry layer allows some evolution of the encoder Af_θ while still retaining the way f_θ learned to separate dissimilar states during pre-training.

$$Iso \quad = \quad \left\|A^\top A - I\right\|_2^2, \tag{2}$$

where I is the identity mapping of the appropriate dimension.

The t-SNE plots in Figure 2 explain why the isometry layer benefits the downstream task. We show the 30 states most similar to the same input in Figure 4, this time after the isometry layer, i.e., the 30 states with the smallest $||Af_\theta(x) - Af_\theta(x')||_2$. In contrast to Figure 4, in the example in Figure 5, ProtoX identifies scenario similarity, as it considers jumping over a hole is similar to jumping over a green pipe. For more analysis of the isometry layer, please see the supplementary material.

3.2 Learning Prototypes

We then add a prototype layer after the isometry layer that automatically learns K prototypes for each action in $\mathcal{A}$, represented by $\{p_k^a\}_{k=1,\cdots,K}^{a \in \mathcal{A}}$. Note that the prototypes are learned via an end-to-end training. Each prototype has a shape of $1 \times C'H'W'$, which is the latent representation of an actual image from training data. Given an input state x, ProtoX computes the distances $\|Af_\theta(x) - p_k^a\|$ for $k = 1, \ldots, K$ and $a \in \mathcal{A}$. It uses these distances to compute similarity scores between the image and prototypes, following the function from [41]:

$$\text{sim}(Af_\theta(x), p_k^a) = \exp(-\beta \|Af_\theta(x) - p_k^a\|_2) \tag{3}$$

where β controls the concentration of similarity scores and is set to 0.05 in all our experiments.

The similarities are then fed into a fully connected layer with a weight matrix $W \in \mathbb{R}^{|\mathcal{A}|K \times |\mathcal{A}|}$ with $W = (w_{a',k}^a)_{k=1,\ldots,K,a' \in \mathcal{A}}^{a \in \mathcal{A}}$. The subscript represents the prototypes and the superscript represents the output actions, and there are $|\mathcal{A}|K$ prototypes in total. The linear layer outputs $|\mathcal{A}|$ evidence scores, each of them linearly combines the contributions from all prototypes:

$$e^a(x, W) = \textstyle\sum_{k=1,\ldots,K,a' \in \mathcal{A}} w_{a',k}^a \cdot \text{Sim}(Af_\theta(x), p_k^{a'}), a \in \mathcal{A} \tag{4}$$

The linear model is chosen for its ease of explanations. It could be replaced by other interpretable models like decision trees, following the idea in [9]. The evidence score is computed for every action and the final prediction (action) is made as $\text{argmax}_{a\in\mathcal{A}} e^a(x, W)$. Given this decision rule, we next design an objective function that encourages the model to associate scenarios with an appropriate action to perform.

3.3 Objective Function

Since π_ϕ is a post-hoc explainer for an expert π^*, we consider two desiderata when training π_ϕ: 1) **fidelity** of imitation, and 2) **interpretability** of the prototypes. We construct the following terms in an objective function for the two purposes.

Fidelity A cross-entropy term, denoted as $\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{CE}(y_i, (e^a(x_i, W))_{a\in\mathcal{A}})$, is included to make sure ProtoX can reproduce the actions of the black-box expert as accurately as possible, which reflects the fidelity in the explanation. Here, e^a is defined in (4).

Separation and Clustering Terms The terms Clst (Clustering) and Sep (Separation) are taken from the ProtoPNet [8]. The Clst term encourages prototypes to be close to the encodings of training instances that have the same action label as the prototypes. The Sep term encourages prototypes to be far from encoded training instances with different action labels as the prototypes. This term enforces consistency with our time-contrastive loss, since both state that image representations should be dissimilar if their corresponding actions differ. The Clst and Sep terms together encourage the model to learn prototype scenarios such that similar scenarios have an associated appropriate action and dissimilar prototype scenarios have different appropriate actions.

$$Sep = -\frac{1}{n}\sum_{i=1}^{n}\min_{k,a \text{ s.t. } a\neq y_i}\|Af(x_i) - p_k^a\|_2,\ Clst = \frac{1}{n}\sum_{i=1}^{n}\min_k\|Af(x_i) - p_k^{y_i}\|_2, \quad (5)$$

Representability term The Rep (representability) term encourages prototypes to be similar to *some* encoded training instances, to insure the representability of the prototypes. Because our pre-training routine already ensures that encodings $f_\theta(s_t)$ are meaningful representations, we use this Rep term to further ensure that the prototypes are representative of the data. By continuously ensuring that prototypes are close to actual encodings of training data, we justify the success of prototype projection steps that will be explained later. This term is inspired by the Evidence term of [41].

$$Rep = \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\min_i\|Af(x_i) - p_k^a\|_2^2, \quad (6)$$

Isometry Term As discussed previously, the isometry term in formula (2) allows some evolution of the encoder Af_θ while still retaining the way f_θ learned to separate dissimilar states during pre-training.

Therefore, the objective function is

$$\mathcal{L}(\{p_k^a\}_{k=1,\ldots,K}^{a\in\mathcal{A}}, A, W) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{CE}(y_i, (e^a(x_i, W))_{a\in\mathcal{A}}) + \lambda_1 Sep + \lambda_2 Clst + \lambda_3 Rep + \lambda_4 Iso, \quad (7)$$

where $\lambda_i \geq 0$, $i = 1, \ldots, 4$, are weighting parameters subject to tuning. The first term encourages high fidelity and the last four terms are for prototype shaping.

3.4 Downstream Training Routine

We train our model with the standard behavior cloning (BC) algorithm, but with the aforementioned surrogate loss $\mathcal{L}$. This training routine is essentially identical to that of the related ProtoPNet[8].

Prototype Projection The standard image classification training pipeline is augmented with a *prototype projection* step every 25 epochs. During this step, each prototype in ProtoX is replaced with the encoding of the training datum it is most similar to, i.e.,

$$\mathbf{p}_k \leftarrow Af_\theta(x),\ \text{where } x = \arg\min_{x\in\mathcal{D}}||Af_\theta(x) - \mathbf{p}_k||_2. \quad (8)$$

Then, the A layer and the prototype vector are frozen, and the fully connected layer is trained by itself for 20 epochs. The representability term in the objective ensures the success of the projection step since each prototype is already close to at least one training instance, thus projecting them does not cause a significant change in the loss function.

Prototype Merging Once a ProtoX model is trained, and we project prototypes into the space of encodings of real states, we often find that there is a significant amount of repetition among the prototypes. To eliminate the redundant prototypes, we create a new prototype layer consisting only of the unique prototypes. We then also create a new fully connected layer whose weights capture the redundancy present in the previous prototype layer. Supposing that prototype $i = i_1$ is the same as prototypes $i_2, ..., i_k$ after projection, the connection between prototype i and action l in the fully connected layer is $w_{il} = \sum_{j=1}^{l} w_{i_j l}$. We find that this prototype merging procedure effectively reduced the number of prototypes while retaining the performance of the network. This also means when we build ProtoX, we can over parameterize it by allowing it to learn a large number of prototypes, but they will eventually be projected to a small number of unique prototypes.

4 Experiments

We evaluate ProtoX on four classic control problems from OpenAI Gym and compare the results with two black-box and one interpretable baseline. We first simulate a near optimal black-box agent and then use different methods to imitate the black-box. We then evaluate the **fidelity** of different methods, measured as the accuracy of their predictions compared to the agent. In addition, we evaluate the sensitivity of ProtoX on flip points, to test whether ProtoX can capture important time steps in a reinforcement learning task. Then we investigate the interpretability of the explanations generated by ProtoX. Finally, we simulate a bad agent and uses ProtoX to diagnose its bad behaviors.

4.1 Experimental Setup

We use four classic control tasks from OpenAI Gym, namely, Pong, Seaquest and two levels from Super Mario Bros[42]. For each game, we generate a near-perfect black-box agent, denoted as π^*. Pong and Seaquest use PPO[43] models from the stable-baselines3[44] package. The experts for Super Mario are also PPO models from [45]. Then we train ProtoX and other methods to imitate each agent. See the Appendix for hyperparameter settings and further details on our experimental design.

We compare ProtoX with three baselines, namely, 1) VIPER[7] which builds a decision tree model to output policies. Since VIPER only works with structured data and does not directly work with images, we extract features from the Pong and Seaquest games. Due to the complexity of Super Mario, we do not benchmark VIPER on the Super Mario games. 2) a GAIL-based method that learns from observational data generated by an agent, called GAIfO[46]. GAIfO is a black-box model. Note that unlike ProtoX, GAIfO still interacts with the environment. 3) ResNet-BC, which follows the same training procedure as ProtoX except using a ResNet18 model to work with the game frames. ResNet-BC is a black-box model that does not explain an agent. We run ResNet-BC to evaluate how much fidelity is lost due to using prototypes; Both ProtoX and ResNet-BC are trained with the behavior cloning algorithm using 30,000 state-action pairs obtained via an expert trained with PPO. VIPER is trained with the DAgger algorithm and queries the same PPO expert during its data aggregation step. GAIFO's discriminator uses the aforementioned PPO expert's state transitions to judge the quality of the generator. All models utilizing CNN encoders used a ResNet18[47] as the encoder. See the supplementary material for hyper-parameter settings.

4.2 Fidelity Evaluation

We first evaluate the fidelity of each method, i.e., how often each method agrees with the PPO agent π^*. For each game, we let the agent generate a test set of $10,000$ state-action pairs $\mathcal{D}_{\text{test}} = \{s_i, \pi^*(s_i)\}_i$. We then apply each method to the states and record their predicted actions $\{\pi(s_i)\}_i$, which is compared against $\{\pi^*(s_i)\}_i$ to compute the accuracy, i.e., $\frac{\sum \mathbb{1}(\pi^*(s_i)=\pi(s_i))}{|\mathcal{D}_{\text{test}}|}$ recorded in Table 1.

Table 1: Fidelity evaluation of different methods.

Datasets	Pong	Seaquest	Super Mario 1-1	Super Mario 8-3
ProtoX (**interpretable**)	95%	96%	93%	95%
VIPER (**interpretable**)	45%	31%	–	–
GAIfO (**non-interpretable**)	19%	11%	33%	25%
ResNet-BC (**non-interpretable**)	95%	98%	99%	99%

ProtoX achieved much better fidelity than the existing interpretable method, VIPER, since VIPER is a decision tree-based model and does not allow image state inputs. We set the maximum tree depth to 10 for both games, which will produce a tree with thousands of nodes for both games. Increasing the maximum tree depth will slightly increase the fidelity by less than 3 percent, but it will significantly increase the number of nodes in the tree. Yet VIPER still performs much worse than ProtoX. In addition, as numerical features (coordinates of a ball, or a submarine, etc) are more difficult to understand than images in video games. GAIfO also performs poorly. GAIL based methods are typically only applied to low dimensional problems like joint dynamics [34, 46]. When applied to high-dimensional Atari problems, GAIL has been shown to achieve poor performance [48]. We arrived at a similar conclusion with our experiments.

The comparison of ProtoX and ResNet-BC, even though ResNet-BC is not interpretable, shows that the interpretability design in ProtoX has small negative impact on the fidelity, a 3% drop on average. When training ProtoX, we purposely over-parameterize the network by using more prototypes than should be necessary. For example, when training ProtoX for Super Mario Bros. World 1-1, we use 350 prototypes initially. However, after the prototype projection and merging step, we are able to significantly reduce the number of prototypes in a trained ProtoX model, thus making each explanation low in complexity. In the case of Super Mario Bros, we can reduce the number of prototypes from 350 to only 17 while retaining relatively high fidelity. We report the number of prototypes for ProtoX models in Table 2. The small number of prototypes ensure that the explanations are easy to follow.

Table 2: The number of prototypes in ProtoX models

Datasets	**# of prototypes**
Pong	33
Seaquest	42
SM 1-1	17
SM 8-3	12

4.3 Sensitivity to Flip Points

For a reinforcement learning agent, usually the most important explanations are for time steps where the action changes from the previous action. This is because there's often no significant pixel change in two consecutive game frames, and thus a change in the action would be worth investigating. Therefore, if a state receives a different action from its previous state, we call it a *flip point*. The set of flip points can be represented as $\mathcal{F} = \{s_t \in \mathcal{S} | \pi^*(s_t) \neq \pi_\phi(s_{t-1})\}$. We show an example of flip point in Figure 6. At time step $t-1$, the agent presses RIGHT + SPEED to run to the right. In the next time step, the agent presses RIGHT + JUMP to begin jumping over the green pipe. We evaluate the sensitivity of ProtoX to flip points, i.e., *when the black-box agent changes its action, can ProtoX catch up?* We define sensitivity as the percentage of successful flips, i.e., sensitivity$(\pi_\phi) = \frac{\sum_{\mathbf{x} \in F} \mathbb{1}(\pi^*(\mathbf{x}) \neq \pi_\phi(\mathbf{x}))}{|\mathcal{F}|}$. We collect a total of 10,000 flip points from the test set and report the fidelity of ProtoX on this set in Table 3[2]. This shows that even though flip points are visually similar, ProtoX realizes that they differ and is able to reproduce the agent's action. For example, Figure 6 shows that ProtoX explains the running action in time step $t-1$ with a prototype of Mario running to the right, unimpeded by any obstacle. In contrast, ProtoX recognizes the jumping action in time step t as being similar to another scenario where the expert agent jumped over a green pipe.

Table 3: Sensitivity Analysis

Datasets	**Sensitivity**
Pong	84%
Seaquest	87%
SM 1-1	94%
SM 8-3	89%

4.4 Prototype Interpretation

To investigate the interpretability of ProtoX, we show an example of the prototypes in a ProtoX model learned from Seaquest. Seaquest is an Atari 2600 video game. The game is an underwater shooter in which the player controls a submarine. Its three goals are 1) to shoot enemies, 2) to rescue divers, and 3) to surface for oxygen when needed.

In Figure 7, we show 2 examples of prototypes learned by ProtoX. In addition, to understand what *scenario* each prototype represents, we overlay 60 most similar states to the prototype by lowering their transparency. States are overlaid by plotting the full-color representations of each state with low transparency over one another.

The first prototype is for the action "FIRE", which means the submarine tries to shoot enemies to gain points. The prototype represents a situation where the player has shot a missile to the right at an

[2]Please refer to the supplementary mateiral for other baselines' sensitivity to flip points.

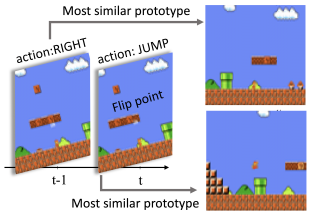
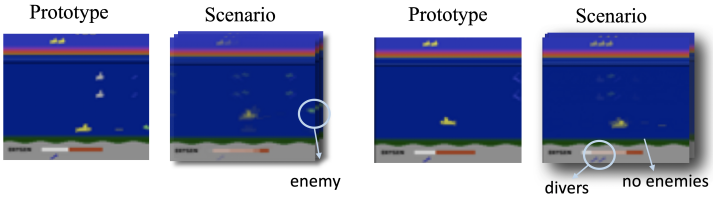


Figure 6: A flip point and the corresponding prototype.



(a) A prototype for action FIRE (b) A prototype for action UP

Figure 7: Prototypes from Seaquest and their interpretation.

enemy coming in from the right screen edge. By aggregating the similar states, we can faintly see that in this scenario, the submarine is located in several places near the bottom center of the screen. But the common feature in this scenario is the enemies to the right of the submarine.

We also show a prototype for the action "UP", which means the submarine will move up for oxygen. By examining the common feature in the scenario, we can understand that the scenario this prototype represents is where the submarine has saved divers and there are no enemies in the front and so it is safe to go up. The most similar states in this scenario show the submarine in the same general position as in the prototype, but with varying air tank level and varying number of rescued divers.

Diagnosing Bad Behaviors The explanations provided by ProtoX can also be used to find why an agent makes mistakes. To simulate this use case, we train a bad agent that never jumps for the Super Mario game, and consequentially is always killed by the first goomba it encounters. We then train a ProtoX on trajectory data from this agent to see whether ProtoX can explain why the agent fails to jump to avoid the goomba. Figure 8 shows that Mario sprinted to the right into the goomba. ProtoX explains this by relating this state to similar prototypes. The most similar prototype with the largest contribution of evidence is the right game frame in Figure 8, which shows that the bad agent views this situation as similar to situations in the beginning of the game when it could sprint to the right because there are no obstacles. In other words, ProtoX is saying that the bad agent fails to distinguish the two general scenarios in which Mario should jump and when Mario should run.

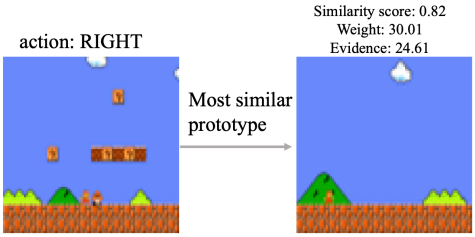


Figure 8: Mario fails to jump over the goomba because he thinks the present state is no different from states where he was running forward.

5 Conclusion

We propose a prototype-based policy explainer, ProtoX, for explaining black-box reinforcement learning agents. ProtoX explains an agent by prototyping its behavior into different scenarios, each represented by a prototype. Then ProtoX explains an action by comparing an input with prototypes, computes the similarity scores, and then calculates accumulated evidence for each action. ProtoX takes into account both visual similarity and scenario similarity. We use self-supervised training and contrastive learning to learn the notion of visual similarity from expert demonstrations, and allow ProtoX to learn scenario similarity by using an isometry layer. Results show that ProtoX achieved high fidelity compared to VIPER, a black-box model with similar network architecture, and also a GAIfO model whose policy network had a similar architecture.

Discussion of Limitations The interpretation of the prototypes requires basic understanding of the task. For example, to interpret the prototype for UP in Figure 7, users need to know the basic logic for playing the game, thus can comprehend why the submarine goes up in that scenario. In addition, we have also noticed in the Seaquest game that the expert anticipates the arrival of enemies and shoots at them before they even enter the screen. Thus, ProtoX learns prototypes showing the expert shooting at nothing. ProtoX's mode of interpretation does not inform the users of such anticipatory behavior. One would not be able to understand such prototypes without a basic understanding of the game. Also, ProtoX may learn similar and redundant prototypes. Some pruning techniques could be created to further improve the quality of prototypes.

References

[1] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[2] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[3] Alex Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo J Rezende. Towards interpretable reinforcement learning using attention augmented agents. *arXiv preprint arXiv:1906.02500*, 2019.

[4] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *International Conference on Machine Learning*, pages 1792–1801. PMLR, 2018.

[5] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. Graying the black box: Understanding dqns. In *International Conference on Machine Learning*, pages 1899–1908. PMLR, 2016.

[6] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[7] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *arXiv preprint arXiv:1805.08328*, 2018.

[8] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition, 2019.

[9] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.

[10] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video, 2018.

[11] Jinghuan Shang and Michael S. Ryoo. Self-supervised disentangled representation learning for third-person imitation learning, 2021.

[12] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.

[13] Murali V. Singh R. Kohli P. Chaudhuri S Verma, A. Programmatically interpretable reinforcement learning. In *PMLR*, pages 5045–5054, 2018.

[14] Hentschel A. Runkler T. Udluft S. Hein, D. Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. *Engineering Applications of Artificial Intelligence*, 65:87–98, 2017.

[15] Xiong C. Socher R. Shu, T. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning, 2017.

[16] Shafti A. Faisal A. A. Beyret, B. Dot-to-dot: Explainable hierarchical reinforcement learning for robotic manipulation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5014–5019, 2019.

[17] Seurin M. Strub F. Pietquin O. Cideron, G. Self-educated language agent with hindsight experience replay for instruction following, 2019.

[18] Wenbo Guo, Xian Wu, Usmann Khan, and Xinyu Xing. Edge: Explaining deep reinforcement learning policies. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Daniel Hein, Alexander Hentschel, Thomas Runkler, and Steffen Udluft. Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies. *Engineering Applications of Artificial Intelligence*, 65:87–98, 2017.

[20] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 414–429. Springer, 2018.

[21] Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, Ann Nowé, Tim Miller, Rosina Weber, and Daniele Magazzeni. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 workshop on explainable artificial intelligence*, pages 1–6, 2019.

[22] Nathan Dahlin, Krishna Chaitanya Kalagarla, Nikhil Naik, Rahul Jain, and Pierluigi Nuzzo. Designing interpretable approximations to deep reinforcement learning with soft decision trees. *arXiv preprint arXiv:2010.14785*, 2020.

[23] J Waa, J van Diggelen, K Bosch, and M Neerincx. Contrastive explanations for reinforcement learning in terms of expected consequences. In *Proceedings of the Workshop on Explainable AI on the IJCAI conference, Stockholm, Sweden., 37*, 2018.

[24] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2493–2500, 2020.

[25] Jung Hoon Lee. Complementary reinforcement learning towards explainable agents. *arXiv preprint arXiv:1901.00188*, 2019.

[26] Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018.

[27] Raghuram Mandyam Annasamy and Katia Sycara. Towards better interpretability in deep q-networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4561–4569, 2019.

[28] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. Efficient saliency maps for explainable ai. *arXiv preprint arXiv:1911.11293*, 2019.

[29] Zoe Juozapaitis, Anurag Koul, Alan Fern, Martin Erwig, and Finale Doshi-Velez. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on Explainable Artificial Intelligence*, 2019.

[30] Pedro Sequeira and Melinda Gervasio. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence*, 288:103367, 2020.

[31] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[32] Brian D Ziebart, Andrew Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 3*, pages 1433–1438, 2008.

[33] Brian D Ziebart, J Andrew Bagnell, and Anind K Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, 2010.

[34] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.

[35] Menghai Pan, Weixiao Huang, Yanhua Li, Xun Zhou, and Jun Luo. xgail: Explainable generative adversarial imitation learning for explainable human decision analysis. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1334–1343, 2020.

[36] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3815–3825, 2017.

[37] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

[38] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics, 2015.

[39] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases, 2022.

[40] Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press., 2018.

[41] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and steerable sequence learning via prototypes. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2019.

[42] Christian Kauten. Super Mario Bros for OpenAI Gym. GitHub, 2018.

[43] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[44] Antonin Raffin, Ashley Hill, Maximilian Ernestus, Adam Gleave, Anssi Kanervisto, and Noah Dormann. Stable baselines3. `https://github.com/DLR-RM/stable-baselines3`, 2019.

[45] Viet Nguyen. [pytorch] proximal policy optimization (ppo) for playing super mario bros. `https://github.com/uvipen/Super-mario-bros-PPO-pytorch`, 2021.

[46] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation, 2018.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[48] Xingrui Yu, Yueming Lyu, and Ivor W. Tsang. Intrinsic reward driven imitation learning via generative model, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See the Section
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5 Conclusion
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] N/A
   (b) Did you include complete proofs of all theoretical results? [N/A] N/A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code, data, and necessary instructions in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All training details are provided in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Experiments were performed using a single seed.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All experiments were done on a system with an RTX 3060Ti GPU, AMD Ryzen 7 3700X 8-Core Processor, with 32GB RAM

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the models and libraries we used.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not use such data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not use such data. All data are from OpenAI Gym
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]