Widening the Pipeline in Human-Guided Reinforcement Learning with Explanation and Context-Aware Data Augmentation

Lin Guan[1], Mudit Verma[1], Sihang Guo[2], Ruohan Zhang[3], Subbarao Kambhampati[1]

[1]School of Computing & AI, Arizona State University
[2]Department of Computer Science, The University of Texas at Austin
[3]Department of Computer Science, Stanford University

lguan9@asu.edu

Human-Advisable RL

An intelligent agent's ability to adjust its behavior according to feedbacks from users in the loop is important:

- Reward design is hard
  - User's personal preferences (value alignment)
- Action advice or demonstrations are expensive to obtain
  - Expertise of end users
  - Bad user experience
  - Complex setup (e.g. teleoperation)
- Learning methods like RL suffer from high sample complexity

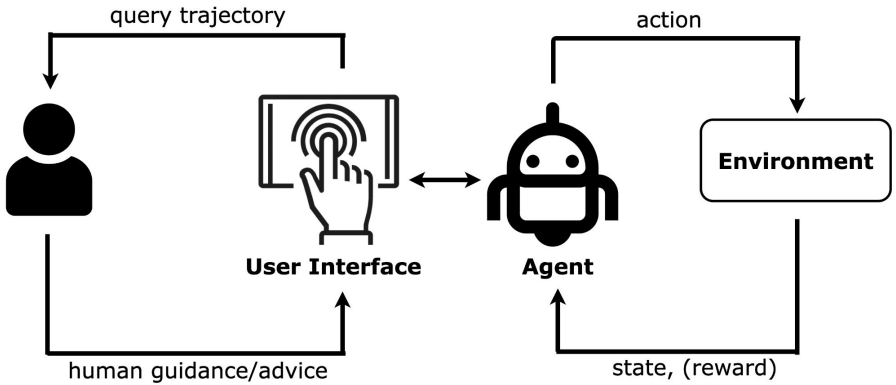
A more flexible solution can be **Human-Advisable RL**

Scenario 1: the passenger wants the robot taxi to slow down when he is drinking coffee.

Scenario 2: the user doesn't want the robot vacuum to clean the room when he/she is having an online meeting.

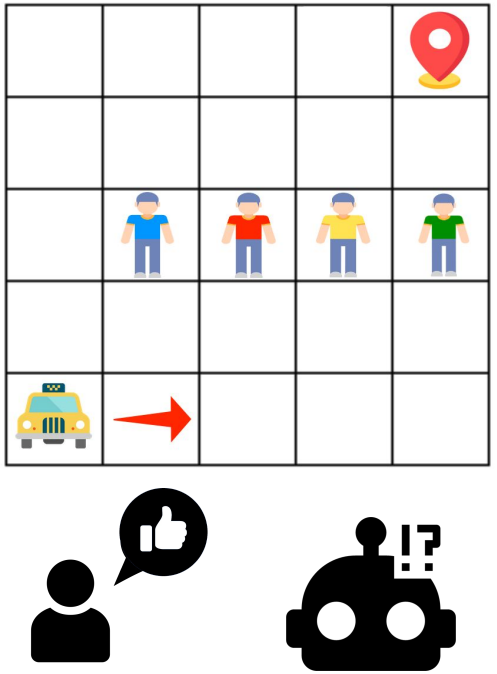
Human-Advisable RL

- A human trainer monitors the learning process of RL
- The agent adjusts its policy according to human advice
- Forms of advice
  - Inexpensive and intuitive to specify.
  - Reduced to TAMER [Knox and Stone, 2009] when advice is binary evaluative feedback
- Human-Advisable RL generalizes from Human-in-the-Loop RL (HIRL) but has separate challenges **beyond** HIRL



Challenges in Human-Advisable RL

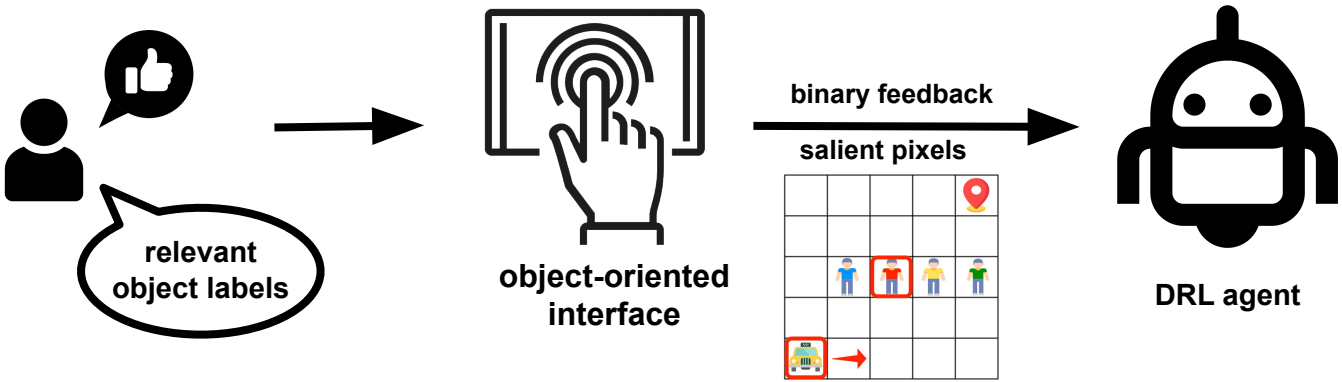
- The Quandary:
  - Human feedbacks are **expensive** and **sparse**
  - DNNs are always **data-hungry**
- Missing **Lingua Franca** (shared vocabulary) between humans and agents
  - Limit the forms of feedback to **simple numerical labels** (e.g. evaluative feedback, binary preference labels)
  - Numerical labels are **not informative** enough
- Communicative Modalities
  - Humans prefer multi-modal communications
  - Easy (**effortless**) to provide
  - The agent can easily understand



Binary feedback doesn't indicate why certain action is good/bad.

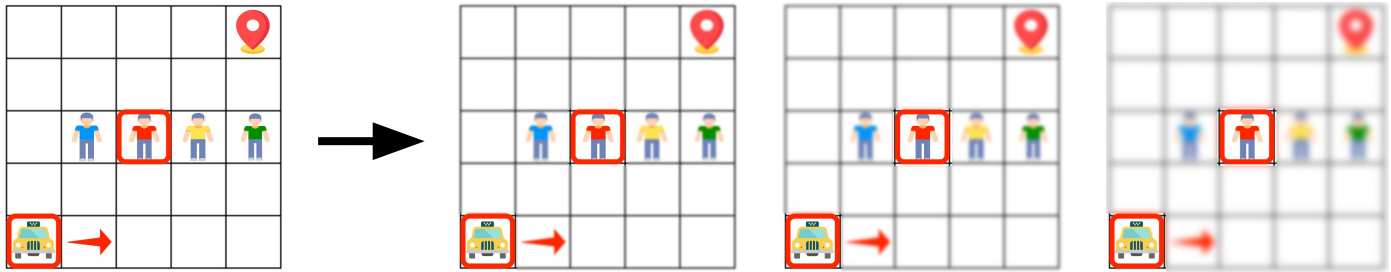
Our Goals

- The Quandary:
  - Improve **human feedback sample efficiency** & **environment sample efficiency**
- Lingua Franca & Multi-Modal Communication
  - Augment binary evaluative feedback with **human visual explanation**
    - Annotations of **task-relevant regions (pixels)** in image
    - Help in “maximally” utilizing each binary feedback
  - **Effortlessly** collect human visual feedback
    - An **object-oriented** middle layer (interface)



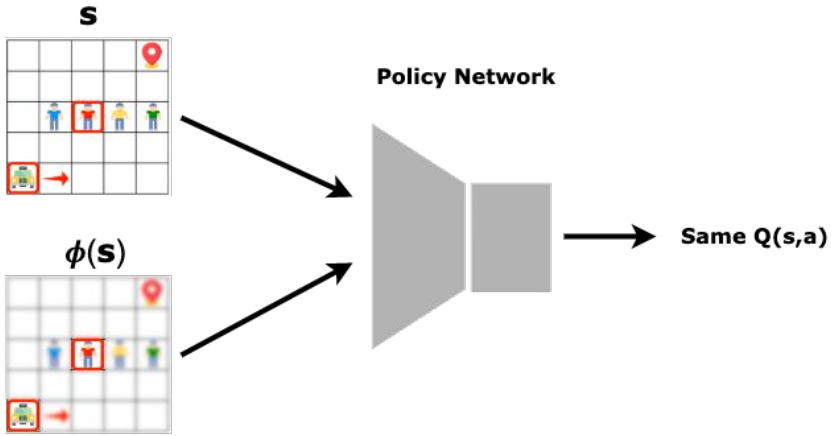
Context-Aware Data Augmentation

- Existing ways to incorporate saliency information into supervised learning systems are not suitable for **less stable** learning systems like deep reinforcement learning
- **Context-Aware Data Augmentation**
  - Intuition: small perturbations on irrelevant regions should not alter the agent's policy
  - Approach:
    - Apply various image transformations to the irrelevant regions, and obtain a set of augmented feedback
    - **Gaussian blurring** with different Gaussian kernels
    - Two loss terms to enforce invariance
  - Examples:



Context-Aware Data Augmentation

- Q-Values Invariance
  - Intuition: the Q-values of state **s** should be the same as Q-values of perturbed state $\phi(\mathbf{s})$
  - Regularization loss: $|| Q(\mathbf{s},\mathbf{a}) - Q(\phi(\mathbf{s}), \mathbf{a})||_2$
  - An **inverse** of perturbation-based Explainable RL



- Feedback Invariance
  - Intuition: the human teacher's judgement on <state **s**, action **a**> remains the same for any other <ϕ (**s**), **a**>
  - Generate **more training samples** for the agent

Efficiently Collecting Visual Explanation

An **object-oriented interface**:

- Observations:
  - Human visual explanations are usually associated with certain **objects** or **regions** in image
  - Salient regions/objects are usually the same in nearby frames
- Use a simple **tracking and detection** module to detect possible salient objects/regions
- **Effortless** communication at the level of **symbols** (e.g. object labels) even though the DRL agent is operating in pixel-space
- User study: collected over 2k feedbacks (binary feedback & visual explanation) in 30 min

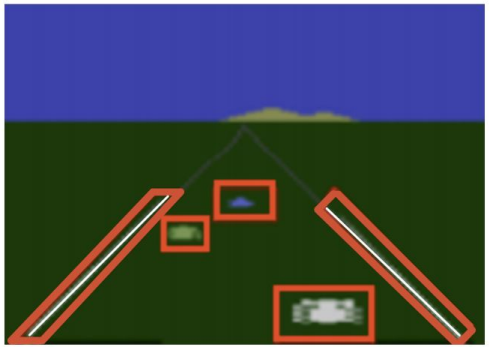


Fig. 3. All the lanes and cars are automatically highlighted and tracked, so the human trainers only need to deselect irrelevant objects in the image.

Learning from Binary Feedback

- We propose a new method that bypasses the need to explicitly approximate human feedback
- We use the **advantage value** to formulate the **feedback loss function**:

We interpret human feedback as the **human's judgment on the optimality** of an action

Penalize if the agent has a different judgment on the action optimality

Advantage value is **the agent's** prediction on the optimality of an action

$A^{\pi}(s, a) = Q^{\pi}(s, a) - Q^{\pi}(s, \pi(s))$:

- $A^{\pi}(s, a)=0$ means the agent predicts action **a** as an optimal action
- $A^{\pi}(s, a)<0$ means the agent predicts action **a** as a sub-optimal action

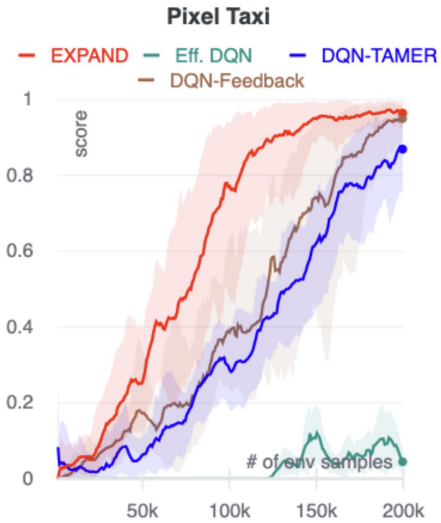
Experimental Evaluation

- We evaluated our method EXPAND with oracles (simulated feedbacks) in five tasks
- Questions to answer:
  - Does human explanation help?
  - Other better ways to use human explanation?

Experimental Evaluation

Whether the use of human explanation improves the environment and feedback sample efficiency?

- Baselines: DQN-TAMER [Arakawa et al., 2018]; DQN-Feedback, an ablated version of EXPAND (no visual explanation, only binary feedback)
- Takeaways:
  - Visual explanation results in a **significant improvement** in both environment sample efficiency and feedback sample efficiency
  - Richer interaction is a right direction

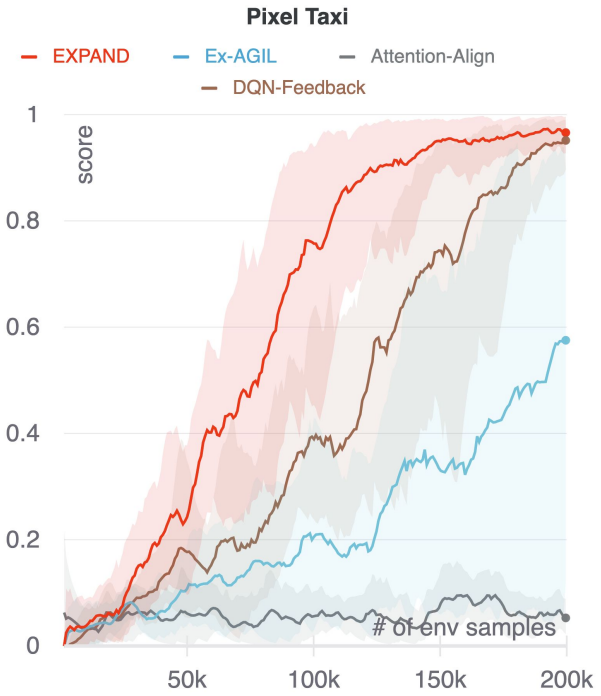


35% reduction in both env samples and feedback samples.

Experimental Evaluation

Does EXPAND utilize the human explanation better than other baselines?

- Baselines: two explanatory interactive learning methods that use auxiliary attention-alignment loss (i.e. Attention-Align) and train a separate attention predictor (i.e. Ex-AGIL)
- Takeaways:
  - Data augmentation is a more stable methodology to regularize Deep RL



Summary

- Human-Advisable RL and challenges
- EXPAND: widening the human-agent interaction pipeline by adding an explanation channel
- Encode task-relevant features through a context-aware data augmentation
- An object-oriented interface to reduce human efforts

Future work:

- Beyond object-oriented interface: concepts or even natural language.
- More sophisticated way to inject saliency prior