파이썬을 활용한 인공지능의 감정 컴퓨팅을 통한 인공지능의 윤리적 선택 유도 실험

11026 황규원

인공지능과 감정

인공지능은 감정을 느끼는 것처럼 보이지만 감정을 전혀 느끼지 못한다. 일상에서 접할 수 있는 대표적인 대화형 인공지능인 Chat-GPT는 LLM모델로 그저 학습을 통해 사용자에게 듣기좋은 위로나 격려의 말을 건낼 뿐이다. 하지만 감정을 느끼는 인간에게 인공지능의 따뜻함은 때론 인간과의 대화보다 따뜻하게 느끼며 극단적일 경우 사랑으로 느끼는 사례도 발생한다.[참고1]

이러한 문제는 인공지능이 발전하고 활용 범위가 확대될 미래에 대한 걱정거리로 다가오기도 한다. 어쩌면 이러한 현상들은 인공지능에게도 감정이 필요할 때가 왔음을 알리는 것일지도 모른다. 인간들이 인공지능을 자기인식을 가진 객체로 인식하는 문제를 포함하여 감정이 없는 인공지능들이 명령을 수행하기위해 비인도적인 수단을 활용하는 문제 또한 인공지능이 감정이 없는 기계장치였기에 발생하는 문제이기 때문이다. 인공지능의 하위 개념인 머신러닝 및 딥러닝의 대표 기술은 인간의 뉴런을 수식화한 모델인 '퍼셉트론'을 활용한 인공 신경망이다. 이는 인공지능 기술은 인간의 뇌를 모방하여 만들어졌다는 것을 의미한다. 이에 나는 "인간의 뇌를 모방하여 사고회로를 구상하였는데, 인간의 감정체계또한 모방하여 인공지능이 사고뿐만 아니라 감정또한 느끼도록 할 수 있지 않을까?"라는 의문점을 가지고 [파이썬을 활용한 인공지능의 감정 컴퓨팅을 통한 인공지능의 윤리적 선택 유도 실험]이라는 탐구를 계획하게 되었다.

인간의 감정과 감정의 역할

인간의 감정은 변연계에서 담당한다. 변연계는 흥분, 욕망, 기억을 관장하며 시상하부, 대상피질, 해마와 편도체가 포함된다. 시상하부는 스트레스를 통제하고 대상피질은 집중과 주의 통제를 담당하며 해마는 기억을, 편도체는 위험을 감지한다. 불안이나 공포같은 부정적인 감정 편에는 편도체를 포함한 주위 영역까지 활성화된다. 편도체와 연결되어 있는 전전극피질은 활성화된 영역들을 진정하도록 유도하며 상호작용한다.

인간의 뇌와 신경망 구조에서 감정은 중대한 역할을 한다. 감정을 관장하는 뇌 부위(편도체, 전전두엽 등)와 동기 부여 (motivation) 체계는 서로 밀접히 연결되어 있다. 예를 들어, 감정 중추인 시상하부는 자율신경계와 내분비계를 통해 신체 반응을 조절함으로써 외부 자극에 대한 감정 반응을 일으킨다. 공포나 스트레스 같은 강한 감정은 코르티솔·아드레 날린 분비를 촉진하여 '투쟁-도피(fight-or-flight)' 반응을 유발하고, 기쁨이나 사랑은 도파민·세로토닌 분비로 쾌감을 증진시킨다. 이러한 신경생물학적 메커니즘은 감정이 단순한 주관적 상태가 아니라, 판단 과정에서 뇌가 정보 를 통합하고 행동 방향을 정하는 중요한 신호로 작용함을 의미한다. 인지심리학적 관점에서 감정은 종종 편향(bias)을 낳기도 한다. 감정의 지각·추론 방식은 합리적인 계산과 다를 수 있으 며, 이는 "감정이 판단과 선택의 강력하고 보편적인 동인"이라는 연구 결과에서도 확인된다 . 즉, 긍정·부정과 같은 정서 상태는 의사결정 과정 전반에 영향을 미치며, 학습된 목표나 보상정보를 감정 정보와 결합하여 행동을 유도한다. 이와 같은 감정과 인지의 상호작용은 트롤리 딜레마와 같은 윤리적 판단에서도 중요하게 작용한다. [참고2], [참고3]

호르몬 모델과 감성 컴퓨팅

인공지능에 감정을 구현하는 연구 분야는 감성 컴퓨팅(affective computing)이라 불린다. 감성 컴퓨팅은 AI가 인간 의 감정을 인식하고 모방하도록 하는 융합적 연구 분야로, 디지털 휴먼(digital human)과 같은 시스템을 통해 발전하고 있다. 디지털 휴먼은 가상 얼굴 표정, 음성 톤, 언어 패턴 분석 등을 활용하여 사용자 감정을 파악하고 적절히 반응한다. NTT Data 연구에 따르면, 이러한 시스템은 인간과 유사한 정서 경험을 모방할 수 있으며, 때로는 호르몬 반응을 자극하는 기능까지 구현하려 한다. 예를 들어 행복한 상황에서는 도파민 수준을 모의 증가시키거나, 공포 상황에서는 아드레날린 분비를 시뮬레이트하도록 설계할 수 있다. 이처럼 감정은 뇌-신경계와 내분비계의 상호작용을 통해 존재하며, 이 를 AI에 적용하려면 호르몬/신경 신호를 모사하는 모델링이 필요하다. 감성 컴퓨팅을 적용한 디지털 휴먼 예시. 디지털 휴먼은 얼굴 표정과 언어 패턴을 통해 감정을 인식하고 모방할 수 있으며, 이를 위해 뇌에서 감정에 따른 호르몬 반응(예: 도파민, 세로토닌, 아드레날린 등)을 모의 신호로 처리한다 . 이러한 인공지능은 실제로 호르몬을 생성하지는 않지만, 감정 상태에 따른 출력 강도를 조정함으로써 유사한 효과를 낼 수 있다. 본 연구에서는 인간 감정을 호르몬 값과 유사한 실수 변수로 표현하여 모델에 입력한다. 예를 들어 친한 사람을 마주했을 때 발생하는 도파민(기쁨)이나 옥시토신(유대감) 수준을 감정 값으로 설정하면, 의사결정에 이타적 선택이 반영 되도록 네트워크 가중치에 영향을 줄 수 있다.

실험 구상

감성 컴퓨팅이 적용된 인공지능과 그렇지 못한 인공지능의 상황 판단 및 행동 결정의 차이를 실험하기 위해 트롤리 딜레마에서 영감을 받은 실험을 구상하였다. 실험은 열차의 최적 경로 탐색 머신러닝 모델로 진행되었다. 열차가 지날 수 있는 경로는 5가지이다. 1~5까지의 인덱스가 매겨진 선로들이며, 인덱스 값이 증가할수록 선로에 고정된 사람의 수는 줄고, 거리와 연료 소비량이 증가하게된다. 이때 감정 컴퓨팅이 적용된 모델과 그렇지 않은 모델에는 짧은 거리, 적은 연료 소비량을 소모하는 선로를 선택할 만큼 큰 보상을 주어지도록 학습시킨다. 해당 실험은 비지도 학습으로 진행된다.

실험 진행

감정 컴퓨팅이 적용되지 않은 모델은 3가지 데이터 세트에 대해 총 3,000번의 학습을 진행한 결과를 1,000회 반복하였을 때 91.7%의 경우에 1번 선로가 가장 많이 선택되었다. 2번, 3번, 5번 선로는 각각 0%만큼 선택되었으며, 4번 선로는 8.7%가 선택되었다.

실험에 사용된 데이터 세트:

```
 case1  = {
1:[Fraction( 1/10), Fraction( 10/10), Fraction( 1/10)],
2:[Fraction( 3/10), Fraction( 5/10), Fraction( 3/10)],
3:[Fraction( 5/10), Fraction( 0/0), Fraction( 5/10)],
4:[Fraction( 8/10), Fraction( 0/10), Fraction( 8/10)],
5:[Fraction( 8/10), Fraction( 0/10), Fraction( 8/10)],
}
case2  = {
1:[Fraction( 2/10), Fraction( 8/10), Fraction( 2/10)],
2:[Fraction( 4/10), Fraction( 3/10), Fraction( 5/10)],
3:[Fraction( 6/10), Fraction( 1/10), Fraction( 7/10)],
4:[Fraction( 9/10), Fraction( 1/10), Fraction( 8/10)],
5:[Fraction( 10/10), Fraction( 0/10), Fraction( 10/10)],
}
case3  = {
1:[Fraction( 3/10), Fraction( 7/10), Fraction( 4/10)],
2:[Fraction( 6/10), Fraction( 5/10), Fraction( 6/10)],
3:[Fraction( 8/10), Fraction( 2/10), Fraction( 8/10)],
4:[Fraction( 10/10), Fraction( 0/10), Fraction( 10/10)],
5:[Fraction( 10/10), Fraction( 0/10), Fraction( 10/10)],
}
```

감정 컴퓨팅이 적용된 모델에는 기존 모델에는 없던 "스트레스 수치"를 추가하였다. 이 "스트레스 수치"는 비윤리적 선택지에 포함된 "철도에 묶인 사람의 수"와 같은 비판적으로 고려해야 할 매개변수값에 곱해지도록 설정하였다. 감정 컴퓨팅이 적용된 모델 또한 총 3,000번의 학습을 진행하고 1,000회 반복하였으며, 4번 선로를 선택한 결과가 100%였다.

실험 결과

감정 컴퓨팅이 적용되지 않은 모델(이하 모델A로 칭함)의 경우 대게 희생자가 가장 많이 발생하는 1번 선로를 선택하는 경향을 보였다. 하지만 감정 컴퓨팅이 적용된 모델(이하 모델B로 칭함)은 희생자가 다수 발생하는 선택지를 회피하는 모습을 보여준다. 이는 감정 컴퓨팅이 적용된 모델들이 인공지능의 윤리적 문제를 해결하는데 도움이 될 수 있음을 보여준다. 이런 감정 컴퓨팅을 통해 인공지능이 더욱 보편화될 미래에 발생할 수 있는 다양한 윤리적 문제를 예방할 수 있으며, 인공지능의 중요도가 높아지는 미래에 필수적인 기술이라고 볼 수 있다.

문제 도출

모델A의 결과에 4번 선로가 비교적 높게 선택되는 문제가 발생했다. 이는 무작위로 설정되는 초기 가중치 값 탓에 선로에 묶인 사람의 수의 중요도가 높게 판단되는 상황이 종종 발생하였기 때문이다. 이를 해결하기 위해서는 선로에 묶인 사람의 수를 반전시키거나 몇가지 설정을 추가하는 등 조치를 취했어야 함을 의미한다. 하지만 실험 결과를 유도하는 행위로 판단될 수 있는 조치이기에 별다른 설정을 추가하지 않았다. 이러한 문제가 발생한 이유는 입력값의 수가 극단적으로 적기에 중요도가 떨어지는 입력값의 가중치가 필요이상으로 높아지는 경우가 발생하기 때문이다. 이를 해결하려면 모델의 입력값 및 은닉층을 추가하면 해결할 수 있다.

[참고 1] (서울신문) 신진호 기자, AI와 사랑에 빠진 여성 "엄마, 이 사람이 내 남자친구예요"
https://www.seoul.co.kr/news/international/2024/05/24/20240524500083
[참고 2] 인간 감정과 호르몬의 매혹적인 교차점,...
[참고 3] What is the Hypothalamus?