Báo cáo tuần 14

Một số thuật toán học máy trong bài toán phân bổ nguồn lực trên các mô hình điện toán.

1. **Deep reinforcement learning** 
   1. **Deep Q-Networks (DQN)**

* Là một trong những mô hình DRL đầu tiên và được sử dụng rộng rãi nhất.
* Sử dụng mạng nơ-ron nhân tạo để ước tính giá trị Q cho mỗi trạng thái-hành động, giúp tác nhân lựa chọn hành động mang lại phần thưởng cao nhất.
* DQN đã được ứng dụng thành công trong nhiều trò chơi điện tử như Atari.
  1. **Deep Deterministic Policy Gradient (DDPG)**
* Cải tiến DQN bằng cách sử dụng mạng nơ-ron để ước tính chính sách hành động trực tiếp.
* DDPG học cách hành động một cách trơn tru và liên tục, phù hợp cho các bài toán điều khiển liên tục.
* Ứng dụng: Robot, xe tự lái.
  1. **Proximal Policy Optimization (PPO)**
* Cải tiến DDPG bằng cách sử dụng thuật toán tối ưu hóa để cập nhật chính sách hành động, giúp ổn định việc học tập và tránh các vấn đề như bùng nổ giá trị (value explosion).
* PPO phù hợp cho các bài toán
* Ứng dụng: Chơi game, Robot.

**1.3.1. PPO là gì?**

PPO (Proximal Policy Optimization) là một thuật toán học máy tăng cường (RL) được thiết kế để giải quyết các bài toán điều khiển liên tục. Nó giúp tìm ra chính sách (hành động) tốt nhất cho một tác nhân trong môi trường mà các hành động có thể có bất kỳ giá trị nào trong một phạm vi cụ thể. Khác với các bài toán điều khiển rời rạc, nơi các hành động bị giới hạn trong một tập các lựa chọn, các bài toán điều khiển liên tục liên quan đến các hành động có thể lấy bất kỳ giá trị nào trong một phạm vi được xác định. Ví dụ như điều khiển xe hơi - góc lái có thể là bất kỳ giá trị nào từ trái hết cỡ đến phải hết cỡ.

**1.3.2. PPO hoạt động như thế nào?**

PPO sử dụng hai phương pháp chính để huấn luyện tác nhân:

1. **Kiến trúc Actor-Critic**: Nó sử dụng kiến trúc actor-critic. Mạng actor đại diện cho chính sách, xác định hành động mà tác nhân thực hiện trong một trạng thái nhất định. Mạng critic, mặt khác, ước tính giá trị (phần thưởng dự kiến ​​tương lai) của việc ở trong một trạng thái cụ thể và thực hiện một hành động cụ thể.
2. **Cập nhật chính sách với clipping**: Đây là khái niệm cốt lõi của PPO:
   * Tác nhân tương tác với môi trường, thu thập dữ liệu (trạng thái, hành động, phần thưởng, trạng thái tiếp theo).
   * Mạng critic ước tính giá trị (V) của các trạng thái gặp phải.
   * Một lượng gọi là lợi thế (A) được tính toán. Nó thể hiện mức độ tốt hơn của một hành động được thực hiện so với hành động trung bình mà chính sách sẽ thực hiện trong trạng thái đó. (A = Q-V, trong đó Q là giá trị ước tính từ critic cho cặp trạng thái-hành động)
   * Mạng chính sách sau đó được cập nhật bằng cách sử dụng lợi thế để ưu tiên các hành động dẫn đến phần thưởng cao hơn.

Tuy nhiên, PPO kết hợp một bước quan trọng gọi là clipping. Trong quá trình cập nhật chính sách, sự thay đổi trong chính sách bị clipping để giữ cho nó gần với chính sách trước đó. Cơ chế clipping này giúp ngăn chính sách thay đổi quá mạnh trong một lần cập nhật duy nhất, dẫn đến việc học ổn định hơn.

**1.3.3. Ưu điểm của PPO**

* **Hiệu quả**: PPO được biết đến với hiệu quả trong việc học các chính sách hiệu quả cho các nhiệm vụ điều khiển liên tục.
* **Ổn định**: Cơ chế clipping trong PPO góp phần vào tính ổn định của nó, khiến nó ít bị dao động lớn trong các bản cập nhật chính sách hơn so với các thuật toán khác.
* **Triển khai**: PPO được coi là tương đối dễ triển khai so với một số thuật toán RL khác.

**1.3.4. Nhược điểm của PPO**

* **Yêu cầu dữ liệu**: Giống như hầu hết các thuật toán RL, PPO có thể yêu cầu một lượng lớn dữ liệu để đạt được hiệu suất tối ưu.
* **Tối thiểu cục bộ**: Vẫn có khả năng bị mắc kẹt trong tối thiểu cục bộ, nơi thuật toán tìm thấy một chính sách tốt nhưng không nhất thiết là tốt nhất.

**1.3.5. Ứng dụng của PPO**

PPO được ứng dụng trong nhiều lĩnh vực mà điều khiển liên tục là rất quan trọng:

* **Robot học**: PPO có thể được sử dụng để huấn luyện robot cho các nhiệm vụ như thao tác hoặc điều hướng.
* **Chơi game**: Nó có thể được sử dụng để huấn luyện các tác nhân chơi các trò chơi có yếu tố điều khiển liên tục.
* **Giao dịch tài chính**: PPO có tiềm năng ứng dụng trong các chiến lược giao dịch tài chính liên quan đến việc ra quyết định liên tục.

**1.3.6. Kết luận**

PPO là một thuật toán RL mạnh mẽ và linh hoạt cho các bài toán điều khiển liên tục. Hiệu quả, độ ổn định và độ dễ triển khai tương đối của nó khiến nó trở thành lựa chọn phổ biến cho nhiều nhiệm vụ học máy tăng cường. Tuy nhiên, giống như các thuật toán RL khác, yêu cầu dữ liệu và khả năng mắc kẹt trong tối thiểu cục bộ vẫn là những vấn đề cần cân nhắc.

1. **Q-learning**
2. **Reinforcement learning**

A diagram of a system

Description automatically generated

**4. các thách thức nghiên cứu mở trong lĩnh vực quản lý tài nguyên và tối ưu hóa hành vi trong các hệ thống máy tính lớn bao gồm:**

* Thách thức về chất lượng dữ liệu: Mọi sự nhiễu hoặc ảnh hưởng trong dữ liệu có thể làm giảm chất lượng học và quản lý hệ thống. Việc theo dõi và ghi lại thông tin có vai trò quan trọng trong việc giải quyết thách thức này .
* Thách thức về sự cố hệ thống: Hệ thống này rất dễ bị sự cố và mặc dù đã có nhiều nỗ lực, tỷ lệ sự cố vẫn khá cao. Điều này có thể gây ra sự gián đoạn trong việc ghi lại thông tin cần thiết cho quản lý hệ thống. Cung cấp các phương pháp thay thế và dự phòng cho các tình huống mà hệ thống gặp sự cố cũng là một thách thức quan trọng .
* Thách thức về quản lý tài nguyên hiệu quả: Quản lý tài nguyên trong các hệ thống máy tính lớn đang trở nên phức tạp hơn do sự đa dạng và quy mô lớn. Để giải quyết thách thức này, cần phải thực hiện lập lịch công việc và quản lý tài nguyên mà không tốn nhiều thời gian, đồng thời mô hình học được tạo ra không nên tốn nhiều thời gian trong giai đoạn suy luận .
* Thách thức về sự phức tạp của hệ thống: Với sự phát triển của công nghệ phần cứng, việc theo dõi và ghi lại thông tin liên quan đến các hệ thống lớn trở nên khả thi. Điều này tạo điều kiện cho việc sử dụng các phương pháp học máy trong quản lý hệ thống, bao gồm việc giải quyết vấn đề quản lý tài nguyên .

Những thách thức này đề cập đến những vấn đề quan trọng cần được giải quyết để cải thiện quản lý tài nguyên và tối ưu hóa hành vi trong các hệ thống máy tính lớn.

**5. kết luận và công việc tương lai được đề cập bao gồm:**

* Tối ưu hóa và tăng cường hiệu suất thông qua các phương pháp trí tuệ nhân tạo mới: Các phương pháp trí tuệ nhân tạo, đặc biệt là học sâu và học tăng cường sâu, đã được sử dụng để giải quyết vấn đề quản lý tài nguyên và tối ưu hóa hành vi trong các môi trường máy tính đa lớp. Sự so sánh toàn diện giữa các phương pháp trí tuệ nhân tạo đã được thực hiện trong các môi trường khác nhau .
* Phát triển các phương pháp quản lý tài nguyên dựa trên học máy và học sâu: Các phương pháp học máy và học sâu đã trở nên ngày càng phổ biến trong việc giải quyết vấn đề quản lý tài nguyên trong các môi trường máy tính. Các phương pháp này có thể được kết hợp với các thuật toán meta-heuristic để tìm ra mô hình tốt nhất với khả năng mở rộng cao và tối ưu hóa tối đa trong các yếu tố như độ trễ thời gian, tốc độ và khả năng tìm ra tối ưu hóa toàn cầu .
* Chuyển đổi từ môi trường Cloud sang Fog và Edge Computing: Trong thập kỷ qua, quản lý tài nguyên trong môi trường Cloud đã được coi trọng, nhưng gần đây, quản lý tài nguyên ở mức độ của các thiết bị thông minh cũng được xem xét. Với sự gia tăng của các thiết bị thông minh, như ô tô tự lái, các thiết bị gia đình thông minh, điện thoại thông minh và robot, dự kiến sẽ có nhiều nghiên cứu hơn về quản lý tài nguyên trong môi trường này trong những năm tới .
* Sự phát triển của các thuật toán học sâu trong mô hình hành vi của các hệ thống: Với sự phức tạp ngày càng tăng của các hệ thống và yêu cầu quản lý dựa trên phản hồi liên tục từ môi trường, việc sử dụng các phương pháp học sâu để mô hình hành vi của các hệ thống sẽ đóng một vai trò quan trọng trong tương lai.