계층적 심층 강화 학습: 시간적 추상화와 내재적 동기 부여의 통합

테자스 D. 쿨카르니*
BCS, MIT
tejask@mit.edu

카틱 R. 나라심한*
CSAIL, MIT
karthikn@mit.edu

아르다반 사에디
CSAIL, MIT
ardavans@mit.edu

조슈아 B. 테넨바움
BCS, MIT
jbt@mit.edu

초록

피드백이 희박한 환경에서 목표 지향적인 행동을 학습하는 것은 강화 학습 알고리즘의 주요 과제입니다. 가장 큰 어려움은 불충분한 탐색으로 인해 에이전트가 강력한 가치 함수를 학습하지 못하기 때문에 발생합니다. 내재적 동기를 가진 에이전트는 문제를 직접 해결하기보다는 스스로 새로운 행동을 탐색할 수 있습니다. 이러한 내재적 행동은 결국 에이전트가 환경이 제시하는 과제를 해결하는 데 도움이 될 수 있습니다. 유니티는 다양한 시간적 규모에서 작동하는 계층적 가치 함수를 내재적 동기 부여 심층 강화 학습과 통합하는 프레임워크인 계층적-DQN(h-DQN)을 소개합니다. 최상위 가치 함수는 내재적 목표에 대한 정책을 학습하고, 하위 수준 함수는 주어진 목표를 충족하기 위한 원자적 행동에 대한 정책을 학습합니다. h-DQN을 사용하면 엔티티 및 관계에 대한 함수와 같은 유연한 목표 사양을 지정할 수 있습니다. 이는 복잡한 환경에서 탐색을 위한 효율적인 공간을 제공합니다. 피드백이 매우 희박하고 지연된 두 가지 문제, 즉 (1) 복잡한 이산 확률적 의사 결정 과정과 (2) 고전적인 아타리 게임 '몬테주마의 복수'를 통해 접근 방식의 강점을 입증합니다.

1 소개

복잡한 환경에서 희박한 피드백으로 목표 지향적 행동을 학습하는 것은 인공지능의 근본적인 과제입니다. 이러한 환경에서 학습하려면 에이전트가 여러 수준의 시공간적 추상화에서 지식을 표현하고 환경을 효율적으로 탐색해야 합니다. 최근에는 강화 학습[21, 28, 37]과 결합된 비선형 함수 근사화 기법을 통해 고차원 상태 공간에서 추상화를 학습할 수 있

게 되었지만, 희박한 피드백으로 탐색하는 작업은 여전히 주요 과제로 남아 있습니다. 볼츠만 탐색 및 톰슨 샘플링[45, 32]과 같은 기존 방법은 ϵ-탐욕에 비해 상당한 개선점을 제공하지만 기본 동작 수준에서 작동하는 기본 모델로 인해 제한적입니다. 이 연구에서는 심층 강화 학습과 계층적 가치 함수(h-DQN)를 통합하여 에이전트가 학습 옵션을 통해 내재적 목표를 해결하도록 동기를 부여하여 탐색을 지원하는 프레임워크를 제안합니다. 이러한 목표는 효율적인 탐색을 지원하고 희박한 피드백 문제를 완화하는 데 도움이 됩니다. 또한 엔티티와 관계의 공간에 정의된 목표는 복잡한 환경에서 데이터 효율적인 학습을 위해 탐색 공간을 크게 제한하는 데 도움이 될 수 있음을 관찰했습니다.

*저자들은 동등하게 기여했으며 가나다순으로 나열되었습니다.

강화 학습(RL)은 제어 문제를 미래의 예상 보상을 최대화하는 정책 π를 찾는 것으로 공식화합니다[46]. 가치 함수 V(s)는 RL의 핵심이며, 에이전트의 전체 목표를 달성하는 데 있어 모든 상태 s의 효용을 캐시합니다. 최근에는 주어진 목표 g ∈ G를 달성하기 위한 상태 s의 효용을 나타내기 위해 가치 함수가 V(s, g)로 일반화되기도 했습니다[47, 34]. 환경이 지연 보상을 제공할 때, 우리는 먼저 내재적으로 생성된 목표를 달성하는 방법을 학습한 다음, 이를 서로 연결하기 위한 최적의 정책을 학습하는 전략을 채택합니다. 에이전트가 목표 상태 g에 도달하면 종료되는 정책을 생성하는 데 각 가치 함수 V(s, g)를 사용할 수 있습니다. 이러한 정책 모음은 세미 마르코프 의사 결정 프로세스의 프레임워크 내에서 학습 또는 계획을 위한 시간 역학으로 계층적으로 배열할 수 있습니다[48, 49]. 고차원 문제에서 이러한 값 함수는 신경망에 의해 V(s, g; θ)로 근사화될 수 있습니다.

저희는 계층적으로 구성된 심층 강화 학습 모듈이 서로 다른 시간 규모에서 작동하는 프레임워크를 제안합니다. 이 모델은 두 가지 수준의 계층 구조에 따라 의사 결정을 내립니다.
(a) 최상위 모듈(*메타 컨트롤러*)이 상태를 가져와 새 목표를 선택하고, (b) 하위 모듈(*컨트롤러*)이 상태와 선택한 목표를 모두 사용하여 목표에 도달하거나 에피소드가 종료될 때까지 작업을 선택합니다. 그런 다음 *메타* 컨트롤러가 다른 목표를 선택하고 단계 (a-b)를 반복합니다. 다양한 시간적 규모에서 확률적 경사 하강을 사용하여 모델을 훈련하여 예상되는 미래의 내재적(*컨트롤러*) 및 외재적 보상(*메타 컨트롤러*)을 최적화합니다. (1) 최적의 외재적 보상을 받기까지 긴 상태 체인이 있는 이산 확률적 의사 결정 과정, (2) 기존의 대부분의 최신 심층 강화 학습 접근 방식이 데이터 효율적인 방식으로 정책을 학습하지 못하는 더 긴 범위의 지연 보상이 있는 고전적인 ATARI 게임('몬테주마의 복수')을 통해 장거리 지연 피드백 문제에 대한 접근 방식의 강점을 입증했습니다.

2 문헌 검토

2.1 시간 추상화를 사용한 강화 학습

다양한 수준의 시간적 추상화를 학습하고 작동하는 것은 장기적인 계획과 관련된 작업에서 핵심 과제입니다. 강화 학습[1]의 맥락에서 Sutton 등[48]은 행동 공간에 대한 추상화를 포함하는 *옵션* 프레임워크를 제안했습니다. 각 단계에서 에이전트는 한 단계의 "기본" 작업 또는 "다단계" 작업 정책(옵션) 중 하나를 선택합니다. 각 옵션은 행동에 대한 정책(기본 또는 다른 옵션)을 정의하며 확률 함수 β에 따라 종료될 수 있습니다. 따라서 옵션을 사용하면 기존의 MDP 설정을 세미 마르코프 의사 결정 프로세스(SMDP)로 확장할 수 있습니다. 최근에는 다양한 보상 함수를 사용하거나[49] 기존 옵션을 구성하여 실시간으로 옵션을 학습하는 여러 방법이 제안되었습니다[42]. 또한 상태와 함께 목표를 고려하도록 가치 함수가 일반화되었습니다 [34]. 이 보편적 가치 함수 V(s, g; θ)는 목표 g를 향한 최적의 행동을 대략적으로 나타내는 보편적 옵션을 제공합니다. 저희의 작업은 이러한 논문에서 영감을 받아 이를 기반으로 합니다.

또한 표 형식의 값 함수 설정에서 옵션 검색에 대한 많은 연구가 진행되었습니다[26, 38, 25, 27]. 보다 최근의 연구에서 Machado 등[24]은 에이전트가 이전에는 도달할 수 없었던 영역을

탐색하도록 권장하는 옵션 발견 알고리즘을 제시했습니다. 그러나 비선형 상태 근사치가 필요한 옵션 탐색은 여전히 미해결 문제로 남아 있습니다.

계층적 공식화에 대한 다른 관련 작업으로는 에이전트의 원자적 행동에 이르기까지 다양한 수준의 세분화에서 의사 결정을 내리는 "관리자"로 구성된 Dayan과 Hinton의 모델[6]이 있습니다. MAXQ 프레임워크[7]는 이 작업을 기반으로 MDP의 가치 함수를 더 작은 구성 MDP의 가치 함수 조합으로 분해하기 위해 구축되었으며, Guestrin 외[17]의 팩터링된 MDP 공식화에서와 마찬가지로 이 프레임워크는 MDP를 더 작은 구성 MDP의 가치 함수 조합으로 분해합니다. 에르난데스-가르디올과 마하데반[19]은 계층적 RL과 높은 수준의 의사결정에 대한 가변 길이 단기 기억을 결합했습니다.

본 연구에서는 심층 강화 학습 환경에서 옵션을 동시에 학습하고 옵션을 구성하는 제어 정책을 포함하는 시간적 추상화 방식을 제안합니다. 우리의 접근 방식은 각 옵션에 대해 별도의 Q 함수를 사용하지 않고 대신 [34]와 유사하게 옵션을 입력의 일부로 취급합니다. 여기에는 두 가지 장점이 있습니다.

학습하고, (2) 이 모델은 잠재적으로 많은 수의 옵션으로 확장할 수 있습니다.

2.2 내재적 동기 부여 RL

'좋은' 내재적 보상 함수의 본질과 기원은 강화 학습에서 미해결 문제입니다. Singh 등 [41]은 다양한 작업에 적용할 수 있는 일반적인 옵션을 학습하기 위해 내재적 보상 구조를 가진 에이전트를 탐색했습니다. 에이전트는 "두드러진 이벤트"라는 개념을 하위 목표로 사용하여 이러한 이벤트에 도달하기 위한 옵션을 학습합니다. 또 다른 논문에서 Singh 등[40]은 진화적 관점에서 에이전트의 보상 함수 공간을 최적화하여 외재적 및 내재적 동기 부여 행동이라는 개념을 도출했습니다. 계층적 RL의 맥락에서 Goel과 Huber[13]는 학습된 정책 모델의 구조적 측면을 사용하여 하위 목표 발견을 위한 프레임워크에 대해 논의합니다. S ,im ,sek 등[38]은 하위 목표 식별을 위한 그래프 분할 접근 방식을 제공합니다.

슈미드후버[36]는 학습 알고리즘에 의해 만들어진 예측 세계 모델의 개선으로 측정되는 내재적 동기에 대한 일관된 공식을 제공합니다. 모하메드와 레젠데[29]는 최근 상호 정보 극대화의 틀 안에서 내재적 동기 부여 학습이라는 개념을 제안했습니다. Frank 등[11]은 휴머노이드 로봇에서 정보 획득 극대화를 사용하여 인공 호기심의 효과를 입증했습니다.

2.3 객체 기반 RL

문제의 기본 구조를 활용할 수 있는 객체 기반 표현[8, 4]이 RL에서 *차원성의 저주를* 완화하기 위해 제안되었습니다. Diuk 등[8]은 객체와 객체의 상호 작용에 기반한 표현을 사용하여 객체 *지향 MDP를* 제안합니다. 각 상태를 객체 간의 가능한 모든 관계에 대한 값 할당 집합으로 정의하여 결정론적 객체 지향 MDP를 풀기 위한 알고리즘을 도입합니다. 이들의 표 현 은 계획의 맥락에서 객체 기반 표현을 설명하는 Guestrin 등[16]의 표현과 유사합니다. 이러한 접근 방식과 달리, 우리의 표현은 객체 간의 관계에 대한 명시적인 인코딩이 필요하지 않으며 확률적 영역에서 사용할 수 있습니다.

2.4 심층 강화 학습

최근 심층 신경망을 이용한 함수 근사화의 발전은 고차원 감각 입력을 처리할 수 있 는 가능성을 보여주었습니다. 딥 큐 네트워크와 그 변형은 아타리 게임[28], 바둑[37] 등 다양한 분야에 성공적으로 적용되었지만, 보상 신호가 희박하고 지연되는 환경에서는 여전히 성능이 좋지 않습니다. 희소한 보상으로부터의 학습 문제를 완화하기 위해 우선순위가 지정된 경험 재생[35] 및 부트스트래핑[32]과 같은 전략이 제안되었습니다. 이러한 접근 방식은 이전 작업보다 상당한 개선을 가져다주지만 보상 신호의 지연 시간이 길면 어려움을 겪습니다. 이는 에이전트가 필요한 피드백을 얻기에 탐색 전략이 충분하지 않기 때문입니다.

2.5 인지 과학 및 신경 과학

인간의 내재적 목표의 본질과 기원은 까다로운 문제이지만 기존 문헌에서 몇 가지 주목할 만한 통찰이 있습니다. 발달 심리학에서는 다양한 문화권의 유아, 영장류, 어린이, 성인이 실체, 행위자 및 행위, 수적 양, 공간, 사회 구조, 직관적 이론 등 특정 인지 체계에 기반하여 핵심 지식을 습득한다는 일관된 증거가 있습니다[43, 23]. 신생아와 영아도 응집성, 연속성, 접촉이라는 시공간적 원리를 중심으로 일관된 시각적 실체라는 관점에서 시각 세계를 표현하는 것으로 보입니다. 또한 에이전트의 행동이 목표 지향적이고 효율적이라는 가정 하에 다른 에이전트를 명시적으로 표현하는 것처럼 보입니다. 또한 유아는 물체의 상대적 크기, 상대적 거리, 물체 크기의 비율과 같은 고차적인 숫자 관계를 구별할 수 있습니다. 호기심 중심 활동을 하는 동안 유아는 이러한 지식을 사용하여 물리적으로 안정적인 블록 구조물을 만드는 것과 같은 내재적 목표를 생성합니다. 이러한 목표를 달성하기 위해 유아는 높은 블록을 만들기 위해 더 무거운 물체를 더 가벼운 물체 *위에 올려놓는* 것과 같은 핵심 지식의 공간에 하위 목표를 구성하는 것으로 보입니다.

공간에 대한 지식은 서로 다른 공간 그룹 사이의 병목 현상이 하위 목표에 해당하는 공간 환경의 계층적 분해를 학습하는 데에도 활용될 수 있습니다. 이는 신경과학에서 예상되는 미래 상태 점유에 대한 가치 함수를 나타내는 후계자 표현을 통해 탐구되어 왔습니다. 후계자 표현을 분해하면 공간 탐색 문제에 대한 합리적인 하위 목표가 산출됩니다[5, 12, 44]. Botvinick 등[3]은 인지 과학 및 신경 과학의 맥락에서 계층적 강화 학습에 대한 일반적인 개요를 작성했습니다.

3 모델

상태 $s \in S$, 액션 $a \in A$, 전이 함수 $T : (s, a) \to s'$ 로 표현되는 마르코프 의사 결정 과정(MDP)을 생각해 보겠습니다. 이 프레임워크에서 작동하는 에이전트는 외부 환경으로부터 상태 s를 수신하고 액션 a를 수행하여 새로운 상태 s' 를 얻을 수 있습니다. 에이전트의 목표는 오랜 기간 동안 이 함수를 최대화 하는 것입니다. 예를 들어, 이 함수는 에이전트의 생존 시간 또는 게임 점수의 형태를 취할 수 있습니다.

에이전트 MDP에서 효과적인 탐색은 좋은 제어 정책을 학습하는 데 있어 중요한 과제입니다. ϵ-욕심과 같은 방법은 국소 탐색에는 유용하지만 에이전트가 상태 공간의 다른 영역을 탐색하도록 자극을 제공하지 못합니다. 이 문제를 해결하기 위해 에이전트에게 내재적 동기를 제공하는 *목표* $g \in G$ 개념을 활용합니다. 에이전트는 누적된 외재적 보상을 극대화하기 위해 일련의 목표를 설정하고 달성하는 데 집중합니다.

옵션의 시간적 추상화[48]를 사용하여 각 목표 g에 대한 정책 π_g 을 정의합니다. 에이전트는 따라야 할 최적의 목표 순서를 학습하는 동시에 이러한 옵션 정책을 학습합니다. 각 π_g 를 학습하기 위해 에이전트에는 에이전트가 목표를 달성할 수 있는지 여부에 따라 *내재적 보상을* 제공하는 비평가도 있습니다(그림 1 참조).

시간적 추상화 그림 1에서 볼 수 있듯이 에이전트는 컨트롤러와 *메타 컨트롤러로* 구성된 2단계 계층 구조를 사용합니다. 메타 컨트롤러는 상태 s_t 를 수신하고 목표 $g_t \in G$를 선택합니다. 여기서 G는 가능한 모든 현재 목표의 집합을 나타냅니다. 그런 다음 컨트롤러는 s_t 및 g_t 를 사용하여 액션 a_t 를 선택합니다. 목표 g_t 는 달성되거나 종료 상태에 도달할 때까지 다음 몇 개의 시간 단계 동안 효과적으로 유지됩니다. 내부 비평자는 목표 도달 여부를 평가하고 컨트롤러에게 적절한 보상 r_t (g)를 제공할 책임이 있습니다. 컨트롤러의 목적 함수는 누적 내재적 보상을 최대화하는 것입니다.

보상: R_t (g) = $\sum_{t'=t}^{\infty} \gamma^{t'-t} r_{t'}$(g). 마찬가지로 메타 컨트롤러의 목표는 다음과 같습니다. 누적 외적 보상 F_t = 환경으로부터 받은 보상 $\sum_{t'=t}^{\infty} \gamma^{t'-t} f_{t'}$, 여기서 f_t 는 보상 신호입니다. 을 최적화합니다.

이러한 설정은 최적 보상 함수의 공간에서 최적화를 통해 체력을 극대화하는 것과 유사하다고 볼 수도 있습니다[39]. 이 경우 보상 함수는 동적이며 목표의 순차적 이력에 따라 시

간적으로 의존합니다. 그림 1은 에이전트가 후속 시간 단계에 걸쳐 계층 구조를 사용하는 모습을 보여줍니다.

시간 추상화를 사용한 심층 강화 학습

딥러닝 프레임워크[28]를 사용하여 컨트롤러와 메타 컨트롤러 모두에 대한 정책을 학습합니다. 구체적으로 컨트롤러는 다음과 같은 Q값 함수를 추정합니다:

$$Q^*(s, a; g) = \max_{\pi_{ag}} E\left[\sum_{t'=t}^{t+1} \gamma^{t'-t} r_{t'} \mid s_t = s, a_t = a, g_t = g, \pi_{ag}\right] \quad (1)$$

$$= \max_a E\left[r_t + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}; g) \mid s_t = s, a_t = a, g_t = g, \pi_{ag}\right]$$

여기서 g는 상태 s에서 에이전트의 목표이고 $\pi_{ag} = P(a|s, g)$는 액션 정책입니다. 마

찬가지로 메타 컨트롤러의 경우도 마찬가지입니다:

$$Q_2^*(s, g) = \max_\pi E\left[\sum_{t'=t}^{t+N} f_{t'} + \gamma \max_{g'} Q^*(s_{t+N}, g') \mid s_t = s, g_t = g, \pi\right] \quad (2)$$

입니다.

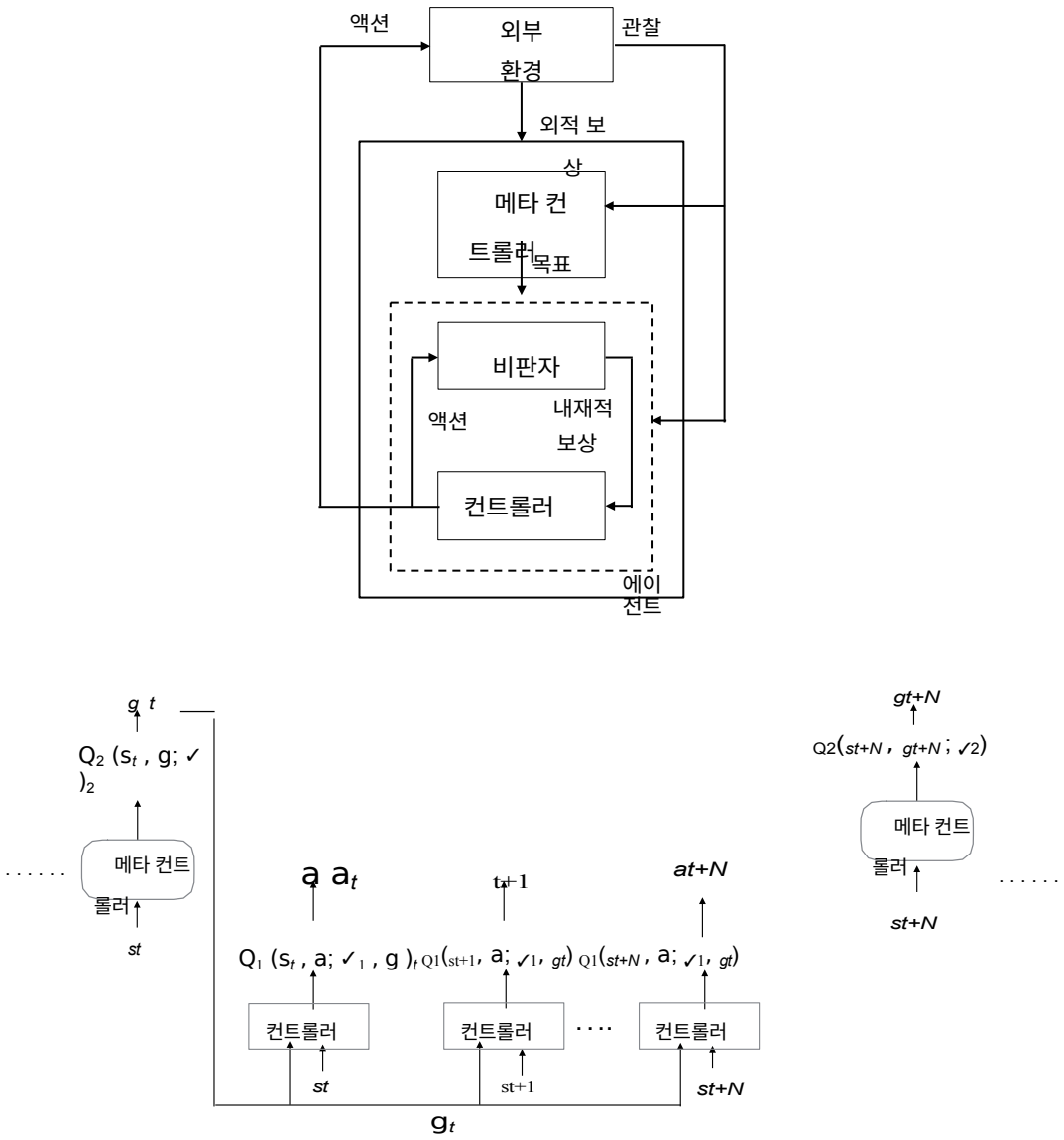


그림 1: **개요:** 에이전트는 동작을 생성하고 감각 관찰을 수신합니다. *메타* 컨트롤러와 *컨트롤러* 내부에는 별도의 딥-Q 네트워크가 사용됩니다. 메타 컨트롤러는 원시 상태를 살펴보고 (예상되는 미래 외적 보상을 최대화하여) 가치 함수 $Q_2(s_t, g_t, \theta_2)$를 추정하여 목표에 대한 정책을 생성합니다. 컨트롤러는 상태와 현재 목표를 입력받고, 예측된 목표를 해결하기 위해 (미래 내재적 보상을 최대화하여) 가치 함수 $Q_2(s_t, a_t; \theta_1, g_t)$를 추정하여 행동에 대한 정책을 생성합니다. 내부 비평자는 목표에 도달했는지 확인하고 컨트롤러에게 적절한 내재적 보상을 제공합니다. 컨트롤러는 에피소드가 종료되거나 g가 달성되면 종료됩니다. 그러면 메타 컨트롤러가 새로운 g를 선택하고 프로세스가 반복됩니다.

여기서 N은 현재 목표가 주어졌을 때 컨트롤러가 중단할 때까지의 시간 단계 수를 나타내고, g′ 는 상태 s_{t+N} 에서 에이전트의 목표이며, $\pi_g = P(g|s)$는 목표에 대한 정책입니다. Q_2 에 의해 생성된 전환(s_t , g_t , f_t , s_{t+N})은 Q_1 에 의해 생성된 전환(s_t , a_t , g_t , r_t , s_{t+1})보다 느린 시간 척도로 실행된다는 점에 유의해야 합니다.

파라미터 θ가 있는 비선형 함수 근사화기를 사용하여 $Q^*(s, g) \approx Q(s, g; \theta)$를 표현할 수 있는데, 이를 심층 Q-네트워크(DQN)라고 합니다. 각 $Q \in \{Q_1 , Q_2 \}$는 해당 손실 함수인 $L_1(\theta_1)$ 및 $L_2(\theta_2)$를 최소화하여 학습할 수 있습니다. Q_2 에 대한 경험 (s_t , g_t , f_t , s_{t+N})과 Q_1 에 대한 경험 (s_t , a_t , g_t , r_t , s_{t+1})을 각각 분리된 메모리 공간 D_1 과 D_2 에 저장합니다. 그러면 Q_1 의 손실 함수는 다음과 같이 표현할 수 있습니다:

$$L_1(\theta_{1,i}) = E_{(s,a,g,r,s')\sim D1}[(y_{1,i} - Q_1(s, a; \theta_{1,i}, g))^2], \quad (3)$$

여기서 i는 훈련 반복 횟수를 나타내고 $y_{1,i} = r + \gamma \max_{a'} Q_1(s', a'; \theta_{1,i-1}', g)$.

28]에 따라 손실 함수를 최적화할 때 이전 반복의 파라미터 $\theta_{1,i-1}$ 는 고정된 상태로 유지됩니다. 파라미터 θ_1 는 기울기를 사용하여 최적화할 수 있습니다:

$$\nabla_{\theta 1,i} L1(\theta 1,i)$$

$$= E_{(s,a,r,s'\sim D)}\Big[\big(r + \gamma \max_{a'} Q_1(s', a'; \theta_{1,i-1}, g) - Q_1(s, a; \theta_{1,i}, {}_{1,i}Q_1(s, a; \theta_{1,i}, g)\big) \nabla_\theta$$

손실 함수 L_2 및 그 기울기는 유사한 절차를 사용하여 도출할 수 있습니다.

학습 알고리즘 컨트롤러의 경험(또는 전환)은 모든 시간 단계에서 수집되지만 메타 컨트롤러의 경험은 컨트롤러가 종료될 때(즉, 목표가 다시 선택되거나 에피소드가 종료될 때)에만 수집됩니다. 각각의 새로운 목표 g는 학습이 진행됨에 따라 (시작 값 1에서) 탐험 확률 ϵ_2 을 어닐링하여 ϵ-탐욕적인 방식으로 그려집니다(알고리즘 1 및 2).

컨트롤러에서는 모든 시간 단계에서 현재 경험적 성공률인 g에 도달하는 확률에 따라 달라지는 탐색 확률 $\epsilon_{1,g}$ 을 사용하여 목표가 있는 동작을 도출합니다. 모델 파라미터(θ_1 , θ_2)는 각각 리플레이 메모리 D_1 및 D_2)에서 경험을 도출하여 주기적으로 업데이트됩니다(알고리즘 3 참조).

4 실험

저희는 지연 보상과 관련된 두 가지 영역에서 실험을 진행했습니다. 첫 번째는 확률적 전환이 있는 이산 상태 MDP이고, 두 번째는 '몬테주마의 리벤지'라는 ATARI 2600 게임입니다.

4.1 이산 확률적 의사 결정 프로세스

게임 설정 현재 상태 외에 방문한 상태의 이력에 따라 외재적 보상이 달라지는 확률적 결정 과정을 고려합니다. 이 과제를 선택한 이유는 임포러

이러한 환경에서 탐험에 대한 내재적 동기를 부여할 수 있습니다.

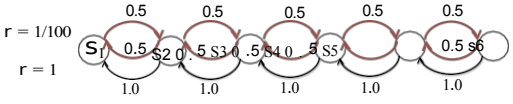


그림 2: 터미널 상태 s_1 의 보상이 s_6 의 방문 여부($r = 1$)에 따라 달라지는 확률적 결정 프로세스($r = 1/100$).

가능한 상태는 6가지이며 상담원은 항상$_2$ 에서 시작합니다. 상담원이 왼쪽으로 이동합니다.

결정론적으로 *왼쪽 동작을* 선택하지만 *오른쪽* 동작은 50%만 성공하며, 그렇지 않으면 왼쪽으로 이동합니다. 터미널 상태는 s_1 이며 에이전트가 처음 s_6 를 방문한 후 s_1 를 방문하면 1의 보상을 받습니다. s_6 를 방문하지 않고 s_1 로 이동하면 보상은 0.01입니다. 이는 [32]의 MDP를 수정한 버전으로, 보상 구조가 작업에 복잡성을 더합니다. 이 과정은 그림 2에 설명되어 있습니다.

알고리즘 1 h-DQN 학습 알고리즘

컨트롤러와 메타 컨트롤러에 대해 각각 경험 리플레이 메모리 $\{D_1, D_2\}$와 파라미터 $\{\theta_1, \theta_2\}$를 초기화합니다.
모든 목표 g에 대한 컨트롤러의 탐색 확률 $\epsilon_{1,g} = 1$, 메타 컨트롤러의 경우 $\epsilon_2 = 1$을 초기화 합니다.
FOR I = 1, NUM EPISODE **DO**
    4: 게임 초기화 및 시작 상태 설명 얻기 s
    g ← EPSGREEDY(s, G, ϵ_2, Q_2) 6: s가 종단이 **아닌 동안** 7:
        F ← 0
        s_0 ← s
        하지 않는 동안 (S는 터미널 **또는** 목표 G에 도달) **DO**
            a ← EPSGREEDY({s, g}, A, $\epsilon_{1,g}$, Q)$_1$
            a를 실행하고 다음 상태 s′ 와 환경으로부터 외적 보상 f를 얻습니다.
            내부 비평가로부터 내재적 보상 r(s, a, s′)을 획득합니다.
            전환({s, g}, a, r, {s′, g})을 D에 저장합니다.$_1$
            UPDATEPARAMS(L_1 ($\theta_{1,i}$), D)$_1$
            UPDATEPARAMS(L_2 ($\theta_{2,i}$), D)$_2$
            F ← F + f
            s ← s′
        동안 종료
        스토어 전환 ($_0$, g, F, s′)을 D에 저장합니다.$_2$
        IF S가 터미널이 **아니라면**
            g ← EPSGREEDY(s, G, ϵ_2, Q)$_2$
        END IF
    while
    $_2$ 을 어닐링하고 목표 g에 도달하는 평균 성공률을 사용하여$_{1,g}$ 을 적응적으로 어닐링합니다.
종료

알고리즘 2: EPSGREEDY(x, B, ϵ, Q)

if random() < ϵ **then**
    세트 B에서 무작위 요소 **반환**
else
    반환 argmaxm∈BQ(x, m)
END IF

알고리즘 3: UPDATEPARAMS(L, D)

D에서 미니 배치를 무작위로 샘플링합니다.
손실 L(θ)에 대해 경사 하강을 수행합니다(참조: (3)).

각 상태를 탐색 가능한 목표로 간주합니다. 이렇게 하면 에이전트가 상태 s_6 (목표로 선택될 때마다)를 방문하여 최적의 정책을 학습하도록 유도합니다. 각 목표에 대해 에이전트는 해당 상태에 도달할 경우에만 양의 내재적 보상을 받습니다.

결과 한 에피소드에서 얻은 평균 외재적 보상의 관점에서 딥 뉴럴 네트워크가 없는 접근 방식과 내재적 보상이 없는 Q러닝의 성능을 비교합니다. 실험에서 모든 ϵ 파라미터는 50,000단계

에 걸쳐 1에서 0.1까지 어닐링되었습니다. 학습률은 0.00025로 설정했습니다. 그림 3은 10회 실행에 걸쳐 평균화된 두 가지 방법의 보상 변화를 보여줍니다. 예상대로 Q-Learning은 200개의 에포크가 지난 후에도 최적의 정책을 찾지 못하고 0.01의 보상을 얻기 위해 상태 s_1 에 직접 도달하는 차선의 정책으로 수렴하는 것을 볼 수 있습니다. 이와 대조적으로 계층적 Q-추정기를 사용하는 접근 방식은 목표 s_4 , s_5 또는 s_6 를 선택하는 방법을 학습하여 에이전트가 통계적으로 s_6 를 방문한 후 다시 s_1 로 돌아가도록 유도합니다. 따라서 에이전트는 약 0.13의 훨씬 높은 평균 보상을 얻게 됩니다.

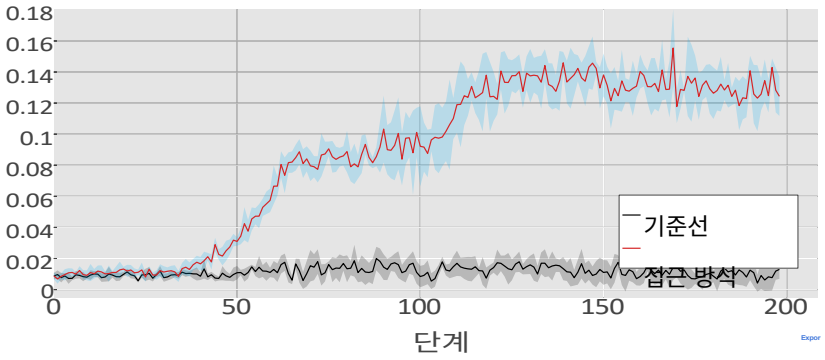


그림 3: Q-러닝과 비교한 당사 접근 방식의 10회 실행에 대한 평균 보상.

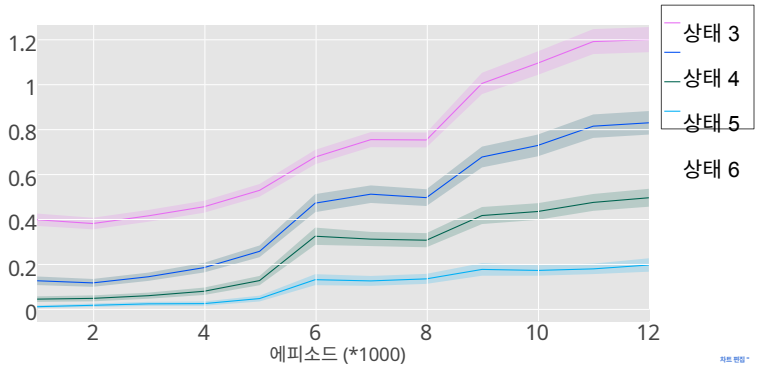


그림 4: 1000회 이상 평균 방문 횟수(상태 $s_3 \sim s_6$)의 경우. 초기 상태는 s_2 이고 터미널 상태는 s_1 입니다.

그림 4는 주 방문 횟수가 훈련 에피소드에 따라 증가하는 것을 보여줍니다. , ,$_{3456}$,. 각 데이터 포인트는 지난 1000개의 에피소드 동안 각 주에 대한 평균 방문 횟수를 보여줍니다. 이는 모델이 임계 상태 s_6 에 더 자주 도달할 수 있도록 목표를 선택하고 있음을 나타냅니다.

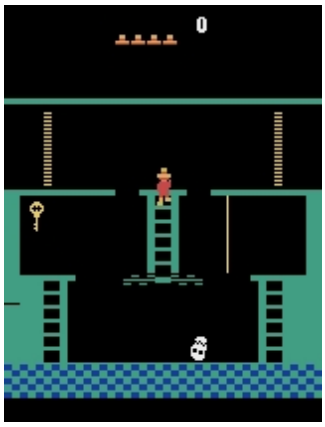
4.2 보상이 지연되는 아타리 게임

게임 설명 보상이 드물고 지연되는 아타리 게임인 '몬테주마의 복수'를 예로 **들 어 보 겠 습 니 다 .** 이 게임(그림 5(a))에서는 플레이어가 탐험가(빨간색)를 통해 여러 방을 탐색하면서 보물을 수집해야 합니다. 그림의 오른쪽 상단과 왼쪽 상단 모서리에 있는 문을 통과하려면 플레이어는 먼저 열쇠를 집어 들어야 합니다. 그런 다음 플레이어는 오른쪽 사다리를 타고 내려와 열쇠를 향해 왼쪽으로 이동해야 하며, 열쇠를 수집한 보상(+100)을 받기까지 긴 일련의 동작을 수행해야 합니다. 그 후 문을 향해 이동하여 문을 열면 또 다른 보상(+300)을 받을 수 있습니다.

기존의 딥러닝 접근 방식은 에이전트가 보상이 0이 아닌 상태에 도달하는 경우가 거의 없기 때문에 이러한 환경에서는 학습에 실패합니다. 예를 들어, 기본 DQN [28]은 0점을 달성하는

반면, 최고 성능의 시스템인 Gorila DQN [30]은 평균 4.16점만 관리합니다.

설정 에 이 전 트 가 스스로 열쇠를 얻을 수 있는 이점을 학습하기 전에 장면의 의미 있는 부분을 탐색할 수 있는 내재적 동기가 필요합니다. 발달 심리학 문헌[43]과 객체 지향 MDP[8]에서 영감을 받아 장면의 엔티티 또는 객체를 사용하여 이 환경에서 목표를 매개변수화합니다. 시각적 장면에서 객체를 비지도 방식으로 감지하는 것은 컴퓨터 비전에서 해결되지 않은 문제이지만, 최근 이미지 또는 모션 데이터에서 직접 객체를 얻는 데 진전이 있었습니다[10, 9, 14]. 이 연구에서는 그럴듯한 객체 후보를 제공하는 맞춤형 객체 감지기를 구축했습니다. 컨트롤러와

(a) 

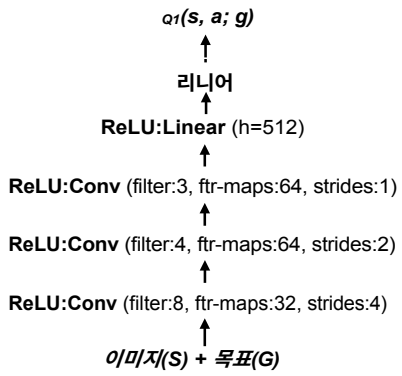
(b) 

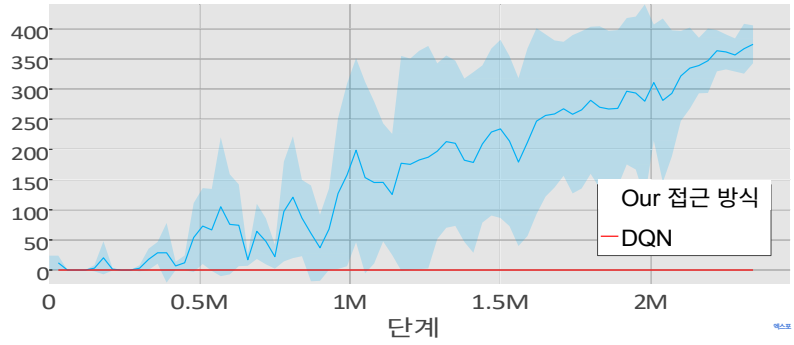
그림 5: (a) '몬테주마의 복수'라는 ATARI 2600 게임의 샘플 화면. (b) **아키텍처**: 컨트롤러용 DQN 아키텍처(Q_1). 유사한 아키텍처가 메타 컨트롤러용 Q_2 를 생성합니다(입력으로 목표가 없음). 실제로 이 두 네트워크는 하위 수준의 기능을 공유할 수 있지만 이를 강제하지는 않습니다.

메타 컨트롤러는 원시 픽셀 데이터로부터 표현을 학습하는 컨볼루션 신경망(그림 5(b) 참조)입니다. 아케이드 학습 환경[2]을 사용하여 실험을 수행합니다.

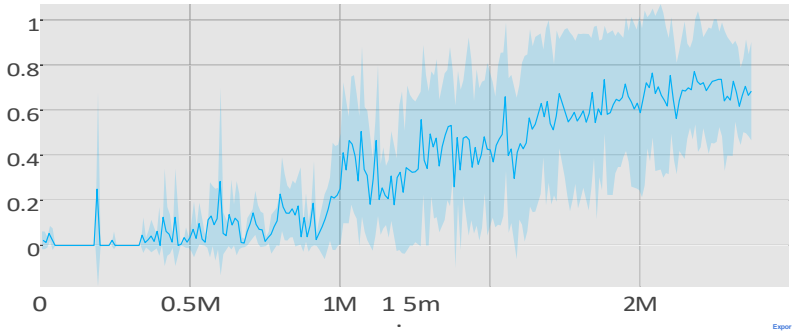
내부 비평자는 ⟨실체$_1$, 관계, 실체$_2$ ⟩의 공간에 정의되며, 여기서 관계는 실체의 구성에 대한 함수입니다. 실험에서 에이전트는 어떤 엔티티$_2$ 를 자유롭게 선택할 수 있습니다. 예를 들어 에이전트 엔티티가 *문과* 같은 다른 엔티티에 *도달하면 에이전트가* 목표를 완료한 것으로 간주하고 보상을 받습니다. 이러한 관계형 내재적 보상 개념은 다른 설정에도 일반화할 수 있습니다. 예를 들어, 아타리 게임 '소행성'에서는 총알이 소행성에 **도달할** 때 에이전트가 보상을 받거나 단순히 우주선이 소행성에 **도달하지** 않을 경우 에이전트가 보상을 받을 수 있습니다. '팩맨' 게임에서는 화면의 펠릿에 **도달하면** 에이전트가 보상을 받을 수 있습니다. 가장 일반적인 경우, 엔티티가 주어지면 모델이 매개변수화된 내재적 보상 함수를 진화시키도록 할 수 있습니다. 이 부분은 향후 작업을 위해 남겨두겠습니다.

모델 아키텍처 및 학습 그림 5b에서 볼 수 있듯이 모델은 정류된 선형 단위(ReLU)가 있는 스택 컨볼루션 레이어로 구성됩니다. 메타 컨트롤러에 대한 입력은 84×84 크기의 연속된 이미지 4개 세트입니다. 메타 컨트롤러의 목표 출력을 인코딩하기 위해 원본 4개의 연속 프레임과 함께 이미지 공간에서 목표 위치의 바이너리 마스크를 추가합니다. 이 증강 입력은 컨트롤러로 전달됩니다. 경험 재생 메모리 D_1 및 D_2 는 각각 1E6 및 5E4로 설정되었습니다. 학습률은 2.5E-4, 할인율은 0.99로 설정했습니다. (1) 첫 번째 단계에서는 메타 컨트롤러의 탐색 파라미터 ϵ_2 를 1로 설정하고 컨트롤러에 동작을 학습시킵니다. 이렇게 하면 컨트롤러가 목표의 하위 집합을 해결하는 방법을 학습할 수 있도록 효과적으로 사전 훈련할 수 있습니다. (2) 두 번째 단계에서는 컨트롤러와 메타 컨트롤러를 공동으로 훈련합니다.

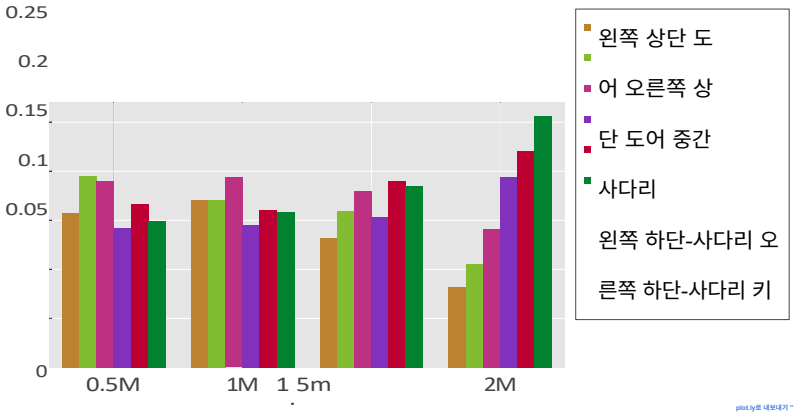
결과 그림 6(a)는 공동 훈련 단계의 보상 진행 상황을 보여 주며, 이 단계부터 모델이 점차적으로 열쇠에 도달하고 문을 열어 에피소드당 약 +400의 보상을 받는 방법을 학습하기 시작했음을 알 수 있습니다. 그림 6(b)에서 볼 수 있듯이 에이전트는 훈련이 진행됨에 따라 열쇠를 더 자주 선택하는 방법을 학습하고 열쇠에 도달하는 데도 성공합니다. 훈련이 진행됨에 따라 에이전트는 먼저 더 간단한 목표(예: 오른쪽 문 또는 중간 사다리 도달)를 수행하는 방법을 학습한 다음 더 높은 보상을 제공하는 열쇠 및 하단 사다리와 같은 '더 어려운' 목표를 천천히 학습하기 시작하는 것을 관찰할 수 있습니다. 그림 6(c)는 선택된 목표의 성공률의 변화를 보여줍니다. 훈련이 끝날수록 '열쇠', '왼쪽 하단 사다리', '오른쪽 하단 사다리'가 점점 더 자주 선택되는 것을 볼 수 있습니다. 전체 게임을 풀 수 있도록 확장하기 위해서는 우리가 고려한 목표 매개변수화를 지원하기 위해 동영상에서 객체를 자동으로 검색하는 기능, 유연한 단기 메모리, 진행 중인 옵션을 간헐적으로 종료하는 기능 등 몇 가지 핵심 요소가 누락되었습니다.



(a) 총 외적 보상



(b) 목표 '핵심' 달성 성공률



(c) 시간 경과에 따른 다양한 목표의 성공률

그림 6: **몬테주마의 복수에 대한 결과:** 이 그림은 모델의 공동 훈련 단계를 보여줍니다. 4.2절에서 설명한 대로 첫 번째 훈련 단계에서는 하위 레벨 컨트롤러를 약 230만 걸음에 대해 사전 훈련합니다. 공동 훈련은 **(a)**에 표시된 것처럼 추가로 2백만 걸음을 걸은 후 지속적으로 높은 보상을 받는 방법을 학습합니다. **(b) 목표 성공률:** 에이전트는 훈련이 진행됨에 따라 키를 더 자주 선택하는 방법을 학습하고 목표 달성에 성공합니다. **(c) 목표 통계:** 공동 훈련의 초기 단계에서는 탐색이 많기 때문에 모든 목표가 똑같이 선호되지만 훈련이 진행됨에 따라 에이전트는 열쇠 및 왼쪽 하단 문과 같은 적절한 목표를 선택하는 방법을 학습합니다.

또한 그림 7에는 에이전트를 사용하여 테스트 실행한 스크린샷(엡실론이 0.1로 설정됨)과 실행 샘플 애니메이션이 나와 있습니다.[1]

[1] '몬테주마의 복수'에서 실행한 샘플 궤적 - https://goo.gl/3Z64Ji

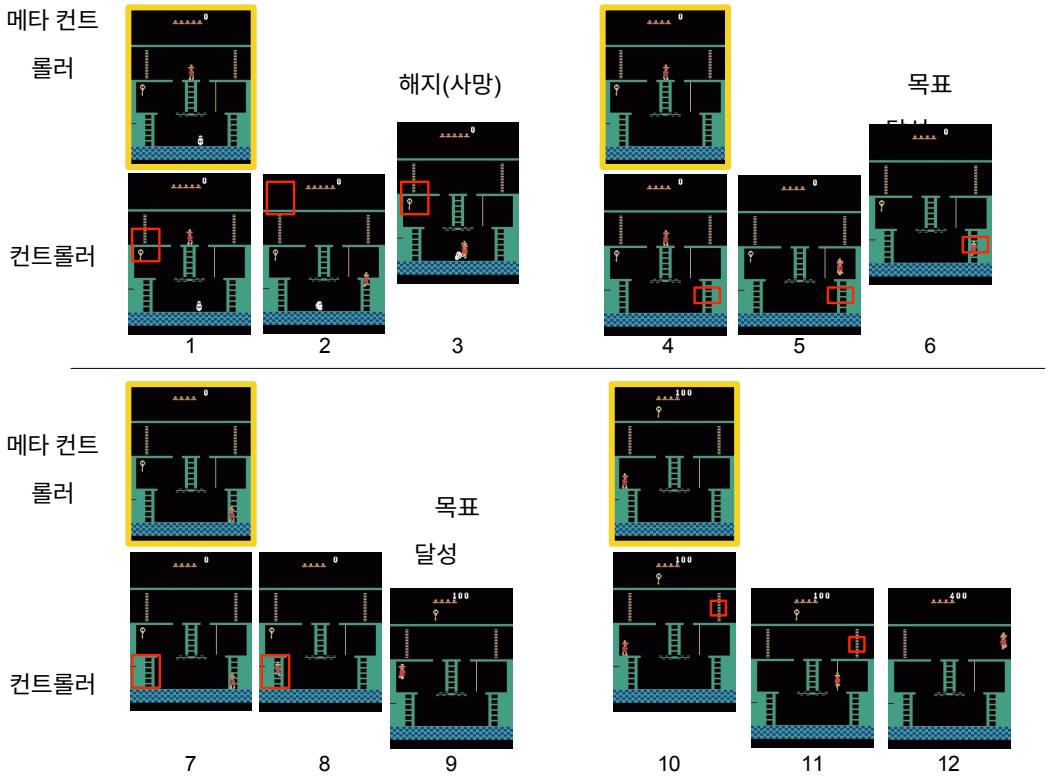


그림 7: **몬테주마의 복수에서 에이전트의 게임플레이 샘플:** 네 개의 사분면이 시간적으로 일관된 방식으로 배열되어 있습니다(왼쪽 상단, 오른쪽 상단, 왼쪽 하단, 오른쪽 하단). 처음에 메타 컨트롤러는 키를 목표로 선택합니다(*빨간색* 그림). 그런 다음 컨트롤러는 일련의 낮은 수준의 동작(일부만 표시됨)을 수행하여 이 목표를 충족하려고 시도하지만 두개골과의 충돌로 인해 실패합니다(에피소드는 여기서 종료됨). 그러면 메타 컨트롤러는 다음 목표로 오른쪽 하단의 사다리를 선택하고 컨트롤러는 사다리에 도달한 후 종료됩니다. 그 후 메타 컨트롤러가 열쇠와 오른쪽 상단 문을 선택하면 컨트롤러는 이 두 가지 목표를 모두 성공적으로 달성할 수 있습니다.

5 결론

유니티는 다양한 시간 척도에서 작동하는 계층적 가치 함수로 구성된 프레임워크인 h-DQN을 제시했습니다. 가치 함수를 시간적으로 분해하면 에이전트가 내재적 동기가 부여된 행동을 수행할 수 있으며, 이는 보상이 지연되는 환경에서 효율적인 탐험을 가능하게 합니다. 또한 엔티티와 관계의 공간에서 내재적 동기를 매개변수화하는 것이 시간적으로 확장된 탐색 기능을 갖춘 에이전트를 구축하는 데 유망한 방법을 제공한다는 것을 관찰했습니다. 또한 향후에는 h-DQN으로 목표의 다른 매개변수화를 탐색할 계획입니다.

현재 프레임워크에는 원시 픽셀에서 객체를 자동으로 분리하는 기능과 단기 메모리 등 몇 가지 누락된 구성 요소가 있습니다. 바닐라 딥-Q 네트워크가 학습한 상태 추상화는 구조화되어 있지 않거나 충분히 구성적이지 않습니다. 최근 심층 생성 모델을 사용하여 픽셀 데이터에서 여러 가지 변형 요소(물체, 포즈, 위치 등)를 분리하는 연구[9, 14, 33, 22, 50, 15, 20]가 진행되었습니다. 저희의 연구가 이미지의 심층 생성 모델과 h-DQN의 결합에 동기를 부여할 수 있기를 바랍니다. 또한 더 긴 범위의 종속성을 처리하기 위해 에이전트는 이전 목표, 동작 및 표현의 이력을 저장해야 합니다. 최근 강화 학습과 함께 순환 네트워크를 사용하는 연구가 일부 진행되었습니다[18, 31]. 마르코비안이 아닌 더 어려운 환경으로 접근 방식을 확장하려면 유연한 에피소드 메모리 모듈을 통합해야 합니다.

감사

중요한 피드백과 토론을 제공해 주신 Vaibhav Unhelkar, Ramya Ramakrishnan, Sam Gershman, Michael Littman, Vlad Firoiu, Will Whitney, Max Kleiman-Weiner, Pedro Tsividis에게 감사의 말씀을 드립니다. 뇌, 기계 및 마음 센터(NSF STC 어워드 CCF - 1231216)와 MIT OpenMind 팀의 지원을 받은 것에 대해 감사드립니다.

참조

[1] A. G. 바르토와 S. 마하데반. 계층적 강화 학습의 최근 발전. *이산 이벤트 동적 시스템*, 13(4):341-379, 2003.

[2] M. G. 벨마레, Y. 나다프, J. 베네스, M. 볼링. 아케이드 학습 환경: 일반 에이전트를 위한 평가 플랫폼. *인공 지능 연구 저널*, 2012.

[3] M. M. Botvinick, Y. Niv, 및 A. C. Barto. 계층적으로 조직된 행동과 그 신경적 기초: 강화 학습의 관점. *인지*, 113(3):262-280, 2009.

[4] L. C. Cobo, C. L. Isbell, 및 A. L. Thomaz. 자율 에이전트를위한 객체 중심 q- 러닝. *2013 자율 에이전트 및 다중 에이전트 시스템에 관한 국제 컨퍼런스 논문집*, 1061-1068 페이지. 자율 에이전트 및 다중 에이전트 시스템을 위한 국제 재단, 2013.

[5] P. Dayan. 시간적 차이 학습을 위한 일반화 개선: 후속 표현. *신경 계산*, 5(4):613-624, 1993.

[6] P. 다얀과 G. E. 힌튼. 봉건 강화 학습. *신경 정보 처리 시스템의 발전*, 271-271 페이지. 모건 카우프만 출판사, 1993.

[7] T. G. Dietterich. maxq 값 함수 분해를 사용한 계층적 강화 학습. *J. Artif. Intell. Res.(JAIR)*, 13:227-303, 2000.

[8] C. Diuk, A. Cohen, and M. L. Littman. 효율적인 강화 학습을 위한 객체 지향 표현. *제 25 회 기계 학습 국제 컨퍼런스 논문집*, 240-247쪽. ACM, 2008.

[9] S. Eslami, N. Heess, T. Weber, Y. Tassa, K. Kavukcuoglu, 및 G. E. Hinton. 참석, 추론, 반복: 생성 모델을 사용한 빠른 장면 이해. *arXiv 사전 인쇄물 arXiv:1603.08575*, 2016.

[10] K. Fragkiadaki, P. Arbelaez, P. Felsen, 및 J. Malik. 비디오에서 움직이는 물체를 분할하는 법 배우기. *컴퓨터 비전 및 패턴 인식(CVPR), 2015 IEEE 컨퍼런스,* 4083-4090페이지. IEEE, 2015.

[11] M. Frank, J. Leitner, M. Stollenga, A. F¨orster, 및 J. Schmidhuber. 휴머노이드의 동작 계획을 위한 호기심 기반 재강화 학습. *동물, 인간, 로봇의 내재적 동기와 개방형 개발*, 245페이지, 2015.

[12] S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, 및 P. B. Sederberg. 후속 표현과 시간적 맥락. *신경 계산*, 24(6):1553-1568, 2012.

[13] S. Goel and M. Huber. 학습된 정책을 사용한 계층적 강화 학습을 위한 하위 목표 발견. *FLAIRS 컨퍼런스*, 346-350페이지, 2003.

[14] K. Greff, R. K. 스리바스타바, 및 J. 슈미드후버. 재구성 클러스터링을 통한 바인딩. *arXiv 사전 인쇄물 arXiv:1511.06418*, 2015.

[15] K. 그레고르, I. 다니헬카, A. 그레이브스, D. 비에르스트라. Draw: 이미지 생성을 위한 순환 신경망. *arXiv 사전 인쇄물 arXiv:1502.04623*, 2015.

[16] C. Guestrin, D. Koller, C. Gearhart, 및 N. Kanodia. 관계형 MDPS에서 새로운 환경에 대한 계획 일반화. *제18회 인공 지능 국제 공동 컨퍼런스 논문집*, 1003-1010페이지. 모건 카우프만 출판사, 2003.

[17] C. Guestrin, D. Koller, R. Parr, 및 S. Venkataraman. 인수분해 MDPS를 위한 효율적인 솔루션 알고리즘. *인공 지능 연구 저널*, 399-468 페이지, 2003.

[18] M. Hausknecht and P. Stone. 부분적으로 관찰 가능한 mdps를 위한 심층 반복 q-학습. *arXiv 사전 인쇄물 arXiv:1507.06527*, 2015.

[19] N. 에르난데스-가르디올과 S. 마하데반. 계층적 메모리 기반 강화 학습.
신경 정보 처리 시스템의 발전, 1047-1053페이지, 2001.

[20] J. Huang and K. Murphy. 이미지의 오클루전 인식 생성 모델에서 효율적인 추론.
arXiv 사전 인쇄물 arXiv:1511.06362, 2015.

[21] J. Koutnık, J. Schmidhuber, 및 F. Gomez. 비전 기반 강화 학습을 위한 진화하는 심층 비지도 컨볼루션 네트워크. *유전 및 진화 계산에 관한 2014 컨퍼런스 논문집*, 541-548 페이지. ACM, 2014.

[22] T. D. 쿨카르니, W. F. 휘트니, P. 콜리, 및 J. 테넨바움. 심층 컨볼루션 역 그래픽 네트워크. *신경 정보 처리 시스템의 발전*, 2530-2538 페이지, 2015.

[23] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, 및 S. J. Gershman. 사람처럼 학습하고 생각하는 기계 구축. *arXiv 사전 인쇄물 arXiv:1604.00289*, 2016.

[24] M. C. 마차도와 M. 볼링. 보상이 없는 상황에서 의도적인 행동 학습. *arXiv 사전 인쇄물 arXiv:1605.07700*, 2016.

[25] S. 매너, I. 메나체, A. 호제, U. 클라인. 클러스터링을 통한 강화 학습의 동적 추상화. *기계 학습에 관한 제 21 회 국제 컨퍼런스 논문집,* 71 페이지. ACM, 2004.

[26] A. 맥거번과 A. G. 바토. 다양한 밀도를 사용한 강화 학습에서 하위 목표의 자동 발견. *컴퓨터 과학과 교수 출판물 시리즈*, 8 페이지, 2001.

[27] I. 메나치, S. 매너, N. 심킨. 강화 학습에서 하위 목표의 Q- 컷 동적 발견. In *Machine Learning: ECML 2002*, 295-306 페이지. Springer, 2002.

[28] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski 외. 심층 강화 학습을 통한 인간 수준의 제어. *Nature*, 518(7540):529-533, 2015.

[29] S. 모하메드와 D. J. 레젠데. 내재적 동기 강화 학습을 위한 변형 정보 최대화. *신경 정보 처리 시스템의 발전*, 2116-2124쪽, 2015.

[30] A. Nair, P. 스리니바산, S. 블랙웰, C. 알시첵, R. 피론, A. 드 마리아, V. 판네르셸밤, M. Suleyman, C. Beattie, S. Petersen 외. 심층 강화 학습을 위한 대규모 병렬 방법. *arXiv preprint arXiv:1507.04296*, 2015.

[31] K. 나라심한, T. 쿨카르니, 및 R. 바질레이. 심층 강화 학습을 이용한 텍스트 기반 게임의 언어 이해. *arXiv 사전 인쇄물 arXiv:1506.08941*, 2015.

[32] I. 오스밴드, C. 블런델, A. 프리첼, B. 반 로이. 부트스트랩된 dqn을 통한 심층 탐색. *arXiv 사전 인쇄물 arXiv:1602.04621*, 2016.

[33] D. J. 레젠데, S. 모하메드, I. 다니헬카, K. 그레고르, D. 비어스트라. 심층 생성 모델에서의 원샷 일반화. *arXiv 사전 인쇄물 arXiv:1603.05106*, 2016.

[34] T. Schaul, D. Horgan, K. Gregor, 및 D. Silver. 범용 가치 함수 근사화. *제32회 국제 기계 학습 컨퍼런스(ICML-15) 논문집*, 1312-1320페이지, 2015.

[35] T. 숄, J. 콴, I. 안토노글루, 및 D. 실버. 우선 순위가 지정된 경험 재생. *arXiv 사전 인쇄물 arXiv:1511.05952*, 2015.

[36] J. 슈미트후버. 창의성, 재미, 내재적 동기에 대한 공식 이론(1990-2010). *자율적 정신 발달, IEEE 트랜잭션 온*, 2(3):230-247, 2010.

[37] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot 외. 심층 신경망과 트리 검색으로 바둑 게임 마스터하기. *Nature*, 529(7587):484-489, 2016.

[38] Ö. Şimşek, A. Wolfe, and A. Barto. 로컬 그래프 분할을 통한 강화 학습에서 유용한 하위 목표 식별. *기계 학습에 관한 국제 컨퍼런스 논문집*, 816-823쪽, 2005.

[39] S. Singh, R. L. Lewis, 및 A. G. Barto. 보상은 어디에서 오는가. *인지과학회 연례 학술대회 논*

문집, 2601-2606쪽, 2009.

[40] S. Singh, R. L. Lewis, A. G. Barto, 및 J. Sorg. 내재적 동기 강화 학습: 진화론적 관점. *자율적 정신 발달, IEEE 트랜잭션,* 2(2):70- 82, 2010.

[41] S. P. Singh, A. G. Barto, 및 N. Chentanez. 내재적 동기 강화 학습. In *신경 정보 처리 시스템의 발전*, 1281-1288페이지, 2004.

[42] J. 소그와 S. 싱. 선형 옵션. *제9회 자율 에이전트 및 다중 에이전트 시스템에 관한 국제 컨퍼런스 논문집: 1권 - 1권*, AAMAS '10, 31-38페이지, 리치랜드, SC, 2010. 자율 에이전트 및 다중 에이전트 시스템을 위한 국제 재단.

[43] E. 스펠케와 킨츨러. 핵심 지식. *발달 과학*, 10(1):89-96, 2007.

[44] K. L. 스타첸펠드, M. 보트비닉, S. J. 거쉬만. 해마 인지 맵의 설계 원리. *신경 정보 처리 시스템의 발전*, 2528-2536 페이지, 2014.

[45] B. C. Stadie, S. Levine, and P. Abbeel. 심층 예측 모델을 통한 강화 학습의 탐색 인센티브. *arXiv 사전 인쇄물 arXiv:1507.00814*, 2015.

[46] R. S. Sutton and A. G. Barto. *강화 학습 소개*, 135 권. MIT Press Cambridge, 1998.

[47] R. S. 서튼, J. 모다일, M. 델프, T. 데그리스, P. M. 필라르스키, A. 화이트, D. 프레컵. Horde: 비지도 감각 운동 상호 작용으로부터 지식을 학습하기 위한 확장 가능한 실시간 아키텍처. *제10회 자율 에이전트 및 다중 에이전트 시스템에 관한 국제 컨퍼런스-제2권*, 761-768페이지. 자율 에이전트 및 다중 에이전트 시스템을 위한 국제 재단, 2011.

[48] R. S. 서튼, D. 프레컵, 및 S. 싱. MDPS와 세미 MDPS 사이: 강화 학습에서 시간적 추상화를 위한 프레임워크. *인공 지능*, 112(1):181-211, 1999.

[49] C. Szepesvari, R. S. Sutton, J. Modayil, S. Bhatnagar 등 . 범용 옵션 모델. In *신경 정보 처리 시스템의 발전*, 990-998페이지, 2014.

[50] W. F. 휘트니, M. 창, T. 쿨카르니, J. B. 테넌바움. 지속적 학습을 통한 시각적 개념 이해. *아카이브 프리프린트 arXiv:1602.06822*, 2016.