Vicky Nguyen, vtn180004
Meinhard Capucao, mdc190005
CS 4395.001

ACL Paper

Title: "Perceiving the World: Question-guided Reinforcement Learning for Text-based Games"
Author list and affiliations: Yunqiu Xu, Meng Fang, Ling Chen, Yali Du, Joey Zhou, and Chengqi Zhang

Introduction

Text-based games provide a secure and engaging environment for studying dialogue systems, commonsense reasoning, and natural language understanding. Even though deep reinforcement learning has proven useful in creating game-playing agents, its inability to be deployed in the real world is still hampered by two significant issues: it's low sample efficiency and large action space. Low sample efficiency, which refers to the fact that training an agent to function at a human level often requires a vast quantity of data, is a key constraint of reinforcement learning. The reason is due to the fact that humans typically possess prior knowledge, saving them the time and effort of having to learn everything from the start. In comparison, the agent must perform language learning and decision-making procedures in a language-informed RL system, where the former can be thought of as prior knowledge and precedes much more slowly than the latter. The other key constraint, large action space, is due to the agent potentially wasting time and training data if the agent attempts to do inferior or irrelevant actions.

To solve the issue, the researchers designed a "world-perceiving module to realize the aforementioned functionalities such as task decomposition and action pruning" [1]. They named their "method as Question-guided World-perceiving Agent (QWA)" [1]. The agent would be guided by a couple of questions, then the agent would decompose the task in order to gain a set of available subtasks. Afterward, the agent would select one of the subtasks and perform action pruning to obtain a refined set of actions based on the selected subtask. They proposed achieving the world-perceiving modules through supervised pre-training in order to further improve the sample efficiency by decoupling language learning from decision-making. The authors achieved the supervised pre-training by designing a two-phased framework to train the QWA. A dataset is created in the first stage so that the world-perceiving modules can be trained. In the second stage, the pre-trained modules are frozen while the agent is deployed in games and trained using reinforcement learning.

2.1: RL agents for text-based games

Based on the style of the observations, RL agents for text-based games can be split up into text-based agents and KG-based agents. Knowledge graphs, as opposed to text-based agents, "provide structural and historical information that helps the agent to handle partial observability,

reduce action space, and improve generalizability across games [1].” In this study, the researchers use pre-training to increase sample effectiveness and decrease action space.

2.2: Data Efficiency

The work in the paper is closely related to task decomposition [2] [3] [4] and hierarchical reinforcement learning [5] [6] [7] from various different sources. One difference in this work, however, is a practical assumption that interaction data is limited, and that subtask termination states can’t be accessed.

2.3: Pre-training methods for RL

There are many works that have studied pre-training methods [8] One important concept is *Imitation Learning,* where agents imitate human interactions before deployment to RL. Another breakthrough in this work is the use of IL [9], and focusing on the agent perceiving the learning environment instead of relying on solutions. This approach favors human review and interaction since it prevails on simpler tasks compared to more complicated ones.

While other works have integrated state representation learning, knowledge graph constructing, and action pruning, “As far as we know, we are the first to incorporate pre-training based task decomposition in this domain. Besides, instead of directly pruning the actions based on the observation, we introduce subtask-conditioned action pruning to further reduce the action space.“

3: Background

Text-based games can be formulated as POMDPs [10] (Partially Observable Markov Decision Processes). This is based on a set of states, where there are one set *s* that represents all possible states, and an action set that are all the actions the agent can take in the game. As is standard with the Markov Process, the goal is to maximize the expected reward of taking certain actions by taking into account previous states and the game's partial observability. Observations can take the form of text, knowledge graphs, or hybrids. Then, with the problem setting, the goal is to think of games with similar themes and tasks, but varying complexities. A game's complexity depends on its number of subtasks; with less, it is considered simple. RL agents struggle with complex games because of training costs and the sheer amount of data needed, so sample efficiency and performance are metrics focused on instead.

“Our objective is to leverage the labeled data collected from simple games to speed up RL training in complex games, thus obtaining an agent capable of complex games.”

4.1: Framework Overview

The framework consists of two world-perceiving modules: one that obtains currently available subtasks with the task selector (T), and one that scores each action with the highest score chosen with the action selector (A). For training, there are two modules: one that collects

human interaction data in simple games and another that trains the action selector in complex games.

4.2: Task Selector

Subtasks are available if they are required to solve the global task and if there are no prerequisites. The task selector is world-perceiving since it can be cast as a binary classification problem where subtask availability is either yes or no. Supervised pre-training and reinforcement learning augment sample efficiency. While previous works consider task decomposition [11] [12], this work accounts for multiple available subtasks at a time step and has no requirement for a demonstration that solves the entire game.

4.3: Action Validator

After acquiring the subtask, action pruning is conducted to reduce the action space, which solves the second problem of large action space. Action pruning can also be cast as a binary classification which introduces another world-perceiving module: the action validator. This checks the relevance of each action with respect to the task.

4.4: Action Selector

The action selector is trained through RL with no human data in this phase. Subtasks will be used until they're not included in the task set. A concern is compound errors, with imperfect training modules affecting the RL training. To alleviate this, a time limit is added to subtasks which increases robustness to errors and improves subtask selection diversity.

5.1 - Experiment Setting

The researchers conducted the experiments using vocabulary-shared culinary games offered by the rl.0.2 game set and the FTWP game set. Their three game sets – 3488 simple games, 280 medium games, and 420 hard games – are designed based on the number of subtasks, which is strongly associated with the number of ingredients and preparation requirements for the cooking. The researchers noted that there are no overlapping games between the simple game sets and the medium-to-hard game sets. They collected human interaction data from the simple game sets for pre-training and because they regarded the medium-to-hard game set as complex, the researchers used them in the reinforcement learning without labeled data.

5.3 - Implementation Details

The researchers trained the task selector and action validator independently because they utilized different QAs and asked human players to participate in the simple games. The human players had to answer yes-or-no questions based on their observations. Afterward, the researchers trained the task selector and action validator, and the modules were trained using Focal loss and Adam optimizer with a learning rate of 0.001 across 10-20 epochs [1].

Since the researchers considered the medium and hard game sets as different experiments compared to the simple game set. They divided the hard game set into 300 training games / 60 validation games / 60 testing games and the medium game set into 200 / 40 / 40. For reinforcement learning, the researchers used the default settings from Adhikari. For training, they set the episode's step limit at 50, and for validation/testing, at 100. Then, set the subtask limit to $\xi = 5$. For 100,000 episodes, the researchers trained the models and optimized them by Double DQN and Pritorized Experience Replay, and would validate the model and report the testing performance for every 1,000 training episodes.

5.4 - Evaluation Metrics

The researchers evaluated the models based on how well they performed in RL testing. They refer to a game's score as the cumulative total of all awards, without any deductions. Since the maximum available scores for various games can vary, they reported the normalized score.

6.1 - Main Results

After 20,000 training episodes, this work's QWA surpasses other methodologies' performance in both medium and hard games. The performance of QWA can be improved with RL but not IL-based models because of observed instability. This model is also more robust against domain shifts from small to large games. There is an observable gap between subtasks available in the domains, since testing harder games inherently causes greater subtask encounters. IL-task selectors don't adapt well to these medium-hard games, since they are trained to identify unique subtasks and can miss other available ones.

6.2 Performance on the Simple Games

The QWS model achieves over 80% of scores on medium and hard games. The "IL w/o FT" performs well on simple pre-trained models but struggles with unseen generalized simple models [1]. These results indicate the QWS model generalizes well on games of all difficulties.

6.3 Ablation Study

Adding time limits to subtasks prevents agents from pursuing difficult ones, which improves subtask diversity. Then, compared to oracle world-perceiving modules, the QWA model improves when assigning the expert task selector or expert action validator (as expected).

6.4 Pre-training on the Partial Dataset

It's considered burdensome for human players to collect data from simple games, so conditions of limited data are investigated. In respect to QWA in consideration of reduced training data, the model performs well with pre-training datasets having 75% and 50% of their original, but with 25% remaining the model displays instability when learning hard games. In general, the work's model is robust to limited training data and "alleviates burdens of human annotations".

7 Conclusion

For solving text-based games using deep reinforcement learning, the researchers addressed the issues of low sample efficiency and large action space in this study. They introduced the world-perceiving modules, which are designed with the ability to automatically decompose tasks and prune actions by responding to environmental queries. A two-phase training framework that the researchers proposed decouples language learning from reinforcement learning. According to their experimental findings, their technique improves the performance while having a high sample efficiency. Moreover, it demonstrates robustness against insufficient pre-training data and compound errors. For future work, they want to provide contrastive learning objectives and KG-based data augmentation to further improve the pre-training performance.

Number of Citations the authors have received on Google Scholar

Yunqiu Xu - 326 or 118, Meng Fang - 2221, Ling Chen - 9313, Yali Du - 568, Joey Zhou - 5697, **Chengqi Zhang- 23831**

We believe the reason why their work is important is that they proposed a novel approach to solving the challenges of text-based games using reinforcement learning and question-guided exploration, which achieved relatively good performance on several benchmark datasets and has significant implications for the development of more advanced AI systems that can interact with humans in NLP. For example, more advanced AI systems may include AI agents, virtual assistants, and powerful robotics.

Reference

[1] Y. Xu, M. Fang, L. Chen, Y. Du, J. Zhou, and C. Zhang, “Perceiving the World: Question-guided Reinforcement Learning for Text-based Games,” *ACLWeb*, May 01, 2022. https://aclanthology.org/2022.acl-long.41/ (accessed Apr. 01, 2023).

[2] Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. In International Conference on Machine Learning (ICML), volume 70, pages 2661–2670.

[3] Suvir Mirchandani, Siddharth Karamcheti, and Dorsa Sadigh. 2021. Ella: Exploration through learned language abstraction. Advances in Neural Information Processing Systems (NeurIPS), 34.

[4] Sungryull Sohn, Junhyuk Oh, and Honglak Lee. 2018. Hierarchical reinforcement learning for zero-shot generalization with subtask dependencies. In Advances in Neural Information Processing Systems (NeurIPS), volume 31, pages 7156–7166.

[5] Peter Dayan and Geoffrey E Hinton. 1992. Feudal reinforcement learning. In Advances in Neural Information Processing Systems (NeurIPS), volume 5.

[6] Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. Advances in Neural Information Processing Systems (NeurIPS), 29:3675– 3683

[7] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. FeUdal networks for hierarchical reinforcement learning. In International Conference on Machine Learning (ICML), volume 70, pages 3540–3549

[8] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is all you need:

[9] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. 2018. Deep q-learning from demonstrations. In Thirtysecond AAAI conference on artificial intelligence (AAAI).

[10] Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for textbased games. arXiv preprint arXiv:1806.11532.

[11] Valerie Chen, Abhinav Gupta, and Kenneth Marino. 2021. Ask your humans: Using human instructions to improve generalization in reinforcement learning. In International Conference on Learning Representations (ICLR).

[12] Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuandong Tian, and Mike Lewis. 2019. Hierarchical decision making by generating and following natural language instructions. Advances in Neural Information Processing Systems (NeurIPS), 32:10025–10034.