**Exploration with Exemplar Models for Deep Reinforcement Learning**

**Tóm tắt**

Các thuật toán Deep reinforcement learning ( học tăng cường sâu) đã được chứng minh để học các công việc phứt tạp sử dụng các lớp chính xác tổng quát cao. Tuy nhiên, vấn đề về reward( Phẩn thường) thưa thớt vẫn là một thách thức quan trọng. Các phương pháp thăm dò dựa trên phát hiện mới lạ đã có kết quả thành công, nhưng thường yêu cầu các mô hình quan sát hoặc tiên đoán của các quan sát, điều đó có thể làm việc huấn luyện trở nên rất khó khi quan sát phứt tạp như trong trường hợp hình ảnh thô

1. Giới thiệu  
   Các công việc gần đây đã cho thấy các phương pháp kết hợp Deep reinforcement learning cùng với hàm xấp xỉ, như là mạng neural, có thể giải quyết một loạt công việc phứt tạp, từ việc chơi trò Atari để điều khiển robot mô phỏng. Mặc dù phương pháp Deep Reinforcement learning cho phép biểu diễn các chính sách phứt tạp, nhưng nó không tự giải quyết được các vấn đề về thăm dò (exploration problem) : Khi tín hiệu phần thưởng hiếm và thưa thớt, các phương pháp như vậy có thể đấu tranh để có được các chính sách có ý nghĩa. Chiến lược thăm dò tiêu chuẩn, chẳng hạn như chiến lược tham lam ( e-greedy strategies) hoặc Gaussian noise, không bị loại trừ và không tìm kiếm một cách rõ ràng các trạng thái thú vị. Một phương pháp đầy hứa hẹn đề điều khiển thăm dò nhiều hơn là ước tính một cách rõ ràng tính mới lạ của một trạng thái. Các khái niệm liên quan như tiền thưởng dựa trên số lượng đã được chứng minh là cung cấp tăng tốc đáng kể trong reinforcement learning cổ điển, và một số công việc gần đây đã đề xuất phương pháp tiếp cận lý thuyết hoặc xác suất thông tin để thăm dò, dựa trên ý tưởng này, bằng cách vẽ các kết quả chính thức trong các hệ thống rời rạc hoặc tuyến tính đơn giản hơn. Tuy nhiên, phương pháp ước lượng mới nhất dựa vào việc xây dựng các mô hình tiên đoán hoặc dự đoán mô hình rõ ràng sự phân bố trên quan sát hiện tại hoặc tiếp theo. Khi quan sát phức tạp và cao, chẳng hạn như trong trường hợp hình ảnh thô, các mô hình này khó đào tạo, vì việc tạo và dự đoán hình ảnh và các vật thể cao khác vẫn là một vấn đề mở, bất chấp tiến độ gần đây. Mặc dù kết quả thành công với các mô hình mới lạ mang tính đa dạng đã được báo cáo với các hình ảnh tổng hợp đơn giản, chẳng hạn như trong các trò chơi Atari, chúng tôi cho thấy trong các thí nghiệm của chúng tôi các phương pháp sinh học như vậy đang phải vật lộn với các quan sát phức tạp và tự nhiên hơn, chẳng hạn như các quan sát hình ảnh trung tâm trong tiêu chuẩn vizDoom. Làm cách nào chúng tôi có thể đánh giá được các trạng thái truy cập mới, và do đó cung cấp sự trình bày tín hiệu có sẵn cho việc học tăng cường, mà không xây dựng một cách rõ rang các mô hình tiên đoán hoặc mang tính tiên đoán của trạng thái hoặc quan sát
2. Công việc liên quan  
   Trong các MDP hữu hạn, các thuật toán thăm dò như E3 và R-MAX đưa ra các đảm bảo tối ưu về mặt lý thuyết. Tuy nhiên, các phương pháp này thường yêu cầu bảo tồn trạng thái – hành động số lượt truy cập, điều này có thể mở rộng chúng lên chiều cao hơn (dimensional) và/hoặc trạng thái liên tiếp rất thách thức . Khám phá trong các không gian trạng thái như vậy thường có các chiến lược liên quan như giới thiệu các số liệu khoảng cách trên không gian trạng thái, và xấp xỉ số lượng được sử dụng trong các phương pháp thăm dò cổ điển. Các công việc trước đây đã sử dụng số liệu xấp xỉ cho số lượt truy cập trạng thái, tăng thông tin, hoặc dự đoán lỗi. Bellemare et al. (2016) cho thấy các phương pháp đếm dựa trên ý nghĩa ràng buộc tiền thưởng được tạo ra bởi các khuyến khích thăm dò dựa trên sự trình bày có sẵn, chẳng hạn như sự không chắc chắn của mô hình hoặc tăng thông tin, làm cho tiền thưởng dựa trên mật độ hấp dẫn và đơn giản.