

MineRL Diamond 2021 Competition: Overview, Results, and Lessons Learned

Anssi Kanervisto*
University of Eastern Finland

ANSSK@UEF.FI

Stephanie Milani*
Carnegie Mellon University

SMILANI@CS.CMU.EDU

Karolis Ramanauskas
Independent

KAROLIS.RAM@GMAIL.COM

Nicholay Topin
Carnegie Mellon University

NTOPIN@CS.CMU.EDU

Zichuan Lin

ZICHUANLIN@TENCENT.COM

Junyou Li

JUNYOU LI@TENCENT.COM

Jianing Shi

JIANINGSHI@TENCENT.COM

Deheng Ye

DERICYE@TENCENT.COM

Qiang Fu

LEONFU@TENCENT.COM

Wei Yang

WILLYANG@TENCENT.COM

Tencent AI Lab

Weijun Hong
Zhongyue Huang
Haicheng Chen
Guangjun Zeng
Yue Lin

HONGWEIJUN@CORP.NETEASE.COM
HUANGZHONGYUE@CORP.NETEASE.COM
CHENHAICHENG@CORP.NETEASE.COM
GZZENGGUANGJUN@CORP.NETEASE.COM
GZLINYUE@CORP.NETEASE.COM

NetEase Games AI Lab

Vincent Micheli
Eloi Alonso
François Fleuret
University of Geneva

VINCENT.MICHELI@UNIGE.CH
ELOI.ALONSO@UNIGE.CH
FRANCOIS.FLEURET@UNIGE.CH

Alexander Nikulin
Yury Belousov
Oleg Svidchenko
Aleksei Shpilman
HSE University, JetBrains Research

HSEHOWUHH@GMAIL.COM
YURY.BELOUSOV@UNIGE.CH
OLEG.SVIDCHENKO@JETBRAINS.COM
ALEKSEI@SHPILMAN.COM

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

* Equal contribution

Abstract

Reinforcement learning competitions advance the field by providing appropriate scope and support to develop solutions toward a specific problem. To promote the development of more broadly applicable methods, organizers need to enforce the use of general techniques, the use of sample-efficient methods, and the reproducibility of the results. While beneficial for the research community, these restrictions come at a cost—increased difficulty. If the barrier for entry is too high, many potential participants are demoralized. With this in mind, we hosted the third edition of the MineRL `ObtainDiamond` competition, MineRL Diamond 2021, with a separate track in which we permitted *any* solution to promote the participation of newcomers. With this track and more extensive tutorials and support, we saw an increased number of submissions. The participants of this easier track were able to obtain a diamond, and the participants of the harder track progressed the generalizable solutions in the same task.

Keywords: Reinforcement Learning, Imitation Learning, Deep Learning, Sample-Efficient Learning, Generalization, Minecraft

1. Introduction

Reinforcement learning (RL) and deep RL (DRL) have successfully been used to tackle many complex sequential decision-making problems, from board games (Silver et al., 2018) to video games (Vinyals et al., 2019; Berner et al., 2019) to stratospheric balloon navigation (Bellemare et al., 2020). Most of these successes required massive amounts of computational resources, mainly because DRL algorithms require millions or billions of environment interactions to train. These algorithms are also sensitive to minor changes in hyperparameters (Henderson et al., 2018; Ibarz et al., 2021) or observations (Cobbe et al., 2020), leading to higher compute and sample requirements for hyperparameter tuning. Furthermore, solutions often overfit to the domain by exploiting inductive biases, so insights do not transfer to other problems.

Previous DRL competitions mostly focus on achieving the highest possible scores in a particular environment. The winning solutions usually include environment-specific, hard-coded rules and action/reward shaping (Kanervisto et al., 2021). Some competitions even include separate tracks for DRL agents (Küttler et al., 2020; Mohanty et al., 2020) because they are often weaker than non-learning-based approaches. Another relatively common way to succeed in a DRL competition is to throw massive amounts of compute at the problem (Bou and De Fabritiis, 2020), which not everyone can afford. Such solutions also do not transfer well to real-world settings, where lightweight simulators are not available.

To promote the development of DRL techniques that generalize to new domains and are more sample-efficient, we ran the third iteration of the MineRL Diamond competition as part of the NeurIPS 2021 Competition Track. The goal of this competition is for agents to obtain a diamond in Minecraft. To encourage sample-efficient and generalizable solutions (Houghton et al., 2020), we limited the number of environment interactions agents can be trained on, retrained submissions for final evaluation, and prohibited manual programming of behaviour. To encourage more general participation outside of the DRL research community, we introduced a novel track, the `intro` track, which follows the standard competition format: no limits on hardcoding, training time, or approaches. To increase accessibility, we developed resources, such as Colab notebooks, to help participants

get started with training agents. Additionally, we improved our community support through office hours and a workshop.

2. Competition Overview

To encourage broader participation and move toward the use of the MineRL challenge as a general research benchmark, the **intro** track acts as a stepping stone towards the main competition. In this track, participants may use any techniques to develop agents, which are evaluated post-training (as is done in standard benchmarks and most other RL competitions). We maintained the original competition rules and constraints Guss et al. (2021) that foster the development of novel techniques in the **research** track. We eliminated the imitation-learning-only track from the 2020 competition (Guss et al., 2021), since participants tended to leverage a combination of imitation learning and reinforcement learning. Focusing on the combined track better reflects the overall goals of the competition.

2.1. Task

The core task, **ObtainDiamond**, was to train an agent to obtain a diamond in Minecraft, which requires obtaining resources and crafting tools to progress through the item hierarchy (Guss et al., 2019). Agents start in a random location in a random world generated by Minecraft. The agent has access to a 64×64 pixel observation from the point-of-view of the player (similar to what a human would see) and its current inventory contents. The agent controls the avatar with actions similar to keyboard buttons (e.g., moving forward or left), selected 20 times per second. The agent receives a reward each time they obtain the next item in the list of required items towards diamond, detailed in Table 1. An episode ends once the agent obtains a diamond, the agent dies, or the episode lasts for 18000 steps (15 minutes of in-game time).

2.2. research Track

Because the **research** track retained sample efficiency, reproducibility, and generalizability constraints as in the 2020 competition, we provide an outline here and refer to Guss et al. (2021) for full details. The competition structure and rules prohibited manual specification of behaviour. We obfuscated the observations and actions to have no human-readable meaning, which prevented semantic hardcoding. The rules prohibited reward shaping, selecting sub-policies based on environment observations, and manually pruning or selecting actions. To enforce these rules, we manually validated the participants’ training and evaluation code during both rounds. In case of uncertain rule-breaking, we contacted participants and asked for clarification. Rule-breaking teams were disqualified.

Although in Round 1 participants submitted trained agents along with the training code, in Round 2 the finalists submitted only their training code. We trained the submissions on the evaluation server with unseen seeds and environment dynamics to promote generalization. To enforce sample efficiency, we limited training to eight million environment interactions on fixed hardware (1 NVIDIA K80 for 4 days).

2.3. intro Track

The **intro** track had the same challenge without the restrictions of the **research** track. Participants were allowed to use any techniques as part of their agent (minus directly modifying the environment), train the agents on any hardware, and only submitted the trained agent. Observations and actions were in the original, human-readable format. However, we still evaluated agents on an unseen set of seeds. To encourage the transition to the **research** track, we scheduled the **intro** track to end a few weeks prior to the end of Round 1 of the **research** track.

2.4. Resources

In addition to the human gameplay dataset (Guss et al., 2019), we provided participants with template code for the submission system for a number of different baseline methods. For the **intro** track, the first baseline used manual behavior specification (a predefined sequence of actions), the second used manual specification in addition to behaviour cloning, and the third used manual specification along with PPO (Schulman et al., 2017). These baselines demonstrated how the more flexible rules of **intro** track can be used advantageously along with learning methods.

For the **research** track, we provided a new baseline using behavioural cloning in addition to the baselines from the 2020 competition (Guss et al., 2021) by Preferred Networks, which were applicable to this competition (four RL methods). We also compiled a list of research articles and projects using MineRL in a centralized place as further background material for the participants.¹

To increase the accessibility of the competition, we provided Colab notebook versions of the **intro** track baselines,² which could be run without any local installations. Together with a video guide on how to make a submission through the browser, the Colab-based baselines enabled people to fully participate in the competition on any device with an internet connection. We also organized a virtual workshop,³ where participants worked on their submissions or a set of tutorial tasks⁴ to get interested people started, regardless of skill.

2.5. Evaluation

As in the previous years, we hosted the competition with AICrowd. Participants created a git repository to which they uploaded their submission code. The evaluation system cloned the repository, built a Docker image according to the participant’s settings, and performed

Table 1: Rewards for obtaining items (once per episode).

| Item | Reward |
|----------------|--------|
| Log | 1 |
| Planks | 2 |
| Stick | 4 |
| Crafting table | 4 |
| Wooden pickaxe | 8 |
| Stone | 16 |
| Furnace | 32 |
| Stone pickaxe | 32 |
| Iron ore | 64 |
| Iron ingot | 128 |
| Iron pickaxe | 256 |
| Diamond | 1024 |

1. <https://minerl.readthedocs.io/en/latest/notes/useful-links.html>

2. For example, https://colab.research.google.com/drive/1qfjHCQkukFcR9w1aPvGJyQx0a-Gv7Gt_

3. <https://minerl.io/workshop>

4. <https://github.com/minerllabs/getting-started-tasks>

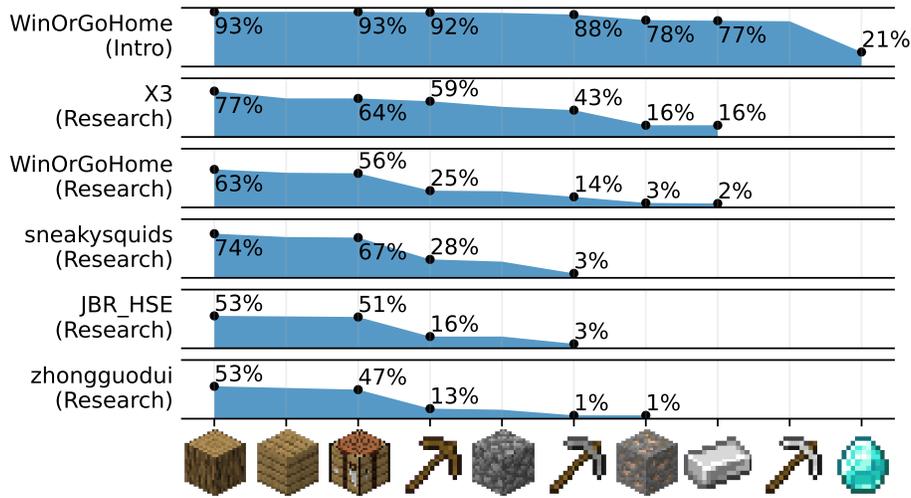


Figure 1: Fraction of episodes agent obtains the item on x-axis, from the Round 2 top submissions for **research** track and the top **intro** submission, measured over 100 episodes.

the training and/or evaluation. The Minecraft environment ran in a separate container to prevent the participants from accessing the environment.

Submissions were evaluated based on the average sum of rewards over 100 episodes. Each relevant item for obtaining a diamond provided an exponentially increasing reward, illustrated in Table 1. This system credited agents that are closer to obtaining a diamond over agents that reliably obtain earlier items to prevent specification gaming (Clark and Amodei, 2016). For participants who submitted multiple solutions, the final score was the highest over all of their submissions.

3. Solutions

In Section 3.1, we describe the performance of the submissions and how they compare to the performance of submissions to the 2020 edition of the competition. The remaining sections summarize the techniques used by the competitors.

3.1. research Track Submissions Overview

Table 3 compares this year’s competition activity with previous ones. This year, we attracted more active participation: despite a decreased number of teams compared to the 2020 edition and the lowest number of individual competitors, the 2021 edition saw the most submissions. This finding is further supported by the higher number of non-zero-score submissions before the final round for both tracks and a comparable number of non-zero-score submissions in the Round 1 **research** track. We believe this increased active participation is due to the inclusion of the **intro** track, which eased the entry to the competition.

Table 2: Top 10 submissions to both tracks, including submissions to the 2020 edition of the competition for **research** track Round 2.

| intro | | research, Round 1 | | research, Round 2 | |
|-------------|--------|-----------------------|-------|-------------------|-------|
| Name | Score | Name | Score | Name | Score |
| WinOrGoHome | 645.55 | X3 | 68.12 | X3 | 76.97 |
| X3 | 442.38 | PA-P | 54.94 | 2020 #1 | 39.55 |
| zhongguodui | 349.97 | SneakySquids | 22.54 | WinOrGoHome | 22.97 |
| MCAgent | 323.82 | MCAgent | 17.70 | SneakySquids | 14.35 |
| ISYAgent | 220.03 | xianyu | 16.71 | 2020 #2 | 13.29 |
| ced | 167.78 | Reforcos_de_Minecraft | 13.36 | 2020 #3 | 12.79 |
| xianyu | 79.06 | zhongguodui | 11.08 | JBR_HSE | 10.33 |
| PA-P | 69.50 | WinOrGoHome | 10.55 | zhongguodui | 8.84 |
| ttom | 45.35 | grimoirez | 8.55 | MikeAI | 6.25 |
| orithu | 38.42 | okkkkkkkk | 6.44 | 2020 #4 | 5.16 |

Table 3: Competition activity statistics.

| Metric | 2019 | 2020 | 2021 |
|--|-------------|-----------|-------------|
| Number of Teams | 47 | 95 | 59 |
| Number of Individuals | 1115 | 707 | 468 |
| Number of Submissions | 662 | 490 | 1190 |
| Non-zero-score Pre-Final-Round Submissions | 38 | 36 | 55 |
| Non-zero-score research Track Submissions | 38 | 36 | 36 |
| Number of Discord Server Messages | 2464 | 2444 | 5556 |
| New Discord Server Users | 445 | 160 | 955 |

In the **intro** track, 16 teams outperformed the baseline. In the **research** track 20 teams outperformed the BC baseline in Round 1. In Round 2 of the **research** track, nine teams achieved a non-zero score. Table 2 shows the final ranking of the **intro** and **research** tracks, respectively. The overall winner of the **research** track outperformed last year’s top solution (Mao et al., 2021) by a large margin.⁵

Solutions in the **intro** track reached notably higher scores (top three average of 479.3 vs. 48.5 for **research** Round 1), due to fewer restrictions on usable techniques. With these less restrictive rules, teams could design useful macro actions and train agents on them (e.g., WinOrGoHome) or hard-code most of the behaviour, with more difficult sub-parts learned by an RL algorithm (e.g., X3).

5. Original #1 team from the 2020 competition with over 70 score was retroactively disqualified for breaking the rules.

3.1.1. RESEARCH FIRST PLACE: TEAM X3

The first-place team (X3) proposed a sample-efficient hierarchical RL approach equipped with representation learning and imitation learning to deal with perception and exploration (Lin et al., 2021). Specifically, their approach included two levels of hierarchy—the high-level meta-policy learned to control over options to switch between different sub-tasks, while the low-level sub-policies learned to solve each sub-task efficiently. To boost the learning of sub-tasks, they used a combination of techniques including 1) action-aware representation learning which captured underlying relations between action and representation, 2) discriminator-based self-imitation learning for efficient exploration, and 3) ensemble behavior cloning with consistency filtering for policy robustness. These techniques significantly improved the sample efficiency of learning agents and the proposed approach achieved the highest score (76.97) in the research track, which outperformed last year’s top solution by a large margin (39.55).

3.1.2. RESEARCH SECOND PLACE: TEAM WINORGOHOME

Team WinOrGoHome based their solution of team NoActionWasted from the previous year (Guss et al., 2021). This solution used k-means to cluster the actions, after which a recurrent neural network was trained with behavioural cloning. Team WinOrGoHome improved this solution by following techniques:

- Add the current reward into the state for helping the agent distinguish between different stages.
- For helping the agent focus on the key frames, weigh the policy loss by the magnitude of the change in the obfuscated observation vector (this contains information of the player’s inventory and item held in hand).
- Predict the discounted return on the `ObtainIronPickaxe` dataset as an auxiliary task, where the target to predict is $\log(1 + \text{value})/8$.
- Add linear learning rate decay, from $6.25 \cdot 10^{-5}$ to $1.00 \cdot 10^{-5}$.

This improved imitation-only agent ranked 2nd place on the final leaderboard, scoring 22.97, which is a notable improvement over team NoActionWasted’s score of 12.79.

3.1.3. RESEARCH THIRD PLACE: TEAM SNEAKYSQUIDS

Language models (Radford et al., 2018, 2019; Brown et al., 2020; Devlin et al., 2019) have demonstrated strong results both in text-based environments (Micheli and Fleuret, 2021; Yao et al., 2020) and in tasks and environments that have seemingly no ties to natural language processing (Lu et al., 2021; Reid et al., 2022). Team SneakySquids followed this line of work by finetuning a pretrained GPT-S model (Radford et al., 2019) with behavioural cloning of expert gameplay in the `Treechop`, `ObtainIronPickaxe` and `ObtainDiamond` environments. During an initial data preprocessing step they removed below average demonstrations and discretized the action space with k-means clustering of expert actions. Then the model was trained to predict the actions taken by expert players conditioned on a sequence of previous observations. An observation consisted of two distinct embeddings: a

frame embedding learnt end-to-end by a convolutional neural network (Zhang et al., 2019) and an obfuscated inventory embedding given by the environment.

With this design, the agent had access to a compressed representation of the last ten seconds of gameplay for deciding the action, which enabled smooth navigation and short bursts of crafting actions. Consistently with recent work (Reid et al., 2022), team SneakySquids observed that a pretrained GPT-2 model learnt faster and achieved higher scores than GPT-2-like architectures initialized from scratch. Even though this approach relied purely on imitation learning and was thus heavily penalized by the scarcity of the training data, their best performing agent achieved an average score of 22.540 on the evaluation platform in Round 1, and 14.35 in the final ranking of Round 2.

3.1.4. RESEARCH FOURTH PLACE: TEAM JBR_HSE

Team JBR_HSE’s approach was based on a composition of hierarchical reinforcement learning and imitation learning. They hypothesized that RL performs well with short-horizon planning and high-level actions, while micromanagement is too complex and requires more time to learn than the competition limit provides. Therefore, micromanagement should be learned through imitation learning.

On a high level, their approach involved two levels, i.e., one high-level policy and many low-level options (Sutton et al., 1999). All trajectories were evenly divided into N chunks by time. Then a separate IL policy was learned as an option for each chunk. After options were trained, they were used in place of direct actions in a Deep Q-Network (DQN) (Mnih et al., 2013). When DQN chose an option, the policy responsible for it performed K steps in the environment with low-level actions from the original action space. In all architectures they used standard ResNet (Zhang et al., 2019) and long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) (for both levels).

Such an approach appeared to be better than pure imitation learning but did not achieve impressive results (10.33 in Round 2). The main reason for this is that larger intervals of training data resulted in better options but as a result, DQN’s training budget was reduced. This work was supported by computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

3.2. intro Track Submissions Overview

With the relaxed rules of the `intro` track, top participants were able to obtain the diamond for the first time in all of the MineRL competitions. While these results are not directly comparable to previous MineRL competitions or `research` track submissions, it demonstrates the effectiveness of domain knowledge when it comes to training agents.

3.2.1. INTRO FIRST PLACE: TEAM WINORGOHOME

Team WinOrGoHome leveraged end-to-end reinforcement learning without any hierarchy training paradigm. For the action space, they designed 20 useful discrete actions with some actions that sometimes get masked according to their comprehension of Minecraft, which largely sped up the exploration. For feature engineering, they used every fourth observation and modified the image and vector observations. They stacked the most recent four non-skipped frames into a $64 \times 64 \times 12$ tensor for the image part. The vector part

included the info of masked actions, the current count of items in the inventory together with their cumulative counts through the episode. They designed a more elaborate reward system helping the agent explore the hierarchical dependency property of different items. Their model follows the structure in IMPALA (Espeholt et al., 2018) but without long LSTM (Hochreiter and Schmidhuber, 1997), while PPO (Schulman et al., 2017) is used to train the model. Their final model achieved an average score of above 560 after consuming 625 million frames (i.e., 2.5 billion frames in total) and ranked 1st for the `intro` track with the best submission score of 645.55. The agent obtained 21 diamonds out of the overall 100 episodes in this submission.

3.2.2. `INTRO` SECOND PLACE: TEAM X3

For the `intro` track, team X3 used a hard-coded behavior tree for all stages, except for tree-chopping, where Soft Q Imitation-Learning (Reddy et al., 2020) was used instead. Several combo-actions were designed to achieve high-level abstract actions, such as “attack-until-inventory-changes-with-timeout”, “reset-yaw-pitch” and “clear-nearby-ore-vein”. In order to avoid death and recognize certain blocks, they deployed a simple yet effective hard-coded color-mask detection. The agent gets an iron-pickaxe in 70% of the runs and gets a diamond in 10% runs.

4. Discussion

Here we discuss the impact of the inclusion of the `intro` track and the progress made in the task over three years of MineRL Diamond competitions.

4.1. Impact of the `intro` Track

While the high-scoring solutions of the `intro` track are not academically as interesting as solutions for `research` track (generalizable), the inclusion of `intro` track along with newcomer-friendly material attracted more newcomers to the field. We saw more activity on our Discord server for the competition along with more new users (see Table 3).⁶ We also received multiple messages thanking us for the more inclusive nature of the competition, noting that getting started was too difficult in the previous years. One participant even reported using the entire inclusivity compute grant and more for this competition.

The overall higher score of the `intro` track solutions compared to the `research` track solutions signals that simple, domain-specific engineering can yield more performant agents. For example, crafting in the MineRL environment requires picking a specific action once requirements are satisfied, which can be achieved with a single if-then statement. However, chopping down a tree requires locating a tree and then navigating to a it, which is not easily programmable. Instead, learning it with RL has proven more effective. This provides a more complete picture of “which approaches work” and provides a fairer comparison between symbolic and learning agents. Similar results were observed in the NetHack

6. Some of the users joined for the MineRL BASALT competition which used the same Discord server, but majority joined for the Diamond competition.

Challenge (Küttler et al., 2020), where symbolic agents outperformed learning agents by a considerable margin.⁷

Given these insights, we highly recommend that competitions aim to include a track similar to the `intro` track, along with more approachable tutorials, instructions, events and submission instructions, to attract people outside the field of AI. However, this easier track should not replace the original, and more difficult, research track, but should be held alongside it. This way, the easier track helps to attract participants and drive the competition by including different parties. It will also create a sense of progression which helps with the morale of the teams in the harder track. The harder track should be encouraged by providing larger prizes or by ending the easier track sooner to allow people to switch tracks. We also advise preparing answers for newcomers to programming, especially if you organize a competition around a video game. We received messages from many users who were familiar with Minecraft but did not have experience with programming but were eager to try out MineRL. We directed these people to the beginner programming tutorials and provided step-by-step code examples for MineRL. We recommend other competitions to do the same, time permitting, to avoid distinguishing the curiosity of interested minds.

4.2. Brief Analysis of the Results

For three years, hundreds of participants have tuned their agents to find the diamond in Minecraft, each year with their own setting. In 2019, the competition outline was designed and focused on the sample efficiency and reproducibility of the training results (Milani et al., 2020). In 2020, this was further encouraged with the obfuscated observations and actions (Guss et al., 2021). In 2021, we noted that the restrictions limit active participants to a select group (highly skilled people in RL) and added a less-constrained track to ramp up newcomers. Only in this easier track and after two competitions, did we finally see agents obtaining diamonds. How come?

Rarity of important events. For example, to obtain a wooden pickaxe (Figure 1, fourth item), the player must craft a crafting table (third item), place it down, look towards it and execute “craft wooden pickaxe” action. If the player has the necessary items in the inventory, they will be replaced with a wooden pickaxe. This only needs to be done once per episode to progress further. Despite getting an immediate reward for crafting the item, this is difficult to learn for the RL agent to due its rarity (once per episode) and the dependencies it requires. However, manually programming this behaviour is simple. This is reflected in Figure 1, where the `research` track agents see a significant drop at obtaining wooden pickaxe, while `intro` track agent does not experience this. The research-oriented submissions have, in the three years, largely utilized some form of hierarchical learning to combat this.

Strict limits. With the compute limit of eight million environment steps and an episode time limit of 18,000 steps, participants had 450 full episodes of data to train on. Sample efficiency is desirable, but this might be too restrictive to obtain high performance, especially with the above limitations. Team WinOrGoHome obtained diamond multiple times by training for longer with crafted actions to avoid the rarity issue in the `intro` track.

7. <https://nethackchallenge.com/report.html>

MineRL’s slow sampling. Because of the relatively slow sampling rate of MineRL (max. 300 steps-per-second per environment which uses multiple cores), extensive hyperparameter tuning and experimentation requires a large amount of compute. By creating a Minecraft-like game from the ground up, [Hafner \(2021\)](#) was able to deeper explore the effectiveness of different algorithms. This highlights how important hyperparameter tuning is for good performance, but at the same time partially breaks the desiderata for sample-efficient algorithms.

5. Related Work

The year 2021 saw a number of Minecraft competitions that were organized alongside this MineRL Diamond competition. A sister competition, MineRL BASALT ([Shah et al., 2021](#)), made use of the same MineRL library to express desired agent goals with human demonstrations instead of the rewards, and then used human evaluators to rate the agents on how well they completed the tasks. The IGLU competition ([Kiseleva et al., 2021](#)) used the building mechanics of Minecraft for two challenges for the participants: build an agent that provides descriptions on what to build, and then build an agent which builds a structure based on the instructions. Prior to 2021, MarLÖ competition ([Perez-Liebana et al., 2019](#)) challenged participants to solve multiple tasks, some unseen, in a multi-agent setting in Minecraft. The Generative Design in Minecraft competition ([Salge et al., 2020](#)) challenges participants to generate settlements that fit into the surrounding environment.

The Progen competition ([Mohanty et al., 2021](#)), the Nethack challenge ([Küttler et al., 2020](#)) and the Unity Obstacle Tower challenge ([Juliani et al., 2019](#)) all made use of the procedural generation of levels as well. However, only the Progen competition also limited the number of environment samples for training and retrained the agents on the evaluation server.

As for providing an easier track, the ViZDoom competitions in 2016 and 2017 included a “limited deathmatch” and “full deathmatch”, where the former took place in known maps and the latter in unknown maps ([Wydmuch et al., 2018](#)). However, participants were not limited by a number of samples or by strict rules, and the top teams often reported using a large amount of compute.

6. Conclusion

We ran the 2021 MineRL Diamond competition to promote the development of general, sample-efficient DRL and imitation learning algorithms. In this work, we described the competition, highlighted changes from the previous competitions, and summarized the performance of the submissions, and contrasted them with the performance from last year. The core changes this year were the inclusion of an easier track and more tutorials to welcome newcomers. We saw an increased number of submissions and active participants, and the top team outperformed the previous best result by a large margin.

These results demonstrate that hosting an easier track of a more challenging competition is beneficial, and we recommend other competitions take this approach. In the future, the core of the MineRL challenge could be redesigned to reap the most benefit of a public competition, and the results over three years could be summarized and analysed in detail.

Acknowledgments

We thank the other organizers for helping to run the competition: William H. Guss, Brandon Houghton, Byron Galbraith, Noboru Sean Kuno, Rohin Shah, Steven H. Wang, and Cody Wild. We also thank the advisory committee. We thank Microsoft and AI Journal for providing financial support. We are grateful to AICrowd members Sharada Mohanty and Shivam Khandelwal for setting up the backbone of the submission, evaluation and ranking system. Last but not the least, we would like to thank all the participants on the MineRL Discord server who helped others by answering questions, sharing thoughts, and reporting their progression.

References

- Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.
- Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.
- Albert Bou and Gianni De Fabritiis. Pytorchrl: Modular and distributed reinforcement learning in pytorch. *arXiv preprint arXiv:2007.02622*, 2020.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- Jack Clark and Dario Amodei. Faulty reward functions in the wild, 2016. URL <https://blog.openai.com/faulty-reward-functions>.
- Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pages 2048–2056. PMLR, 2020.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.

- Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures, 2018.
- William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: a large-scale dataset of minecraft demonstrations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2442–2448, 2019.
- William Hebggen Guss, Stephanie Milani, Nicholay Topin, Brandon Houghton, Sharada Mohanty, Andrew Melnik, Augustin Harter, Benoit Buschmaas, Bjarne Jaster, Christoph Berganski, et al. Towards robust and domain agnostic reinforcement learning competitions: Minerl 2020. In *NeurIPS 2020 Competition and Demonstration Track*. PMLR, 2021.
- Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *Deep RL Workshop NeurIPS 2021*, 2021.
- Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- Brandon Houghton, Stephanie Milani, Nicholay Topin, William Guss, Katja Hofmann, Diego Perez-Liebana, Manuela Veloso, and Ruslan Salakhutdinov. Guaranteeing reproducibility in deep learning competitions. In *The 23rd Conference on Neural Information Processing Systems, Challenges in Machine Learning (CiML) Workshop*, 2020.
- Julian Ibarz, Jie Tan, Chelsea Finn, Mrinal Kalakrishnan, Peter Pastor, and Sergey Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.
- Arthur Juliani, Ahmed Khalifa, Vincent-Pierre Berges, Jonathan Harper, Ervin Teng, Hunter Henry, Adam Crespi, Julian Togelius, and Danny Lange. Obstacle tower: A generalization challenge in vision, control, and planning. *arXiv preprint arXiv:1902.01378*, 2019.
- Anssi Kanervisto, Christian Scheller, Yanick Schraner, and Ville Hautamäki. Distilling reinforcement learning tricks for video games. In *2021 IEEE Conference on Games (CoG)*, pages 01–04. IEEE, 2021.
- Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hove, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. Neurips 2021 competition IGLU: Interactive grounded language understanding in a collaborative environment. *arXiv preprint arXiv:2110.06536*, 2021.

- PS Kostenetskiy, RA Chulkevich, and VI Kozyrev. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, page 012050. IOP Publishing, 2021.
- Heinrich Küttler, Nantas Nardelli, Alexander H Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The nethack learning environment. In *NeurIPS*, 2020.
- Zichuan Lin, Junyou Li, Jianing Shi, Deheng Ye, Qiang Fu, and Wei Yang. Juewu-mc: Playing minecraft with sample-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:2112.04907*, 2021.
- Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *CoRR*, abs/2103.05247, 2021. URL <https://arxiv.org/abs/2103.05247>.
- Hangyu Mao, Chao Wang, Xiaotian Hao, Yihuan Mao, Yiming Lu, Chengjie Wu, Jianye Hao, Dong Li, and Pingzhong Tang. Seihai: A sample-efficient hierarchical ai for the minerl competition. In *International Conference on Distributed Artificial Intelligence*, pages 38–51. Springer, 2021.
- Vincent Micheli and Francois Fleuret. Language models are few-shot butlers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9312–9318, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.734. URL <https://aclanthology.org/2021.emnlp-main.734>.
- Stephanie Milani, Nicholay Topin, Brandon Houghton, William H Guss, Sharada P Mohanty, Keisuke Nakata, Oriol Vinyals, and Noboru Sean Kuno. Retrospective analysis of the 2019 minerl competition on sample efficient reinforcement learning. In *NeurIPS 2019 Competition and Demonstration Track*. PMLR, 2020.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.
- Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, et al. Flatland-rl: Multi-agent reinforcement learning on trains. *arXiv preprint arXiv:2012.05893*, 2020.
- Sharada Mohanty, Jyotish Poonganam, Adrien Gaidon, Andrey Kolobov, Blake Wulfe, Dipam Chakraborty, Gražvydas Šemetulskis, João Schapke, Jonas Kubilius, Jurgis Pašukonis, et al. Measuring sample efficiency and generalization in reinforcement learning benchmarks: Neurips 2020 procgen benchmark. *arXiv preprint arXiv:2103.15332*, 2021.
- Diego Perez-Liebana, Katja Hofmann, Sharada Prasanna Mohanty, Noburu Kuno, Andre Kramer, Sam Devlin, Raluca D Gaina, and Daniel Ionita. The Multi-Agent Reinforcement Learning in MalmÖ (MARLÖ) Competition. *arXiv preprint arXiv:1901.08129*, 2019.

- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Siddharth Reddy, Anca D Dragan, and Sergey Levine. Sqil: Imitation learning via reinforcement learning with sparse rewards. In *International Conference on Learning Representations*, 2020.
- Machel Reid, Yutaro Yamada, and Shixiang Shane Gu. Can wikipedia help offline reinforcement learning? *arXiv preprint arXiv:2201.12122*, 2022.
- Christoph Salge, Michael Cerny Green, Rodrigo Canaan, Filip Skwarski, Rafael Fritsch, Adrian Brightmoore, Shaofang Ye, Changxing Cao, and Julian Togelius. The ai settlement generation challenge in minecraft. *KI-Künstliche Intelligenz*, 34(1):19–31, 2020.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- Rohin Shah, Cody Wild, Steven H Wang, Neel Alex, Brandon Houghton, William Guss, Sharada Mohanty, Anssi Kanervisto, Stephanie Milani, Nicholay Topin, et al. The minerl basalt competition on learning from human feedback. *arXiv preprint arXiv:2107.01969*, 2021.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362, 2018.
- Richard S. Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, 1999. ISSN 0004-3702. doi: [https://doi.org/10.1016/S0004-3702\(99\)00052-1](https://doi.org/10.1016/S0004-3702(99)00052-1). URL <https://www.sciencedirect.com/science/article/pii/S0004370299000521>.
- Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michael Mathieu, Andrew Dudzik, Junyong Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019.
- Marek Wydmuch, Michał Kempka, and Wojciech Jaśkowski. Vizdoom competitions: Playing doom from pixels. *IEEE Transactions on Games*, 11(3):248–259, 2018.
- Shunyu Yao, Rohan Rao, Matthew Hausknecht, and Karthik Narasimhan. Keep CALM and explore: Language models for action generation in text-based games. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8736–8754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.704. URL <https://aclanthology.org/2020.emnlp-main.704>.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *CoRR*, abs/1901.09321, 2019. URL <http://arxiv.org/abs/1901.09321>.