Explanation Augmented Feedback in Human-in-the-Loop Reinforcement Learning

Lin Guan*, Mudit Verma*, Subbarao Kambhampati

Binary Evaluations predominantly used in conventional HRL setting is not enough.

Why not let the human explain how to act desirably by selecting important regions in observation?

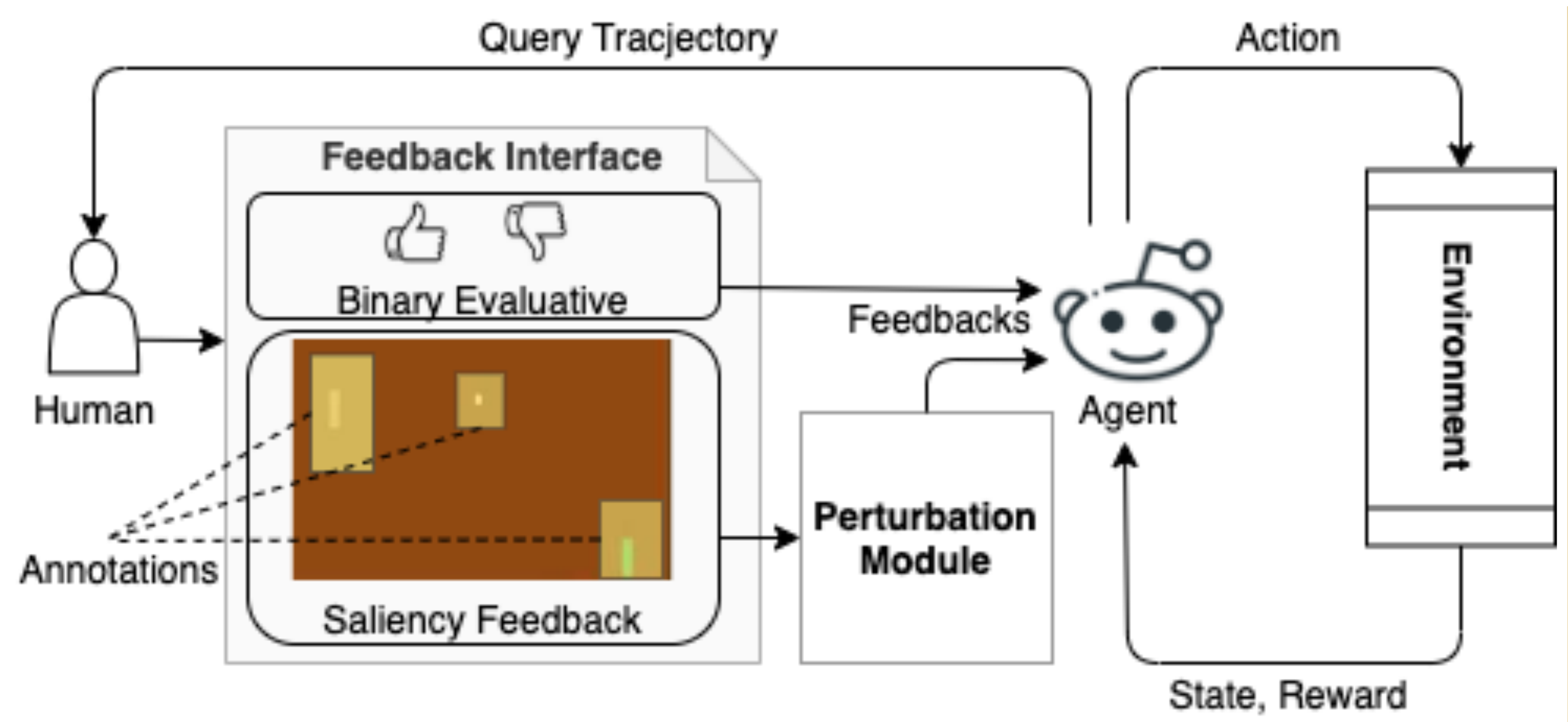
Human explanations along with Binary feedbacks can provide significant performance gains.

We can exploit Human explanations by perturbing irrelevant regions.

Aim is to improve Feedback and Environment Sample efficiency.

QUICK FACTS

- > **Integrate** Human Explanation with Binary Evaluations for Human in the loop RL
- > **Explanation Augmented Feedback (EXPAND)** outperforms HRL baseline
- > **25% improvement** in both Feedback Sample efficiency and Environment Sample efficiency
- > **Focusing on relevant regions helps.**
- > Results confirm that agent is indeed focusing on relevant regions.
- > Employing multiple perturbations is helpful
- > **Advantage Loss** on Binary Evaluations
- > **Policy & Value Invariant losses** on Human Explanations



BINARY FEEDBACK & EXPLANTIONS

Use Environment Reward with L_{DQN} *loss.*

Using the Binary Evaluations Only => Optimality of Action

Advantage Loss :

Agent's judgement on optimality of an action

$$L_A(s,a,h) = L_A^{Good}(s,a,h) + L_A^{Bad}(s,a,h)$$

$$L_A^{Good}(s,a,h;b^h=1) = \begin{cases} 0 & ; A_{s,a}=0 \\ Q^\pi(s,\pi(s)) - Q^\pi(s,a) & ; \text{otherwise} \end{cases}$$

Where,

$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s) = Q^\pi(s,a) - Q^\pi(s,\pi(s))$$

$$L_A^{Bad}(s,a,h;b^h=-1) = \begin{cases} 0 & ; A_{s,a}<0 \\ Q^\pi(s,a) - (\max_{a'\neq a} Q^\pi(s,a') - l_m) & ; A_{s,a}=0 \end{cases}$$

Using Human Explanation (Saliency Information)

Policy Invariant Loss :

Under perturbations "good" Action is always good "bad" Action is always bad.

$$L_P = \frac{1}{g}\left(\sum_g L_A(\tilde{s}^{rg}, a, h)\right)$$

Helps distinguish the optimal action

Value Invariant Loss :

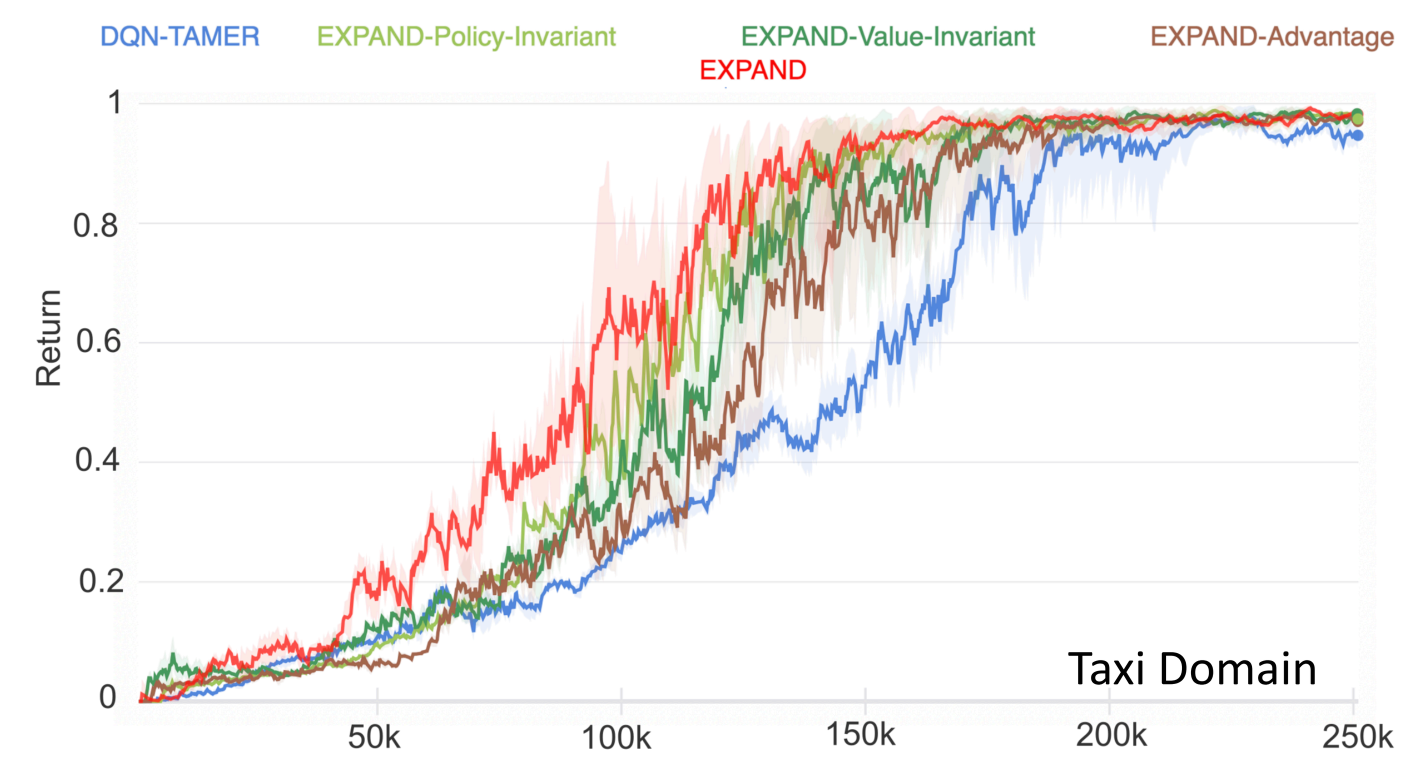
Q-values of the original state should be similar to the Q-values of Perturbed states

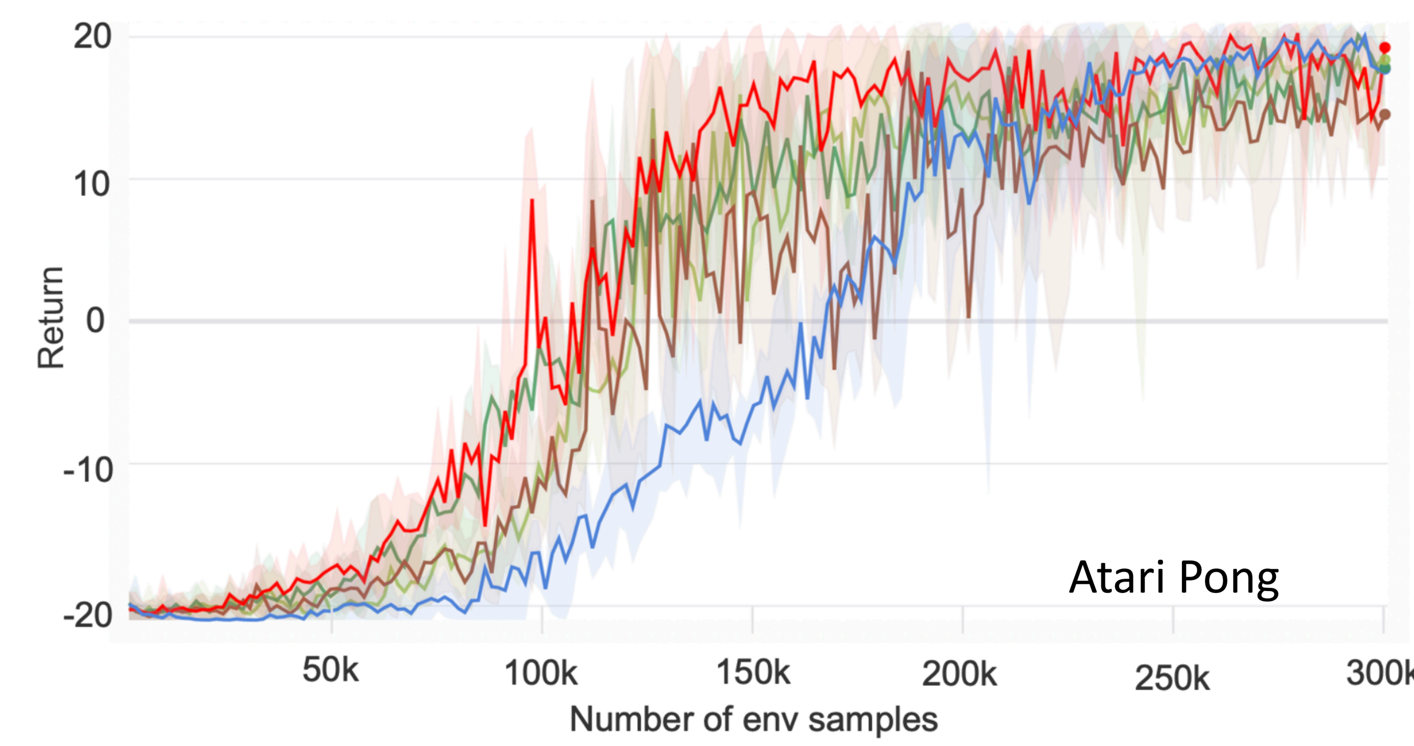
$$L_V = \frac{1}{g}\sum_{i=1}^{g}\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}(Q^\pi(s,a) - Q^\pi(\tilde{s}^{ri},a))$$

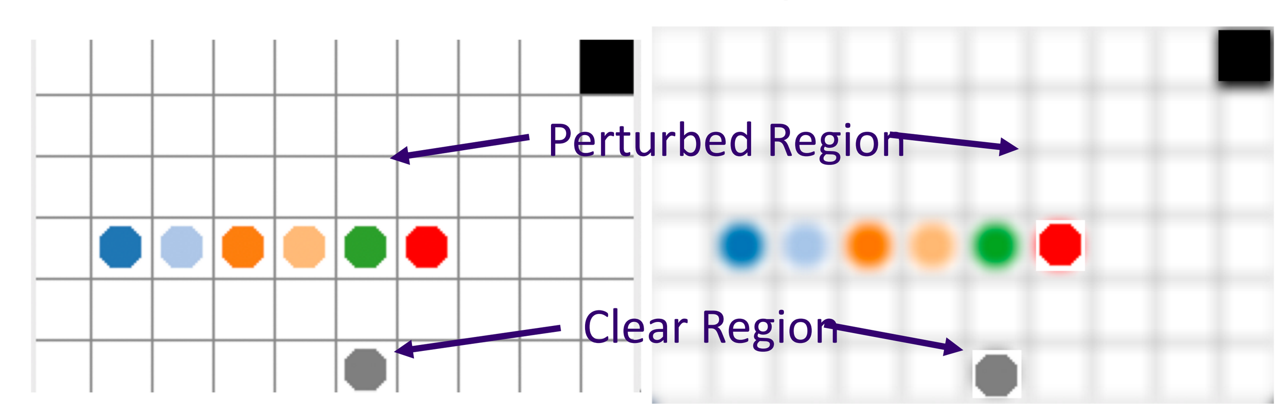
Regularize the internal representation

RESULTS

- > **Domains :** Taxi & Atari-Pong
- > **Metrics :**
  - Score : Able to reach optimal score w.r.t HRL baseline
  - Environment Sample efficiency : 25% improvement
  - Human Feedback Sample Efficiency : 25% improvement
- > **Ablation :** Confirms use of binary evaluation not enough. Confirms the importance of each loss term.







Perturbing Irrelevant State Information

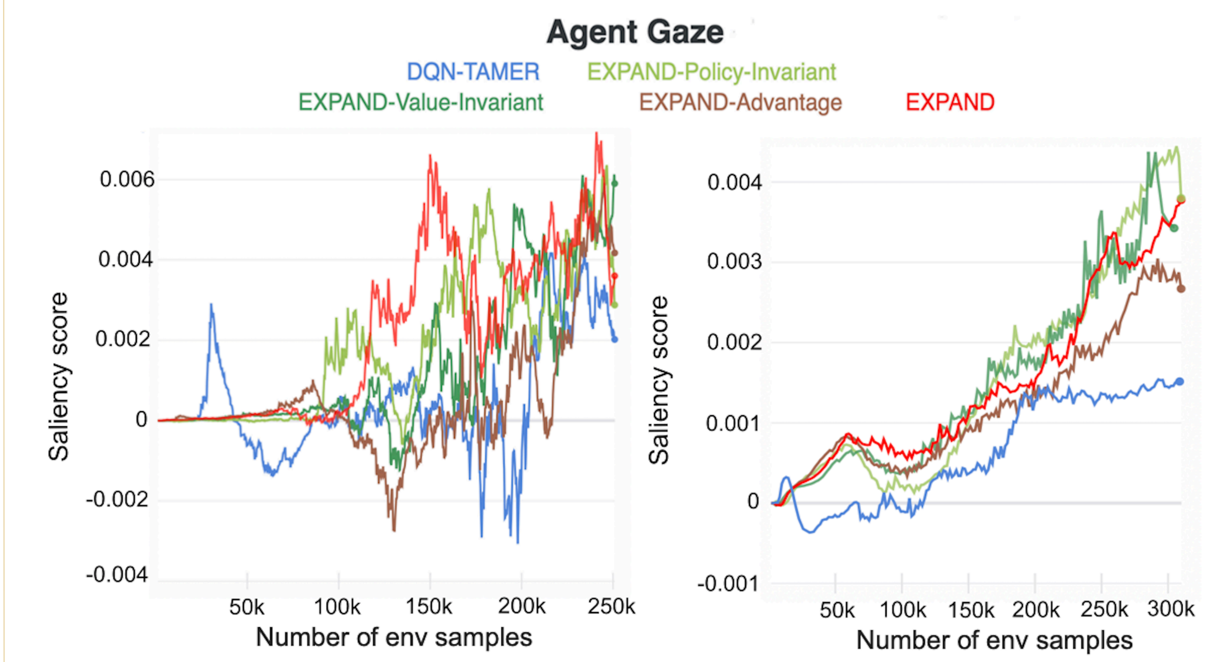
We apply Gaussian Perturbations over marked irrelevant regions In hope that such changes should not affect agent's decision making.

AGENT-GAZE & PERTURBATIONS

We verified our intuition :

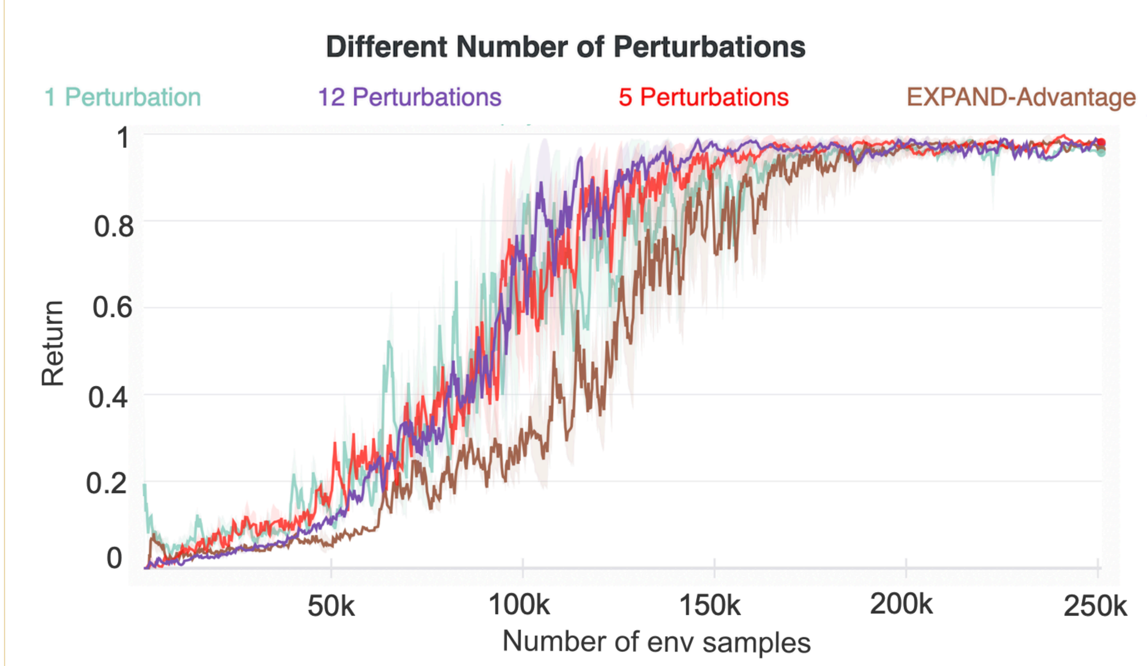
Focusing attention towards "relevant" state regions improves performance.

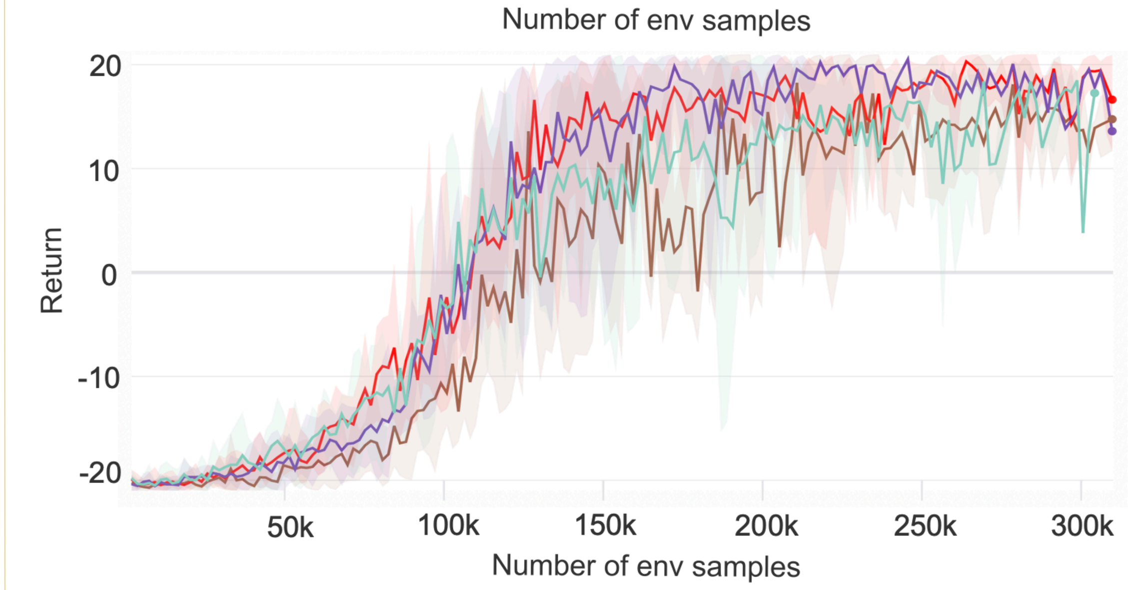
$$S_F = \frac{1}{|\mathcal{A}_g|}\sum_{s_g^r\in S_g^r, a_g\in\mathcal{A}_g} SARFA(s_g^r, a_g) - \frac{1}{|\mathcal{A}_b|}\sum_{s_b^r\in S_b^r, a_b\in\mathcal{A}_b} SARFA(s_b^r, a_b)$$



We tested for the best perturbation setting :

5 perturbations per feedback is adequate





*Equal Contribution

Acknowledgements:
Kambhampati's research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, N00014-18-1-2840, N00014-19-1-2119, AFOSR grant FA9550-18-1-0067, DARPA SAIL-ON grant W911NF-19-2-0006, NSF grants 1936997 (C-ACCEL), 1844325, and a NASA grant NNX17AD06G.