MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

Chương 1

Giới thiệu

Những năm gần đây, **học tăng cường** (Reinforcement learning) liên tục đạt được những thành tựu quan trọng trong lĩnh vực Trí tuệ nhân tạo (Artificial Intelligence). Những đóng góp nổi bật của phương pháp này bao gồm: tự động điều khiển robot di chuyển, điều khiển mô hình máy bay trực thăng, hệ thống chơi cờ vây... Trong số các thành tựu này, hệ thống chơi cờ vây với khả năng chiến thắng những kỳ thủ hàng đầu thế giới là một cột mốc quan trọng của lĩnh vực Trí tuệ nhân tạo. Dù vậy, học tăng cường không phải là một phương pháp mới được phát triển gần đây. Nền tảng lý thuyết của học tăng cường đã được xây dựng từ những năm 1980.

Được xây dựng nhằm mô phỏng quá trình học của con người, ý tưởng chính của học tăng cường là tìm cách lựa chọn hành động *tối ưu* để nhận được **nhiều nhất giá trị điểm thưởng** (Reward). Giá trị điểm thưởng này có ý nghĩa tương tự cảm nhận của con người về môi trường. Khi một đứa trẻ bắt đầu "học" về thế giới xung quanh của mình, những cảm giác như đau đớn (ứng với điểm thưởng thấp) hay vui sướng (điểm thưởng cao) chính là mục tiêu cần tối ưu của việc học. Một điểm quan trọng của học tăng cường là nó được xây dựng với ít giả định nhất có thể về môi trường xung quanh. Hệ thống sử dụng học tăng cường (Agent) không cần biết cách thức hoạt động của môi trường để hoạt động. Ví dụ như để điều khiển robot tìm được đi trong mê cung, hệ thống không cần biết mê cung được xây dựng thế nào hay kích thước là bao nhiêu. Việc hạn chế tối đa những ràng buộc về dữ liệu đầu vào của bài toán học tăng cường giúp cho phương pháp này có thể áp dụng vào nhiều bài toán thực tế.

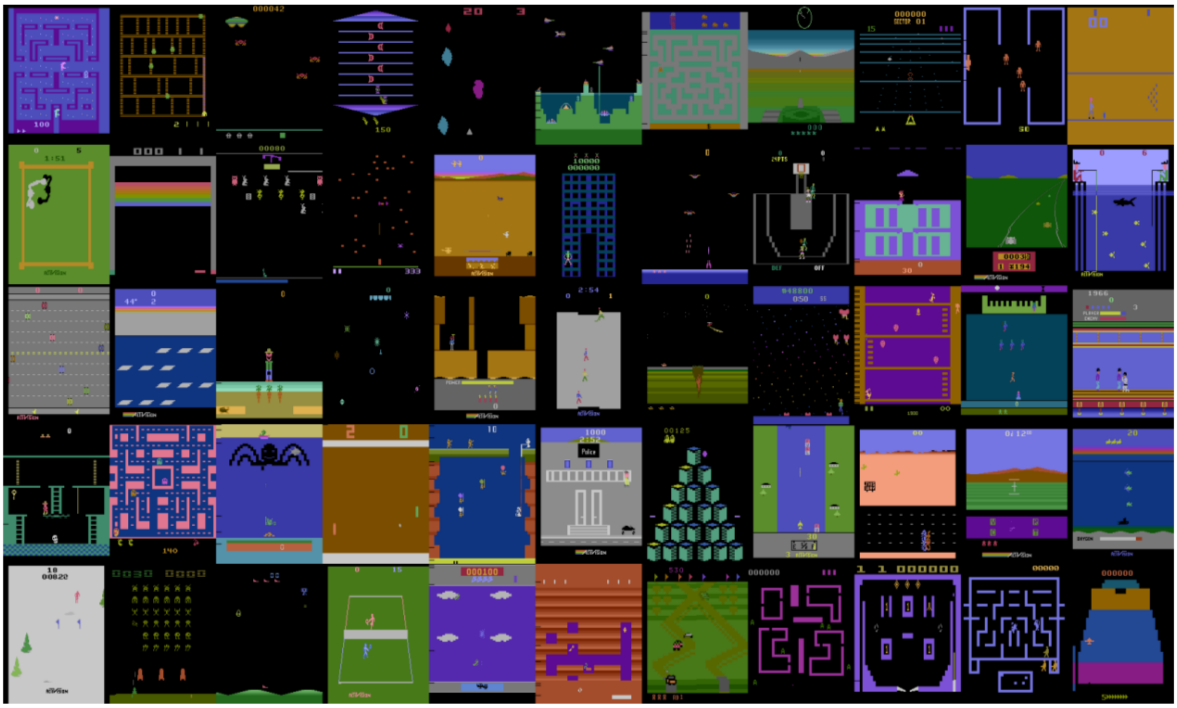
Học tăng cường được xem là một nhánh trong lĩnh vực máy học ngoài hai nhánh: học có giám sát và học không có giám sát. Trong bài toán học có giám sát, dữ liệu

thường được gán nhán thủ công sẵn và việc chủ yếu của hệ thống là làm sao dự đoán chính xác các nhãn đó với dữ liệu mới. Các nhãn này có thể xem như là sự hướng dẫn trong quá trình học; tính đúng sai của việc học lúc này có thể được xác định dựa vào kết quả dự đoán của hệ thống và nhãn đúng của dữ liệu. Tiếp theo đối với những bài toán học không có giám sát, dữ liệu học thường không được gán nhãn nên công việc của việc học là phải tự tìm ra được cấu trúc "ẩn" bên dưới dữ liệu đó. Khác với hai loại bài toán vừa nêu, trong bài toán học tăng cường, hệ thống *không nhận được nhãn thực sự* (tức hành động tối ưu của tình huống hiện tại) mà chỉ nhận được điểm thưởng từ môi trường. Điểm thưởng lúc này chỉ thể hiện mức độ "tốt/xấu" của hành động vừa chọn chứ không nói lên hành động đó có phải là hành động tối ưu hay không. Điểm thưởng này thông thường rất thưa: ta có thể chỉ nhận được điểm thưởng có ý nghĩa (khác không) sau hàng nghìn hành động. Ngoài ra, giá trị điểm thưởng thường là không đơn định và rất nhiễu: cùng một hành động tại cùng một trạng thái, ta có thể nhận được điểm thưởng khác nhau vào hai thời điểm khác nhau. Đây cũng chính là những khó khăn cơ bản của bài toán học tăng cường.

Các trò chơi điện tử thường hay có điểm số mà người chơi cần phải tối ưu hoá. Đặc điểm này trùng với yêu cầu của bài toán học tăng cường, vì vậy các trò chơi này cũng chính là những ứng dụng tự nhiên nhất của phương pháp học tăng cường. Trong luận văn này, chúng em áp dụng phương pháp học tăng cường nhằm xây dựng **hệ thống tự động chơi các game** trên hệ máy Atari. Dữ liệu đầu vào của hệ thống chỉ bao gồm các frame ảnh RGB cùng với điểm số hiện tại. Từ hình ảnh thô này, hệ thống cần tìm cách chơi sao cho điểm số cuối màn chơi (Episode) là lớn nhất có thể. Hệ thống hoàn toàn không biết quy luật của game trước khi bắt đầu quá trình học mà phải tự tìm hiểu quy luật và chiến thuật chơi tối ưu. Lý do luận văn sử dụng game của máy Atari là vì các game này có quy luật chơi tương đối đơn giản nhưng lại rất đa dạng. Mỗi màn chơi thường có độ dài vừa phải (từ 2 - 15 phút) và số hành động có ý nghĩa không quá nhiều (18 hành động). Ngoài ra, các trò chơi này có thể được giả lập trên máy vi tính với tốc độ cao, giúp quá trình học được tăng tốc.

Một số khó khăn trước mắt có thể thấy ở bài toán tự động chơi game bao gồm:

- Hệ thống không được cung cấp luật chơi của game. Chính vì thế nó cũng không thể biết được hành động nào nên làm hoặc không nên làm ứng với từng tình huống cụ thể.



Hình 1.1: Hình ảnh các game trên hệ máy Atari

- Dữ liệu đầu vào là hình ảnh RGB có kích thước 210×160. Để học được một chiến thuật chơi đơn giản thì hệ thống cũng phải chơi "thử và sai" một số lượng lớn màn chơi (có thể lên đến 10000 frame). Vì vậy, lượng dữ liệu đầu vào cần phải xử lý là rất lớn.

- Các game có hình ảnh, nội dung rất khác nhau. Để có thể học cách chơi của nhiều game khác nhau thì thuật toán học phải mang tính tổng quát cao, không sử dụng các tính chất riêng biệt của từng game.

- Để đạt được điểm số cao (ngang hoặc hơn điểm số của con người) thì phải tìm được chiến thuật chơi mang tính lâu dài. Những phương pháp tham lam, lựa chọn hành động để đạt điểm tối đa trong tương lai gần thường không tối ưu.

[TODO: Thêm hướng tiếp cận liên quan + các thực nghiệm + Reference]

Trong những năm gần đây, học sâu đạt đươc nhiều bước đột phá trong nhiều lĩnh vực như Thị giác máy tính (Computer Vision), Nhận diện giọng nói (Speech Recognition), ... Việc kết hợp giữa học sâu và học tăng cường đã dẫn đến một hướng tiếp cận mới cho bài toán tự động chơi game: học tăng cường sâu (Deep reinforcement learning) [1]. Với học sâu, ta có thể học được những đặc trưng cấp cao (high level

features) từ hình ảnh thô mà không cần phải tự thiết kế đặc trưng bằng tay (hand-designed features). Khi kết hợp với học tăng cường, ta có một hình "**End-to-end**": việc học đặc trưng và học chiến thuật chơi được liên kết chặt chẽ với nhau. Trong luận văn này, chúng em thực hiện việc cài đặt lại phương pháp học tăng cường sâu và thử nghiệm mô hình với những tham số khác nhau. Cùng với đó, luận văn thử nghiệm kỹ thuật học chuyển tiếp (Transfer learning) nhằm giảm thời gian huấn luyện cho nhiều game.

Chương 2

Kiến thức nền tảng

Trong chương này sẽ trình bày những kiến thức nền tảng của học tăng cường. Trong phần đầu tiên chúng em sẽ trình bày định nghĩa của các thành phần cơ bản trong học tăng cường. Tiếp đó sẽ đề cập đến mô hình Markov Decision Processes được áp dụng trong việc đánh giá lý thuyết một số thành phần của bài toán học tăng cường. Cùng với đó sẽ trình bày qui trình tổng quát để đánh giá và cải thiện chính sách trong bài toán. Cuối cùng chúng em sẽ trình bày một số phương pháp phổ biến thường đượcAgent áp dụng để đánh giá cũng như cải thiện giúp hệ thông có cách giải tốt hơn cho bài toán trên.

2.1 Các thành phần cơ bản của học tăng cường

2.1.1 Agent và môi trường

Trong học tăng cường, đối tượng học và đưa ra quyết định được gọi chung là *agent*. Nó tương tác trực tiếp tới một đối tượng được gọi là *môi trường*. Sự tương tác này được diễn ra liên tục. Agent lựa chọn hành động dựa trên những gì nó nhận được từ môi trường. Môi trường cung cấp giá trị điểm thưởng (reward) cho hành động vừa được thực hiện và những quan sát (observation) tiếp theo cho agent. Từ những quan sát này, agent có thể xây dựng ra các *trạng thái* (state) dựa vào đó để ra quyết định chọn hành động với mục tiêu cố gắng đạt được nhiều điểm thưởng nhất.

Cụ thể hơn, agent và môi trường tương tác theo một chuỗi tuần tự các time-steps,

$t = 0, 1, 2, ...$ Tại mỗi time step t, agent nhận những mô tả trạng thái của môi trường, $S_t \in \mathscr{S}$, với $\mathscr{S}$ là tập các trạng thái có thể có. Dựa vào những mô tả trạng thái nhận được, agent chọn một hành động, $A_t \in (S_t)$, trong đó (S_t) là tập các hành động có thể thực hiện tại trạng thái S_t. Tại time step sau đó, agent nhận được giá trị điểm thưởng, $R_{t+1} \in \mathbb{R}$, cùng với trạng thái tiếp theo S_{t+1} Quá trình tương tác giữa agent và môi trường được mô tả trong hình []

2.1.2 Các thành phần của agent

Để đạt được mục tiêu được nhiều điểm thưởng nhất, agent cần có một*chính sách* chọn lựa hành động mỗi khi gặp một trạng thái. Hay nói cách khác, chính sách, π, xác định khả năng chọn một hành động khi agent nhận được một trạng thái s. Chính xác tại time step t được xác định $\pi_t(a|s) = \mathbb{P}[A_t = a|S_t = s]$. Những phương pháp học tăng cường thường

2.2 Mô hình Markov Decision Processes (MDP)

- Các thành phần MDP
- Ví dụ cho mô hình MDP
- Phương trình Bellman
- Qui trình đánh giá chính sách: Kỹ thuật qui hoạch động
- Qui trình cải thiện chính động: Kỹ thuật qui hoạch động

2.3 Những phương pháp đánh giá và cải thiện chính sách

- Dẫn nhập: Trên thực tế ta không có thông tin về môi trường
- Qui trình đánh giá chính sách
    + Dựa trên hàm giá trị trạng thái: Monte-Carlo, TD(0), n-step TD, TD(λ)

+ Dựa trên hàm giá trị hành động: Monte-Carlo, Sarsa(0), n-step Sarsa, Sarsa(λ)

- Qui trình cải thiện chính sách: Phương pháp greedy

Chương 3

Kết hợp học sâu với học tăng cường

Học sâu đạt được nhiều kết quả khả quan trong nhiều lĩnh vực như xử lý ngôn ngữ tự nhiên, nhận diện giọng nói, nhận dạng đối tượng ... Ưu điểm của học sâu là xấp xỉ hàm tốt, rút trích đặc trưng ở mức high-level.
Trong chương này sẽ trình bày chi tiết những phương pháp mà chúng em áp dụng để giải quyết bài toán "tự động chơi game". Chúng em sử dụng phương pháp Q-Learning, một phương pháp đánh giá được sử dụng phổ biến trong nhiều công trình nghiên cứu về học tăng cường gần đây, để đánh giá hàm giá trị của hành động cho bài toán này. Ngoài ra, chúng em cũng thực hiện xấp xỉ hàm giá trị của bài toán bằng mạng nơ-ron và áp dụng hai kỹ thuật Experience Replay và Fixed Q-targets để tăng tính ổn định cho quá trình học.

Chương 4

Kết quả thực nghiệm

Trong chương này sẽ trình bày chi tiết về cấu trúc mô hình mà chúng em đã thiết lập. Đồng thời đề cập đến những phương pháp để đánh giá mô hình học và những kết quả thực nghiệm đã nhận được. Qua đó so sánh với các phương pháp đã đề xuất trước đây để thấy được tính hiệu quả của mô hình này.

4.1 Giới thiệu Arcade Learning Environment

4.2 Giới thiệu cấu trúc mạng và các siêu tham số đã chọn

4.3 Kết quả thực nghiệm

Chương 5

Kết luận và hướng phát triển

TÀI LIỆU THAM KHẢO

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, pp. 529–533, 2015. 3