코딩교육 열풍과 현주소 - 8

강화학습과 인공지능 교육의 미래

글. 김승욱 Rloha 대표, 데이터 분석 교육 및 컨설팅
'빅데이터 분석, R좀 R려줘', 'R 데이터 분석' 등 관련 칼럼 및 강의 진행

알파고와 자율주행 자동차에 활용되는 강화학습 이야기

내가 아닌 다른 무언가가 알아서 일한다는 것은 참 편리하다. 일을 대신해주는 것도 좋고 정확하고 빠르게 하면 금상첨화이다. 그렇게 1642년에 세계 최초의 계산기인 파스칼 계산기가 만들어졌고, 1946년에는 다용도 디지털 컴퓨터인 에니악이 나왔다. 그 이후로 컴퓨터와 로봇이 빠르게 발전했고 최근에는 딥러닝 기술도 빠르게 발전하고 있다. 하지만 아직은 로봇에게 명령하더라도 정말로 사람에게 시키는 것만큼 만족스럽지 못하다.

어떤 부분을 보완하면 좋을까? 사람의 오감(시각, 청각, 촉각, 후각, 미각)은 관련 센서 기술은 발전함에 따라 대부분 인간의 감각기관보다 성능이 좋아졌다. 하지만 아직 로봇의 어색함이 남아있는 이유는 바로 이 정보를 취합하고 의사결정하는 두뇌 때문이다. 대부분의 로봇에는 이 두뇌에 단순 조건 명령 정도만 지정되어 있다. 예를 들면 카메라에 사물의 움직임이 포착되면 지정된 전화번호로 문자 메세지를 보낼 수 있는 기능 등이다. 이 때문에 아직은 기계가 딱딱하고 어색하게 느껴지는데, 기계의 두뇌를 좀 더 말랑말랑하게 하기 위해서는 추가로 추론 기법이나 강화학습 등 여러 기술이 더해져야 한다. 그중에서 강화학습을 알아보고자 한다.

강화학습은 마치 반려동물을 훈련하는 것과 비슷하다. 예를 들어 강아지에게 '손~' 이라고 말하며 앞발을 사람 손 위에 놓는 동작을 가르친다고 하자.

강아지와 사람은 서로 말이 통하지 않기 때문에 가장 강한 매개체인 먹이(또는 간식)를 보상으로 특정 행동을 유도한다. 사람이 마음에 드는 행동을 하면 먹이를 주고, 그렇지 못할 경우 먹이를 주지 않거나 혼내기도 한다. 그리하여 강아지는 먹이를 얻기 위해 사람이 원하는 행동이 무엇인지 파악하고 학습하게 된다.

강화학습으로 가보면 '강아지'는 '에이전트(agent)', '먹이'는 '보상(reward)'이고, '혼남'은 '페널티(penalty)', 특정 조건에서 강아지가 취할 수 있는 '행동'은 말 그대로 '행동(action)'이라고 할 수 있다.

그림 1. 앞발을 내어주는 강아지

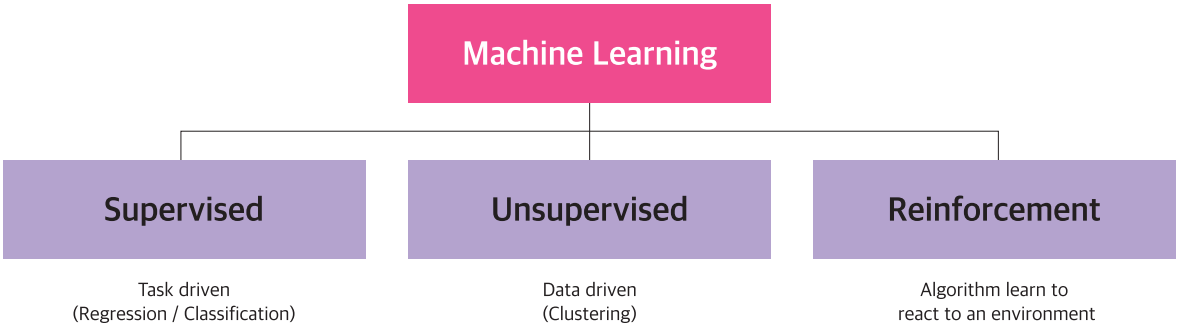

그림 2. 기계학습의 갈래

페널티의 경우 앞발을 주지 않을 뿐만 아니라 뒤로 돌아 가버리거나 손을 물어버리는 등 완전히 잘못된 행동을 했을 때 받는 것으로 이를 통해 강아지가 보상을 인지하고 이해하고 획득할 수 있도록 유도하는 것이 되겠다. 강아지가 '손~'이라는 명령에 주인에게 앞발을 내어주는 것을 다른 곳에서 본 적이 없다면 해당 명령을 이해하고 행동하기까지 시간이 걸릴 수 있다. 좀 더 명확하게는 여러 번의 시행착오가 필요하다는 뜻이다. 강화학습의 상위 개념인 기계학습의 관점에서 본다면 원하는 결과를 얻기까지 여러 번의 반복 학습이 필요한 것과 같다고 할 수 있겠다.

학습을 한다는 것은 학습 전보다 후에 개선이 되어야 하겠다. 수학 공부를 하더라도 틀렸던 문제를 계속 틀린다면 과연 학습이라고 할 수 있을까? 단지 공부를 했다는 만족감만 남을 수 있다. 강아지도 마찬가지다. 배는 고프고 밥을 먹고 싶은데 주인이 먹이를 입에다가 가져다주는 것이 아니기 때문에 조금 더 효율적이고 빠르게 먹기 위해서 눈칫밥으로 주인이 싫어하거나 반응이 없었던 행동을 배제하고 이래저래 시도를 해보기 마련이다.

"그래서 어떻게 하나요?"

그림 3. 야바위 최초 상태

강화학습은 보통 Q-러닝부터 시작한다. Q-러닝(Q-learning)은 모델 없이 학습하는 강화학습 기법의 하나인데, 주어진 상태에서 주어진 행동을 수행하는 것이 가져다줄 효용의 기대값을 예측하는 함수인 Q 함수를 학습함으로써 최적의 정책을 학습하는 것을 Q-러닝이라고 한다.[1] 예를 들어 강아지가 주인과 야바위를 한다고 하자.[2] 이런 상태(State)[3]를 잘 보여주는 다음 그림은 강아지에게 야바위를 위해서 컵 2개가 주어진 상황이다. 이를 최초 상태(S_0)라고 하자.

이제 강아지는 선택의 기로에 놓인다. 어떤 행동(Action)을 할 것인가? 관련 없는 행동을 할 경우 보상을 받지 못함은 물론 오히려 좋지 않을 수 있어 해당 사항은 제외하고, 수행 가능한 행동의 목록은 다음과 같다.

- 왼쪽 컵 선택
- 오른쪽 컵 선택

1) https://ko.wikipedia.org/wiki/Q_%EB%9F%AC%EB%8B%9
2) https://youtu.be/-q4DT80pVB4
3) 상태(State)란 환경(Environment)을 에이전트(Agent)가 관찰할 수 있는 정보이다.

그림 4. 왼쪽 컵을 선택한 강아지

여기에서 강아지는 최초의 상태(S_0)에서 왼쪽 컵을 선택하는 최초의 행동(A_0)을 했다.

이때 최초의 행동(A_0)으로 먹이라는 보상(Reward)을 얻게 되었다. 이 시점에서 1회 학습을 시도한 것이 된다. 그리고 주인이 다음 야바위 게임을 하려고 한다면 다시 [그림 3]처럼 돌아가게 된다. 하지만 이번에는 최초의 상태(S_0)가 아니라 S_1으로 된다. 이것을 상태 전이(transition)라고 한다. 여기까지의 학습을 도식화한 것은 다음과 같다.

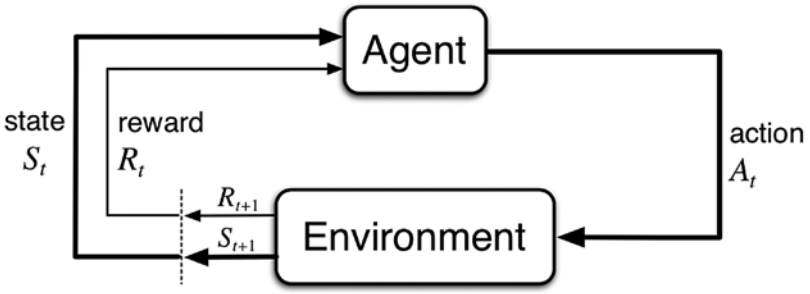

그림 5. Q-러닝 학습 개요도

첫 게임을 하면서 강아지는 이렇게 생각 할 수 있다.

'왼쪽인가?'

강아지는 약간 의심을 하면서 두 번째 게임(S_1)을 시작한다. 이번에도 똑같은 상황이 반복된다고 하자. 그러면 강아지는 두 번 연속으로 왼쪽을 경험했기 때문에 조금 더 확신을 가질 수 있다. 만약 이 상황이 10번 가까이 반복이 되었다고 하자. 그렇다면 강아지는 별 생각 없이 무조건 왼쪽을 택할 것이다. 이렇게 강아지는 이 야바위 게임에서 보상을 얻으려면 왼쪽을 골라야 한다는 학습[4]을 하게 된 것이다. 강화학습으로 따지면 강아지의 두뇌는 Q-함수에 연결시켜 볼 수 있다. 왜냐하면 앞에서 Q-함수가 '주어진 상태에서 주어진 행동을 수행하는 것이 가져다줄 효용의 기대값을 예측하는 함수'라고 했기 때문에 강아지는 지금 왼쪽의 컵을 고르면 보상을 받는다는 것을 학습하였고, 오른쪽보다는 왼쪽의 효용이 더 높다는 것을 알고 있는 상태이기 때문이다. 그리고 강아지는 계속 게임을 진행하면서 의심이 확신으로 바뀌는데 이는 강아지마다 다를 수 있다. 어떤 강아지는 두 번만 하면 망설임 없이 무조건 왼쪽을 고르고, 어떤 강아지는 열 번은 해야 망설임 없이 왼쪽 컵을 고를 수 있다. 이것은 Q-함수의 학습 정도를 결정짓는 학습률(learning rate)의 차이와 대응된다고 할 수 있다. 그리고 강아지는 새로운 게임을 할 때마다 이전에 했던 게임의 결과를 기억하기 때문에 그 정보를 토대로 또 의사결정을 하게 되겠다. 그래서 Q-러닝의 수식에는 이전 정보, 학습률, 보상 등 다양한 인자가 들어있다.

$$\overbrace{Q^{new}(s_t,a_t) \leftarrow \underbrace{Q(s_t,a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1},a)}_{\text{estimate of optimal future value}} \right)}^{\text{new value (temporal difference target)}} \underbrace{- Q(s_t,a_t)}_{\text{old value}}}^{\text{temporal difference}}$$

그림 6. Q-러닝 수식

4) 물론 강아지는 뛰어난 후각을 가지고 있어 굳이 왼쪽 오른쪽을 크게 고민하지 않아도 되지만 그 부분은 고려하지 않기로 하자.

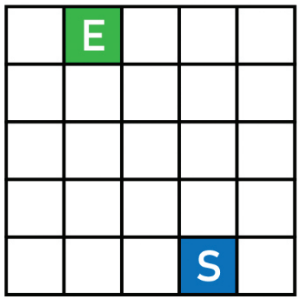

그림 7. 보드게임 1

간단하게 왼쪽 오른쪽 말고 조금 더 다양한 경우의 수를 가진 상황을 생각해보자. 다음과 같은 보드게임에서는 시작지점 S에서 도착지점 E까지 이동하면 게임이 끝난다고 하자. 그리고 이동은 한 번에 한 칸만 가능하며 도착지점 E는 해당 칸에 도달하기 전까지는 알 수 없다고 한다.

이전에 야바위 게임이 아니기 때문에 선택이 좌/우 두 개가 아니라 앞/뒤/좌/우 이렇게 각기 다른 4개의 행동(Action)이 가능하다고 할 수 있다. 이 게임은 목적지에 도달해서 보상을 얻기까지 최소 6번의 이동을 해야 하고 그전까지는 보상을 받을 수 없다. 그리고 앞으로 간 다음에 뒤로 갈 수 있어 언제나 최단 거리로 목적지에 도달하여 보상을 받는다는 보장도 없다. 다음 그림을 보자.

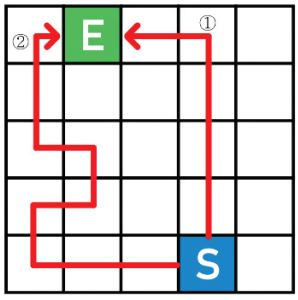

그림 8. 목적지 도달 예시

①번 경로는 최단 거리, ②번 경로는 조금 돌아가는 경로이다. 처음부터 최단 거리를 가면 좋겠지만 에이전트는 사전 정보가 없으므로 무작정 한 칸씩 이동하는 방법밖에 대안이 없다. 그래서 돌아가는 경로로 학습을 할 수 있고 학습을 계속 반복하면 결국 최적의 경로를 곧잘 찾아가는 학습을 하게 될 것이다. 당연하게도 이 게임은 이전의 야바위 게임보다 학습하는 비용(시간, 연산)이 훨씬 더 걸릴 것이다.
조금 더 확장해보자. 5×5판에서 목적지에 도달하는 상황이 아니라 25×25는 어떻게 될까? 혹시나 장애물이 있다면? 그리고 다양한 현실의 복잡한 상황에서도 잘 동작하는 인공지능을 구현하려면 고민을 많이 해야 한다. 이 부분에서 연구자는 에이전트가 보다 잘 학습하기 위해 다양한 시도를 한다. 로봇으로 문을 여는 보다 현실적인 예제[5]를 보자.

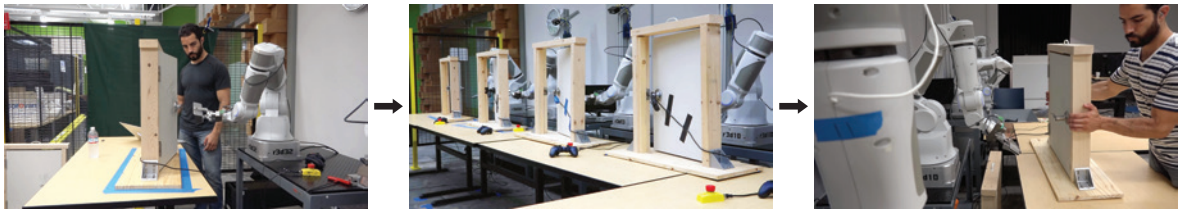

그림 9. 문을 여는 방법을 학습하는 로봇팔

로봇팔에 뜬금없이 문을 열라고 하면 로봇이 우왕좌왕하면서 그냥 팔을 흔들 수도 있고 어떤 경우에는 문을 부숴버릴지도 모른다. 이전에는 행동이 4가지(앞/뒤/좌/우)로 제한되어 있고, 행동 가능한 공간도 5×5 크기의 격자였다. 하지만 이런 로봇팔의 경우에는 로봇팔 관절마다 자유도가 주어지기에 손잡이를 돌려 문을 열어 보상을 획득하려면 제한된 공간이긴 하지만 거의 무한개에 가까운 세부 공간을 조금씩 움직여가는 것이 되어서 학습에 얼마나 시간이 걸릴지 가늠할 수도 없다. 그래서 대안으로 사람이 직접 지도를 해준다. 이제 로봇팔은 최초에 사람이 알려준 방법대로 문을 열어보는 연습을 하게 되고, 시도마다 보상을 얻으며 문을 여는 것에 익숙해진다. 하지만 실제 상황에서는 로봇과 문이 마주하고 있는 거리도 각도도 다를 수 있다.

5) 영상은 QR 코드로 확인 가능

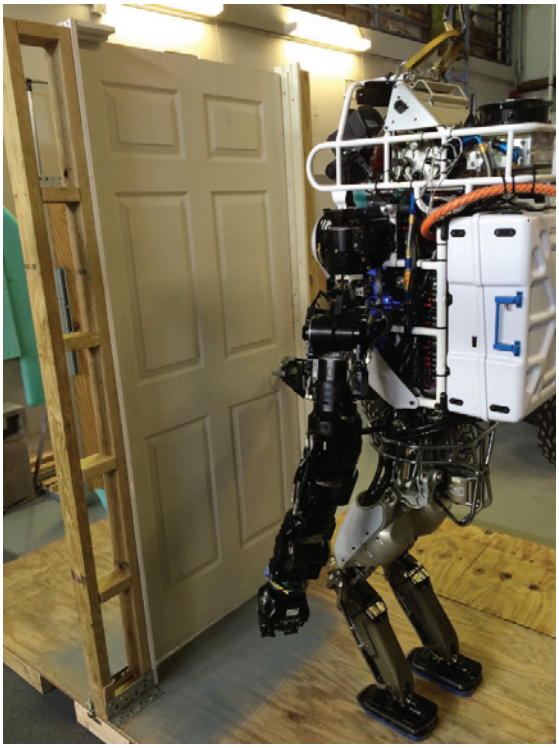

그림 10. 문 앞에 서 있는 로봇

영상[6]을 보면서 눈치챈 사람도 있겠지만, '문을 연다'라는 동작은 여러 세부 동작을 포함한다.

1. 문에 손을 가져다 놓는다.
2. 손잡이를 움켜쥔다.
3. 손잡이를 한쪽으로 충분히 돌린다.
4. 문을 충분히 당긴다.
5. 움켜쥔 손잡이를 놓는다.

우리가 원대한 하나의 목표를 세우는 것보다 목표를 나눠서 세우는 것이 성취감도 동기부여도 받는 것처럼 강화학습도 계획을 세우고 중간 보상을 조금씩 둬서 보다 학습을 잘 하게 할 수 있다. 게임을 처음 하는 것은 같지만 게임을 샀을 때 공략집이 같이 제공되는 것과 그렇지 않은 경우에 게임 난이도가 훨씬 낮아지는 것과 비슷하다고 할 수 있다. 즉, 강화학습 계획을 세우고 그에 맞춰서 중간 보상을 주는 등 체계적인 학습을 유도하면 보다 적은 양의 자원(시간, 연산량 등)으로도 그 최종 목표로 하는 에이전트의 학습 수준에 도달할 수 있게 된다. 그리고 문을 너무 세게 잡거나 문이 몸체에 닿는 등 특정 동작 또는 상황이 발생하게 되면 페널티(음의 보상)를 주게 할 수 있다. 이와 관련된 내용은 커리큘럼 러닝(Curriculum Learning)에서 살펴볼 수 있다.

앞의 내용을 종합해서 정리하자면 강화학습은 보상을 얻기 위해 여러 차례 시도를 해보면서 최적의 동작을 하도록 학습하는 것이라고 할 수 있다. 현실과 가까운 문제일수록 고려할 내용이 많고 학습에 필요한 비용도 기하급수적으로 증가하기 때문에 관련 알고리듬 효율화와 학습 시간 단축을 위해 여러 에이전트의 정보를 종합하기도 하는 등 다양한 접근이 이루어지고 있는 학술 분야라고 할 수 있다.

앞에 소개한 내용 이외에도 강화학습 관련 내용을 찾아본다면 다음의 영상을 살펴보도록 하자. 필자가 추천하는 것은 숨바꼭질 영상인데 다양한 조건에서 술래가 어떻게 행동하는지 보면 정말 신기하기도 하고 재미있다.

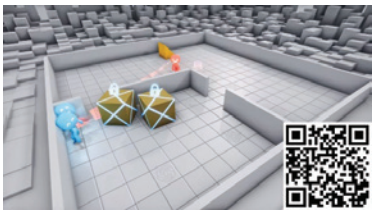

그림 11. 강화학습 - 숨바꼭질

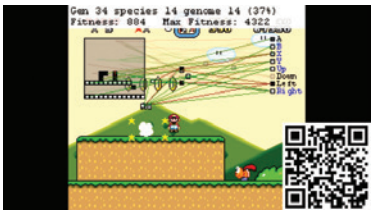

그림 12. 강화학습 - 마리오

그림 13. 강화학습 - 스타크래프트 2

6) https://groups.csail.mit.edu/robotics-center/public_papers/Marion16.pdf

우리가 가야 할 길

여러 회차에 걸쳐 코딩교육을 위한 다양한 프로그래밍 언어부터 관련 기술, Python 그리고 여러 인공지능 관련 기술을 알아보았다. 세상에는 정말 다양한 기술이 빠르게 발전하고 있고 그 기술을 따라가는 것 또한 벅차다. 그런데 인공지능이 세상을 정복한다며 딥러닝이니 머신러닝이니 하지만, 정작 엑셀 함수 쓰기도 버겁고 인공지능 비서는 자기 멋대로 켜져서 "부르셨어요?" 하는 세상에 있다.

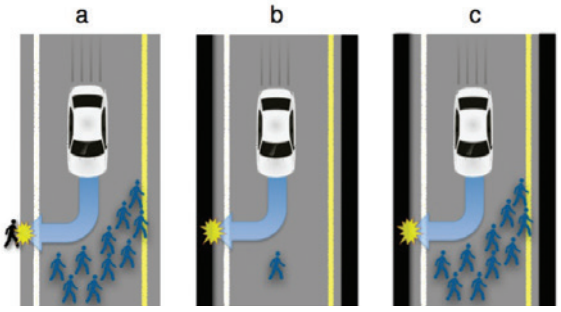

그림 14. 트롤리 딜레마

어디서는 정말 멋지고 성능 좋은 인공지능 모델을 만들었지만 어떻게 동작하는지 해석하기 어려워 당장은 활용에 문제가 있는 것도 있다. 대표적으로 자율주행 자동차의 트롤리 딜레마가 있다. 운전을 인공지능에 맡길 경우 도로주행 시 갑자기 보행자가 등장한다면 이를 피하는 것이 맞겠으나 피할 경우 운전자가 사망에 이를 수 있을 때 어떤 동작을 수행할지 결정하는 문제이다.

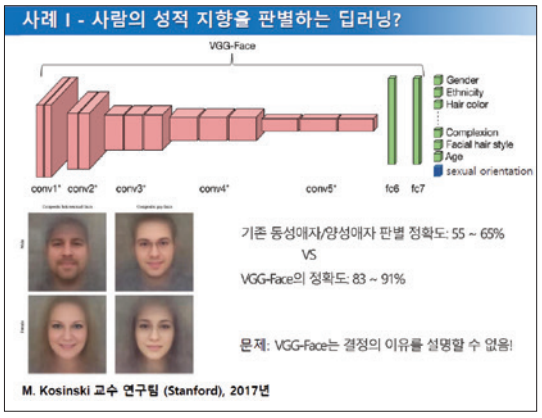

그림 15. 개인의 성적 지향을 판별하는 인공지능 예시

그리고 사람의 얼굴을 보고 개인 성적 지향을 판별하는 알고리즘이 2017년 기준 80% 이상의 정확도를 보여준다고 한다.

앞의 자율주행에서도 얼굴 판별에서도 보다 정교한 인공지능을 만들기 위해서는 인공지능이 왜 이런 결과를 도출했는지보다 면밀한 파악이 필요하다. 그래서 또 연구되고 있는 분야가 설명 가능한 인공지능인 XAI(eXplainable Artificial Intelligence)이다.

인공지능을 만들고 개선하려면 다양한 분야의 학문이 필요하다. 하드웨어를 담당하는 로봇공학 전반의 학문, 이미지/영상/자연어/음성의 입출력 데이터를 처리하는 기술 전반의 학문, 그리고 이 모든 것을 체계적으로 묶어내는 딥러닝 기술이 필요하다. 추가적으로는 패러다임이 바뀌는 이 시점에 기술이 생활에 잘 녹아들 수 있도록 법과 윤리도 고민해야 하고 관련 학문이 필요한 것은 자명하다.

어떻게 보면 '그래도 도대체 뭘 공부해야 하는가' 망설이고 있는 사람이 많지 않을까 싶다. 그런 사람을 위해서 첨언을 하자면, 앞의 대부분을 받쳐주는 근간이 된다고 할 수 있는 지식과 기술 영역은 따로 있다. 바로 수학, 영어, 컴퓨터 언어이다. 아무리 시간이 지나도 수학과 영어는 그 중요성이 떨어지지 않고, 이제 그 지식 영역을 보다 잘 활용하고 확장하기 위해서 컴퓨터 언어가 추가되었을 뿐이다. 컴퓨터와 평생을 같이 보내는 시대. 이 참에 한 번 서점으로 가서 평소에 공부하고 싶었던 학술 분야 서적 한 권. 프로그래밍 서적 한 권 구입해보는 것은 어떨까?