대규모 멀티모달 모델 학습용 데이터의 구축 현황 분석

이진용, 박정하, 이상복
한국정보통신기술협회
wlsdyd620@tta.or.kr, parkjh516@tta.or.kr, jangpo@tta.or.kr

Analysis of the Building Status of Large Multimodal Model Training Data

Lee Jin Yong, Park Jeong Ha, Lee Sang Bok
Telecommunications Technology Association

요 약

본 논문은 2024년 과학기술정보통신부와 한국지능정보사회진흥원(NIA)에서 추진하는 '초거대AI 확산 생태계 조성 사업(1차)'에서 대규모 멀티모달 모델 학습용 데이터의 형식, 라벨링 유형 및 구축 목적을 기준으로 데이터를 분석하였다. 데이터 형식의 경우 텍스트, 영상·이미지, 3D, 음성, 수치형으로 정의하여 분류했으며, 라벨링 유형은 바운딩박스(B-Box)·세그멘테이션(Segmentation), 폴리곤(Polygon)·폴리라인(Polyline), 캡션(Caption), 전사(Transcription), 태깅(Tagging), 분류(Classification), 질의응답(Question/Answer), 번역(Translation)으로 분류하였다. 구축 목적은 텍스트 생성, 이미지 생성, 음성 인식/음성 합성, 객체/행위에 대한 탐지, 수치 예측, 클래스 분류, 로봇 동작 및 작업 생성, 경로 최적화로 분류하여 분석하였다. 이를 다각도로 분석한 결과, 대규모 멀티모달 모델 학습용 데이터의 다양성과 활용 가능성을 확인하였으며, 여러 응용 분야에서의 기술 혁신을 이끌어내는데 기여할 것으로 기대된다.

Ⅰ. 서 론

최근의 인공지능 기술은 컴퓨터 비전, 자연어 처리, 음성 인식 등 다양한 분야에서 혁신적인 발전을 이루어냈다. 특히 딥 러닝 기술의 발전은 인공지능 모델의 성능을 지속적으로 향상시키고 있으며, 이러한 발전은 대규모 언어 모델(LLM)의 등장과 함께 더욱 빛을 발하고 있다. GPT-3, LaMDA, PaLM과 같은 텍스트 기반의 대규모 언어 모델은 놀라운 성과를 보여주며, 다양한 분야에서의 활용이 이루어졌다.[1] 그러나 이러한 모델은 텍스트 데이터에 대한 처리에만 제한되어 있어, 멀티모달 데이터의 처리와 이해에 한계를 가지고 있다.

이러한 점을 극복하기 위해 최근에는 대규모 멀티모달 모델(LMM)이 등장하였다. 대규모 멀티모달 모델은 텍스트 뿐만 아니라 이미지, 음성, 영상 등 다양한 형태의 데이터를 처리하고 이해할 수 있다. 이러한 발전은 인공지능 기술의 새로운 지평을 열어주며, 다양한 응용 분야에서의 혁신을 도모할 수 있을 것으로 기대된다.

대규모 멀티모달 모델 학습용 데이터의 형식, 구축 목적, 그리고 라벨링 유형에 대해 분석하고자 한다. 이를 통해 멀티모달 데이터의 다양성과 활용성을 탐구하고, 이를 통해 다양한 응용 분야에서의 기술 혁신을 이끌어내는데 기여하고자 한다.

Ⅱ. 본 론

과학기술정보통신부와 한국지능정보사회진흥원에서 주관하는 '초거대AI 확산 생태계 조성 사업(1차)'의 대규모 멀티모달 모델 학습용 데이터 37종을 데이터 형식, 라벨링 유형, 구축 목적으로 분석하였다.

1. 데이터 형식

멀티모달 데이터는 다양한 형태의 데이터 형식이 결합되어 표현되며, 이러한 다양성은 인공지능 모델이 실제 세계의 복잡한 정보를 정확하게 이해하고 처리할 수 있도록 돕는 중요한 역할을 한다. 따라서, 멀티모달 데이터의 형식을 분석함으로써 특징을 파악하고, 이를 통해 다양한 분야에서의 활용 가능성을 탐구한다.

본 논문에서는 대규모 멀티모달 모델 학습용 데이터 37종의 데이터 형식을 텍스트, 영상·이미지, 3D, 음성, 수치형 총 5가지로 분류하였다.

- 텍스트 데이터: 자연어로 이루어진 정보를 포함하고 있으며, 대화, 요약, 캡션 등 다양한 형태로 표현될 수 있다.
- 영상·이미지 데이터: 시각적인 정보를 포함하고 있으며, 2D 이미지의 경우 정적인 이미지를 표현하고, 영상은 시간의 흐름에 따라 이미지가 연속적으로 변화하는 형태이다.
- 3D 데이터: 공간적인 정보를 포함하고 있으며, 3D 포인트 클라우드(PCD)와 3D 모델링 데이터 등으로 구성된다.
- 음성 데이터: 사람의 발음하는 말소리를 포함하고 있으며, 음성 인식 및 음성 합성 분야에서 주로 활용된다.
- 수치형 데이터: 숫자로 이루어진 정보를 포함하고 있으며, 센서 데이터 및 시계열 데이터 등의 형태로 구성될 수 있다.[2]

멀티모달 데이터는 단일 데이터에 다수의 데이터 형식이 포함될 수 있으며, 37종의 데이터에서 각 데이터 형식의 개수와 비율은 표 1과 같다.

표 1. 대규모 멀티모달 데이터 37종의 데이터 형식 분석 결과

번호	형식	개수	비율
1	영상·이미지	31	83.78%
2	텍스트	29	78.38%
3	수치형	14	37.84%
4	음성	7	18.92%
5	3D	5	13.51%

데이터 형식 간 조합을 10가지 유형으로 분류하였으며, 표 2와 같다.

표 2. 대규모 멀티모달 데이터 37종의 데이터 형식 조합 분석 결과

번호	형식	개수	비율
1	텍스트, 영상·이미지	16	43.24%
2	영상·이미지, 수치형	8	21.62%
3	텍스트, 음성	3	8.11%
4	텍스트, 영상·이미지, 3D	3	8.11%
5	텍스트, 음성, 수치형	2	5.41%
6	텍스트, 음성, 영상·이미지, 3D	1	2.70%
7	텍스트, 영상·이미지, 음성, 수치형	1	2.70%
8	텍스트, 영상·이미지, 3D, 수치형	1	2.70%
9	텍스트, 수치형	1	2.70%
10	텍스트, 영상·이미지, 수치형	1	2.70%
합계		37	100%

2. 라벨링 유형

멀티모달 데이터는 여러 가지 형태의 정보가 결합되어 있기 때문에, 단일 데이터 유형에 대한 라벨링만으로는 충분하지 않을 수 있다. 데이터 라벨링은 주어진 데이터에 대한 추가 정보를 부여하여 해당 데이터의 의미를 명확히 하는 과정으로, 인공지능 모델의 학습 및 활용을 확장시키는데 중요한 역할을 한다.

본 논문에서는 라벨링 유형을 바운딩박스·세그멘테이션, 폴리곤·폴리라인, 캡션, 전사, 태깅, 분류, 질의응답, 번역 총 8가지로 분류하였다.[3]

- 바운딩박스·세그멘테이션: 객체의 위치와 크기를 객체를 인식하고 분류하는 데 사용되는 방식이다
- 폴리곤·폴리라인: 객체의 경계를 정확하게 지정하기 위해 다각형이나 라인으로 객체를 정의하는 라벨링 방식이다.
- 캡션: 이미지나 영상에 대한 설명 또는 해석을 제공하는 텍스트로, 시각적 콘텐츠를 이해하는 데 도움이 되는 정보다.
- 전사: 전사는 음성 데이터를 텍스트 형태로 변환하는 작업을 의미한다.
- 태깅: 데이터에 대한 추가 정보를 부착하여 객체나 속성을 식별하고 분류하는 라벨링 방식이다.
- 분류: 데이터를 특정 카테고리에 할당하거나 분류하여 해당 데이터의 특성을 이해하는 라벨링 방식이다.
- 질의응답: 질문에 대한 정확한 답변을 생성하거나 추론하는 작업이다.
- 번역: 텍스트를 한 언어에서 다른 언어로 변환하는 작업이다.[4]

멀티모달 데이터는 단일 데이터에 다수의 라벨링 유형이 적용될 수 있으며, 37종의 데이터에서 각 라벨링 유형의 개수와 비율은 표 3과 같다.

표 3. 대규모 멀티모달 데이터 37종의 라벨링 유형 분석 결과

번호	라벨링 유형	개수	비율
1	캡션	25	67.57%
2	태깅	10	27.03%
3	바운딩박스·세그멘테이션	7	18.92%
4	전사	6	16.22%
5	폴리곤·폴리라인	2	5.41%
6	QA	2	5.41%
7	번역	2	5.41%
8	분류	1	2.70%

3. 구축 목적

데이터 유형을 효과적으로 구축하기 위해서는 해당 데이터가 가지고 있는 다양한 정보를 적절하게 표현하고, 인공지능 모델이 이를 이해하고 활용할 수 있도록 데이터 구축 방식을 고려해야 한다.

본 논문에서는 구축 목적을 텍스트 생성, 이미지 생성, 음성 인식/음성 합성, 객체/행위에 대한 탐지, 수치 예측, 클래스 분류, 로봇 동작 및 작업 생성, 경로 최적화 총 8가지로 분류하였다.

- 텍스트 생성: 입력 정보나 문맥을 바탕으로 새로운 텍스트를 생성하는 작업을 말한다.
- 이미지 생성: 조건에 따라 새로운 이미지를 생성하는 작업을 의미한다.
- 음성 인식/음성 합성: 음성 데이터를 텍스트로 변환하여 인식하는 작업과 텍스트를 음성으로 변환하여 생성하는 작업이다.
- 객체/행위에 대한 탐지: 이미지나 영상 데이터에서 객체나 특정 행위를 감지하고 식별하는 작업을 의미한다.
- 수치 예측: 패턴이나 특성을 기반으로 수치 값을 예측하는 작업을 말한다.
- 클래스 분류: 정의된 클래스 또는 카테고리로 분류하는 작업을 의미한다.
- 로봇 동작 및 작업 생성: 로봇이 특정 환경에서 움직이고 작업을 수행하는 방법을 결정하는 작업을 의미한다.[5]
- 경로 최적화: 특정 목적지로 이동하는 데 가장 효율적인 경로를 계산하는 작업을 말한다.

37종의 데이터에서 각 데이터 구축 목적의 개수와 비율은 표 4와 같다.

표 4. 대규모 멀티모달 데이터 37종의 데이터 구축 목적 분석 결과

번호	데이터 구축 목적	개수	비율
1	텍스트 생성	12	32.43%
2	클래스 분류	8	21.62%
3	객체/행위에 대한 탐지	6	16.22%
4	이미지 생성	3	8.11%
5	수치 예측	3	8.11%
6	로봇 동작 및 작업 생성	2	5.41%
7	경로 최적화	2	5.41%
8	음성 인식/음성 합성	1	2.70%
합계		37	100%

Ⅲ. 결 론

본 논문에서는 '24년도 초거대AI 확산 생태계 조성 사업의 대규모 멀티모달 데이터의 현황을 분석하였다. 다양한 형식으로 구축될 데이터는 각각 텍스트 생성, 이미지 생성, 음성 인식/합성, 객체/행위 탐지, 수치 예측, 클래스 분류, 로봇 동작 및 작업 생성, 경로 최적화 등 여러 형태의 목적과 응용 분야를 대상으로 구성되어 있음을 확인하였다. 또한 라벨링 유형의 다양성은 데이터의 활용성과 효율성을 높이는데 중요한 역할을 한다. 따라 데이터의 특성에 따라 적절한 정보를 부여할 필요가 있다. 구축된 데이터는 각 분야에서의 인공지능 모델 학습과 응용을 통해 대규모 멀티모달 모델의 발전에 기여할 것으로 기대된다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업 (2100-2131-305, 2024년 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참 고 문 헌

[1] 윤여찬. “멀티모달 생성형 AI 기술 동향.” 한국정보과학회, 2024

[2] 최종원, 이정현. “멀티모달 센서를 이용한 스마트기기 사용자 인증 기술 동향.”, 정보보호학회지, 2014

[3] 정보통신단체표준(TTAS). “컴퓨터 비전 분야 라벨데이터의 의미적 정확성 및 유효성 품질검증 방법.” TTAK.KO-10.1420, 2023.06.30.

[4] 배장성, 황현선, 이창기. “이미지 정보를 이용한 영어-한국어 자동 번역.”, 한국정보과학회, 2019

[5] 이준기, 박상준, 김낙우, 김에덴, 고석갑. “거대언어모델 기반 로봇 인공지능 기술 동향.”, 전자통신동향분석 제39권 제1호 pp 95-105