멀티모달 데이터의 학습 목적별 품질검증 방안 고찰

안지혜, 박정하, 이상복

한국정보통신기술협회

zihye29295@tta.or.kr, parkjh516@tta.or.kr, jangpo@tta.or.kr

A Study on the Quality Evaluation Method for Learning purpose of Multimodal Data

An Ji Hye, Park Jeong Ha, Lee Sang Bok

Telecommunications Technology Association

요 약

비디오, 이미지, 음성, 텍스트 등 다양한 데이터 유형을 학습한 멀티모달 모델의 발전이 빠르게 진행되고 있으며, 모델이 높은 성능을 도출하기 위해서는 학습 목적에 따른 품질검증 방법이 필수적이다. 본 논문에서는 멀티모달 데이터의 품질을 평가하기 위해 표현성, 일치성, 정렬성 등 3가지 품질특성을 정의하고, 다양한 학습 목적에 따라 데이터 조합을 구분하여 이를 적용하는 방법을 제시하였다. 향후 본 논문에서 제안된 방법을 실제 멀티모달 데이터의 품질검증에 적용하여 구체적인 검증 기준 및 방안을 모색하고자 한다.

Ⅰ. 서론

인공지능 분야에서는 OpenAI의 GPT-3, 메타의 LLama와 같은 대형 언어 모델에서 GPT-4o, CLIP, Gemini와 같은 다양한 데이터 유형을 학습한 멀티모달 모델로의 발전이 빠르게 진행되고 있다.[1] 비디오, 이미지, 음성, 텍스트 등을 동시에 학습시킨 모델은 기존의 텍스트 중심의 학습 모델 성능을 뛰어넘어 인간의 사고방식을 이해하는 것과 비슷한 능력을 보인다. 멀티모달 모델의 성능을 평가하는 리더보드를 확인하면 다양한 벤치마크 데이터 셋을 통해 높은 성능을 기록하고 있는 것을 알 수 있다. 이때, 높은 성능을 보장하기 위해서는 데이터의 품질이 결정적인 역할을 한다. 특히 멀티모달 데이터 특성상 다양한 이기종의 데이터 조합으로 구성되어 있어서 멀티모달 간 매핑의 정확성이 부각되므로, 부정확하거나 일관성이 없는 데이터는 모델이 잘못된 패턴을 학습하게 만들어 성능 저하와 예측 오류로 이어질 수 있다. 따라서 본 논문에서는 학습 목적에 맞는 멀티모달 데이터의 품질을 검증하는 방안을 고찰하고자 한다.[2]

Ⅱ. 본론

멀티모달 모델의 성능 향상을 위해 양질의 데이터를 구축하기 위해서는 데이터 조합에 따라 의미적인(semantic) 요소에 대한 특성을 세분화하여 품질검증을 할 필요가 있다.[6][7] 특히 멀티모달 모델은 두 개 이상의 여러 가지 유형의 데이터를 함께 입력하여 학습하므로, 유니모달 데이터의 어노테이션 정확도, 멀티모달 데이터 간 일치 여부, 멀티모달 데이터 간의 정렬성을 확인해야 하는 특성이 있다.[5] 이 특성들을 표 1과 같이 표현성, 일치성, 정렬성으로 정의하였다.

표 1. 멀티모달 데이터 품질특성

품질특성	정의
표현성	각 유니모달별 라벨링 정보와 실제 참값이 부합하는지를 확인하는 특성으로, 각 유니모달의 개체 식별 가능 여부, 라벨링과 참값의 부합 여부를 판단한다.
일치성	멀티모달간의 매핑이 정확한지를 판단하기 위해, 각 유니모달별 식별 대상(entity)이 같은 객체인지 등 일치요소를 추출하여 요소들간 의미적으로 동일한지를 판단한다. 원천 데이터의 단위는 하나의 이미지 장면이나 컨텍스트를 가지는 텍스트 문장/문단, 발화 문장을 단위로 한다. 이미지 캡셔닝과 같이 캡션의 단어나 문구에 해당하는 이미지 영역을 찾아 매칭의 정확성을 평가한다.
정렬성	멀티모달간 시간/공간적 요소가 존재하는 데이터의 품질특성으로, 둘 이상의 양식에서 인스턴스의 하위 구성 요소 간의 관계와 대응이 정확한지를 평가한다. 비디오 설명과 같이 비디오를 기반으로 한 대본이나 책의 챕터에 맞춰 시간적 정렬이 정확한지를 평가한다.

표현성은 유니모달의 어노테이션, 라벨링 정확도를 검증하는 특성이므로, 본 논문에서는 여러 모달리티의 조합에 대한 매핑 정확도를 검증할 수 있는 특성인 일치성과 정렬성에 한정하여 기술한다. 그리고 멀티모달 모델의 대표적인 6가지의 학습 목적(시각적 음성인식, 객체 및 행동 분류, 이벤트 탐지 및 이상 탐지, 비디오 및 이미지 설명, 시각적 질의응답, 비디오 및 이미지 생성)에 따라 멀티모달 데이터의 조합을 분류하고, 각 학습 목적에 대한 설명과 데이터 조합 유형의 예시를 제시한 후, 앞에서 정의한 품질특성 항목을 적용하였다.

1. 시각적 음성인식

시각적 음성인식은 비디오에서 입 모양과 같은 시각적 신호와 음성 신호를 분석하여 말하는 사람이 무엇을 말하고 있는지를 파악하는 것을 목표로 한다. 이는 청각 정보가 불분명하거나 소음이 있는 환경에서 유용하며,

시각적 정보를 통해서 음성인식의 정확도를 높이는데 기여할 수 있다.[3] 데이터 유형은 비디오 내 입 모양을 식별할 수 있는 어노테이션(Bounding Box, Keypoint 등), 그에 해당하는 음성, 전사 텍스트 등으로 이루어져 있다. 또한, 다양한 발음, 얼굴 각도 등으로 이미지 데이터 셋의 다양성을 고려할 수 있다. 시각적 음성인식을 효과적으로 수행하기 위해서는 음성, 비디오 내 이미지 프레임, 전사 텍스트 간의 일치성을 검증해야 하며, 음성과 이미지 프레임의 정렬성을 검증하는 것이 필요하다.

2. 객체 및 행동 분류
객체 및 행동 분류는 해당 비디오 또는 이미지의 특정 객체나 행동을 식별하고 하나의 카테고리로 분류하는 것을 목표로 한다. 데이터 유형으로는 비디오, 이미지(3D, 2D), 수치형 데이터 등이 포함된다. 2D 이미지와 3D 이미지에서 동일한 객체의 일치성 검증 및 비디오와 이미지 간의 일치성 검증을 하는 작업이 필요하며, 비디오 데이터에서는 객체가 시간에 따라 이동하는 경로를 추적하여 수치형 데이터와 함께 객체 추적의 정렬성을 확인하는 것이 필요하다.[4]

3. 이벤트 탐지 및 이상 탐지
이벤트 탐지 및 이상 탐지는 데이터 내 특정 사건이나 이상 상황을 탐지하고 식별하는 목적으로 한다. 데이터 유형은 비디오, 이미지, 수치형 데이터, 음성, 텍스트 등의 여러 가지 데이터 조합으로 구성되어 있다. 예를 들어, 사고 예방을 위한 기상 정보 데이터에서는 텍스트(기상 정보), 이미지, 센서 데이터의 일치성을 검증하고, 이미지와 센서 데이터 간의 정렬성을 검증하는 것이 필요하다. 또, 열화상 데이터에서는 열화상 비디오에서 감지된 온도 변화와 텍스트 설명과 일치하는지를 확인하는 일치성을 검증해야 한다. 또한, 비디오와 온도의 변화 정렬성을 검증하여 비디오에서 나타나는 시각적 변화와 실제 온도가 일관되게 나타나는지 확인할 필요가 있다.

4. 비디오 및 이미지 설명
비디오 및 이미지 설명은 비디오나 이미지 내용을 정확하게 설명하거나 요약하는 것을 목표로 한다. 데이터 유형은 비디오, 이미지, 텍스트 등으로 이루어져 있다. 비디오 설명의 경우, 비디오의 시각적 내용과 설명문의 일관성을 유지하기 위해 비디오 데이터, 주요 객체, 주요 사건의 일치성을 검증하는 것이 필요하고, 비디오 데이터와 설명문 간의 시간적 흐름이 정확한지에 대한 정렬성을 검증해야 한다. 이미지 설명의 경우, 이미지 데이터의 시각적 내용을 정확하게 반영하는지를 확인하기 위해 이미지 데이터와 설명문의 일치성을 검증한다. 이미지 설명은 하나의 이미지 데이터에 대한 장면을 설명하므로, sequence 정보가 존재하지 않아 정렬성을 검증하기에 부적합하다.

5. 시각적 질의응답
시각적 질의응답은 비디오나 이미지에 대한 자연어 질문에 답변을 생성하는 것이다. 이를 통해 시각적 데이터를 이해하고 해석하는 능력을 향상시키는 것을 목표로 한다. 데이터 유형으로는 비디오, 이미지, 텍스트 등으로 이루어져 있다. 이미지와 질의응답 텍스트로 이루어져 있는 데이터의 경우, 이미지와 질의응답의 일치성을 검증할 필요가 있으며, 이미지 내 sequence 정보가 존재하지 않아 정렬성을 검증하기에 부적절하다. 비디오, 질의응답 텍스트, 번역 텍스트로 이루어져 있는 데이터는 비디오 데이터와 질의응답의 일치성 검증과 질의응답, 번역문의 일치성을 검증할 필요가 있다. 질의응답의 구성에 비디오 순서를 고려한 경우에는 비디오 내용과 질의응답에 대한 정렬성 검증이 필요하다.

6. 비디오 및 이미지 생성
비디오 및 이미지 생성은 인공지능 모델을 사용하여 여러 가지 방식으로 데이터 증강, 콘텐츠 창작, 시뮬레이션 등을 목적으로 하는 시각적 데이터를 생성한다. 데이터 유형은 비디오, 이미지(3D, 2D), 음성, 텍스트 등으로 이루어져 있다. 예를 들어, 발화 얼굴 생성 데이터의 경우 이미지(3D, 2D), 음성, 텍스트로 구성되어 있는데, 3D와 2D 이미지의 객체 일치성 검증 및 음성과 이미지(3D)의 입 모양 일치성 검증이 필요하다.
또, 수어 영상 데이터에서는 비디오 데이터와 수어 동작의 일치성 및 비디오 데이터의 동작과 수어 설명 텍스트의 동작 일치성을 검증해야 한다. 그리고 동일한 화자 객체를 확인하기 위해서는 특정 시간 단위 프레임 기준으로 프레임별 동일 화자 여부를 확인하는 정렬성을 검증해야 한다.

Ⅲ. 결론

본 논문에서는 멀티모달 데이터의 학습 목적별 데이터 조합을 분석하고 멀티모달 데이터 특성으로 정의한 일치성, 정렬성의 품질특성을 적용하여 학습 목적에 맞는 멀티모달 데이터의 품질을 검증하는 방안을 제시했다. 멀티모달 모델의 대표적인 학습 목적을 6가지로 구분하고 그 학습 목적에 따라 데이터 조합 유형과 그에 따른 품질검증 방법을 상세히 분석하여 멀티모달 모델의 성능에 도움이 될 수 있도록 방법을 제시했다. 향후 본 논문에서 제안한 품질검증 방안을 실제 멀티모달 데이터에 적용하여 세부적인 검증 항목을 도출하고, 품질특성 별 구체적인 검증 기준 및 방안을 모색하고자 한다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] 윤여찬, “멀티모달 생성형 AI 기술 동향”, 정보과학회지, 2024.01, pp. 42-47
[2] 박정원 외, “사전학습단계 초거대 인공지능 학습용 데이터의 품질 요소 및 검증방안 고찰”, 한국통신학회, 2023.09, pp. 123-124
[3] 황정욱, “Audio-Visual Speech Recognition Based on Joint Training with Audio-Visual Speech Enhancement”, 서강대학교, 2022
[4] 김현서 외, “대형 멀티모달 모델을 활용한 비디오에서의 타겟 물체 출입 탐지“, 한국정보과학회 학술발표논문집, 2024.06
[5] Tadas Baltrusaitis at al., Multimodal Machine Learning: A Survey and Taxonomy, arXiv:1705.09406v2, 2017
[6] Guo, Wenzhong et al., Deep Multimodal Representation Learning: A Survey, IEEE Access 7 63373-63394, 2019
[7] Jiayang Wu at al., Multimodal Large Language Models: A Survey, arXiv:2311.13165v1, 2023