거대언어모델 기반 실시간 로봇 제어 기술에 대한 연구

박진수, 신수용*

국립금오공과대학교

jp@kumoh.ac.kr, *wdragon@kumoh.ac.kr

Real-time robot control system based on Large-Language-Model

Jin Su Park, Soo Young Shin*

Kumoh national institute of technology.

요 약

거대언어모델(LLM: Large Language Model) 기술의 발전과 함께 이를 활용한 다양한 시스템 제어 기법이 연구되고 있다. LLM을 기반으로 하는 제어 기법은 사용자와의 상호작용과 맥락 도출을 통해 다양한 환경에 적용할 수 있고, 자연어 기반 입력을 통해 사용자 친화성이 뛰어나다는 장점이 있다. 최근에는 시각적 정보와 언어를 함께 다루는 비전언어모델(VLM: Vision Language Model)을 결합하여 주변 환경을 인식하여 처리하는 방법이 함께 연구되어 활용되고 있다. 본 논문은 LLM과 VLM을 기반으로 하는 로봇 제어 기술의 구조와 구현에 필요한 핵심 요소를 실시간 제어에 초점을 두어 분석하고 소개한다.

Ⅰ. 서 론

인공지능의 발전으로 다양한 산업에서 인공지능을 활용한 기술이 활발히 적용되고 있다. 특히 LLaMA와 ChatGPT와 같은 거대언어모델(LLM)은 자연어 처리 기술을 통해 복잡한 시스템, 예를 들어 로봇 제어와 같은 작업을 사용자의 자연어 입력을 기반으로 수행할 수 있어 큰 주목을 받고 있다[1][2]. 그러나 이러한 LLM 기반 제어 시스템은 자연어 입력의 구조적 한계로 인해 사전에 학습된 정보에만 의존하거나, 사용자가 환경 정보를 직접 입력하거나 인식하기 위한 별도의 과정이 필요하다는 단점이 있다[3]. 이에 따라 시스템의 실시간성이 저하되는 문제도 발생한다. 최근에는 카메라와 같은 센서를 활용하여 주변 환경을 인식하고, 이를 처리해 자연어로 변환하는 비전언어모델을 결합한 연구가 활발히 진행되고 있다[4][5]. 본 연구에서는 실시간 제어가 가능한 거대언어모델 기반 로봇 제어 시스템의 구성을 제안하고, 각 핵심 요소를 설명한다.

Ⅱ. 본론

그림 1은 제안하는 거대언어모델(LLM) 기반 실시간 로봇 제어 기술의 시스템 구성이다. 제안하는 시스템은 사용자의 입력을 로봇의 제어를 위한 입력값으로 변환하는 LLM, 주변 환경을 인식하고 이를 자연어로 변환하여 LLM에 입력하기 위한 비전언어모델(VLM)이 사용된다. 제안하는 시스템 구성에서 LLM은 크게 3가지의 입력을 가진다. 첫 번째는 사용자로부터 입력받는 명령, 두 번째는 시스템 구성 시 사전에 입력하는 로봇의 기본적인 동작에 대한 정의 및 제어 코드가 포함된 API, 마지막은 VLM을 통해 자연어로 변환된 주변 환경 정보다.

사용자의 입력은 LLM을 통해 로봇이 실제로 수행할 구체적인 동작으로 변환된다. 먼저 사용자의 입력에서 핵심이 되는 키워드를 추출하고, 이를 바탕으로 명령의 의도를 해석한다. 예를 들어, 제어할 대상이 로봇팔일 때, “휴지를 책상 밖 휴지통에 버려줘”라는 명령이 입력되면, 대상을 나타내는 “휴지”, 위치정보를 나타내는 “책상 밖 휴지통”, 그리고 동작을 나타내는 “버려줘”가 추출된다. 최종적으로 이 입력을 통해 로봇이 “휴지”라는 객체를 “책상 밖 쓰레기통”이라는 지정된 위치로 “이동시키는” 동작을 추론한다.

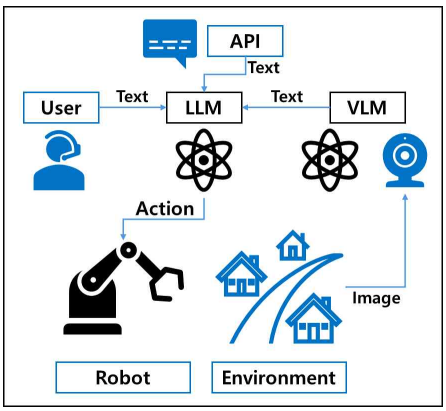

그림 1 제안하는 거대언어모델 기반 실시간 로봇 제어 시스템 구성

사용자 입력을 처리하는 과정에서 시스템 성능을 향상하는 방법의 하나로 프롬프트 엔지니어링이 있으며, 대표적인 기법으로 사고 사슬(CoT: Chain of Thought)이 있다[6]. 사고 사슬 기법은 여러 단계의 사고가 필요한 입력을 단계별로 나누어 제시하는 방법이다. 표1은 사슬 사고 기법을 활용해 사용자의 입력을 5단계로 나눈 것이다. 이를 통해 주어진 환경에 있는 객체의 유무, 위치정보를 인지하여 명령 수행 정확성을 향상하고 오류를 줄일 수 있다.

1. “휴지가 어디에 있는지 찾아”
2. “휴지가 있다면 그 위치를 인식해”
3. “휴지를 집고 책상 밖으로 이동해”
4. “쓰레기통이 어디에 있는지 찾아”
5. “휴지를 쓰레기통에 버려”

표 1 사슬 사고로 구분한 사용자 입력 예시

LLM은 사용자의 명령에서 추출된 키워드를 기반으로 사용자의 의도를 파악하고, 이를 바탕으로 로봇이 수행할 동작을 정의한 후, 로봇 제어를 위한 코드를 생성한다. 코드 생성에는 시스템 구성 시 입력한 API의 정보를 활용한다. API는 로봇의 기본적인 동작 및 제어와 관련된 함수와 명령어들이 미리 정의되어 있는 인터페이스로, LLM이 사용자 명령을 구

체적인 동작을 실행하는 명령어로 변환하는 데 필요한 정보를 제공한다.

LLM은 로봇이 수행할 동작에 맞게 사용자의 입력에서 추출한 키워드와 사전에 정의된 로봇 제어 API를 매칭하여, 자연어 명령을 API 함수 호출로 변환한다. 그림 2는 표 1의 사용자 입력을 받아 생성한 로봇 제어 코드의 예시이다. LLM을 사용해 추출한 키워드를 호출한 함수의 입력으로 사용하여 순서에 맞게 함수를 실행한다.

Algorithm 1 로봇팔을 이용한 휴지 버리기

```
function main()
  # 1. 휴지가 어디에 있는지 찾아
  tissue_position = detect_object("휴지")
if recognize_position(tissue_position) then
    # 2. 휴지가 있다면 그 위치를 인식해
    move_to(tissue_position)                    ▷ 휴지 위치로 이동
    grip_object()                               ▷ 휴지 집기
  # 3. 휴지를 집고 책상 밖으로 이동해
    move_to(outside_desk_position)              ▷ 책상 밖으로 이동
  # 4. 쓰레기통이 어디에 있는지 찾아
    trash_bin_position = detect_object("쓰레기통")
  if recognize_position(trash_bin_position) then
        move_to(trash_bin_position)             ▷ 쓰레기통 위치로 이동
        release_object()                        ▷ 휴지를 쓰레기통에 버림
  else
        print("쓰레기통 위치를 찾을 수 없습니다.")
  end if
else
    print("휴지의 위치를 찾을 수 없습니다.")
end if
```

그림 2 API를 활용해 LLM으로 생성한 로봇 제어 코드 예시

사용자의 자연어 입력을 코드로 변환할 때 시스템의 성을 향상시키는 방법으로 API와 함께 작업에 관련된 시나리오나 서식을 추가하는 기법이 있다. 시나리오와 서식이 주어지면 복잡한 작업을 단계별로 체계화하고 표준화하여 로봇이 일관된 동작을 실행할 수 있게 해 오동작을 방지한다. 또한 LLM의 추론 능력을 통해 새로운 환경에서 기존의 시나리오나 서식을 확장하는 방식으로 유연한 대응을 가능하게 한다.

LLM 기반 실시간 로봇 제어 기술에서, 시스템은 외부 센서 데이터에 의존하여 주변 환경을 인식한다. LLM만을 활용한 시스템의 경우 카메라로 주변 환경을 이미지로 획득하여 Yolo, OpenCV 등의 컴퓨터 비전 알고리즘을 사용해 물체를 탐지한다. 이때 물체의 종류를 클래스로 분류하고, 위치, 크기, 거리 등의 사전에 정의한 정보를 계산하여 LLM의 입력으로 활용한다. 하지만 이러한 방법은 시각적 정보를 언어 데이터와 결합하는 능력이 떨어져 획득한 정보 간의 상관관계를 도출하기 어렵고, 변화하는 동적 환경에서 적응성이 떨어진다.

그림 3은 대표적인 VLM 모델인 CLIP의 학습 구조를 나타낸다[7]. CLIP은 이미지와 텍스트를 대조 학습으로 연결하여, 두 가지 데이터를 함께 처리할 수 있는 멀티모달 모델로 텍스트로 입력된 설명에 맞는 이미지를 찾거나, 이미지를 보고 해당하는 텍스트 설명을 생성할 수 있다.

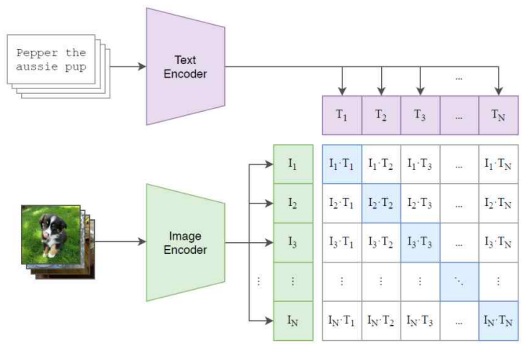

그림 3 VLM 모델 CLIP의 대조 학습 구조

제안하는 시스템은 LLM과 VLM을 결합해 로봇이 실시간으로 주변 환경을 인식하고 인식한 정보를 LLM으로 전달해 보다 정확한 명령 수행을 할 수 있게 한다. 예를 들어 표1의 경우 VLM을 사용해 로봇 주변 환경을 인식해 휴지, 쓰레기통이 어디에 있는지 위치를 도출하고, 쓰레기를 정확하게 집었는지, 도중에 놓쳤는지 인지할 수 있으며, 휴지가 책상 밖으로 이동했는지, 객체 간의 상관관계를 통해 쓰레기통에 휴지가 들어갔는지 판단 할 수 있다.

Ⅲ. 결론

본 논문은 거대언어모델(LLM)과 비전언어모델(VLM)을 결합한 실시간 로봇 제어 시스템의 구성을 제안한다. 제안하는 시스템은 로봇의 기본적인 정보와 로봇 제어를 위한 API만 있다면, LLM을 통해 사용자의 자연어 명령을 이해하여 프로그래밍 없이 로봇 제어를 할 수 있으며, VLM을 통해 실시간으로 주변 환경을 인식하며 피드백을 지원할 수 있을 것으로 기대된다. 추후 ChatGPT와 CLIP을 사용해 제안하는 시스템을 실제 로봇 환경에 구현하여 다양한 산업 및 실생활에서의 적용 가능성을 테스트하고, 실시간 작업 수행 능력과 효율성을 검증하고자 한다.

ACKNOWLEDGMENT

이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업" (IITP-2024-RS-2024-00437190, 50%)과 과학기술정보통신부 및 정보통신기획평가원의 ICT혁신인재4.0의 연구결과로 수행되었음" (IITP-2024-RS-2022-00156394, 50%)

참 고 문 헌

[1] Brown, T. B., Mann, B. “Language Models are Few-Shot Learners.” 2020 arXiv preprint arXiv:2005.14165, https://arxiv.org/abs/2005.14165

[2] Touvron, H., Lavril, T. “LLaMA: Open and Efficient Foundation Language Models.” 2023 arXiv preprint arXiv:2302.13971, https://arxiv.org/abs/2302.13971

[3] Vemprala, S., Bonatti, R. “ChatGPT for Robotics: Design Principles and Model Abilities.” 2023 Microsoft Research Technical Report MSR-TR-2023-8, https://www.microsoft.com/en-us/research/publication/chatgpt-for-robotics-design-principles-and-model-abilitie

[4] Brown, T. B., Mann, B. “Language Models are Few-Shot Learners.” 2020 arXiv preprint arXiv:2005.14165, https://arxiv.org/abs/2005.14165

[5] OpenAI, Achiam, J. “GPT-4 Technical Report.” 2024 arXiv preprint arXiv:2303.08774, https://arxiv.org/abs/2303.08774

[6] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." 2023, arXiv preprint arXiv:2201.11903, https://arxiv.org/abs/2201.11903

[7] Radford, A., Kim, J. W. “Learning Transferable Visual Models From Natural Language Supervision.” 2021 arXiv preprint arXiv:2103.00020, https://arxiv.org/abs/2103.00020