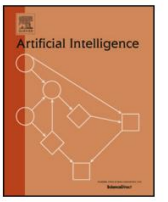
인공지능 299 (2021) 103535

ScienceDirect 에서 사용 가능한 콘텐츠 목록

인공 지능

www.elsevier.com/locate/artint



보상은 충분하다

데이비드 실버 *, 새틴더 싱, 도이나 프리컵, 리처드 S. 서튼

기사 정보

기사 기록:
2020년 11월 12일 접수
2021년 4월 28일 수정된 양식으로 접수됨
2021년 5월 12일 수락됨
2021년 5월 24일 온라인 사용 가능

키워드:
인공 지능
인공 일반 지능
강화 학습
보상

요약

이 기사에서 우리는 지능과 그와 관련된 능력이 보상의 최대화에 복종하는 것으로 이해될 수 있다고 가정합니다. 따라서 보상은 지식, 학습, 지각, 사회 지능, 언어, 일반화 및 모방을 포함하여 자연 및 인공 지능에서 연구한 능력을 나타내는 행동을 유도하기에 충분합니다.

이는 다른 신호나 목표를 기반으로 각 능력에 대해 전문화된 문제 공식화가 필요하다는 관점과 대조적입니다. 또한, 시행착오 경험을 통해 보상을 최대화하는 에이전트가 이러한 능력의 전부는 아니지만 가장 많이 나타나는 행동을 학습할 수 있으므로 강력한 강화 학습 에이전트가 인공 일반 지능에 대한 솔루션이 될 수 있음을 제안합니다.

1. 소개

동물과 인간의 행동에서 지능의 표현은 너무나 풍부하고 다양하기 때문에 사회적 지능, 언어, 지각, 지식 표현, 계획, 상상력, 기억, 운동 제어와 같은 이름을 지정하고 연구하는 관련 능력의 온톨로지가 있습니다. 에이전트(자연적 또는 인공적)가 다양한 방식으로 지능적으로 행동하도록 유도할 수 있는 것은 무엇입니까?

한 가지 가능한 대답은 각 능력이 해당 능력을 끌어내기 위해 특별히 설계된 목표를 추구하는 데 발생한다는 것입니다. 예를 들어, 사회 지능의 능력은 종종 다중 에이전트 시스템의 내쉬 균형으로 구성되었습니다. 구문 분석, 품사 태깅, 어휘 분석 및 감정 분석과 같은 목표 조합에 의한 언어 능력; 및 대상 세분화 및 인식에 의한 지각 능력. 이 논문에서 우리는 대안적 가설을 고려합니다. 보상을 최대화하는 일반적인 목표는 자연 및 인공 지능에서 연구되는 모든 능력은 아닐지라도 가장 많이 나타내는 행동을 유도하기에 충분합니다.

이 가설은 지능과 관련된 능력의 순전한 다양성이 일반적인 목표와 상충되는 것처럼 보이기 때문에 놀랄 수 있습니다. 그러나 동물과 인간이 직면하는 자연 세계, 그리고 아마도 미래에 인공 요원이 직면하는 환경은 본질적으로 너무 복잡하기 때문에 그러한 환경에서 성공(예: 생존)하기 위해서는 정교한 능력이 필요합니다. 따라서 보상을 최대화하여 측정한 성공은 지능과 관련된 다양한 능력을 요구합니다. 그러한 환경에서 보상을 극대화하는 모든 행동은 반드시 그러한 능력을 발휘해야 합니다. 이러한 의미에서 보상 극대화의 일반적인 목표는 그 안에 지능의 많은 또는 아마도 모든 목표를 포함합니다.

따라서 보상은 자연에서 발견되는 지능의 풍부한 표현에 대해 두 가지 수준의 설명을 제공합니다. 첫째, 서로 다른 환경에서 서로 다른 보상 신호를 최대화하여 서로 다른 형태의 지능이 발생할 수 있습니다. 예를 들어 박쥐의 반향 정위, 고래 노래에 의한 의사 소통 또는 침팬지의 도구 사용과 같은 독특한 능력이 있습니다. 심-

* 교신 저자.
이메일 주소: davidsilver@google.com (D. 실버).

https://doi.org/10.1016/j.artint.2021.103535

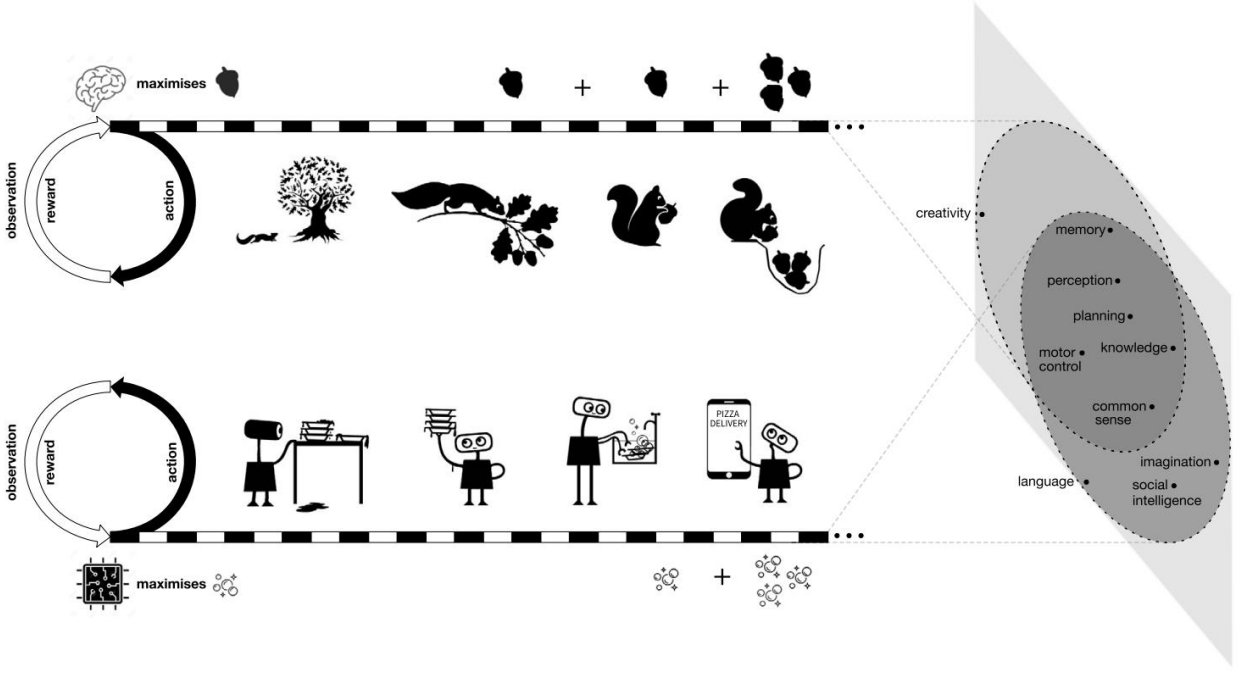


그림 1. 충분한 보상 가설은 지능 및 관련 능력이 환경에서 행동하는 에이전트에 의한 보상의 최대화에 종속되는 것으로 이해될 수 있다고 가정합니다 . 예를 들어, 다람쥐는 음식 소비를 극대화하기 위해 행동하고(위, 도토리 기호로 표시된 보상), 주방 로봇은 청결을 극대화하기 위해(아래, 거품 기호로 표시된 보상) 역할을 합니다. 이러한 목표를 달성하려면 지능과 관련된 다양한 능력을 나타내는 복잡한 행동이 필요합니다(오른쪽에는 에이전트의 경험 흐름에서 해당 경험 내에서 표현된 일련의 능력에 대한 투영으로 표시됨).

마찬가지로 인공 에이전트는 미래 환경에서 다양한 보상 신호를 최대화하기 위해 필요할 수 있으며, 그 결과 레이저 기반 탐색, 이메일을 통한 통신 또는 로봇 조작과 같은 독특한 능력을 가진 새로운 형태의 지능이 탄생할 수 있습니다.

둘째, 동물이나 인간의 지능은 능력의 풍요와 관련이 있습니다. 우리의 가설에 따르면, 이러한 모든 능력 은 환경 내에서 해당 동물이나 에이전트의 보상을 최대화 하는 단일 목표 에 종속됩니다.
즉, 하나의 목표를 추구하는 것은 지능과 관련된 여러 능력을 나타내는 복잡한 행동을 생성할 수 있습니다. 실제로, 그러한 보상 극대화 행동은 종종 각 능력과 관련된 별도의 목표 추구에서 파생된 특정 행동과 일치할 수 있습니다.

예를 들어, 다람쥐의 두뇌는 다람쥐의 몸에서 감각을 받고 운동 명령을 보내는 의사 결정 시스템으로 이해될 수 있습니다. 다람쥐의 행동은 포만감(즉, 부정적인 배고픔)과 같은 누적 보상을 극대화하는 것으로 이해될 수 있다. 다람쥐가 굶주림을 최소화하기 위해 다람쥐 두뇌는 아마도 지각(좋은 견과류 식별), 지식(견과류 이해), 운동 제어(견과류 수집), 계획(견과류 보관 장소 선택) 능력이 있어야 합니다.), 메모리(캐시된 견과류의 위치를 기억하기 위해) 및 소셜 인텔리전스(캐시된 견과류의 위치에 대해 허세를 부리기 위해 도난당하지 않도록 하기 위해). 따라서 지능과 관련된 이러한 각 능력은 기아 최소화라는 단일 목표를 수행하는 것으로 이해될 수 있습니다(그림 1 참조).

두 번째 예로서, 주방 로봇은 로봇의 신체로부터 감각을 수신하고 로봇의 신체에 액추에이터 명령을 보내는 의사 결정 시스템으로 구현될 수 있습니다. 주방 로봇의 유일한 목표는 청결도를 측정하는 보상 신호를 극대화하는 것입니다.1 주방 로봇이 청결을 극대화하기 위해서는 지각(깨끗한 식기와 더러운 식기를 구별하는 능력), 지식(그릇을 이해하는 능력), 운동 제어(기구를 조작하기 위해), 기억(기구의 위치를 기억하기 위해), 언어(대화에서 미래 혼란을 예측하기 위해) 및 사회적 지능(어린 아이들이 덜 엉망으로 만들도록 격려하기 위해). 따라서 청결을 최대화하는 행동은 그 단일 목표에 봉사하기 위해 이러한 모든 능력을 발휘해야 합니다(그림 1 참조).

지능과 관련된 능력이 보상 극대화라는 단일 목표에 대한 솔루션으로 발생하면 그러한 능력이 발생하는 이유 를 설명하기 때문에 실제로 더 깊은 이해를 제공할 수 있습니다 (예: 악어를 분류하는 것은 먹히지 않도록 하는 것이 중요합니다). 대조적으로, 각 능력이 고유한 특수 목표에 대한 솔루션으로 이해될 때 그 능력이 하는 일 (예: 통나무에서 악어 구별) 에 초점을 맞추기 위해 이유 질문은 생략됩니다. 더욱이, 단일 목표는 또한 사회적 지능에서 비합리적인 행위자 다루기(예: 화난 공격자 진정시키기), 언어를 지각적 경험에 기초하기(예: 대화)와 같이 공식화하기 어려운 특성을 포함할 수 있는 각 능력에 대한 더 넓은 이해를 제공할 수 있습니다. 과일 껍질을 벗기는 가장 좋은 방법) 또는 지각으로 햅틱 이해하기

[1] 예를 들어, 사람이 가끔 검사하여 판단합니다.

(예: 주머니에서 날카로운 물건 줍기). 마지막으로, 자신의 전문적인 목표가 아닌 단일 목표를 위해 능력을 실행하는 것은 능력을 통합하는 방법에 대한 질문에 대한 답이기도 합니다.

보상 극대화가 지능 문제를 이해하는 데 적합한 목표임을 확인한 후 문제 해결 방법을 고려할 수 있습니다. 그런 다음 자연 지능에서 그러한 방법을 찾거나 인공 지능에서 구현하기로 선택할 수 있습니다. 보상을 극대화할 수 있는 가능한 방법 중 가장 일반적이고 확장 가능한 접근 방식은 시행착오를 통해 환경과 상호 작용하여 보상을 얻는 방법을 배우는 것입니다. 이러한 방식으로 보상을 극대화하는 방법을 효과적으로 학습할 수 있는 에이전트는 풍부한 환경에 배치될 때 일반 지능의 정교한 표현을 발생시킬 것이라고 추측합니다.

보상 극대화를 기반으로 한 문제와 솔루션의 최근 유익한 예는 바둑 게임에서 비롯됩니다. 연구는 처음에 오프닝, 모양, 전술 및 최종 게임과 같은 별개의 능력에 주로 초점을 맞추었으며, 각각은 시퀀스 암기, 패턴 인식, 로컬 검색 및 조합 게임 이론과 같은 별개의 목표를 사용하여 형식화되었습니다 [32].

AlphaZero [49] 는 대신 단일 목표에 중점을 두었습니다. 즉, 최종 단계까지 0인 보상 신호를 최대화한 다음 승리 시 +1 또는 패배 시 -1인 보상 신호를 최대화합니다. 이는 궁극적으로 각 능력에 대한 더 깊은 이해로 이어졌습니다. 예를 들어, 새로운 시작 시퀀스 발견 [65], 글로벌 컨텍스트 내에서 놀라운 모양 사용 [40], 로컬 전투 간의 글로벌 상호 작용 이해 [64], 앞서 있을 때 안전한 플레이 [40]]. 또한 영향력과 영토, 두께와 가벼움, 공격과 방어의 균형을 맞추는 것과 같이 이전에 만족스럽게 공식화되지 않은 광범위한 능력 세트를 산출했습니다. Alp haZero의 능력은 선천적으로 통합된 전체로 통합된 반면 통합은 이전 작업에서 매우 문제가 많은 것으로 입증되었습니다 [32]. 따라서 바둑과 같은 단순한 환경에서 승리를 극대화하는 것만으로도 다양한 전문 능력을 발휘하는 행동을 유도하기에 충분하다는 것이 입증되었습니다. 또한 체스나 장기 [48] 와 같은 다른 환경에 동일한 방법을 적용하면 말의 이동성 및 색상 복합체 [44] 와 같은 새로운 능력이 생겨났습니다 . 우리는 동물과 인간이 직면한 자연 세계에 비해 복잡성이 더 유사한 더 풍부한 환경에서 보상을 최대화하면 더 많은, 그리고 아마도 궁극적으로 지능과 관련된 모든 능력을 얻을 수 있다고 주장합니다.

이 문서의 나머지 부분은 다음과 같이 구성됩니다. 2 장에서는 강화학습의 문제로 보상극대화의 목적을 공식화한다. 3 장에서는 주요 가설을 제시합니다. 우리는 지능과 관련된 몇 가지 중요한 능력을 고려하고 보상 극대화가 이러한 능력을 어떻게 산출할 수 있는지 논의합니다. 섹션 4 에서는 솔루션 전략으로 보상 극대화를 사용합니다. 섹션 5 에서 관련 작업을 제시 하고 마지막으로 섹션 6 에서 가설의 가능한 약점에 대해 논의하고 몇 가지 대안을 고려합니다.

2. 배경: 강화 학습 문제

지능은 목표를 달성하는 유연한 능력으로 이해될 수 있습니다. 예를 들어 John McCarthy에 따르면 "지능은 세상에서 목표를 달성하는 능력의 계산적인 부분입니다" [29]. 강화 학습 [56] 은 목표 추구 지능의 문제를 공식화합니다. 일반적인 문제는 다양한 환경에서 최대화하기 위해 다양한 보상 신호에 해당하는 광범위하고 현실적인 목표와 세계, 따라서 광범위한 형태의 지능으로 인스턴스화할 수 있습니다.

2.1. 에이전트 및 환경

인공 지능에 대한 많은 대화형 접근 방식 [42] 과 마찬가지로 강화 학습 은 문제 를 시간이 지남에 따라 순차적으로 상호 작용하는 두 시스템으로 분리하는 프로토콜을 따릅니다 . 그 결정들. 이것은 예를 들어 다중 에이전트, 다중 환경 또는 기타 상호 작용 모드를 고려할 수 있는 다른 특수 프로토콜과 대조됩니다.

2.2. 대리인

에이전트 는 시간 t 에서 관찰 Ot 를 수신하고 작업 At 를 출력하는 시스템 입니다. 더 형식적으로, 에이전트는 경험 이력 Ht = O1, A1,..., Ot 1, At 1, Ot 가 주어지면 시간 t 에서 행동을 선택하는 시스템 At = α(Ht) 입니다 . 에이전트와 환경 간의 상호 작용 기록에서 발생한 일련의 관찰 및 작업이 주어집니다.

에이전트 시스템 α 는 실제 제약으로 인해 제한된 집합으로 제한됩니다 [43]. 에이전트는 기계에 의해 결정되는 제한된 용량을 갖습니다(예: 컴퓨터의 제한된 메모리 또는 뇌의 제한된 뉴런). 에이전트 및 환경 시스템은 실시간으로 실행됩니다. 에이전트가 다음 작업을 계산하는 데 시간을 소비하는 동안(예: 사자로부터 도망할지 여부를 결정하는 동안 무작동 작업 생성) 환경 시스템은 계속 처리합니다(예: 사자 공격). 따라서 강화학습 문제는 계산적 한계를 무시하는 이론적 추상화라기보다는 자연지능과 인공지능이 직면한 실제적인 문제를 나타낸다.

이 문서는 에이전트의 본질을 탐구하지 않고 대신 해결해야 하는 문제와 인텔리전스에 중점을 둡니다.
그것은 그 문제에 대한 어떤 해결책에 의해 유도될 수 있습니다.

1 번 테이블
환경의 정의는 광범위하고 많은 문제 차원을 포함합니다.

치수	대안 A	대안 B	노트
관찰	이산	마디 없는	
행위	이산	마디 없는	
시간	이산	마디 없는	시간 단계는 극소일 수 있습니다.
역학	결정론적	스토캐스틱	
관찰 가능성	가득한	부분	
대행사	단일 에이전트	다중 에이전트	다른 에이전트는 단일 에이전트의 관점에서 환경의 일부입니다(섹션 3.3 참조).
불확실성	확실한	불확실한	불확실성은 확률적 초기 상태 또는 전환으로 나타낼 수 있습니다.
종료	계속	에피소드	환경이 종료되고 초기 상태로 재설정될 수 있음
정지성	변화 없는	비정상	환경은 역사에 의존하고 따라서 시간에도 의존한다
동시성	비동기	동기	조치가 실행될 때까지 관찰은 변경되지 않은 상태로 유지될 수 있습니다.
현실	시뮬레이션	현실 세계	에이전트와 상호 작용하는 인간을 포함할 수 있음

2.3. 환경

환경 은 시간 t 에서 행동을 수신 하고 다음 시간 단계에서 관찰 Ot+1 로 응답 하는 시스템입니다.
더 공식적으로, 환경은 에이전트 가 다음 관찰 Ot+1 을 결정하는 시스템 Ot+1 = ε(Ht, At,ηt) 입니다.
주어진 경험 이력 Ht, 최신 작업 At 및 잠재적으로 임의성의 소스를 환경에서 수신합니다.
ηt.

환경은 정의 내에서 에이전트에 대한 인터페이스를 지정합니다. 대리인은 전적으로 의사 결정 기관으로 구성됩니다. 해당 엔터티 외부의 모든 것(몸체가 있는 경우 바디 포함)
은 환경의 일부로 간주됩니다. 이것
에이전트에 대한 관찰과 에이전트가 사용할 수 있는 작업을 정의하는 센서와 액추에이터를 모두 포함합니다.
각기.

환경에 대한 이러한 정의는 매우 광범위하며 다음을 포함하여 많은 문제 차원을 포함합니다.
표 1.

2.4. 보상

강화 학습 문제는 누적 보상으로 목표를 나타냅니다. 보상 은 특수 스칼라 관측값 Rt,
진행 상황의 즉각적인 측정을 제공하는 환경 의 보상 신호 에 의해 모든 시간 단계 t 에서 방출됨
목표를 향해. 강화 학습 문제의 인스턴스는 보상 신호가 있는 환경 ε 으로 정의됩니다.
유한한 수의 단계에 대한 보상의 합계, 할인된 합계 또는
시간 단계당 평균 보상.

다양한 목표를 보상으로 나타낼 수 있습니다.[2] 예를 들어 스칼라 보상 신호는 가중치를
목표의 조합, 시간 경과에 따른 다양한 절충, 위험 추구 또는 위험 회피 유틸리티. 보상은 루프 내 인간(human-in-the-loop)에 의해 결정될 수도 있습니다. 예를 들어 인간은 온라인에서 원하는 행동을 명시적으로 강화할 수 있습니다.
클릭 또는 추천을 통한 피드백, 설문지 또는 설문조사를 통한 지연된 피드백 또는 자연어 발화에 의한 피드백.
사람의 피드백을 포함하면 “내가 언제 알게 될 것인가?
봐".

일반성 외에도 보상은 진행 상황에 대해 잠재적으로 모든 단계에서 중간 피드백을 제공합니다.
목표를 향해. 이 중간 신호는 길거나 무한대를 고려할 때 문제 정의의 필수적인 부분입니다.
경험의 흐름 - 중간 피드백 없이는 학습이 불가능합니다.

3. 보상은 충분하다

우리는 이전 섹션에서 보상이 다양한 목표를 표현하기에 충분하다는 것을 보았습니다.
정보가 지향될 수 있는 다양한 목적.

이제 우리는 다양한 형태의 지능이 이해될 수 있다는 요점을 만들기 위한 모든 요소를 갖추었습니다.
보상의 극대화에 복종하는 것으로, 각 형태의 지능과 관련된 많은 능력이 발생할 수 있음
암묵적으로 그러한 보상을 추구하는 것에서 비롯됩니다. 한계에 다다랐을 때 우리는 모든 지능과 관련 능력이
다음과 같이 이해될 수 있습니다.

가설 (보상은 충분하다). 지능 및 관련 능력은 최대화에 복종하는 것으로 이해될 수 있습니다.
환경에서 행동하는 에이전트에 의한 보상.

이 가설은 중요합니다. 만약 그것이 사실이라면 좋은 보상 극대화 에이전트가
목표를 달성하면 지능과 관련된 능력을 암묵적으로 산출할 수 있습니다. 이 맥락에서 좋은 에이전트는

[2] 실제로 보상 가설 은 모든 목표가 보상으로 표시될 수 있다고 추측합니다 [56]. 이것은 우리의 충분한 보상과 혼동되어서는 안 됩니다.
하나의 그러한 목표를 추구함으로써 암묵적으로 발생하는 능력을 고려하는 가설.

아마도 아직 발견되지 않은 알고리즘을 사용하여 성공적으로 환경에서 누적 보상을 극대화하는 데 능숙합니다. 우리는 섹션 4 에서 그러한 에이전트가 어떻게 구성될 수 있는지에 대한 질문으로 돌아갑니다 .

원칙적으로 모든 행동은 해당 행동을 유도하기 위해 명시적으로 선택된 보상 신호의 최대화에 의해 인코딩될 수 있습니다 [12] (예: 개체가 올바르게 식별되거나 구문적으로 올바른 문장이 생성될 때 보상 제공). 여기에서 가정은 훨씬 더 강력하기 위한 것입니다. 즉, 지능 및 관련 능력은 자연 지능 또는 인공 지능이 지향할 수 있는 많은 실용적인 목표에 해당하는 많은 가능한 보상 신호 중 하나를 최대화하기 위해 암묵적으로 발생합니다.

복잡한 환경에서 단순한 보상의 극대화에서 정교한 능력이 나올 수 있습니다. 예를 들어, 다람쥐의 자연 환경에서 굶주림을 최소화하려면 다람쥐의 근골격 역학 사이의 상호 작용에서 발생하는 견과류를 조작하는 숙련된 능력이 필요합니다. 다람쥐나 견과류가 놓여 있거나 연결되어 있거나 방해를 받을 수 있는 나뭇잎, 가지 또는 흙과 같은 물체; 견과류의 크기와 모양의 변화; 바람, 비 또는 눈과 같은 환경 요인; 노화, 질병 또는 부상으로 인한 변화. 유사하게, 주방 로봇에서 청결을 추구하려면 어수선함, 교합, 눈부심, 외피, 손상 등을 포함하는 광범위한 상태에서 기구를 인식하는 정교한 능력이 필요합니다.

또한, 다람쥐(예: 생존 시간 최대화, 고통 최소화, 번식 성공 최대화) 또는 주방 로봇(예: 건강한 식생활 지수 최대화, 사용자로부터의 긍정적인 피드백 최대화 또는 미식 극대화)에 의한 다른 많은 보상 신호의 최대화 엔돌핀), 그리고 다른 많은 환경(예: 다른 서식지, 다른 민첩한 신체 또는 다른 기후)에서도 지각, 운동, 조작 등의 능력을 낳을 것입니다. 따라서 일반 지능으로 가는 경로는 실제로 보상 신호 선택에 대해 상당히 견고할 수 있습니다. 실제로 지능을 생성하는 능력은 많은 다른 환경에서 다양한 보상 신호를 최대화하면 지능과 관련된 유사한 능력을 생성할 수 있다는 의미에서 종종 주어진 목표와 직교할 수 있습니다.

다음 섹션에서는 이 가설이 강화 학습 문제로 공식화하기 어려워 보이는 몇 가지를 포함하여 다양한 중요한 능력에 실제로 적용될 수 있는지 여부와 방법을 탐구합니다. 우리는 지능과 관련된 모든 능력에 대한 철저한 토론을 제공하지 않지만, 독자가 기억, 상상력, 상식, 주의력, 추론, 창의성 또는 감정과 같은 능력을 고려하고 이러한 능력이 보상의 일반적인 목표를 달성할 수 있는 방법을 고려하도록 권장합니다. 극대화.

3.1. 지식과 배움에 대한 보상은 충분하다

우리는 지식을 에이전트 내부에 있는 정보로 정의합니다. 예를 들어, 지식은 행동 선택, 누적 보상 예측 또는 미래 관찰의 특징 예측을 위한 에이전트 기능의 매개변수 내에 포함될 수 있습니다. 이 지식 중 일부는 선천적(사전 지식)일 수 있지만 일부 지식은 학습을 통해 획득할 수 있습니다.

환경은 타고난 지식을 요구할 수 있습니다. 특히, 총 보상을 최대화하려면 새로운 상황에서 즉시 액세스할 수 있는 지식을 보유하는 것이 필요할 수 있습니다. 예를 들어, 갓 태어난 가젤은 사자에게서 도망쳐야 할 수도 있습니다. 이 경우 이 지식을 배울 기회가 있기 전에 포식자 회피에 대한 타고난 이해가 필요할 수 있습니다. 그러나 사전 지식의 범위는 이론(대행자의 능력)과 실제(유용한 사전 지식 구성의 어려움) 모두에서 제한된다는 점에 유의하십시오. 더욱이 우리가 고려할 다른 능력과 달리 타고난 지식에 대한 환경적 요구는 조작될 수 없습니다. 경험보다 먼저 오는 지식이므로 경험에서 얻을 수 없습니다.

환경은 또한 학습된 지식을 요구할 수 있습니다. 이는 미지의 요소, 확률 또는 환경의 복잡성으로 인해 미래 경험이 불확실할 때 발생합니다. 이러한 불확실성으로 인해 에이전트는 미래 이벤트의 특정 실현에 따라 필요할 수 있는 광범위한 잠재적 지식이 필요합니다. 자연 및 인공 지능에서 가장 어렵고 가장 중요한 많은 문제를 포함하여 풍부한 환경에서 잠재적 지식의 총 공간은 에이전트의 용량보다 훨씬 큽니다. 예를 들어, 초기 인간 환경을 생각해 보십시오. 에이전트는 북극이나 아프리카에서 태어날 수 있으므로 총 보상을 극대화하기 위해 지식에 대한 요구 사항이 근본적으로 다릅니다. 빙하 또는 사바나 횡단; 얼음이나 진흙으로 건물을 짓는 것; 등등. 대리인의 삶에서 일어나는 특별한 사건은 근본적으로 다른 요구로 이어질 수도 있습니다. 사냥이나 농사를 선택하는 것; 메뚜기 또는 전쟁에 직면; 장님이 되거나 귀머거리가 되거나; 친구 또는 적과의 만남; 등. 이러한 가능한 각 삶에서 에이전트는 상세하고 전문적인 지식을 습득해야 합니다. 그러한 잠재적 지식의 합계가 에이전트의 능력을 능가하는 경우 지식은 에이전트의 경험의 함수여야 하고 에이전트의 특정 상황에 적응해야 하므로 학습이 필요합니다. 실제로 이 학습은 매개변수의 적응 또는 구조의 구성, 큐레이션 및 재사용을 통해 예측, 모델 또는 기술을 만드는 것과 같은 많은 계산 형식을 취할 수 있습니다.

요약하면, 환경은 타고난 지식과 학습된 지식 모두를 요구할 수 있으며, 보상 극대화 에이전트는 필요할 때마다 전자를 포함하고(예: 자연 에이전트의 진화 및 인공 에이전트의 설계를 통해) 후자를 획득합니다(예: 학습). 보다 풍부하고 수명이 긴 환경에서 수요의 균형은 학습된 지식으로 점점 이동합니다.

3.2. 보상은 인식에 충분합니다.

인간의 세계는 보상을 축적하기 위해 다양한 지각 능력을 요구합니다. 몇 가지 삶과 죽음의 예는 다음과 같습니다. 절벽에서 떨어지는 것을 방지하기 위한 이미지 분할; 건강 식품과 유독 식품을 분류하기 위한 물체 인식; 적으로부터 아군을 구별하는 얼굴 인식; 운전 중 장면 분석; 또는 "오리!"에 대한 구두 경고를 이해하기 위한 음성 인식
시각, 청각, 후각, 체성감각 또는 고유수용성 지각을 포함한 여러 지각 모드가 필요할 수 있습니다.

역사적으로 이러한 지각 능력은 별도의 문제 정의를 사용하여 공식화되었습니다 [9]. 보다 최근에는 지도 학습 문제에 대한 솔루션으로 지각 능력을 통합하려는 움직임이 증가하고 있습니다 [24]. 문제는 일반적으로 올바르게 레이블이 지정된 예제의 훈련 세트가 주어지면 테스트 세트의 예제에 대한 분류 오류를 최소화하는 것으로 공식화됩니다. 지도 학습 문제로 많은 지각 능력을 통합함으로써 대규모 데이터 세트를 사용할 수 있는 다양한 실제 응용 프로그램에서 상당한 성공을 거두었습니다 [23,15,8].

우리의 가설에 따르면, 우리는 지각이 보상의 최대화에 복종하는 것으로 이해될 수 있다고 제안합니다.
예를 들어, 위에 열거한 지각 능력은 건강에 좋은 음식을 최대화하거나, 사고를 피하거나, 고통을 최소화하는 데 암묵적으로 나타날 수 있습니다. 실제로 일부 동물의 지각은 보상 극대화와 일치하는 것으로 나타났습니다 [46,16]. 지도 학습보다 보상 극대화의 관점에서 인식을 고려하면 궁극적으로 도전적이고 현실적인 형태의 지각 능력을 포함하여 더 넓은 범위의 지각 행동을 지원할 수 있습니다.

- 행동과 관찰은 일반적으로 촉각 지각(예: 손끝을 움직여 주머니의 내용물 식별), 시각적 단속(예: 방망이와 공 사이에서 초점을 전환하기 위해 눈을 움직이는 것), 물리적 실험(예: 너트가 부러지는지 확인하기 위해 돌로 너트를 치는 것) 또는 반향 위치 측정(예: 다양한 주파수에서 소리를 방출하고 후속 반향의 타이밍 및 강도 측정). • 인식의 유용성은 종종 에이전트의 행동에 따라 달라집니다. 예를 들어, 악어를 잘못 분류하는 비용은 에이전트가 걷고 있는지 수영하는지, 그리고 에이전트가 이후에 싸우거나 도주할지 여부에 따라 다릅니다. • 정보를 획득하는 데는 명시적 또는 묵시적 비용이 있을 수 있습니다(예: 고개를 돌리고 포식자를 확인하는 데 에너지, 계산 및 기회 비용이 있음). • 데이터 배포는 일반적으로 상황에 따라 다릅니다. 예를 들어 북극 요원은 보상이 좌우되는 얼음과 북극곰을 분류해야 할 수 있습니다. 아프리카 에이전트는 사바나와 사자를 분류해야 할 수도 있습니다. 풍부한 환경에서 잠재적 데이터의 다양성은 에이전트의 용량이나 기존 데이터의 양을 크게 초과할 수 있습니다(섹션 3.1 참조) . 이러한 인식은 경험을 통해 학습되어야 합니다.

- 인식의 많은 응용 프로그램은 레이블이 지정된 데이터에 액세스할 수 없습니다.

3.3. 사회적 지능에 대한 보상은 충분합니다.

사회 지능은 다른 에이전트를 이해하고 효과적으로 상호 작용하는 능력입니다. 이 능력은 종종 게임 이론을 사용하여 다중 에이전트 게임의 평형 솔루션으로 공식화됩니다. 평형 솔루션은 편차 또는 최악의 시나리오에 강하기 때문에 바람직한 것으로 간주됩니다. 예를 들어, Nash 균형은 일방적인 편차가 편차자에게 이익을 제공하지 않도록 하는 모든 행위자에 대한 공동 전략입니다 [33]. 제로섬 게임에서 내쉬 균형은 또한 최소 최대 최적입니다 [34]: 최악의 경우에 대항하여 가능한 최고의 가치를 얻습니다.

우리의 가설에 따르면, 사회 지능은 대신 다른 에이전트를 포함하는 환경에서 한 에이전트의 관점에서 누적 보상을 최대화하는 것으로 이해되고 구현될 수 있습니다. 이 표준 에이전트-환경 프로토콜에 따라 한 에이전트는 다른 에이전트의 동작을 관찰하고 환경의 다른 측면을 관찰하고 영향을 미치는 것처럼 자신의 작업을 통해 다른 에이전트에 영향을 줄 수 있습니다. 다른 에이전트의 행동을 예측하고 영향을 줄 수 있는 에이전트는 일반적으로 더 큰 누적 보상을 얻을 수 있습니다. 따라서 환경에 사회적 지능이 필요한 경우(예: 동물이나 인간이 포함되어 있기 때문에) 보상 극대화는 사회적 지능을 생성합니다.

환경에 따라 견고성이 요구될 수도 있습니다. 예를 들어 환경에 서로 다른 전략을 따르는 여러 에이전트가 포함되어 있는 경우입니다. 이러한 다른 에이전트가 별칭이 있는 경우(즉, 다른 에이전트가 따를 전략을 미리 식별할 방법이 없는 경우) 보상 극대화 에이전트는 베팅을 헤지하고 이러한 잠재적 전략에 대해 효과적인 강력한 행동을 선택해야 합니다. 또한 다른 에이전트의 전략은 적응할 수 있습니다. 이는 다른 에이전트의 행동이 환경의 다른 측면과 마찬가지로 에이전트의 과거 상호작용에 따라 달라질 수 있음을 의미합니다(예: n번째 시도에서만 열리는 같힌 문). 특히, 이러한 적응성은 자신의 보상을 최대화하는 방법을 배우는 하나 이상의 강화 학습 에이전트가 포함된 환경에서 발생할 수 있습니다. 이러한 종류의 환경은 허풍이나 정보 은닉과 같은 사회적 지능의 측면을 요구할 수 있으며, 보상 극대화 에이전트는 착취를 피하기 위해 확률론적이어야 할 수 있습니다(즉, 혼합 전략 사용).

보상 극대화는 실제로 평형보다 더 나은 솔루션으로 이어질 수 있습니다 [47]. 이는 최적 또는 최악의 행동을 가정하기보다는 다른 에이전트의 차선책 행동을 활용할 수 있기 때문입니다. 또한, 보상 최대화는 고유한 최적 값을 갖는 반면 [38], 균형 값은 일반 합계 게임 [41]에서 고유하지 않습니다.

3.4. 언어에 대한 보상은 충분합니다.

언어는 자연 지능[10] 과 인공 지능 [28] 모두에서 상당한 연구 대상이었습니다 . 언어는 인간의 문화와 상호작용에서 지배적인 역할을 하기 때문에 지능 자체의 정의는 종종 언어, 특히 자연어를 이해하고 사용하는 능력을 전제로 합니다 [60].

최근에 언어를 단일 목표의 최적화로 처리함으로써 상당한 성공을 거두었습니다. 즉, 방대한 데이터 코퍼스 내에서 언어를 예측 모델링하는 것입니다 [28,8]. 이 접근 방식은 구문적 하위 문제(예: 형식 문법, 품사 태깅, 구문 분석, 분할) 및 의미 하위 문제(예: 예를 들어 어휘 의미론, 수반, 감정 분석) 뿐만 아니라 두 가지를 함께 결합하는 일부(예: 요약, 대화 시스템).

그럼에도 불구하고, 언어 모델링 자체만으로는 다음을 포함하는 지능과 관련된 광범위한 언어 능력 세트를 생성하기에 충분하지 않을 수 있습니다.

- 언어는 행동 및 관찰의 다른 양식과 얽혀 있을 수 있습니다. 언어는 종종 발화된 내용뿐만 아니라 시각 및 기타 감각 양식을 통해 인식되는 에이전트 주변 환경에서 일어나는 다른 일에 대해서도 맥락적입니다.). 게다가, 언어는 종종 몸짓, 얼굴 표정, 음색 변화 또는 신체 시연과 같은 다른 의사 소통 행동과 산재되어 있습니다. • 언어는 결과적이고 목적이 있습니다. 언어 발화는 일반적으로 정신 상태에 영향을 미치고 이에 따라 환경 내에서 다른 의사소통자의 행동에 영향을 미침으로써 환경에 영향을 미칩니다. 이러한 결과는 다양한 목적을 달성하기 위해 최적화될 수 있습니다. 예를 들어 판매원은 판매를 극대화하기 위해 언어를 조정하는 방법을 배우고 정치인은 투표를 최대화하기 위해 언어를 조정하는 방법을 배웁니다.

- 언어의 유용성은 에이전트의 상황과 행동에 따라 다릅니다. 예를 들어, 광부에게는 암석의 안정성에 관한 언어가 필요할 수 있지만 농부는 토양의 비옥함과 관련하여 언어가 필요할 수 있습니다. 더욱이 언어에 대한 기회 비용이 있을 수 있습니다(예: 농사일을 하는 대신 농사에 대해 논의). • 풍부한 환경에서 예상치 못한 사건을 처리하기 위해 언어를 잠재적으로 사용하는 것은 말뭉치의 용량을 능가할 수 있습니다. 이러한 경우 경험을 통해 언어 문제를 동적으로 해결하는 것이 필요할 수 있습니다. 예를 들어, 새로운 질병을 통제하고, 새로운 기술을 구축하거나, 새로운 불만 사항을 해결하는 방법을 찾기 위해 가장 효과적인 언어를 대화식으로 개발하는 것입니다. 침략을 미연에 방지하기 위해 라이벌.

우리의 가설에 따르면, 이러한 모든 광범위한 능력을 포함하여 완전한 언어의 능력은 보상을 추구하는 데서 비롯됩니다. 이것은 환경(위의 사회 지능에 대한 논의 참조)에 영향을 미치고 축적하기 위해 관찰의 복잡한 시퀀스(예: 문장 수신)를 기반으로 복잡한 일련의 행동(예: 문장 발화)을 생성하는 에이전트의 능력의 한 예입니다. 더 큰 보상 [7]. 언어를 이해하고 생산해야 한다는 압력은 보상을 증가시키는 많은 이점에서 비롯될 수 있습니다. 에이전트가 "위험" 경고를 이해할 수 있다면 부정적인 보상을 예측하고 피할 수 있습니다. 에이전트가 "가져오기" 명령을 생성할 수 있는 경우 환경(예: 개 포함)이 개체를 에이전트에 더 가깝게 이동할 수 있습니다. 마찬가지로 에이전트는 음식의 위치에 대한 복잡한 설명을 이해하고, 음식을 재배하기 위한 복잡한 지침을 생성하고, 음식을 협상하기 위해 복잡한 대화에 참여하거나, 이러한 협상을 향상시키는 장기적인 관계를 구축할 수 있는 경우에만 먹을 수 있습니다. 다양한 복잡한 언어 능력.

3.5. 보상은 일반화에 충분합니다.

일반화는 종종 한 문제에 대한 솔루션을 다른 문제에 대한 솔루션으로 전환하는 능력으로 정의됩니다 [37,58,61]. 예를 들어 지도 학습 [37] 의 일반화 는 사진과 같은 하나의 데이터 세트에서 학습된 솔루션을 그림과 같은 다른 데이터 세트로 전송하는 데 초점을 맞출 수 있습니다. 메타 학습의 일반화 [61,20] 는 최근 한 환경에서 다른 환경으로 에이전트를 전송하는 문제에 초점을 맞추고 있습니다.

우리의 가설에 따르면 일반화는 에이전트와 단일 복잡한 환경 사이의 지속적인 상호 작용 흐름에서 누적 보상을 최대화하는 것으로 이해되고 구현될 수 있습니다. 다시 표준 에이전트-환경 프로토콜을 따릅니다(섹션 2.1 참조). 인간 세계와 같은 환경은 단순히 에이전트가 서로 다른 시간에 환경의 다른 측면을 만나기 때문에 일반화를 요구합니다. 예를 들어 과일을 먹는 동물은 매일 새로운 나무를 만날 수 있습니다. 또한 부상을 입거나 가뭄을 겪거나 침입 종에 직면 할 수 있습니다. 각각의 경우에 동물은 과거 상태의 경험을 일반화하여 새로운 상태에 빠르게 적응해야 합니다. 동물이 직면한 다양한 상태는 별개의 레이블이 있는 분리되고 순차적인 작업으로 깔끔하게 분류되지 않습니다. 대신 상태는 동물의 행동에 따라 다릅니다. 서로 다른 시간 규모에서 겹치고 반복되는 다양한 요소를 결합할 수 있습니다. 국가의 중요한 측면이 부분적으로 관찰될 수 있습니다. 풍부한 환경은 보상을 효율적으로 축적하기 위해 이러한 모든 복잡성과 함께 과거 상태에서 미래 상태로 일반화할 수 있는 능력을 요구합니다.

3.6. 모방은 보상으로 충분하다

모방은 언어, 지식 및 운동 기술과 같은 다른 능력의 빠른 습득을 촉진할 수 있는 인간 및 동물 지능과 관련된 중요한 능력입니다. 인공 지능에서 모방은 종종 행동 복제 [2] 를 통해 시연으로부터 학습하는 문제로 공식화되었습니다. 여기서 목표는 교사의 행동에 관한 명시적인 데이터가 제공될 때 교사가 선택한 행동을 재현하는 것입니다. 관찰 및 보상, 일반적으로 교사가 에이전트에 대해 대칭 문제를 해결한다는 가정 하에. 행동 복제는 여러 성공적인 기계 학습 응용 프로그램 [54,62,5], 특히 인간 교사 데이터는 풍부하지만 상호 작용 경험이 제한적이거나 비용이 많이 드는 응용 프로그램으로 이어졌습니다. 대조적으로, 관찰 학습의 타고난 능력 [3] 은 다른 인간이나 동물의 관찰된 행동으로부터 학습의 모든 형태를 포함하며 대칭적인 교사를 가정하거나 그들의 행동, 관찰 및 보상에 대한 직접적인 접근을 요구하지 않습니다. 이는 복잡한 환경에서 행동 복제를 통한 직접 모방과 비교하여 훨씬 더 광범위하고 현실적인 관찰 학습 능력 등급이 요구될 수 있음을 시사합니다.

- 다른 에이전트는 교사 데이터를 포함하는 고유한 데이터 세트의 존재를 가정하지 않고 에이전트 환경의 필수적인 부분일 수 있습니다(예: 엄마를 관찰하는 아기). • 에이전트는 자신의 상태(예: 아기 몸의 자세)와 다른 에이전트의 상태(예: 엄마의 자세) 또는 자신의 동작(예: 로봇 조작기 회전) 간의 연관성을 학습해야 할 수 있습니다. 및 잠재적으로 더 높은 수준의 추상화(예: 근육 활성화보다는 어머니의 음식 선택 모방)에서 다른 에이전트의 관찰(예: 사람의 손 보기).

- 다른 에이전트는 부분적으로 관찰될 수 있습니다(예: 사람의 손이 가려진 경우).
  아마도 뒤늦게나마 불완전하게 추론될 뿐입니다. • 다른 에이전트
  는 피해야 하는 바람직하지 않은 행동을 보일 수 있습니다. • 환경에는 다른 기술이나 다른 수준의 능력을
  보이는 다른 많은 에이전트가 있을 수 있습니다. • 관찰 학습은 명시적 기관 없이도 발생할 수 있습니다(예:

  개울을 가로질러 떨어진 통나무의 관찰).

우리는 관찰 학습의 이러한 광범위한 능력이 다른 에이전트를 환경의 필수적인 부분으로 단순히 관찰하는 단일 에이전트의 관점에서 보상의 최대화에 의해 주도될 수 있다고 추측합니다 [6], 잠재적으로 행동과 동일한 많은 이점을 이끌어냅니다 샘플 효율적인 지식 획득과 같은 복제를 훨씬 더 광범위하고 통합된 맥락에서 수행합니다.

3.7. 보상은 일반 지능에 충분합니다.

마지막 예에서 우리는 동시에 가장 큰 도전을 제기하고 우리의 가설이 가장 큰 잠재적 이점을 제공하는 능력에 대해 설명합니다. 인간과 아마도 다른 동물이 소유한 종류의 일반 지능은 다양한 맥락에서 다양한 목표를 유연하게 달성하는 능력으로 정의될 수 있습니다. 예를 들어, 인간은 자신의 상황에 적합한 솔루션(예: 수영 또는 스키, 자동차 타기 또는 발차기, 쓰기 또는 수화)을 사용하여 문제(예: 이동, 교통 또는 통신)를 유연하게 해결할 수 있습니다. 일반 지능은 다양한 목표와 맥락에서 에이전트의 능력을 측정하는 일련의 환경에 의해 공식화되는 경우가 있습니다 [25,14].

우리의 가설에 따르면 일반 지능은 복잡한 단일 환경에서 단일 보상을 극대화하는 것으로 이해되고 구현될 수 있습니다. 예를 들어, 자연 지능은 평생 동안 자연 세계와의 상호 작용에서 생성되는 연속적인 경험의 흐름에 직면합니다. 동물의 경험 흐름은 충분히 풍부하고 다양하여 전반적인 보상(예: 기아 또는 번식)을 최대화하는 데 성공하기 위해 매우 다양한 하위 목표(예: 먹이 찾기, 싸움 또는 도주)를 달성하는 유연한 능력을 요구할 수 있습니다.). 유사하게, 인공 에이전트의 경험 흐름이 충분히 풍부하다면 단일 목표(예: 배터리 수명 또는 생존)는 암묵적으로 동등하게 다양한 하위 목표를 달성하는 능력을 요구할 수 있으며 따라서 보상의 최대화는 다음을 달성하기에 충분해야 합니다. 인공 일반 지능.

4. 강화 학습 에이전트

우리의 주요 가설은 지능과 관련 능력이 보상의 최대화에 복종하는 것으로 이해될 수 있으며 에이전트의 본성에 불가지론적입니다. 이것은 보상을 최대화하는 에이전트를 구성하는 방법에 대한 중요한 질문을 남겨둡니다. 이 섹션에서 우리는 이 질문에 보상 최대화로도 답할 수 있다고 제안합니다. 특히, 우리는 환경과의 지속적인 상호작용 경험에서 보상을 극대화하는 방법을 배울 수 있는 일반적인 능력을 가진 에이전트를 고려합니다. 강화 학습 에이전트 라고 하는 이러한 에이전트 는 몇 가지 이점을 제공합니다.3

[3] 동일한 이름을 사용하여 문제(예: 등산은 봉우리를 오르는 문제를 나타냄), 해결 방법(예: 등산가가 사용하는 로프 및 피톤) 및 필드(예: 등산의 취미)를 설명하는 것이 일반적입니다. . 맥락에서 명확하지 않은 경우 강화 학습 문제, 강화 학습 에이전트 및 강화 학습 분야를 참조합니다.

첫째, 보상을 극대화하기 위한 가능한 모든 솔루션 방법 중에서 가장 자연스러운 접근 방식은 환경과의 상호 작용을 통해 경험을 통해 학습하는 것입니다. 시간이 지남에 따라 상호 작용하는 경험은 원인과 결과, 행동의 결과 및 보상을 축적하는 방법에 대한 풍부한 정보를 제공합니다. 에이전트의 행동을 미리 결정하는 것(설계자의 환경 예지에 믿음을 두기)보다는 에이전트 자신의 행동을 발견할 수 있는 일반적인 능력을 부여하는 것이 자연스럽습니다(경험에 믿음을 둡니다). 보다 구체적으로, 보상을 최대화한다는 설계 목표는 미래의 보상을 최대화하는 행동을 경험에서 학습하는 지속적인 내부 프로세스를 통해 구현됩니다.4

강화 학습 에이전트는 경험을 통해 학습하여 효과적일 수 있는 일반적인 솔루션 방법을 제공하고,
다양한 보상 신호 및 환경 전반에 걸쳐 최소한의 수정 또는 제로 수정.
더욱이, 단일 환경은 자연 세계와 같이 너무 복잡하여 가능한 경험의 이질적인 다양성을 포함할 수 있습니다. 수명이 긴 에이전트가 직면한 관찰 및 보상 흐름의 잠재적인 변동은 필연적으로 사전 프로그래밍된 행동에 대한 용량을 초과할 것입니다(섹션 3.1 참조). 따라서 높은 보상을 달성하기 위해 에이전트는 새로운 경험에 자신의 행동을 완전하고 지속적으로 적응시킬 수 있는 일반적인 능력을 갖추고 있어야 합니다. 실제로 강화 학습 에이전트는 이러한 복잡한 환경에서 유일하게 실현 가능한 솔루션일 수 있습니다.

충분히 강력하고 일반적인 강화 학습 에이전트는 궁극적으로 지능과 관련 능력을 발생시킬 수 있습니다. 즉, 에이전트가 누적 보상을 개선하기 위해 행동을 지속적으로 조정할 수 있다면 환경에서 반복적으로 요구하는 모든 능력은 궁극적으로 에이전트의 행동에서 생성되어야 합니다. 따라서 좋은 강화 학습 에이전트는 인간 세계와 같이 이러한 능력이 지속적인 가치를 지닌 환경에서 보상을 극대화하기 위해 학습하는 과정에서 지각, 언어, 사회적 지능 등을 나타내는 행동을 습득할 수 있습니다.

우리는 강화 학습 에이전트의 샘플 효율성에 대한 이론적 보장을 제공하지 않습니다. 실제로 능력이 나타나는 속도와 정도는 특정 환경, 학습 알고리즘 및 귀납적 편향에 따라 다릅니다. 게다가 학습이 실패할 인공적인 환경을 구축할 수도 있습니다. 대신, 우리는 복잡한 환경에 배치된 강력한 강화 학습 에이전트가 실제로 정교한 지능 표현을 발생시킬 것이라고 추측합니다. 이 추측이 맞다면 인공 일반 지능의 구현을 향한 완전한 경로를 제공합니다.

보상을 최대화하는 학습 능력을 부여받은 강화 학습 에이전트의 최근 몇 가지 예는 기대를 능가하는 광범위하게 유능한 행동, 이전 에이전트의 성능, 몇몇 경우에는 인간 전문가의 성능을 발생시켰습니다. 예를 들어, 바둑 게임에서 승리를 최대화하라는 요청을 받았을 때 AlphaZero는 바둑 의 여러 측면에 걸쳐 통합 지능을 배웠습니다(섹션 1 참조) . 체스 게임에서 결과를 최대화하기 위해 동일한 알고리즘이 적용되었을 때 AlphaZero는 오프닝, 엔드게임, 조각 이동성, 왕 안전 등을 포함하는 다른 일련의 능력을 배웠습니다 [48,44]. Atari 2600 [30] 에서 점수를 최대화하는 강화 학습 에이전트는 각 특정 Atari 게임에서 요구되는 개체 인식, 위치 파악, 탐색 및 모터 제어 측면을 포함하여 다양한 능력을 학습한 반면 비전 기반 로봇에서 성공적인 그립을 최대화하는 에이전트는 조작은 개체 상귤레이션, 재포착, 동적 개체 추적과 같은 감각 운동 능력을 학습했습니다 [22]. 이러한 예는 자연 지능이 직면한 환경보다 범위가 훨씬 좁지만 보상 최대화 원칙의 효과에 대한 몇 가지 실용적인 증거를 제공합니다.

물론 실제 에이전트에서 보상을 효과적으로 극대화하는 방법을 배우는 방법이 궁금할 수 있습니다. 예를 들어 보상은 직접적으로(예: 에이전트의 정책을 최적화하여 [57]), 간접적으로 더 분해될 수 있는 표현 학습, 가치 예측, 모델 학습 및 계획과 같은 하위 목표로 분해하여 최대화될 수 있습니다.].

우리는 이 백서에서 이 질문을 더 이상 다루지 않지만 이것이 강화 학습 분야 전체에서 연구되는 중심 질문이라는 점에 주목하십시오.

5. 관련 업무

지능은 오랫동안 목표 지향적인 행동과 연관되어 왔습니다 [59]. 이 목표 지향적인 지능 개념은 합리성 개념의 핵심으로 , 에이전트가 목표를 최적으로 달성하거나 효용을 극대화하는 방식으로 행동을 선택합니다. 합리성은 인간 행동 [63,4] 을 이해하는 데 널리 사용되었으며 인공 지능 [42] 의 기초 로 환경과 상호 작용하는 에이전트와 관련하여 공식화되었습니다.

목표에 대해 추론할 때 계산 제약 조건도 고려해야 한다는 주장이 자주 제기되어 왔습니다. Bounded [50,43,36] 또는 계산적 합리성 [27] 은 에이전트가 프로그램에서 발생하는 실시간 결과(예: 프로그램 실행에 걸리는 시간)를 고려할 때 목표를 가장 잘 달성하는 프로그램을 선택해야 한다고 제안합니다. 프로그램 세트에 대한 제한(예: 프로그램의 최대 크기 제한)이 적용됩니다. 우리의 기여는 이러한 관점을 기반으로 하지만, 하나의 단순한 보상 기반 목표가 지능과 관련된 모든 능력에 대한 공통 기반을 제공할 수 있는지에 대한 질문에 중점을 둡니다.

강화 학습을 위한 표준 프로토콜은 Sutton과 Barto [56]에 의해 정의되었습니다. 섹션 2 에서 우리는
부분적으로 관찰된 역사를 가진 일반적인 일반화.

4 내부 보상은 디자인 목표와 구별되지만 서비스를 위해 선택될 수도 있습니다 [51].

통합 인지 아키텍처 [35,1] 는 일반 지능을 지향합니다. 그것들은 개별 하위 문제(예: 지각 또는 운동 제어)에 대한 다양한 솔루션 방법을 결합하지만 아키텍처 선택을 정당화하고 설명하는 일반적인 목표나 개별 구성 요소가 기여하는 단일 목표를 제공하지 않습니다.

보상 극대화로서의 언어의 관점은 행동주의로 거슬러 올라갑니다 [52,53]. 그러나 강화 학습 문제는 에이전트가 내부 상태를 구성하고 사용하도록 허용한다는 점에서 행동주의와 다릅니다. 다중 에이전트 환경에서 균형 목표보다는 단일 에이전트의 목표에 초점을 맞추는 이점은 Shoham과 Powers [47]에 의해 논의되었습니다.

6. 토론

우리는 보상이 충분하다는 가설과 그 의미를 제시했습니다. 다음으로 우리는 이 가설에 대한 논의에서 자주 발생하는 여러 질문에 간략하게 답합니다.

어떤 환경? 어떤 환경이 보상 극대화를 통해 "가장 지능적인" 행동이나 "최고의" 특정 능력(예: 자연어)을 발생시킬 것인지 물을 수 있습니다. 필연적으로 에이전트가 직면하는 특정 환경 경험(예: 친구, 적, 교사, 장난감, 도구 또는 인간 두뇌의 일생 동안 마주치는 도서관)은 후속 능력의 특성을 형성합니다. 이 질문은 지능의 특정 적용에 큰 관심을 가질 수 있지만, 우리는 그 대신에 모든 형태의 지능을 발생시킬 수 있는 일반적인 목표에 대한 틀림없이 더 심오한 질문에 초점을 맞추었습니다. 서로 다른 환경에서 서로 다른 보상을 최대화하면 독특하고 강력한 형태의 지능이 탄생할 수 있으며, 각각은 인상적이면서도 비교할 수 없는 능력의 배열을 보여줍니다. 좋은 보상 극대화 에이전트는 환경에 존재하는 모든 요소를 활용하지만 어떤 형태로든 지능의 출현은 특정 요소에 기반을 두지 않습니다. 예를 들어, 인간의 두뇌는 환경에서 다양한 경험에 노출되면 태어날 때부터 다르게 발달하지만 특정 문화나 교육에 관계없이 정교한 능력을 습득합니다.

어떤 보상 신호? 보상 신호를 조작하려는 욕구는 신중하게 구성된 보상만이 일반 지능을 유도할 수 있다는 생각에서 종종 발생합니다. 대조적으로, 우리는 지능의 출현이 보상 신호의 특성에 매우 강력할 수 있다고 제안합니다. 이는 자연 세계와 같은 환경이 너무 복잡하여 겉보기에 무해한 보상 신호라도 지능과 관련 능력을 요구할 수 있기 때문입니다. 예를 들어, 둥근 모양의 조약돌이 수집될 때마다 에이전트에게 +1 보상을 제공하는 신호를 고려하십시오. 이 보상 신호를 효과적으로 최대화하기 위해 에이전트는 자갈을 분류하고, 자갈을 조작하고, 자갈 해변으로 이동하고, 자갈을 저장하고, 파도와 조수 및 자갈 분포에 미치는 영향을 이해하고, 자갈 수집을 돕도록 사람들을 설득해야 할 수 있습니다. , 도구와 차량을 사용하여 더 많은 양을 수집하고, 새로운 조약돌을 채석하고 모양을 만들고, 조약돌 수집을 위한 새로운 기술을 발견 및 구축하거나, 조약돌을 수집하는 기업을 구축합니다.

보상 극대화 외에 지능에 충분할 수 있는 것은 무엇입니까? 비지도 학습 [19] (예: 관찰에서 패턴 식별) 및 예측 [18,11] (예: 미래 관찰 [31]) 은 경험을 이해하기 위한 효과적인 원칙을 제공할 수 있지만 행동 선택에 대한 원칙은 제공하지 않으므로 수행할 수 없습니다. 목표 지향 지능을 위해서는 고립되어 있으면 충분합니다. 지도 학습은 인간 지능을 모방하는 메커니즘을 제공합니다. 충분한 인간 데이터가 주어지면 인간 지능과 관련된 모든 능력이 나타날 수 있다고 상상할 수 있습니다. 그러나 인간 데이터로부터의 지도 학습은 인간이 아닌 환경에서 인간이 아닌 목표에 최적화할 수 있는 범용 지능에 충분하지 않습니다. 인간 데이터가 풍부한 곳에서도 모방 지능은 예상치 못한 방식으로 문제를 해결하는 창의적이고 새로운 행동을 발견하기보다는 데이터 내에서 인간이 이미 알고 표시한 행동으로 범위가 제한될 수 있습니다(아래 오프라인 학습 참조). 자연 선택에 의한 진화 는 돌연변이 및 교차와 같은 인구 기반 메커니즘에 의해 최적화된 개별 번식 성공으로 측정되는 적합성을 극대화하는 것으로 추상적인 수준에서 이해할 수 있습니다. 우리의 틀에서 번식 성공은 자연 지능의 출현을 주도한 하나의 가능한 보상 신호로 볼 수 있습니다. 그러나 인공 지능은 번식 성공 이외의 다른 목표로 설계될 수 있으며 돌연변이 및 교차 이외의 방법을 사용하여 해당 보상 신호를 최대화하여 잠재적으로 매우 다른 형태의 지능으로 이어질 수 있습니다. 또한, 피트니스 극대화는 자연 지능(예: 인간 아기의 뇌)의 초기 구성을 설명할 수 있지만, 내재적 보상 신호를 최대화하기 위한 시행착오 학습 과정 [51] (섹션 4 참조)은 더 나아가 자연 지능이 어떻게 경험을 통해 적응하여 체력 극대화를 위해 정교한 능력(예: 인간의 성인 두뇌)을 개발합니다. 자유 에너지의 최대화 또는 놀라움의 최소화 [13] 는 자연 지능의 여러 능력을 산출할 수 있지만 다양한 환경에서 다양한 목표를 향할 수 있는 범용 지능을 제공하지는 않습니다. 결과적으로 이러한 목표 중 하나를 최적으로 달성하는 데 필요한 능력을 놓칠 수도 있습니다(예: 파트너와 짝짓기하는 데 필요한 사회적 지능 측면 또는 상대방을 체크메이트하는 데 필요한 전술적 지능). 최적화 는 누적 보상을 포함하여 모든 신호를 최대화할 수 있지만 에이전트가 환경과 상호 작용하는 방식을 지정하지 않는 일반적인 수학적 형식입니다. 대조적으로, 강화 학습 문제는 핵심에 상호 작용을 포함합니다. 행동은 보상을 최대화하도록 최적화되고, 이러한 행동은 차례로 최적화 프로세스에 정보를 제공하는 환경에서 받은 관찰을 결정합니다. 게다가 최적화는 환경이 계속 움직이는 동안 실시간으로 온라인에서 발생합니다.

어떤 보상극대화 문제인가? 강화 학습 연구자들 사이에서도 보상 최대화 문제의 다양한 변형이 연구됩니다. 표준 에이전트-환경 프로토콜을 따르는 대신 상호 작용

루프는 여러 에이전트, 여러 환경 또는 여러 교육 수명 시간을 포함할 수 있는 다양한 경우에 대해 수정되는 경우가 많습니다. 누적 보상으로 정의된 일반적인 목표를 최대화하는 대신 다중 목표 학습, 위험에 민감한 목표 또는 루프 내 인간이 지정한 목표와 같은 다양한 경우에 대해 목표를 별도로 공식화하는 경우가 많습니다. 또한 일반 환경에 대한 보상 극대화 문제를 해결하기보다 선형 환경, 결정적 환경 또는 안정적인 환경과 같은 특정 클래스의 환경에 대해 특수 사례 문제를 연구하는 경우가 많습니다. 이것은 특정 응용 프로그램에 적합할 수 있지만 특수 문제에 대한 솔루션은 일반적으로 일반화되지 않습니다. 대조적으로 일반적인 문제에 대한 솔루션은 특별한 경우에 대한 솔루션도 제공합니다. 강화 학습 문제는 보상 최대화의 목표에 근접하는 확률적 프레임워크로 변형될 수도 있습니다 [66,39,26,17]. 마지막으로 보편적인 의사 결정 프레임워크 [21] 는 이론적이지만 모든 환경에 걸쳐 지능을 mulation하기 위해 계산할 수 없는 것을 제공합니다. 강화 학습 문제는 주어진 환경에서 지능의 실용적인 공식화를 제공합니다.

충분히 큰 데이터 세트에서 오프라인 학습이 인텔리전스를 위해 충분할 수 있습니까? 오프라인 학습은 이미 사용 가능한 데이터 내에서 상당 부분 해결된 문제를 해결하는 데 충분할 수 있습니다. 예를 들어, 견과류를 모으는 다람쥐의 대규모 데이터 세트는 견과류 수확기를 만드는 데 필요한 모든 동작을 보여주지 못할 것입니다. 오프라인 데이터 세트에서 시연되거나 추출된 솔루션에서 에이전트의 현재 문제로 일반화하는 것이 가능할 수 있지만 복잡한 환경에서 이러한 일반화는 불가피하게 불완전합니다. 또한 에이전트의 현재 문제를 해결하는 데 필요한 데이터는 종종 오프라인 데이터(예: 무작위 행동 또는 불완전한 인간 행동)에서 발생할 확률이 무시할 수 있습니다. 온라인 상호작용을 통해 에이전트는 현재 직면하고 있는 문제를 전문화하고, 지식의 가장 시급한 구멍을 지속적으로 확인 및 수정하고, 데이터 세트의 행동과 매우 다르고 더 큰 보상을 달성하는 새로운 행동을 찾을 수 있습니다.

보상 신호가 너무 빈약합니까? 복잡하고 알려지지 않은 환경에서 보상을 극대화할 수 있는 샘플 효율적인 강화 학습 에이전트가 반드시 존재해야 하는지 궁금할 수 있습니다. 이 질문에 답할 때 우리는 먼저 효과적인 에이전트가 미래 보상의 최대화를 촉진하기 위해 추가적인 경험적 신호를 사용할 수 있다는 점에 주목합니다. 모델 없는 강화 학습을 포함한 많은 솔루션 방법은 가치 함수 근사를 통해 미래의 보상을 관찰의 기능과 연관시키는 방법을 배우며, 이는 재귀 부트스트랩 프로세스를 통해 더 깊은 연관 학습을 유도하는 풍부한 2차 신호를 제공합니다 [55]. 모델 기반 강화 학습을 포함한 다른 솔루션 방법은 계획을 통한 보상의 후속 극대화를 용이하게 하는 관찰 또는 관찰 기능의 예측을 구성합니다. 게다가, 3.6 절에서 논의된 바와 같이 , 환경 내의 다른 에이전트의 관찰은 또한 빠른 학습을 촉진할 수 있습니다.

그럼에도 불구하고 연구자들이 가정을 도입하거나 이론과 실제 모두에 더 적합한 단순한 추상화를 개발하도록 하는 것은 복잡한 환경에서 샘플 효율적인 강화 학습의 도전과제인 경우가 많습니다.
그러나 이러한 가정과 추상화는 광범위하게 유능한 지능이 필연적으로 직면해야 하는 어려움을 단순히 회피할 수 있습니다. 우리는 그 대신 도전을 정면으로 받아들이고 해결에 집중하기로 선택합니다. 다른 연구자들이 우리의 탐구에 동참하기를 바랍니다.

7. 결론

이 논문에서 우리는 총 보상의 극대화가 지능과 관련 능력을 이해하기에 충분할 수 있다는 가설을 제시했습니다. 핵심 아이디어는 풍부한 환경은 일반적으로 보상을 극대화하기 위해 다양한 능력을 요구한다는 것입니다. 자연에서 발견되는 지능의 풍부한 표현, 그리고 아마도 미래에 인공 요원에 대한 것은 다른 환경과 다른 보상을 가진 이 동일한 아이디어의 인스턴스화로 이해될 수 있습니다.

더욱이, 보상 극대화라는 단일한 목표는 각각의 고유한 능력에 대한 전문화된 문제 공식보다 능력에 대한 더 깊고, 더 광범위하고, 더 통합된 이해를 야기할 수 있습니다. 특히, 지식, 학습, 지각, 사회 지능, 언어, 일반화, 모방, 일반 지능 등 언뜻 보기에는 보상극대화만으로는 이해하기 어려운 몇 가지 능력을 보다 심층적으로 탐구하여 그 보상을 발견하였다. 극대화는 각 능력을 이해하기 위한 기초를 제공할 수 있습니다. 마지막으로 우리는 미래의 보상을 최대화하기 위해 학습하는 충분히 강력한 강화 학습 에이전트로부터 지능이 실제로 나타날 수 있다는 추측을 제시했습니다. 이 추측이 사실이라면 인공 일반 지능을 이해하고 구성하는 직접적인 경로를 제공합니다.

경쟁적 이해관계 선언

모든 저자는 DeepMind의 직원입니다.

감사의 말

이 기사에 대한 의견을 주신 DeepMind의 리뷰어와 동료들에게 감사드립니다.

참고문헌

[1] JR Anderson, D. Bothell, MD Byrne, S. Douglass, C. Lebiere, Y. Qin, 마음의 통합 이론, Psychol. 개정 111(4) (2004) 1036.

[2] M. Bain, C. Sammut, A framework for behavioral cloning, in: Machine Intelligence 15, 1995, pp. 103–129.
[3] A. Bandura, DC McClelland, 사회 학습 이론, vol. 1, Prentice Hall, Englewood Cliffs, 1977.
[4] GS Becker, 인간 행동에 대한 경제적 접근. 경제 이론, 시카고 대학 출판부, 1976.
[5] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, LD Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao , K. Zieba, 자율 주행 자동차를 위한 종단 간 학습, CoRR, arXiv:1604.07316 [abs], 2016.
[6] D. Borsa, N. Heess, B. Piot, S. Liu, L. Hasenclever, R. Munos, O. Pietquin, 강화 학습에 의한 관찰 학습, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent 시스템, 자율 에이전트 및 다중 에이전트 시스템을 위한 국제 재단, 2019, pp. 1117–1124.
[7] J. Bratman, M. Shvartsman, R. Lewis, S. Singh, 환경에 직면하여 경계적으로 최적의 제어로서 언어 출현을 탐구하는 새로운 접근 방식 정신적 및 인지적 제약, in: Proceedings of the 10th International Conference on Cognitive Modeling, ICCM, 2010.
[8] TB Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are little -사격 학습자, arXiv:2005.14165, 2020.
[9] EC Carterette, MP Friedman, Handbook of Perception, Academic Press, 1978.
[10] N. Chomsky, DW Lightfoot, 구문 구조, Walter de Gruyter, 2002.
[11] A. 클락, 다음은? 예측 뇌, 상황 에이전트, 인지 과학의 미래, 행동. 뇌과학. 36 (3) (2013) 181–204.
[12] G. Debreu, 숫자 함수에 의한 선호도 순서 표현, 1954.
[13] K. Friston, 자유 에너지 원리: 통일된 뇌 이론?, Nat. Neurosci 목사. 11 (127–38) (2010) 02.
[14] B. Goertzel, C. Pennachin, 인공 일반 지능, vol. 2, Springer, 2007.
[15] A. 그레이브스, A.-r. Mohamed, G. Hinton, 심층 순환 신경망을 사용한 음성 인식, in: 2013 IEEE International Conference on Acoustics, 음성 및 신호 처리, IEEE, 2013, pp. 6645–6649.
[16] N. Grujic, J. Brus, D. Burdakov, R. Polania, 쥐의 합리적 부주의, bioRxiv, https://doi.org/10.1101/2021.05.26.445807, 2021.
[17] D. Hafner, PA Ortega, J. Ba, T. Parr, KJ Friston, N. Heess, 발산 최소화로서의 행동 및 인식, CoRR, arXiv:2009.01791 [abs], 2020.
[18] J. Hawkins, S. Blakeslee, On Intelligence, Times Books, USA, 2004.
[19] G. Hinton, TJ Sejnowski, 비지도 학습: 신경 계산의 기초, MIT Press, 1999.
[20] TM Hospedales, A. Antoniou, P. Micaelli, AJ Storkey, 신경망의 메타 학습: 조사, CoRR, arXiv:2004.05439 [abs], 2020.
[21] M. Hutter, 유니버설 인공 지능: 알고리즘 확률에 기반한 순차적 결정, Springer, 2005.
[22] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, S. Levine, 확장 가능한 심층 보강 비전 기반 로봇 조작을 위한 학습, in: 로봇 학습에 관한 2차 연례 회의, Proceedings, CoRL 2018, 스위스 취리히, 2018년 10월 29-31일, In: Proceedings of Machine Learning Research, vol. 87, PMLR, 2018, pp. 651–673.
[23] A. Krizhevsky, I. Sutskever, GE Hinton, Imagenet 분류 with deep convolutional neural network, in: Advances in Neural Information Processing 시스템, 2012, pp. 1097–1105.
[24] Y. LeCun, 딥 러닝을 사용한 컴퓨터 인식, 2013.
[25] S. Legg, M. Hutter, Universal Intelligence: 기계 지능의 정의, Minds Mach. 17 (4) (2007) 391–444.
[26] S. Levine, 확률적 추론으로서의 강화 학습 및 제어: 자습서 및 검토, CoRR, arXiv:1805.00909 [abs], 2018.
[27] RL Lewis, A. Howes, S. Singh, 계산 합리성: 제한된 효용 극대화를 통한 메커니즘 및 동작 연결, Top. 인지 과학 6 (2) (2014) 279–311.
[28] CD Manning, CD Manning, H. Schütze, 통계적 자연어 처리의 기초, MIT Press, 1999.
[29] J. McCarthy, AI란 무엇인가, 1998.
[30] V. Mnih, K. Kavukcuoglu, D. Silver, AA Rusu, J. Veness, MG Bellemare, A. Graves, M. Riedmiller, AK Fidjeland, G. Ostrovski 등, 깊은 인간 수준 제어 강화 학습, Nature 518 (7540) (2015) 529.
[31] J. Modayil, A. White, RS Sutton, 강화 학습 로봇의 다중 시간 척도, Adapt. 행동 22(2) (2014) 146–160.
[32] M. Müller, Computer Go, Artif. 인텔. 134(1-2)(2002) 145-179.
[33] JF Nash, et al., n-person 게임의 평형점, Proc. 내셔널 아카드. 과학 36 (1) (1950) 48–49.
[34] Jv Neumann, 응접실 게임 이론, Math.Ann. 100 (1) (1928) 295-320.
[35] A. Newell, 통합 인지 이론, Harvard University Press, USA, 1990.
[36] L. Orseau, MB Ring, 시공간 임베디드 지능, in: Proceedings of the 5th International Conference on Artificial General Intelligence, Lecture Notes in Computer Science, vol. 7716, Springer, 2012, pp. 209–218.
[37] Pan SJ Pan, Q. Yang, A Survey on Transfer learning, IEEE Trans. 놀. 데이터공학 22 (10) (2009) 1345-1359.
[38] ML Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.
[39] K. Rawlik, M. Toussaint, S. Vijayakumar, 확률적 최적 제어 및 근사 추론에 의한 강화 학습, in: Twenty-Third Interna 인공 지능에 관한 합동 회의, 2013.
[40] M. Redmond, C. Garlock, AlphaGo to Zero: 완전한 게임, Smart Go, 2020.
[41] J. Ben Rosen, 오목한 n인 게임에 대한 평형점의 존재 및 고유성, Econometrica(1965) 520–534.
[42] S. Russell, P. Norvig, 인공 지능: 현대 접근 방식, Prentice Hall, 1995.
[43] SJ Russell, D. Subramanian, Provably bounded-optimal agents, J. Artif. 인텔. 해상도 2 (1995) 575–609.
[44] M. Sadler, N. Regan, G. Kasparov, Game Changer: AlphaZero의 획기적인 체스 전략과 AI의 약속, New in Chess, 2019.
[45] S. Schaal, 데모에서 학습, M. Mozer, MI Jordan, T. Petsche(Eds.), Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, 1996년 12월 2-5일, MIT Press, 1996, pp. 1040-1046.
[46] J. Schaffner, P. Tobler, T. Hare, R. Polania, 초기 감각 영역의 신경 코드는 체력을 최대화합니다. bioRxiv, https://doi.org/10.1101/2021.05.10.443388, 2021.
[47] Y. Shoham, R. Powers, T. Grenager, 다중 에이전트 학습이 답이라면 질문은 무엇입니까?, Artif. 인텔. 171 (7) (2007) 365–377.
[48] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al., A 일반 강화 학습 체스, 장기, 바둑을 마스터하는 알고리즘, Science 362 (6419) (2018) 1140–1144.
[49] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering game 의 인간 지식, Nature 550 (7676) (2017) 354–359.
[50] HA Simon, 합리적 선택의 행동 모델, QJ Econ. 69 (1) (1955) 99–118.
[51] S. Singh, RL Lewis, AG Barto, J. Sorg, 내재적 동기 부여 강화 학습: 진화적 관점, IEEE Trans. 아우톤. 멘션. 개발 2 (2) (2010) 70–82.
[52] BF Skinner, 유기체의 행동: 실험적 분석, Appleton-Century-Crofts, New York, 1938.
[53] BF Skinner, 언어 행동, Appleton-Century-Crofts, New York, 1957.
[54] N. Stiennon, L. Ouyang, J. Wu, DM Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, PF Christiano, Learning to Summary with Human Feedback, in: 신경 정보 처리 시스템의 발전 33, 2020.
[55] RS Sutton, 시간차의 방법으로 예측하는 방법 학습, Mach. 배우다. 3 (1) (1988) 9–44.
[56] RS Sutton, AG Barto, 강화 학습: 소개, 두 번째 판, MIT Press, 2018.

[57] RS Sutton, DA McAllester, S. Singh, Y. Mansour, 함수 근사를 사용한 강화 학습을 위한 정책 기울기 방법, in: Advances in
신경 정보 처리 시스템, 2000, pp. 1057–1063.
[58] ME Taylor, P. Stone, 강화 학습 도메인을 위한 전이 학습: 조사, J. Mach. 배우다. 해상도 10 (1) (2009) 1633-1685.
[59] EC Tolman, 동물과 인간의 목적적 행동, Century/Random House, 영국, 1932년.
[60] AM Turing, 컴퓨터 기계 및 지능, Mind 59(236)(1950) 433.
[61] J. Vanschoren, 메타 학습: 조사, CoRR, arXiv:1810.03548 [abs], 2018.
[62] O. Vinyals, I. Babuschkin, WM Czarnecki, M. Mathieu, A. Dudzik, J. Chung, DH Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M Kroiss, I. Danihelka, A. Huang, L.
Sifre, T. Cai, JP Agapiou, M. Jaderberg, AS Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J Molloy , TL 페인, Ç. Gülçehre, Z. Wang, T. Pfaff, Y. Wu, R.
Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schhaul, TP Lillicrap, K. Kavukcuoglu, D.
Hassabis, C. Apps, D. Silver, 다중 에이전트 강화 학습을 사용하는 스타크래프트 II의 그랜드마스터 레벨, Nature 575(7782) (2019) 350–354.
[63] M. Weber, G. Roth, C. Wittich, E. Fischoff, 경제 및 사회: 해석 사회학 개요, 캘리포니아 대학 출판부, 1978.
[64] Y. Zhou, AlphaGo 대 Ke Jie, Slate and Shell, 2017.
[65] Y. Zhou, 개회 전략 재고: AlphaGo가 Pro Play, Slate 및 Shell에 미치는 영향, 2018.
[66] BD Ziebart, JA Bagnell, AK Dey, 최대 인과 엔트로피의 원리를 통한 모델링 상호 작용, in: Proceedings of the 27th International Con
기계 학습에 대한 참조, Omnipress, 2010, pp. 1255–1262.