Gain-based prioritisation for explanation awareness

Agent-centred explanations are a useful mechanism for humans to learn.
Can we apply on machines the same strategies useful for humans?
With humans, if we are able to evaluate the quality of the agent "mental model" (understanding what is wrong or correct in the model) with respect to the agent task, then it is possible to produce agent-centred explanations accordingly.
Is it the same for machines?

For humans, we can define explanations as a path in an explanatory space.
In RL, an example of explanatory space might be the experience buffer. Thus explanations are trajectories in the experience buffer.
Agent-centred explanations are those trajectories that might meet the needs of the machine to optimally learn how to solve a given task.
In this sense, we can see prioritised experience replay as an attempt to produce agent-centred explanations.

First of all, if we are going to perform experience replay with PPO [4], we need to consider strategies to reduce the bias and variance of experience replay. One of these strategies is the V-trace advantage estimation algorithm of IMPALA [1].
Considering that IMPALA has some known issues on vis-a-vis continual learning, another strategy might be using the GAE [2] combined with V-trace, in short GAE-V [3].
Thus, both V-trace and GAE-V have been implemented, together with some other strategies to minimise the KL distance [3] between old and new policies during on-policy updates.

To produce agent-centred explanations via prioritised experience replay the prioritisation metric should sort trajectories according to their relevance with respect to the agent needs.
A relevant trajectory is a trajectory that improves the agent "mental model", helping the agent to solve its task better.
How to measure such relevance metric for trajectories?

Naive idea: We make the agent to predict state transitions (new state

and reward, given a state and an action to perform), and the prediction error is an estimate of the quality of the agent mental model. Problems of this idea:

- The evaluation is not with respect to the agent task. The agent might find more efficient to not include, in its mental model, rules that are not useful to reach its objectives.
- Training the transition-predictor jointly with the actor and the critic causes an undesired policy entropy minimisation.

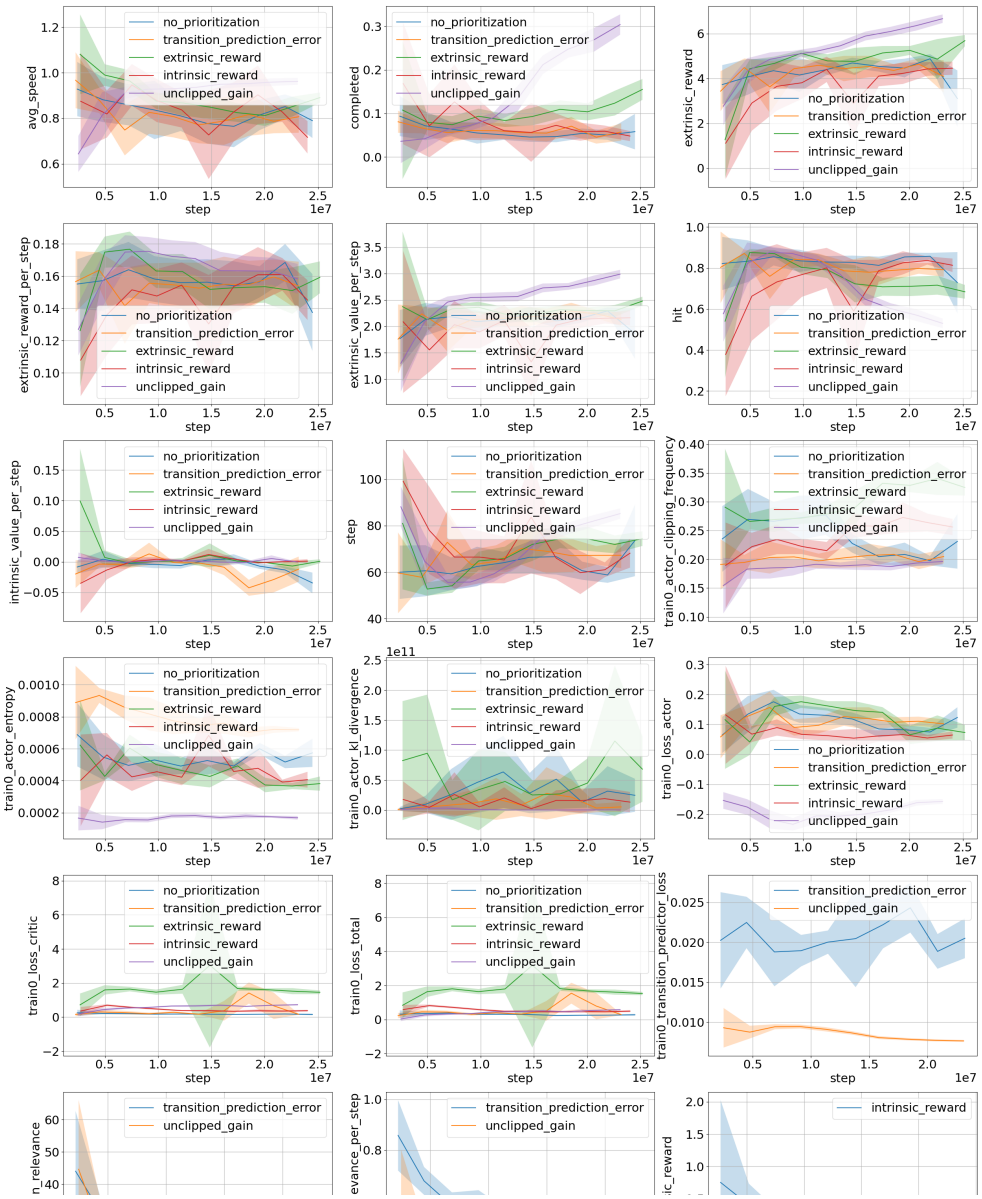
An even more naive solution might be to train the transition predictor separately from the actor and the critic.
The transition prediction was meant to give an estimate of the quality of the agent "mental model", but if the agent does not train together with the transition predictor, then maybe the transition predictor error does not follow the agent "mental model" error!
Another idea: On the other hand, the advantage multiplied by the policy cross entropy ratio (between old and new policy) might be an estimate of the quality of the agent "mental model".
That's because the advantage is different from zero when the critic gives wrong values, and the cross entropy ratio:

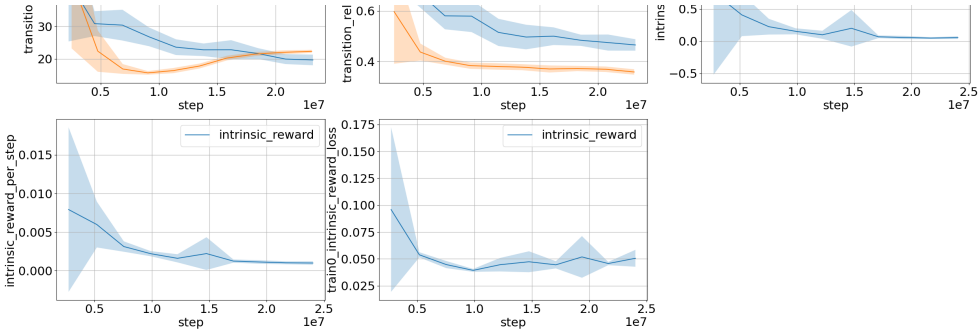
1. Is close to 1 when the new policy is similar to the old policy, or the action is distant from both the new and old policy.
2. Is close to 0 when the action is close to the new policy and it is distant from the old policy.
3. Tends to infinite when the action is distant from the new policy and it is close to the old policy.

In the 1st case the relevance is given by the advantage, in the 2nd case the relevance is unknown and thus 0, in the 3rd case the relevance sign is given by the advantage.
The product of these two values gives what we call the "gain", and it is used by the agent (more or less, ratio clipping is missing in the case of PPO) to train.
Thus, the "gain" might be a mechanism to evaluate the quality of the user "mental model", if combined with a proper experience buffer structure.
In fact, the objective of the (PPO-based) RL agent is to maximise the "gain". Using the "gain" as it is, as prioritisation scheme, the machine would tend to ignore (not replay; not reinforce on) negative gains given by a bad "mental model" rather than a bad action distant from the policy, thus ignoring possibly relevant trajectories.
A solution to this issue might be to keep two prioritised buffers using the "gain" prioritisation scheme, one for the trajectories leading to

positive cumulative returns, while the other one for trajectories leading to negative cumulative returns.
Sampling homogeneously from those two buffers would allow the agent to receive task-oriented/agent-centred explanations regarding both mistakes in the “mental model” and ways to improve the “mental model”.

The whole two-headed prioritisation scheme would work only if the critic does not significantly overestimate the state value, thus producing wrong advantages.
Preliminary results are shown here:





In the graph a comparison of different prioritisation strategies is shown. The purple is the gain based prioritisation strategy.

Clearly, the results are only partial, because there is a lot to do before proving the whole explanation-awareness thing. Furthermore, the results seem promising but the strategy has not been benchmarked properly.

1. Espeholt, Lasse, et al. "Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures." *arXiv preprint arXiv:1802.01561* (2018).
2. Schulman, John, et al. "High-dimensional continuous control using generalized advantage estimation." *arXiv preprint arXiv:1506.02438* (2015).
3. Han, Seungyul, and Youngchul Sung. "Dimension-Wise Importance Sampling Weight Clipping for Sample-Efficient Reinforcement Learning." *arXiv preprint arXiv:1905.02363* (2019).
4. Schulman, John, et al. "Proximal policy optimization algorithms." arXiv preprint arXiv:1707.06347 (2017).