Curiosity Driven Learning

Group 01: Daniel Derkacz, Xinpeng Wang

***Abstract*—A count based intrinsic reward based on the work of Bellemare et al. [3] was implemented into the large scale study of intrinsic reward from Burda et al. [1]. The prediction based and count based rewards were compared to each other and an analysis of mixing extrinsic and intrinsic reward was conducted.**

I. Motivation

Count based methods for intrinsic motivation showed good results in playing Atari 2600 games and even could achieve state of the art results in challenging environments like *MONTEZUMA'S REVENGE* [2], [3]. In contrast to the prediction based approach used in the large scale study of intrinsic motivation by Burda et al. [1], count based methods use state visitations counts as the measure of surprise. The key idea is to give less visited states a higher reward and thus motivate exploration.

Given a density model $\rho_n(x)$ which was trained on n occurrences of the input x, we can the define a *prediction gain*

$$\mathrm{PG}_n(x) = \log \rho'_n(x) - \log \rho_n(x)$$

which is the difference of the prediction of our model $\rho_n(x)$ for x at time n and the prediction if we would train the model one more time on the input x, namely $\rho'_n(x)$. A pseudo-count $\hat{N}_n(x)$ can now be approximated

$$\hat{N}_n(x) \approx \left(e^{\mathrm{PG}_n(x)} - 1\right)^{-1}$$

and a corresponding reward

$$r(x) = \hat{N}_n(x)^{-1/2}$$

For a deeper analysis of this count based method we refer the reader to the original work of Bellemare et al. [2].

Our contribution is the implementation of a count based method into the work of Burda et al. and comparing the performance to the prediction based approach. We used a *Gated PixelCNN* implementation[1] with 2 gated convolutional layers and 10 quantization levels. A key difference between [2], [3] and our work is the usage of a PPO agent instead of a DQN agent.

II. Prediction based evaluation

First, to evaluate the performance of different feature extractors for prediction based method, we train the model on Breakout and BeamRider with Inverse Dynamic Features (IDF) and Random CNN Features (RF), as shown in Fig 1. Both of them perform quite well, and sometimes the fixed RF performs better than learned features. As shown in Fig 1 and Fig 2, IDF performs unstable in these two environments, as we see a quick drop of performance after training for 60M and 120M frames. Therefore we use RF for further experiments. In the game of MontezumaRevenge, we see a sudden increase at around 90M frames in intrinsic reward and game score, as shown in Fig 3. By looking at the game play before and after the increase, we see the agent goes to next room and receives new observation after 85M frames, therefore it gets very high curiosity score.

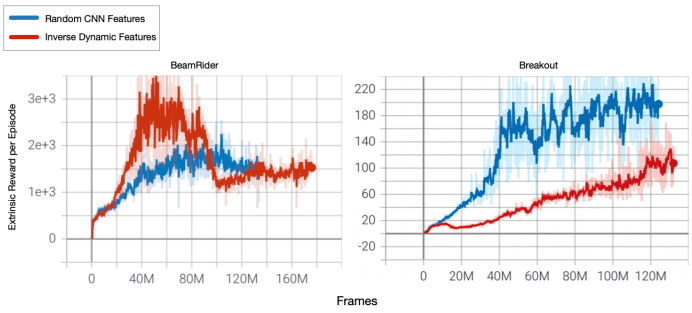

Fig. 1. Comparison of feature learning methods on 2 Atari games: BeamRider and Breakout. The curves show the mean reward of the agents trained purely by curiosity.

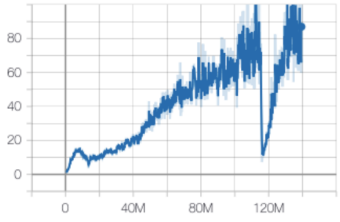

Fig. 2. Extrinsic reward of a second run of training on Breakout with IDF.

III. Comparison between prediction and count based

We compare the count based and prediction based method on 4 different Atari games environments: Breakout, SpaceInvaders, Riverraid and Beamrider. The training result is shown in Fig 4. It turns out that the prediction based method outperforms the count based method on all 4 environments. The agent with count based curiosity performs well on environment SpaceInvader. But for other environments we see a drop of performance after longer training.

[1] https://github.com/jakebelew/gated-pixel-cnn

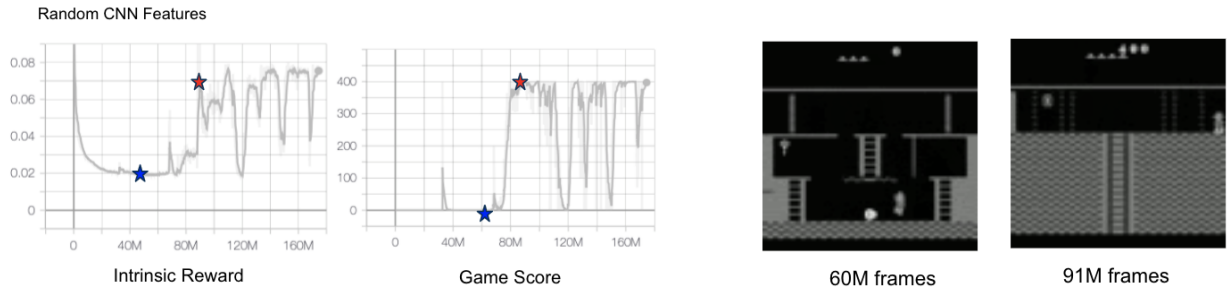

Fig. 3. **Left:**Intrinsic and extrinsic reward of agent in MontezumaRevenge. **Right:** Corresponding game play at 60M and 91M frames.

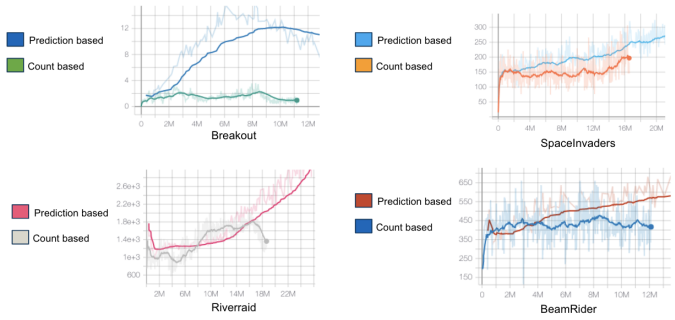

Fig. 4. Comparison of prediction based and count based on 4 different atari game environments.Curves shows the extrinsic reward of the agent trained with different curiosity method.

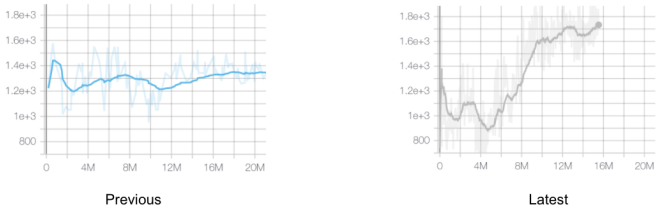

Fig. 5. Comparison of two different approach of calculating intrinsic reward. **Left:** Use observation of single time step. **Right:** Use observation of whole replay buffer.

IV. Mixing the Rewards

Pure intrinsic agents can learn to play different environments quite well as shown in our work and by [1] but so do purely extrinsic driven agents using methods like PPO. To further improve the performance of the agent we mixed the two reward functions and compared the improvement between the count based reward and the prediction based reward.

In Fig. 8 we can see the game score of a prediction based agent in the environments Riverraid and Breakout. Over the small time period we evaluated the methods on, a pure extrinsic driven agent performs the best and the purely intrinsic driven agent the worst. This is an expected result because Burda et al. found that intrinsic learning takes longer to achieve higher scores. A interesting result is the fact that a 50:50 split between the two different reward strategies does not lead to a better performance and is, in the case of the prediction based method, bound by the extrinsic reward.

In the count based setting these mixture of reward functions behaves quite differently, where the function is mostly influenced by the intrinsic reward, as seen in Fig. 9.
We can further investigate this behaviour and see that the count based intrinsic reward has a high influence even if the contribution is only 10% (Fig. 7) which is not the case for the prediction method. In Fig. 6 we can see that the original method is mostly governed by the extrinsic reward which is for small contributions of the intrinsic reward function an expected result.

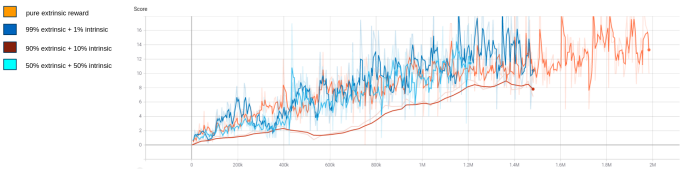

Fig. 6. Score evaluation of the prediction based method for difference mixtures of intrinsic and extrinsic reward. In contrast to the count based reward the intrinsic reward does not have a negative impact on the received score.

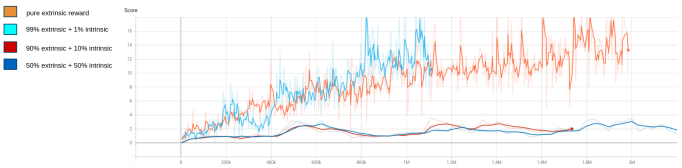

Fig. 7. Score evaluation of difference mixtures of reward functions for the count based reward on the environment Breakout. We can see that the intrinsic reward has a negative impact on the game score the agent can achieve.

These results imply that the count based method may not align that well with the overall goal of the game. In contrast the prediction based reward is mostly aligned with the game score and thus works to better game scores.

V. PixelCNN Training Problems

In early experiments of the count based method, the agent performs poorly and doesn't learn from playing the game.

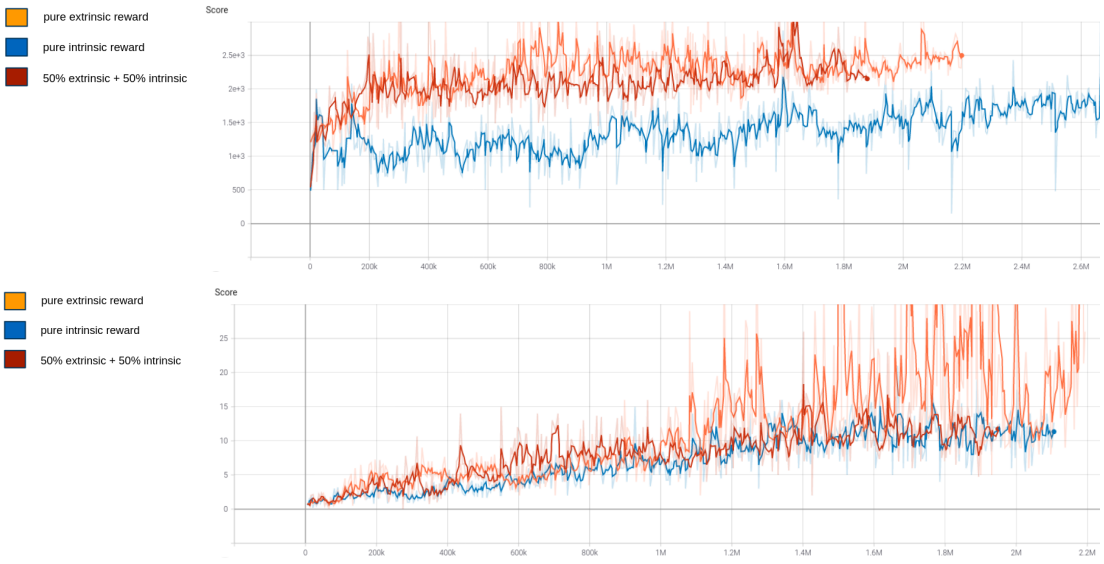

Fig. 8. Score evaluation of the prediction based reward. **Top:** The score achieved by the agent in Riverraid and at the **Bottom** on Breakout.

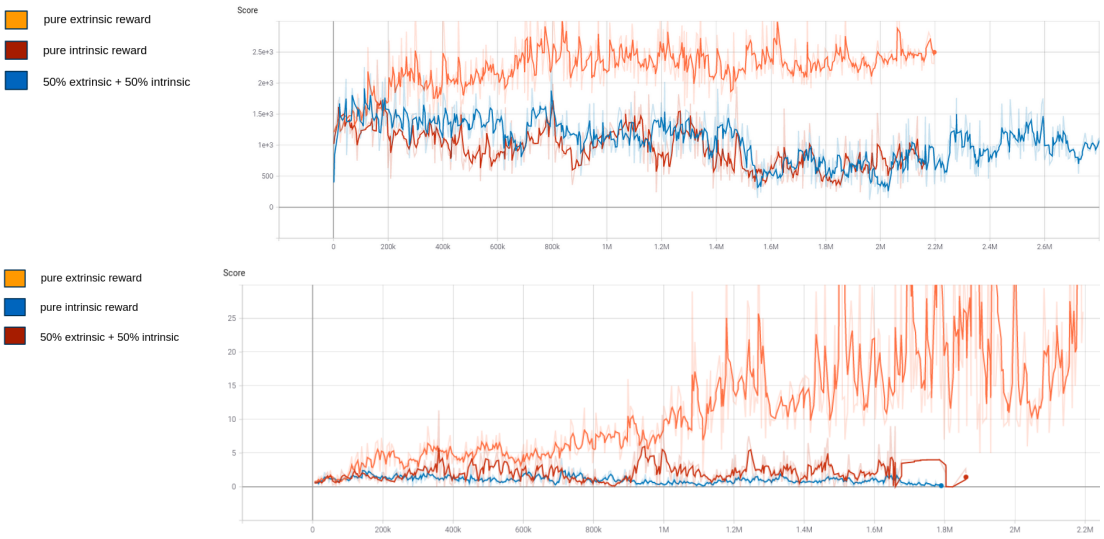

Fig. 9. Score evaluation of the count based reward. **Top:** The score achieved by the agent in Riverraid and at the **Bottom** on Breakout.

We then change the approach of calculating intrinsic reward. Instead of only using single observation for calculating the pseudo-count, we switch to using the whole replay buffer. After the change, the agent starts to learn from the playing, although the learning is not stable. Fig 5 shows the different performance between the two approaches.

VI. Conclusions

We conclude that we could not reproduce the good performance achieved by Ostrovski et al. [3] or Bellemare et al. [2]. It needs to be further investigated if the general implementation of our approach may be the reasons for these results or the fact that a PPO agent was used instead of the DQN agent. It also needs further investigation in which settings the different intrinsic reward strategies perform the best.

The simple idea of mixing both reward strategies to find better strategies does not seem to really work in the tested cases. It would be interesting to see how these results change if we do not use a fixed weighting of the rewards but a learned weighting, which may lead to a performance increase because the agent can pick it's preferred reward.

References

[1] Burda, Yuri, et al. "Large-scale study of curiosity-driven learning." arXiv preprint arXiv:1808.04355 (2018).
[2] Ostrovski, Georg, et al. "Count-based exploration with neural density models." International conference on machine learning. PMLR, 2017.
[3] Bellemare, Marc G., et al. "Unifying count-based exploration and intrinsic motivation." arXiv preprint arXiv:1606.01868 (2016).