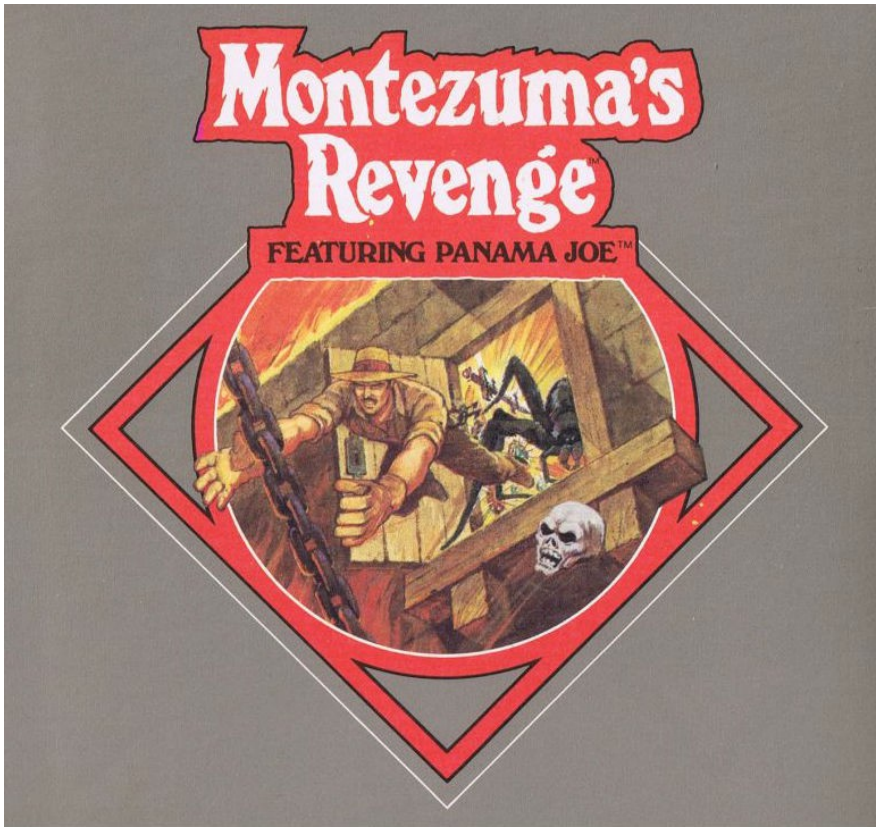
Arthur Juliani Follow
Deep Learning @Unity3D
Jul 13 · 10 min read

On "solving" Montezuma's Revenge

Looking beyond the hype of recent Deep RL successes



In recent weeks DeepMind and OpenAI have each shared that they developed agents which can learn to complete the first level of the Atari 2600 game Montezuma's Revenge. These claims are important because Montezuma's Revenge is important. Unlike the vast majority of the games in the Arcade Learning Environment (ALE), which are now easily solved at superhuman level by learned agents, Montezuma's Revenge has been hitherto unsolved by Deep Reinforcement Learning methods and was thought by some to be unsolvable for years to come.

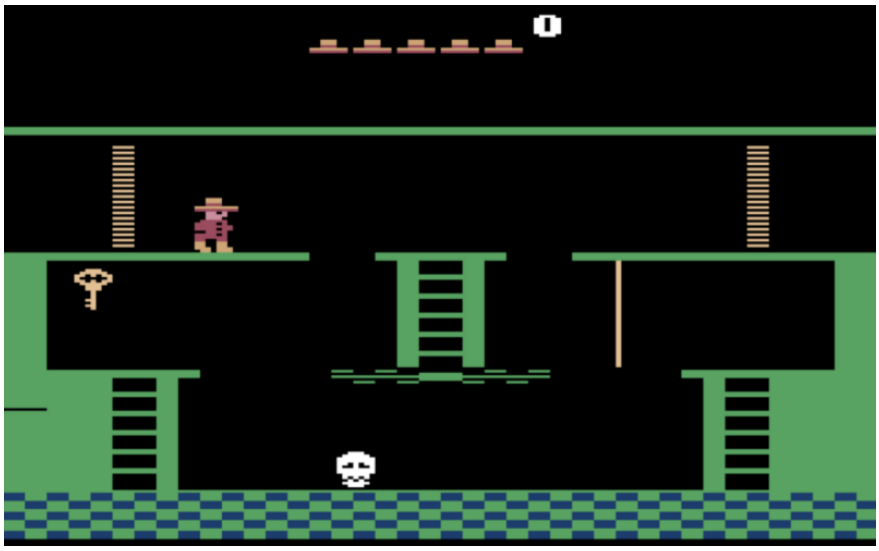


Figure 1. The first room of Montezuma's Revenge.

What distinguishes Montezuma's Revenge from other games in the ALE is its relatively sparse rewards. For those unfamiliar, that means that the agent only receives reward signals after completing specific series of actions over extended periods of time. In the case of the first room of Montezuma's Revenge (see Figure 1 above), this means descending a ladder, jumping across an open space using a rope, descending another ladder, jumping over a moving enemy, and then finally climbing another ladder. All of this is just to get the very first key in the very first room! In the first level there are 23 more such rooms for the agent to navigate through in order to complete the level (see Figure 2 below). To complicate this further, the failure conditions in the game are fairly strict, with the agent's death happening due to any number of possible events, the most punishing of which is simply falling from too high a place. I encourage those who are unfamiliar with the game to try playing it and see how long it takes you to even get past the first room, let alone the first level. You can find an online version of the game here: https://www.retrogames.cz/play_124-Atari2600.php?language=EN

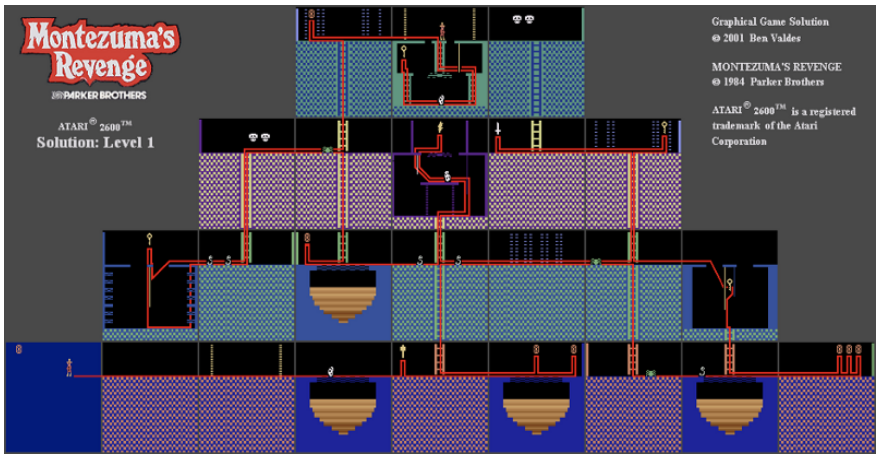


Figure 2. Solution to first level of Montezuma's Revenge.

Due to this notorious difficulty, the game has been seen as a kind of grand-challenge for Deep RL methods. In fact, the game has inspired the development of some of the more interesting approaches to augmenting or reworking the traditional Deep RL algorithm, with approaches utilizing novel methods for hierarchical control, exploration, and experience replay. So, it was big news (at least in certain circles) when DeepMind and OpenAI each claimed to have developed algorithms capable of playing the game so well. To give you a sense of how much better, consider that the previous state of the art in the game was 2600 points, with these new methods achieving scores in the tens of thousands of points. All three proposed methods are impressive efforts from both an engineering and theoretical perspective, all with things to learn from. Unfortunately, the claims of solving Montezuma's Revenge with Deep Reinforcement Learning are not quite what they seem. In all three cases (two papers by DeepMind, and one blog post by OpenAI), the use of expert human demonstrations was an integral part of the algorithm, changing fundamentally the nature of the learning problem.

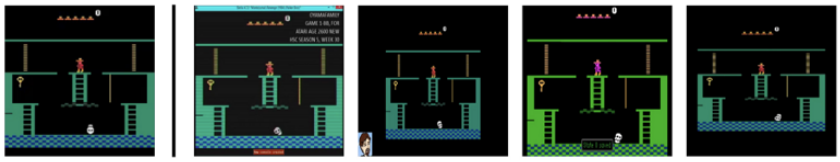
In this post, I want to discuss what these methods do in order to solve the first level of Montezuma's Revenge, and why in the context of the game, and long-term goals for Deep RL, this approach isn't as interesting or meaningful as it might seem. Finally, I will briefly discuss what I would see as truly impressive results on the notorious game, one which would point the way forward for the field.

DeepMind's Results

Learning from YouTube

Sporting the eye-catching title " Playing hard exploration games by watching YouTube," DeepMind offers the most interesting of the three

approaches to solving Montezuma's Revenge. As the title suggests, the research group devised a method by which videos of expert players completing the first level of the game can be used to aid the learning process. The problem of learning from videos is in itself an interesting challenge, completely outside of the additional challenges posed by the game in question. As the authors point out, videos found on YouTube contain a various arrangement of artifacts which can prevent the easy mapping between what is happening in the video, and what an agent playing in the ALE might observe. In order to get around this "gap," they create a method which is able to embed observations of the game state (visual and auditory) into a common embedding space.



(a) ALE frame (b) Frames from different YouTube videos

Figure 3. Comparison of different demonstrations videos to emulator image.

This embedding space is then utilized to provide a kind of bread-crumb reward to the learning agent as it progresses. Instead of only receiving the sparse rewards provided by the original game, the agent has access to intermediate rewards which correspond to reaching checkpoints along a path which are provided by expert players. In this way the agent has access to a much stronger learning signal and is able to ultimately complete the first level of the game with a score of 41,000.

Q-Learning from Demonstrations

Around the same time that the YouTube paper was released, DeepMind shared results on another set of experiments with a somewhat less conspicuous title: " Observe and Look Further: Achieving Consistent Performance on Atari." In the paper they propose a set of interesting algorithmic improvements to Deep Q-learning which are able to increase the stability and capability of the algorithm. The first of these is a method for increasing the discount factor in the Q-update, such that longer-term temporal dependencies can be learned, without the typical downsides of higher discount factors. The second is a means of enabling Deep Q-learning to account for reward signals at varying orders of magnitude, and thus enable their algorithm to solve tasks in which an optimal policy involves learning about these varying rewards.

Alongside these two improvements however, they also propose the use of human demonstrations as a means of augmenting the exploration process, by automatically providing information to the network about trajectories through state space that an expert player would follow. With all of these three improvements combined, the authors arrive at an agent which is able to learn to complete the first level of Montezuma's Revenge with a score of 38,000. Noticeably however, the first two improvements without the expert demonstrations are not enough to lead to compelling performance on the game, scoring only in the 2,000 point range.

OpenAI's Results

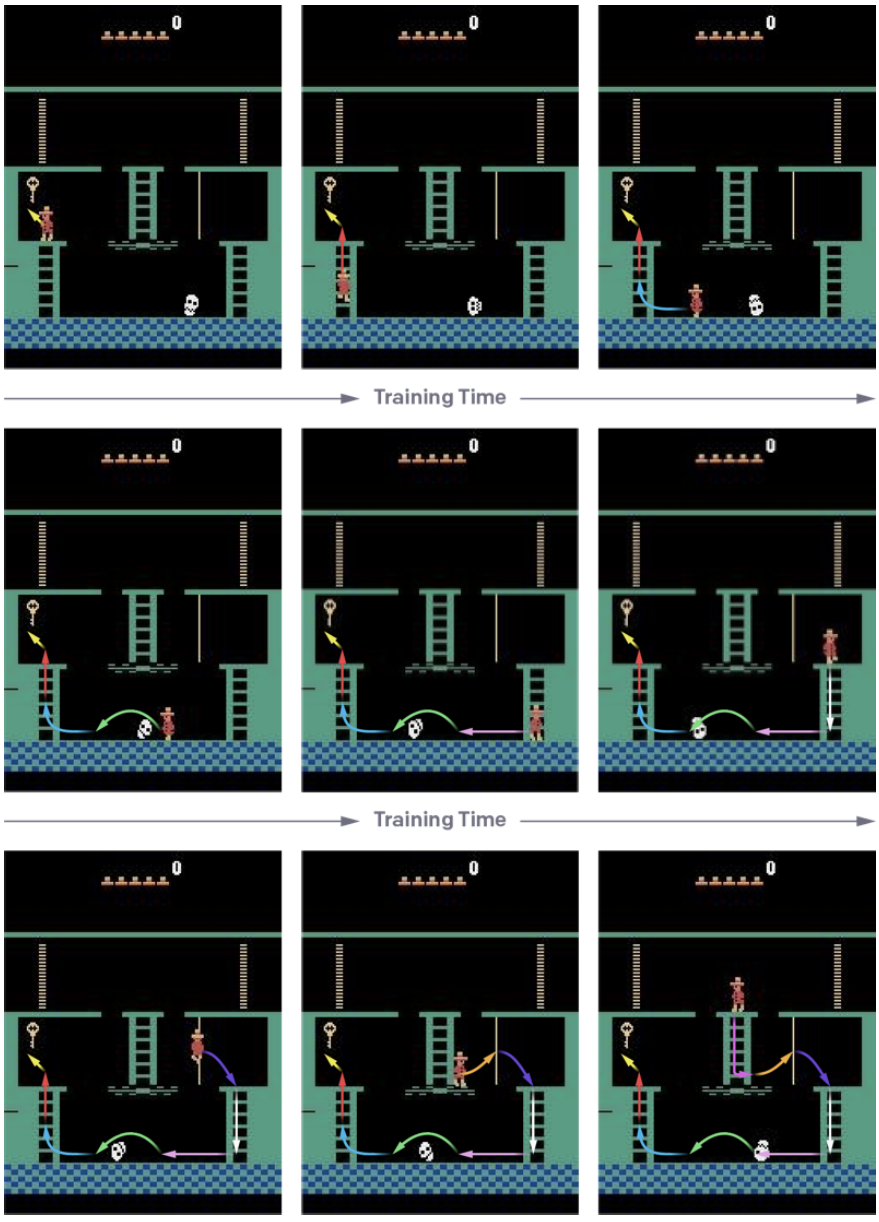


Figure 4. Restarts used during training over time.

A few weeks after the DeepMind results, OpenAI shared a blog post in which they describe a method by which they also are able to train an agent to complete the first level of Montezuma's Revenge. This one also relies on human demonstrations but utilized them in a slightly different way than the DeepMind approach. Rather than using demonstrations as part of the reward or learning signal, demonstrations are used as a means of intelligently restarting the agent. Given an expert human trajectory through the game, the agent is started near the end of the game, and then slowly moved backwards through the trajectory on every restart as learning progresses. This has the effect of exposing the agent to only parts of the game which a human player has navigated through, and only widening the scope as the agent itself becomes more competent. With this method there is no change to the actual learning algorithm, as a default Proximal Policy Optimization (PPO) is used. Simply starting the agent in the "right" place is enough to ensure it stumbles onto the correct solution, earning an impressive score of 74,500.

Limitations of Imitation

The one thing all of the approaches described above have in common is that they utilize a set of expert human demonstrations. The first approach utilized the demonstrations in order to learn a reward signal, the second utilized them in order to learn more accurate Q-values, and the third utilized them to more intelligently restart the agent. In all three cases the demonstrations were critical to the learning process. In general the use of demonstrations is a compelling way to provide agents with meaningful knowledge about a task. Indeed, it is how many humans learn countless tasks. The key though to the humanness of our ability to learn from demonstrations is our ability to abstract and generalize a single demonstration to new situations. In the case of Montezuma's Revenge, rather than developing a general-purpose solution to game playing (as the two DeepMind papers titles suggest), what has really been developed is an intelligent method for exploiting the game's key weakness as an experimental platform: its determinism.

Every time a human or agent plays Montezuma's Revenge, they are presented with the exact same set of rooms, each containing the exact same set of obstacles and puzzles. As such, the simple memorization of the movements through each room is enough to lead to a high-score, and the ability to complete the level. While this wouldn't necessarily be a meaningful flaw if the agents were forced to learn from scratch, it becomes one when expert demonstrations enter into the situation. All three solutions exploit the deterministic nature of the game to allow the agent to more easily learn the solution path through the game.

What is ultimately learned is not how to play difficult platformers, but how to execute a pre-determined set of actions in order to complete a specific game.

The OpenAI blog post briefly mentions the issue of determinism but does so at the level of the Atari emulator itself, rather than the specific game. Their solution is to use a randomized frame-skip to prevent the agent from memorizing the trajectory. While this prevents the agent from literally memorizing a sequence of actions, it does not prevent the memorization of the general trajectory through state space.

In all cases Montezuma's Revenge no longer serves its original purpose of being a hard problem of sparse reward problem solving, and rather becomes an easier problem of learning a trajectory through a fixed state space. This is a shame, because in its original formulation the game still has the potential to provide one of the more compelling challenges to Deep Reinforcement Learning researchers.

Solving Montezuma's Revenge, the hard way

I have personally kept an eye on Montezuma's Revenge results for a few years now because I have seen them as a litmus test for the ability of Deep Reinforcement Learning agents to begin to show signs of more general reasoning and learning. Many results have shown that given enough computational ability Deep Reinforcement Learning, or even random search are able to solve naïve optimization problems. The human-level intelligence so many researchers are interested in however does not involve simple optimization. It involves learning and reasoning about concepts over multiple levels of abstraction. It involves then generalizing that learned conceptual knowledge from one problem space to many in an adaptable way.

When you present any person the first room of Montezuma's Revenge, and ask them what they need to do, they will quickly begin to describe to you a series of actions and observations which suggest a complex understanding of the likely dynamics of the game. The most obvious manifestation of this will be the recognition of the key as a desirable object, the skull as something to be avoided, and the ladders as having the affordance of movement. Keys then suggest the ability to open locked doors, and suddenly complex multi-step plans begin to emerge as to how to complete the level. This reasoning and planning doesn't just work on one fixed level of the game, but works on any possible similar level or game we are presented with. It is these kinds of skills which are essential to human intelligence, and are of interest to those trying to push Deep Reinforcement Learning beyond the realm of a set

of a simple optimization algorithms. However, utilizing human demonstrations in a deterministic environment completely bypasses the need for these exact skills.

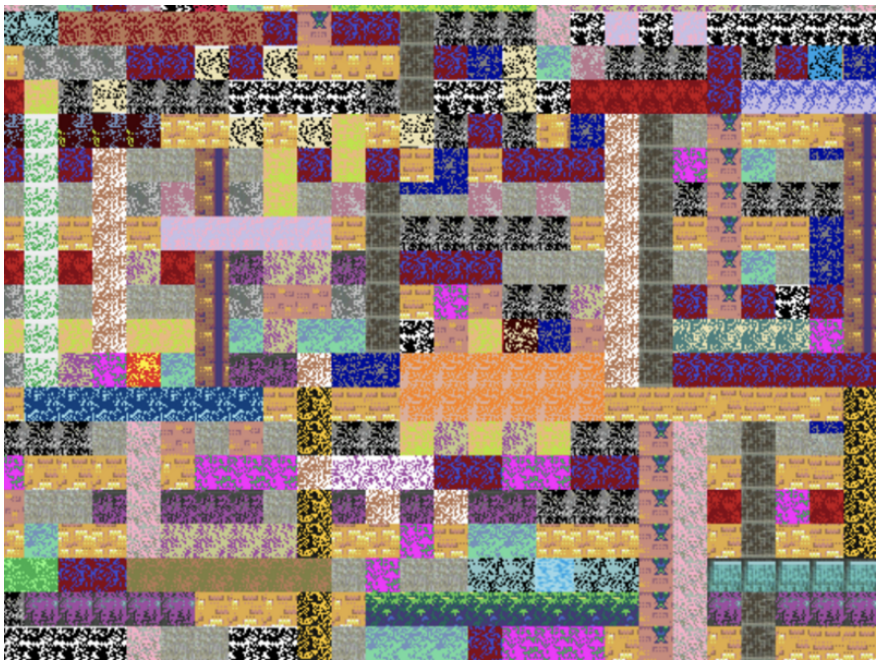


Figure 5. An example of what a game like Montezuma's Revenge might look like to us without the priors we typically rely on to interpret visual scenes.

Of course, these skills are also the ones that are most difficult to put into algorithmic form, especially when they are still not fully understood in their human manifestations. Especially in the case of conceptual learning, outside general knowledge often needs to be brought to bear on a new problem. As a group at Berkeley have pointed out, without our human priors (either biological or learned throughout life), many of the video games which we take for granted as being simple turn out to be much more complex. To demonstrate this, they even have interactive browser game which you can play that simulates what a randomly initialized pixel-based agent might "experience."

The problem then becomes how can agents naturalistically learn the priors required to make sense of a game like Montezuma's Revenge. Furthermore, how can these learned priors be used to learn to play not just one fixed level of the game, but any level of any similar game. There is interesting work being done in the areas of representation learning, and conceptual grounding which I think will be essential to tackling these kinds of problems. There is also work being done to develop more stochastic environments which better test the generalization ability of agents, most compelling among these

approaches being the GVGAI competition. This research direction is still in its early stages, but shows a lot of promise.

I eagerly look forward to the day we can say without a doubt that an agent can learn to play Montezuma's Revenge from scratch. When that day comes there will be a lot to be excited about.

. . .

Feel free to reply in the comments with your thoughts and opinions. What are presented here are just my personal thoughts on the topic, and would love to hear from others, especially if you work in Deep RL and have experience with Montezuma's Revenge.