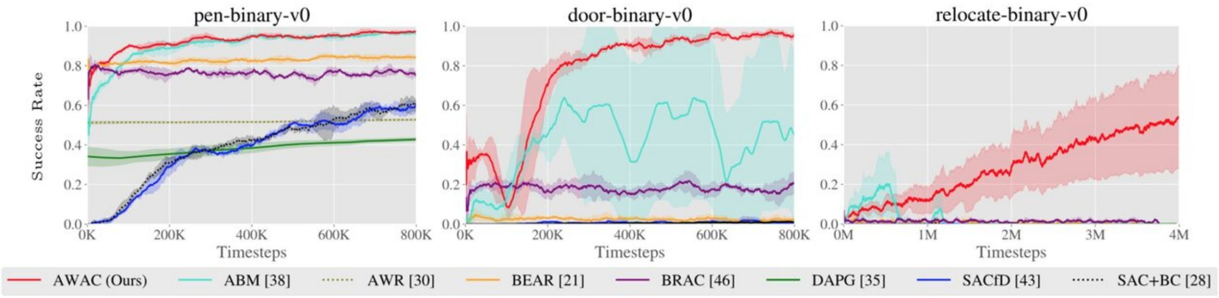
TechXplore

An algorithm that merges online and offline reinforcement learning

July 16 2020, by Ingrid Fadelli



Learning curve for the dexterous manipulation tasks. Credit: Nair et al.

In recent years, a growing number of researchers have been developing artificial neural network (ANN)- based models that can be trained using a technique known as reinforcement learning (RL). RL entails training artificial agents to solve a variety of tasks by giving them "rewards" when they perform well, for instance, when they classify an image correctly.

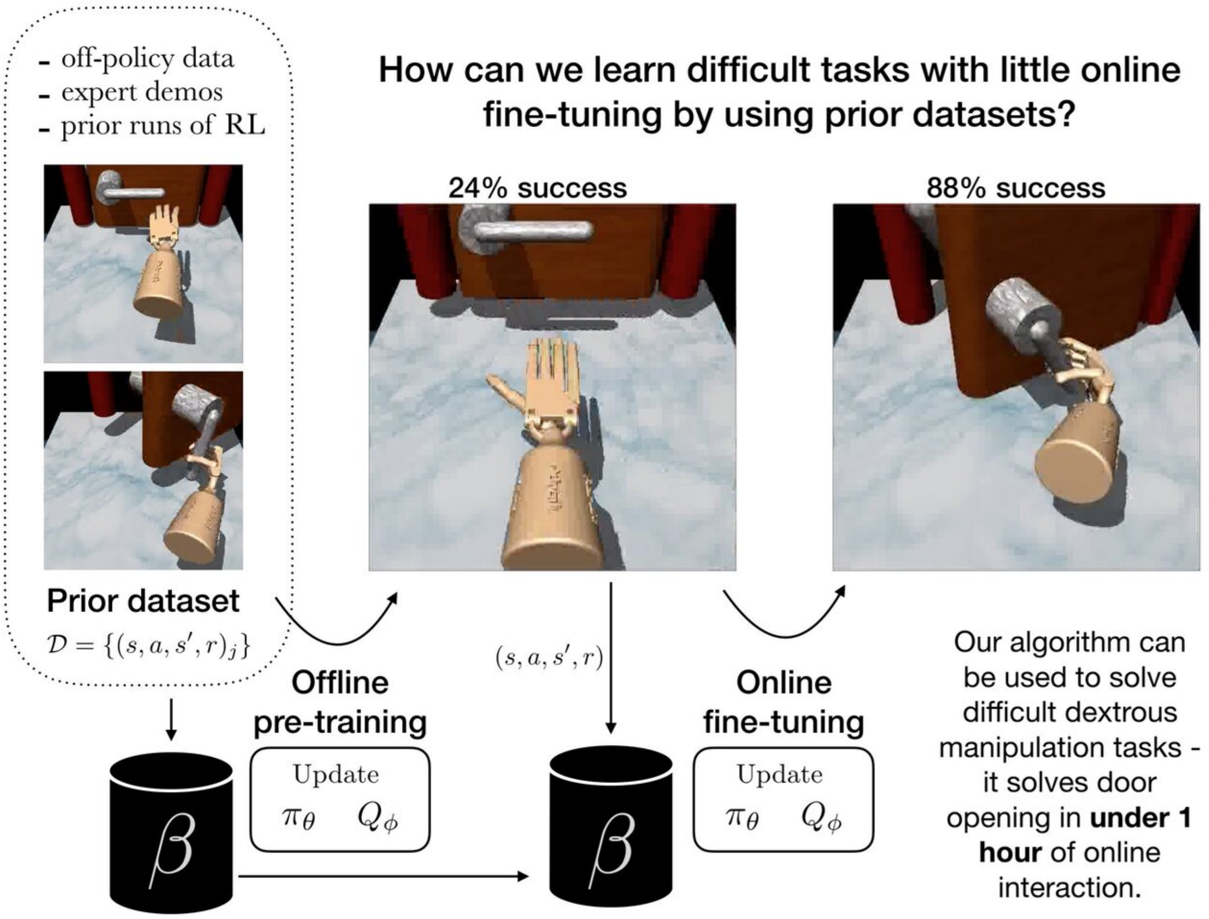
So far, most ANN-based models were trained employing online RL methods, where an agent that was never exposed to the task it is designed to complete learns by interacting with an online virtual environment. However, this approach can be quite expensive, time-consuming and inefficient.

More recently, some studies explored the possibility of training models offline. In this case, an artificial agent learns to complete a given task by analyzing a fixed dataset, and thus does not actively interact with a virtual environment. While offline RL methods have achieved promising results on some tasks, they do not allow agents to learn actively in real time.

Researchers at UC Berkeley recently introduced a new algorithm that is trained using both online and offline RL approaches. This algorithm, presented in a paper pre-published on arXiv, is initially trained on a large amount of offline data, yet it also completes a series of online training trials.

"Our work focuses on a scenario that that lies between two cases that we face constantly in real-world robotics settings," Ashvin Nair, one of the researchers who carried out the study, told TechXplore. "Often, when trying to solve robotics problems, researchers have some prior data (for instance, a few expert demonstrations of how to solve the task or some data from the last experiment you performed) and want to leverage the prior data to solve the task partially, but then be able to fine-tune the solution to master it with a small number of interactions."

While reviewing past RL literature, Nair and his colleagues realized that previously developed models did not perform well when they were first trained offline and then fine-tuned online. This was typically because they learned too slowly or did not make the best use of offline datasets during training.


- off-policy data
- expert demos
- prior runs of RL
Prior dataset
$\mathcal{D} = \{(s, a, s', r)_j\}$
How can we learn difficult tasks with little online fine-tuning by using prior datasets?
24% success
88% success
Offline pre-training
Update π_θ Q_ϕ
(s, a, s', r)
Online fine-tuning
Update π_θ Q_ϕ
β
β
Our algorithm can be used to solve difficult dextrous manipulation tasks - it solves door opening in **under 1 hour** of online interaction.

Schematic of offline learning + online fine-tuning tasks. Credit: Nair et al.

In their study, the researchers studied the limitations of existing models in depth and then devised an algorithm that could overcome these issues. The algorithm they created can achieve satisfactory performance when pre-trained on large quantities of data offline. This allows it to quickly master the task it is designed to complete at a later stage, when it is actively trained in a virtual online environment.

"Our paper addresses a common problem that was stalling our progress: that we were always making robots learn tasks from scratch rather than being able to use existing datasets for RL," Nair explained. "It actually came about as a result of realizing that our experiment cycles for a separate idea was taking too long and too much effort to evaluate running on a robot in the real world, and we needed a way to evaluate the idea by pre-training on data we already had and doing only a small amount of extra real-world interaction."

Nair and his colleagues identified three key limitations of previously developed models trained via RL. First, they observed that on-policy techniques such as advantage weighted regression (AWR) and demonstration augmented policy gradient (DAPG), which are often used to fine-tune models online, typically learn quite slowly compared to off-policy methods.

In addition, the researchers observed that off-policy methods, such as soft actor critic (SAC) approaches, often did not improve much when trained on offline datasets. Finally, they found that techniques to train models offline, such as bootstrap error accumulation reduction (BEAR), behavior regularized actor critic (BRAC) and advantage behavior models (ABM) typically worked well in the offline pre-training stage, but their performance did not improve much when they were trained online. This is primarily because they rely on behavior models, which work well when trying to outline the general distribution of data and learning policies accordingly, but not as well when fine-tuning models in online environments.

"Confronted with these challenges, we developed advantage weighted actor critic (AWAC), which is an off-policy actor-critic algorithm that does not rely on a behavior model to stay close to the data distribution," Nair said. "Instead, we show that we can derive an algorithm that implicitly stays close to the data by sampling."

The AWAC algorithm developed by Nair and his colleagues can be pre-trained just as well offline as techniques that are specifically designed for offline training. However, its performance improves further and by a significant margin when it is trained online.

The researchers evaluated their algorithm's performance on different dexterous manipulation tasks characterized by three key aspects, namely complex discontinuous contacts, very sparse binary rewards and the control of 30 joints. More specifically, their algorithm was trained to control a robot's movements, allowing it to rotate a pen in its hand, open a door or pick up a ball and move it to a desired location. For each task, Nair and his colleagues trained the algorithm on an offline dataset containing 25 human demonstrations and 500 trajectories of off-policy data, attained using a technique known as behavioral cloning.

"The first task, pen rotation, is relatively simpler and many methods eventually solve the task, but AWAC is the fastest," Nair said. "Only AWAC solves the second and third task. Prior methods fail, for a myriad of reasons, centered around their inability to obtain a reasonable initial policy to collect good exploration data, or their inability to learn online from interaction data."

Nair and his colleagues compared their AWAC algorithm to eight other methods trained via offline or online RL and found that it was the only one that could consistently solve the difficult manipulation tasks they tested it on. Their algorithm could also solve simpler MuJoCo benchmark tasks and a pushing task faster than previously developed methods, learning from suboptimal, randomly generated data.

In the future, the algorithm could enable the use of RL to train models on a far wider range of tasks. Other research teams could also draw inspiration from their work and devise similar RL approaches that combine offline and online training.

"Going forward, we plan to use AWAC to speed up experiments and stabilize training on new tasks by taking advantage of existing data," Nair said. "The direction we are really excited about is to scale up the amount of data that we use for RL so that can start seeing significant generalization across tasks and visual and physical characteristics of objects."

More information: Nair et al., Accelerating online reinforcement learning with offline datasets. arXiv:2006.09359 [cs.LG]. arxiv.org/abs/2006.09359

Project software repository: awacrl.github.io/

Citation: An algorithm that merges online and offline reinforcement learning (2020, July 16) retrieved 13 January 2026 from https://techxplore.com/news/2020-07-algorithm-merges-online-offline.html