Dreaming Pac-Man: a reinforcement learning method with replays

Abstract

Sufficient sleep time is proved to be crucial for the formation of learning. Dreaming, in particular, is sometimes interpreted as a way to "replay" events to solidify memories and learning [1]. On a machine learning perspective, the mechanism of dreaming could be roughly equate with re-training with similar set of data with bootstrapping [2]. We propose to simulate the dreaming process in training a Pac-Man game with a reinforcement learning model. This process is done by first training the model on some given environments (awake), and re-training it on environments obtained by bootstrapping the previous seen environments with noise, as an analogy to replaying events in dreams. Our hypothesis is that Pac-Man will perform better after re-training, comparing to training without replaying.

In addition, we will also test the fear extinction [3] hypothesis, that exposure to traumatic experiences in dream help to disassociate feared conditioned stimulus and irrational fear. This experiment is carried out by first conditioning Pac-Man to a harmless cue in the first training phase (awake), and to attempt to disassociate it in the second training phase (dreaming).

Finally, we also explore a Human brain-like sleep and replay method for continual learning in deep neural networks.

Method

A Reinforcement Learning Model with Pac-Man

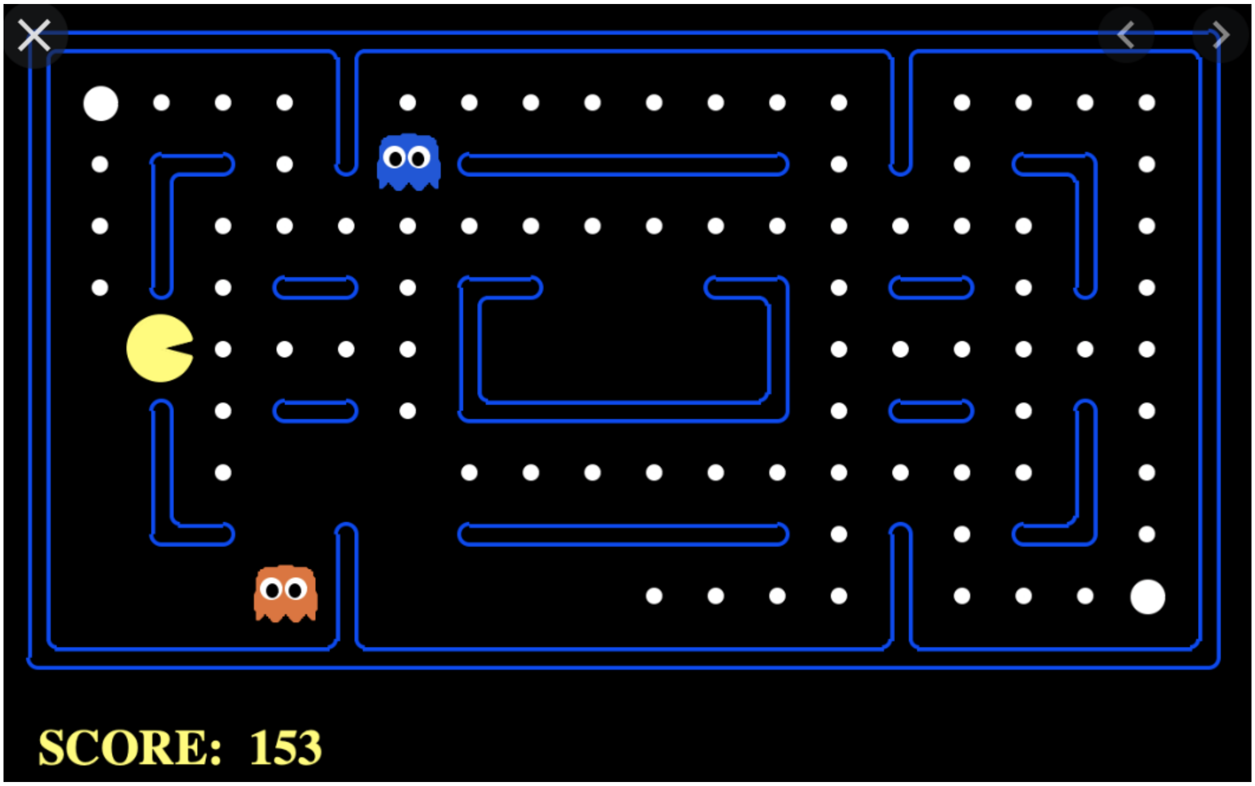
Environment

We will use the open-sourced Berkeley Pac-Man project as the basic architecture of the game [3]. The environment of the game is a 2-D grid world with walls. States $s \in \{(i,j) : i = 0, 1, \ldots, width, j = 0, 1, \ldots, height\}$ are square-shape area that occupy the paths. A *start point* and an *end point* are on two opposite of the map. There are white *dots* scattered along the paths. The end point and each white dots is associated with a reward r.

Agent

There are two kinds of competing agents, *Pac-Man* and *ghosts*. The agents can make *action* $a \in \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$. A move is possible when the resulting new state s' of the state-action pair

(s, a) is not a wall. Pac-Man knows the location of all the dots and the end point. Pac-Man and each ghost know the distance between them: $d = (i_p - i_g)^2 + (j_p - j_g)^2 + 2^{30}\mathbf{1}_{\{(i_p-i_g)^2+(j_p-j_g)^2>threshold\}}$, where (i_p, j_p) and (i_g, j_g) are the states Pac-Man and a ghost currently occupies. When their bee-line distance is far enough, this value is set to infinity.



Policy Making - approximate Q-learning

For faster learning and better generality, we will use an approximate Q-learning method. Q-value is the expected reward for the rest of the game when an agent take an action a when it is on a state s. We will define $Q(s, a) = w_1 f_1(s, a) + \ldots + w_n f_n(s, a)$ where f_i describe hand-picked features that would influence the decision-making process, such as Pac-Man's distance to the closest ghost, distance to the closest dot, is a tunnel, etc. The goal of training is to find appropriate weights w_i:

$$difference = [r + \gamma max_{a'} Q(s', a')] - Q(s, a)$$
$$w_i = w_i + \alpha \times difference \times f_i(s, a)$$

where α is the learning rate and γ is a discount factor [4].

Training Phase 1 - "Awake"

Train the model on grid worlds TRAIN_MAPS_AWAKE specified by human programmers. Training ends when the weights converge, that is, when $\sum_i w_i - w'_i < \epsilon$.

Testing Phase 1

We will test the model on a set of grid worlds different from the training set called TEST_MAPS specified by human programmers and collect the scores for later comparisons.

Training Phase 2 - "Dream"

We will create a new set of grid worlds NOISE_MAPS. The grid worlds for this new training phase is generated by bootstrapping on TRAIN_MAPS_AWAKE and NOISE_MAPS. The purpose for bootstrapping is to simulate a model of dreaming where we replay memories accompany by noise. Bootstrapping can be done by randomly select and manually mutate part of a map by replacing it with a part from another map. A more automatic method is likely and will be explored.

Testing Phase 2

We will first train a Pac-Man on some specified maps, and re-train it with maps that are generated by bootstrapping from the previous scenarios. If the hypothesis were true, we should see a noticeable increase in success rate as the "dreaming" time increase

Question: since the weights already converge, would the second training phase change/improve anything? Should we only do training phase 1 for a fixed number of epochs and stop before converging, and then do two separate trails: in trail one we use replay to train for a certain number of epochs, in trail two (controlled) we train with new maps for the same number of epochs.

Extension: dreaming as fear extinction therapy?

"Fear extinction is defined as a decline in conditioned fear responses following nonreinforced exposure to a feared conditioned stimulus." [4] We propose that "bad dreams", dreams of traumatic memories function similarly as fear extinction therapy. In the dream, the brain replays cues that the agent is conditioned to associate with trauma without letting the agent experience the real consequences. For example, a person who has been bitten by a rattlesnake might become scared of its the sound of running water because its resemblance with the rattling sound. In this sense, she is conditioned to associate the sound with being bitten and hurt. She might frequently report dreaming of the sound of running water. In our hypothesis, her brain expose her to the "feared conditioned stimulus" by replaying the traumatic memories in an attempt to disassociate the cue to actual danger.

A challenge with fear extinction is that we only want to erase irrational fear while maintaining rational avoidant response to actual danger. In the previous example, we would want her to accept the sound of water because more often than not it is harmless. We don't want her to become fearless to rattlesnake because fear is a rational response critical for survival. In order to enable this difference, we will make some modifications based on the set-up in the previous model.

Benign and Evil Ghosts

In addition to Pac-Man and ghosts, we will introduce a third kind of agent, benign ghosts. These benign ghosts act similarly as the normal ghosts (for clarity, we will call them evil ghosts for the rest of the proposal), but they do not cause harm when they occupy the same state as Pac-Man. Pac-Man can discern whether a ghost is benign or evil when they are closed by (distance < $\epsilon_{threshold}$), but not when they are far apart. We will describe how the two kinds of ghosts will be perceived differently by Pac-Man in the feared conditioned stimulus section.

Feared Conditioned Stimulus

We will consider a new feature $f_{fear}(s,a) = 1/((i_p - i_s)^2 + (j_p - j_s)^2)$ which is negatively proportional to the distance between Pac-Man and a source of a feared conditioned stimulus. This signal can be analogically understood as a sound that any ghost makes. Note that this feature looks similar to the feature that categorizes the distance between Pac-Man and a ghost, which was defined in the previous model: $f_d(s,a) = \frac{1}{d} = 1/((i_p - i_g)^2 + (j_p - j_g)^2 + 2^{30}\mathbf{1}_{\{(i_p-i_g)^2+(j_p-j_g)^2>\epsilon_{threshold}\}})$. It will be helpful to think of the f_{fear} as negatively proportional to the sound that both kinds of ghosts can make.

And to think of f_d as categorizing seeing an evil ghost — the closest the ghost is, the higher is f_d. When the ghost is far enough, Pac-Man would lose sight of them and hence f_d goes to nearly zero. We do not need a separate feature that describes the distance between Pac-Man and a benign ghost since within $\epsilon_{threshold}$, Pac-Man can tell it is harmless, and beyond $\epsilon_{threshold}$, the ghost "goes out of sight" and f_d becomes nearly zero.

Training Phase 1 - conditioning harm with stimulus

This will be the same as training phase 1 in the previous example except that we will include f_{fear}. Namely, Pac-Man is trained on TRAIN_MAPS_AWAKE with only evil ghosts who make noises. Pac-Man is expected to "grow fear" to this noise and act to decrease f_{fear}.

Testing Phase 1

We will use the exact same setting as the first testing phase in the previous model except that we will turn half of the ghosts into benign ghosts. Namely we will test Pac-Man on TEST_MAPS with half of the ghosts set to harmless.

Training Phase 2 - fear extinction in dream by exposing to stimulus

We will use the same set up as the second training phase in the previous model, with the exception that all ghosts are set to benign. This is to model that in a bad dream, a brain replays traumatic memories to expose one to the feared conditioned stimulus (noise a ghost makes) without causing actual consequences.

Testing Phase 2

We will test Pac-Man on TEST_MAPS again with half of the ghosts benign and half of them evil. Pac-Man's improvement is expected to improve because it learns that the noise might belong to benign, harmless ghosts. This would demonstrate a disassociation of a cue and certain fear.

A Neural Network Alternative

Sleep and Replays play a critical role in a Human's ability to learn/remember concepts and tasks sequentially without a degradation of performance in previous tasks. On the other hand, a deep neural network fails to learn in continual learning settings, which is known as catastrophic forgetting. We are planning to explore the Human brain like sleep and replay method for continual learning in deep neural networks. We will start with the implementation of methods suggested by [6] and [7]. If time permits, we will try to explore a different reply technique to improve the performance.

Reference

[1] Human Replay Spontaneously Reorganizes Experience. Liu, Y; Dolan, RJ; Kurth -Nelson, Z; Behrens, TEJ; (2019) https://www.cell.com/cell/fulltext/S0092-8674(19)30640-3?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0092867419306403%3Fshowall%3Dtrue

[2] Cazé R, Khamassi M, Aubin L, Girard B. Hippocampal replays under the scrutiny of reinforcement learning models. J Neurophysiol. 2018 Dec 1;120(6):2877-2896. doi: 10.1152/jn.00145.2018. Epub 2018 Oct 10. PMID: 30303758. https://pubmed.ncbi.nlm.nih.gov/30303758/

[3] http://ai.berkeley.edu/project_overview.html

[4] Russell, Stuart J. (Stuart Jonathan). Artificial Intelligence : a Modern Approach. Upper Saddle River, N.J. :Prentice Hall, 2010.

[5] Myers KM, Davis M. Mechanisms of fear extinction. Mol Psychiatry. 2007 Feb;12(2):120-50. doi: 10.1038/sj.mp.4001939. Epub 2006 Dec 12. PMID: 17160066. https://pubmed.ncbi.nlm.nih.gov/17160066/

[6] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continuallearning with artificial neural networks.Nature communications, 11(1):1–14, 2020.

[7] Yogesh Balaji, Mehrdad Farajtabar, Dong Yin, Alex Mott,and Ang Li. The effectiveness of memory replay in largescale continual learning.arXiv preprint arXiv:2010.02418,2020.