An Off-Policy Approach to Learning in the Game of Pokemon

Anonymous Author(s)
Affiliation
Address
email

Abstract: The game of Pokemon is suprisingly complex and difficult to model. As such, applications of reinforcement learning on the game have seen limited success. Certain approaches have also not been feasible due to the inability for certain RL approaches to overcome the adversary they train against, particularly when trained against high-performing opponents. This paper attempts to solve this issue by making learning bi-directional and thus agnostic of the behavior policy's performance.

Keywords: Pokemon, Learning

1 Introduction

Pokemon is the largest grossing media franchise in the world [1], with a variety of mediums which it is present. The Pokemon series of video games feature a turn-based combat system that is easy to learn, but incredibly hard to master.

In this battle system, players each have up to six pokemon each, and (in a competitive setting) can choose for their active pokemon to either use one of four moves, or to switch into one of their benched pokemon. Moves can do damage, inflict status effects, change stats of yourself or the enemy, or even change variables about the battlefield itself, like the weather or the force of gravity.

Given all of these variables, attempts at applying traditional RL techniques to a Pokemon battle are

2 Pokemon Battles

This section will cover the peculiarites of Pokemon battles as a game to apply reinforcement learning to, as well as why certain decisions were made in this particular paper.

We will be making some simplifications to the game of Pokemon here to make reinforcement learning more applicable. Namely, we will be assuming a ”competitive” setting for a pokemon battle. This means items cannot be used and difficulty-reducing features like the ability to switch pokemon following a player knockout are not enabled. Furthermore, we can assume all pokemon present are reasonably powerful such that every pokemon present has some amount of competitive viability.

We will further assume the following to further simplify our model:

- We'll be only considering battles in the first generation of Pokemon. This means newer additions to the series, like EVs, Pokemon Abilities, Held Items, Weather Effects, and other complex moves do not need to be considered.

- Teams are randomly generated using the Pokemon Showdown random team generator. This is used because team selection is a large part of human competitive pokemon, which this paper does not attempt to address. Random battling has its own competitive scene which lends itself much better to an AI.

The first peculiarity is the Pokemon themselves; while in a game like chess pieces can only move in a particular pattern (with a few exceptions) such as rooks always moving cardinally and bishops diagonally; in a Pokemon battle each Pokemon can know up to four of a large number of moves. These moves cannot be changed in the middle of battle, but this means a Pikachu in one battle may not have the same moves as a Pikachu in another battle.

An interesting approach could be to model an enemy's potential moves as a distribution of moves this pokemon has used in training with moves the enemy uses then being "confirmed" for that battle. However, this approach has two major caveats:

- It is heavily reliant on training data representative of real-world application, which has its own caveats when considering an off-policy approach to learning.

- In a competitve setting, most Pokemon aren't present on the field long enough to use all of their moves, so it's most likely that a distribution of an enemy Pokemon's moves will be the main reliance. Therefore, one can simply ignore enemy move choice altogether since the move the enemy will use is for all intents and purposes stochastic, and thus can be taken into account just by that Pokemon being on the field.

Another peculiarity of Pokemon battles is the order in which "turns" are executed; both players choose moves simultaneously, with the turn order then being chosen by a complex series of "priorities". Generally speaking, however, switching Pokemon will occur before moves are executed, and the pokemon with the higher speed stat will move first. If a Pokemon is knocked out between the time a command is issued and it's able to use the move, the knocked out Pokemon will not use its move and the turn is essentially wasted.

We can avoid mis-attributions of no reward to a move because it didn't get executed for some reason by only counting a move as "used" when the Pokemon actually gets to use the move, rather than issuing the command to use a particular move.

A third peculiarity is that moves don't always have a clear nor immediate effect. For example, the move "Mirror Move" will use the move the enemy previously used.

3 Related Work

Citations can be made using either \citep{} or \citet{}, depending from the appropriateness. To avoid the citation moving to the next line, it is often a good practice to replace the space before with a tilde (˜) character. Example 1: "CoRL is the best conference ever [2]." Example 2: "Gauss and Davis [2] proved, both theoretically and numerically, that CoRL is the best conference ever."

4 Agents

Two agents are used in this paper, with each using slightly different information for the state:

- SingleAgent uses the enemy pokemon and the move being used as the state.

- DoubleAgent uses the player's pokemon, the enemy pokemon, and the move being used as the state.

The inclusion of the player's active pokemon in the state has a few notable implications. Firstly, the "Same Type Attack Bonus" or STAB, which increases the power of a move by 50% if the user is the same type as the move, would be apparent based on the state. This would not be captured in SingleAgent. Furthermore, DoubleAgent's value estimates would be more finely tuned to the situation at hand, whereas SingleAgent's would be a composite of any pokemon with a particular move against a given enemy pokemon.

The drawback to including the player's pokemon in the state space is, of course, that the state space grows much larger than before. Specifically, it grows by a factor of n, where n is the number of possible pokemon.

The goal in setting up the states in this way between the two agents is to measure the tradeoff between expanding the state space and its effect on coverage, performance, and training time:

- coverage is the percentage of states that have been visited at least once before. If a state has not been seen yet, the model will default to giving that state a value of zero.

- performace is the agent's ability to win battles, which is distinct from the reward function, which is described below.

- training time is how long it takes for the agent to train. Comparing the two agents with the same number of iterations of training may not be fair because DoubleAgent may take longer to train up to that number of iterations, so we can instead cut off training time at a particular threshold to make performance estimates more fair.

The reward function is simply the difference between the enemy's HP before and after the turn ends. Unfortunately, this will be unable to reliably attribute effects that don't deal damage, such as status effects or stat boosts; as well as any moves with delayed or damage over time effects, such as damage from the poison or burn status conditions. Such an addition may be a feature of future work.

Rewards are taken at the end of a battle, so learning will occur all at once between iterations. Learning will take place with gamma = 0; in other words, only the move directly preceding the reward is taken into account. This gamma is chosen because, with the exception of the aforementioned status effects and stat changes carrying over between turns, which we cannot reliably capture; everything that occurs in a given turn is the result of an action taken (i.e. a move used) in that turn. Therefore, using a gamma greater than 0 would only mis-attribute the cause of rewards.

5 Experiments

To train both agents, 1,000 iterations of training are used. In each iteration, teams are randomly generated, followed by five battles with the teams to help remove variability in the reward estimates (e.g. moves can miss or critically hit, multiplying the reward by a factor of 0 or up to 2 respectively).

The Agents are trained in an off-policy manner; namely, the "Reinforcement Learning" agent will always pick random moves when it is training to ensure 100% coverage given an infinite number of iterations, regardless of the enemy it's training against.

6 Experimental Results

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis.

Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa. Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

7 Conclusion

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris. Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tin- cidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa. Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

References

[1] C. F. Gauss and C. H. Davis. Theory of the motion of the heavenly bodies moving about the sun in conic sections. *Gauss's Theoria Motus*, 76(1):5–23, 1857.

[2] C. F. Gauss and C. H. Davis. Theory of the motion of the heavenly bodies moving about the sun in conic sections. *Gauss's Theoria Motus*, 76(1):5–23, 1857.