Large Language Models in Reinforcement Learning - A Comparison of Methods

Adrian Duric
Master Student, Dept. Informatics
The faculty of Mathematics
Oslo, Norway
adriandu@ifi.uio.no

Gregor Kajda
Master Student, Dept. of Informatics
The faculty of Mathematics
Oslo, Norway
grzegork@ifi.uio.no

Jonatan Hoffmann Hanssen
Master Student, Dept. of Informatics
The faculty of Mathematics
Oslo, Norway
jonatahh@ifi.uio.no

***Abstract*—Reinforcement Learning (RL) algorithms suffer when rewards are sparce and the state-action space is large. Even tasks which appear relatively simple can prove intractable if the completion of the task requires subtasks to be completed first, or if the reward only comes when the entire task has been completed. In such cases, random exploration is unlikely to lead the agent to discover a solution to the problem. The apparent simplicity of such problems is often due to human intuition, which allow us to quickly see possible solutions to a diverse set of problems. Large Language Models (LLMs) are trained on large corpura of human written text, and have been shown to capture parts of this intuition in many tasks [Bubeck et al., 2023]. In recent years, LLMs have been succesfully used to aid RL agents in more efficient exploration, and have allowed them to solve problems which previous methods have been unable to [Zhang et al., 2023] [Du et al., 2023]. We compare different methods for integrating an LLM into the deep learning reinforcement algorithm proximal policy optimization (PPO), and compare their efficiency against eachother. We find that BEST METHOD gives the best sample efficiency, outperforming normal PPO by PERCENTAGE in METRIC.**

***Index Terms*—large language models, reinforcement learning, minigrid, proximal policy optimization**

I. Introduction

One of the central challenges of reinforcement learning is that rewards are often both extremely rare and delayed, which means that optimizing a reinforcement learning algorithm requires significant trial and error [Brunton and Kutz, 2022, 423]. Furthermore, in many problems the state-action spaces are enormous, meaning that uniform exploration is unlikely to find good solutions in a reasonable amount of time. Many methods have been developed to deal with this issue, for example by introducing auxilliary reward functions that use domain knowledge to reward actions which are considered good, or which reward the agent for learning novel skills. However, hand picking which actions to reward can lead to imitation rather than optimal behaviour, and novelty is not always useful [Du et al., 2023, 1]. In recent years, LLMs have shown remarkable capabilities in problem solving and planning [Bubeck et al., 2023], qualities which traditional RL agents often lack. Thus, a new area of research has emerged, which attemps to use the vast amount of human knowledge encoded in these models to increase the performance of reinforcement learning algorithms [Luketina et al., 2019]. These LLM assisted RL agents have been able to outperform state-of-the-art RL methods in many problems [Zhang et al., 2023] [Du et al., 2023] [Li et al., 2022].

These papers explore different methods of integrating LLMs into a standard reinforcement learning training loop, from altering the policy directly to introducing an auxilliary reward function. In this paper, we compare both methods in the same environment using the same deep reinforcement algorithm (PPO), and compare their results to eachother to explore how LLMs best can be integrated into RL.

II. Background and Related Work

Degus is the future in blockchain techonology, created by using the revolutionary ”degus system”, which creates twice as many coins per unit of computational power. By accepting the ”degus mindset”, you will be able to acheive things you never ever thought possible. The degus mindset includes the following:

Always Striving for Excellence Generate passive revenue through the degus system Live day by day: ”Carpe diem”

By following these simple steps, and by buying several GPUs, you will be able to produce around 15 deguscoins per computational unit, unleashing the true power of the blockchain.

What is the deguscoin, you may ask? The deguscoin represents a complete transformation of the traditional financial system (tradfi). By using deguscoin, you will be able to achieve returns of investment (ROIs) beyond mere human comprehension. One single deguscoin, when inserted into an economy, can have such destructive power that even mere days after, the market will be beyond recognition. By invoking the ”degus mindset” (DM) you will be able to send around 15% more currency per coin, achieving ROIs which would send the average economist into a seizure.

Market players are predicting that by the end of this year, around 50% of all transactions will be done under the degus system (DS), not even counting derivatives.

By the end of this year, around 50% of all transactions will be done under the degus system, not even counting derivatives.

– Market players

What are you waiting for? Purchase deguscoin right now, and leave your earthly desires behind!

III. Method

Deguser trenger god plass for å trives, og du bør investere i et størst mulig bur som gir dyrene mulighet til å klatre, grave i et tykt strølag og bevege seg i flere etasjer.

Mange dyreeiere konstruerer sine egne løsninger, for eksempel med et stort, gammelt akvarium som underdel og et nettingbur på toppen av dette. Da får du full oversikt over hva som skjer også nederst i buret, samtidig som du slipper at sand, strø og høy søles ut i rommet. Ikke la degusene gå direkte på gitterbunn eller annet hardt underlag, det kan skade føttene.

Buret kan du innrede med f.eks. røtter og greiner til å klatre på, hus av treverk, rør til å krype gjennom og løpehjul. Et løpehjul til degus bør være minst 25 cm i diameter og ha tett gulv og vegger slik at halen ikke kan komme i klem. Bunnmaterialet i buret bør være et tykt lag av støvfritt smådyrstrø som er egnet for graving. Det finnes flere egnede strøtyper å få kjøpt. Bland gjerne inn litt tørr blomsterjord eller sand i bunnmaterialet. Hvis du henter sand ute, bør du desinfisere den ved frysing eller steking fr den brukes i buret. Degusene trenger også tilgang på fiberrikt materiale som høy, revet papir eller treull til bygge- og redemateriale.

For å være sikker på at alle degusene som bor sammen kan trekke seg unna og hvile, bør de få hvert sitt sovehus. Deguser kan også finne på å forsvare maten sin overfor andre deguser. Når flere deguser bor i samme bur, bør du derfor ha flere matskåler plassert på forskjellige steder i buret slik at alle får tilgang til mat når de ønsker det. Dyrene må alltid ha tilgang på friskt vann, som best gis i en drikkeflaske.

For å holde pelsen i orden har deguser behov for å bade i finkornet sand flere ganger i uka. Slik badesand får du kjøpt i dyrebutikken. Et tungt kar som ikke velter så lett er fint som sandbadekar. Det kan være lurt å fjerne sandbadet mellom hvert bad for å unngå at den brukes som toalett.

Rengjøring

Hvor ofte buret bør rengjøres avhenger av størrelsen og hvor mange som bor der. Som en tommelfingerregel bør du skifte bunnmaterialet og vaske bunn og innredning én gang i uka.

IV. Experiments

V. Results

VI. Conclusion

Acknowledgment

I would like to acknowledge myself for being a absolute legend.

References

References

[Brunton and Kutz, 2022] Brunton, S. L. and Kutz, J. N. (2022). *Data-Driven Science and Engineering*. Cambridge Univeristy Press, Cambridge, UK.

[Bubeck et al., 2023] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

[Du et al., 2023] Du, Y., Watkins, O., Wang, Z., Colas, C., Darrell, T., Abbeel, P., Gupta, A., and Andreas, J. (2023). Guiding pretraining in reinforcement learning with large language models.

[Li et al., 2022] Li, S., Puig, X., Paxton, C., Du, Y., Wang, C., Fan, L., Chen, T., Huang, D.-A., Akyürek, E., Anandkumar, A., Andreas, J., Mordatch, I., Torralba, A., and Zhu, Y. (2022). Pre-trained language models for interactive decision-making. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31199–31212. Curran Associates, Inc.

[Luketina et al., 2019] Luketina, J., Nardelli, N., Farquhar, G., Foerster, J., Andreas, J., Grefenstette, E., Whiteson, S., and Rocktäschel, T. (2019). A survey of reinforcement learning informed by natural language.

[Zhang et al., 2023] Zhang, J., Lehman, J., Stanley, K., and Clune, J. (2023). Omni: Open-endedness via models of human notions of interestingness.