# 비전-언어 멀티모달 모델 학습용 데이터 품질검증 방법 및 적용 사례

김아름, 이상복*

한국정보통신기술협회

ark5139@tta.or.kr, *jangpo@tta.or.kr

# A Method and Application Case of Quality Verification of Vision-Language Multimodal Model Learning Data

Kim A Rum, Lee Sang Bok*

Telecommunications Technology Association

요 약

본 논문은 비전-언어 멀티모달 모델을 학습시키기 위한 학습용 데이터의 품질검증 방법을 제안한다. 또한 비전-언어 멀티모달 데이터 중 이미지-캡션 및 비디오-캡션 데이터에 적용할 수 있는 세부 품질 특성을 구분하고 각 세부 품질 특성에 따라 검증 기준을 설명한다. 그리고 해당 세부 품질 특성 및 검증 기준을 활용하여 텍스트-3D 이미지 쌍 데이터, 영상 교육자료 기반 문제 생성 데이터의 품질검증 수행 사례를 소개한다.

## Ⅰ. 서 론

최근 대규모 언어 모델(LLM, Large Language Model)의 발전은 텍스트뿐만 아니라 이미지, 비디오, 오디오와 같은 다양한 형태의 데이터를 동시에 이해하고 처리할 수 있는 대규모 멀티모달 모델(LMM, Large Multimodal Model)의 부상으로 이어지고 있다. LMM은 단일 모달리티의 한계를 넘어서 다양한 데이터를 통합적으로 처리함으로써 더 복잡하고 의미 있는 예측과 응용을 가능하게 한다[1]. 이러한 LMM을 효과적으로 학습시키기 위해서는 다양한 모달리티의 데이터가 정확하고 일관되게 결합된 데이터가 필요하다.

따라서, 멀티모달 데이터셋의 구축은 단일 모달리티 데이터의 품질뿐만 아니라 멀티 모달리티 간의 품질까지 고려해야 한다. 특히, 비전과 언어의 조합을 다루는 멀티모달 데이터는 이미지, 비디오와 텍스트 간의 의미적 일치성이 매우 중요하며, 이는 LMM 성능에 직접적으로 영향을 미치기 때문에 데이터셋의 품질이 무엇보다 중요하다[2]. 그러므로 멀티모달 데이터의 평가 방안은 단일 모달리티 데이터의 평가 관점에서 더 나아가 멀티모달 데이터의 특성을 반영한 새로운 평가 기준이 필요하다. 이에 따라, 본 논문에서는 비전-언어 멀티모달 모델 학습용 데이터의 품질검증 방법을 살펴보고 실제 데이터에 이를 적용하여 평가 방안의 실효성과 한계점을 고찰하고자 한다.

## Ⅱ. 본론

### 1. 비전-언어 멀티모달 데이터 품질검증 방법

본 논문에서의 품질검증 대상은 이미지와 비디오, 그에 해당하는 캡션 형식의 조합인 이미지-캡션, 비디오-캡션 멀티모달 데이터로 정의한다. 이미지-캡션, 비디오-캡션 멀티모달 데이터 품질검증의 품질 특성은 기본 품질 특성인 다양성, 구문 정확성, 의미 정확성, 유효성을 기반으로 세부 품질 특성을 고려한다[3]. 품질 특성 중 의미 정확성의 세부 품질 특성을 표 1의 3가지로 구분하였다.

표 1. 비전-언어 멀티모달 데이터 의미 정확성 세부 품질 특성

| 세부 품질 특성 | 설명 |
|---|---|
| 단일 모달 표현성 | 이미지, 비디오, 텍스트 등 단일 모달리티 데이터가 그 자체로 이해 및 식별 가능한지 확인하는 특성 |
| 의미적 일치성 | 이미지-캡션, 비디오-캡션 모달리티 간의 내용이 일치하는지 확인하는 특성 |
| 시간적 정렬성 | 비디오-캡션 모달리티 간의 시간적 요소 정렬이 일치하는지 확인하는 특성 |

단일 모달 표현성(uni-modal expression)은 멀티모달 데이터를 구성하는 각 단일 모달리티의 표현이 적절한지 검증하는 특성이다. 이미지-캡션의 경우 이미지가 흐릿하여 이해 및 식별이 불가능하거나 구축 목적에 부합하지 않는 이미지가 구축되어 있다면 단일 모달 표현성의 오류에 해당한다.

의미적 일치성(semantic consistency)은 멀티모달 데이터를 구성하는 각 단일 모달리티의 내용이 일치하는지 검증하는 특성이다. 멀티모달 모델의 학습 목적이 캡션을 통해 이미지를 생성하는 'text to image'라면, 이미지와 캡션의 내용 일치 여부는 중요한 학습 요소가 되므로 의미적 일치성 검증이 필요하다[4].

시간적 정렬성(temporal alignment)은 멀티모달 데이터를 구성하는 각 단일 모달리티의 시간적 요소의 정렬이 일치하는지 검증하는 특성이다. 멀티모달 모델의 학습 목적이 비디오를 기반으로 캡션을 생성하는 것이라면, 비디오의 time stamp에 해당하는 캡션의 시간 정렬 일치 여부는 중요한 학습 요소이다.

위 3가지 품질특성은 의미 정확성 하위 세부 품질특성으로써 각 항목마다 세부 검사 기준을 마련해야 한다. 데이터 구축 목적 및 학습의 임무에 따라 검사 기준이 달라져야 하므로 다양한 검사 기준을 고려할 수 있다. 각 세부 품질 특성마다 검사 기준을 정리한 예시는 표 2와 같다.

표 2. 비전-언어 멀티모달 데이터 세부 품질 특성별 검사 기준 예시

| 세부 품질 특성 | 검사 기준 |
|---|---|
| 단일 모달 표현성 | - 이미지, 비디오 단일 모달 오류 존재 여부<br>- 캡션 문법적 오류 존재 여부 |
| 의미적 일치성 | - 캡션 내용이 이미지, 비디오를 기반으로 눈에 보이는(visible) 정보만을 제공하는지 여부<br>- 캡션 내용이 이미지, 비디오를 기반으로 주관적인(subjective) 평가나 반응이 포함되는지 여부<br>- 캡션 내용이 이미지, 비디오를 기반으로 알 수 없는 정보를 통해 독립적인(independent) 설명을 제공하는지 여부 |
| 시간적 정렬성 | - 비디오 시간적 흐름에 따른 장면 정보 캡션 반영 및 시간적 동기화(temporal synchronization) 여부 |

2. 적용 사례

첫 번째 적용 사례는 한국에서 자주 쓰이는 객체가 포함된 문장에 맞는 3D 객체 생성을 위한 텍스트-3D 객체 쌍 데이터, 즉 이미지-캡션 데이터이다. 구축된 데이터의 편향성을 검증하기 위해서 다양성 항목으로 이미지의 다양성과 캡션의 다양성을 먼저 고려해야 한다. 이미지 다양성 항목으로는 '3D 객체의 클래스 분포'를, 캡션의 다양성 항목으로는 '어절 수', '유사성(중복성)'을 확인해야 한다. 1절에서 설명한 의미 정확성 세부 품질 특성을 적용하여 첫 번째 단일 모달 표현성에서 이미지 표현성 항목으로는 '3D 모델링 정확성'을 설정해야 한다. 3D 모델링 정확성의 검사 기준은 mesh와 texture가 존재하는지, mesh 내 hole이 존재하는지, 2D 이미지와 3D 모델링 데이터의 특징이 유사한지 등을 기준으로 검사를 진행한다. 캡션 표현성 항목에 해당하는 캡션 문법 적정성과 두 번째 세부 품질 특성인 의미적 일치성 항목에 해당하는 3D-캡션 내용 일치성은 '이미지 캡션 적정성'이라는 항목을 통해 함께 검사를 진행할 수 있다. 마지막 유효성 모델 검증은 AI 임무인 텍스트 입력을 통해 한국형 3D 객체를 생성하는 것, 즉 텍스트 기반 3D 생성(3D Generation)을 검증하기 위해 CLIP ViT-B/32 모델을 활용하여 이미지와 텍스트를 동일한 벡터 공간에서 표현하고 이들의 유사성을 정량화하는 방법을 통해 검증 가능하다. 각 품질 특성에 대한 결과는 표3과 같다.

표 3. 한국형 텍스트-3D 객체 쌍 데이터 품질검증 결과

| 품질 특성 | 항목명 | 측정 지표 | 결과값 |
|---|---|---|---|
| 다양성(통계) | 이미지캡션 유사성 | 구성비 | 분포 확인 |
| 다양성(요건) | 3D 객체 자연물 분포 | 구성비 중첩률 | 90.34% |
| | 3D 객체 인공물 분포 | 구성비 중첩률 | 96.01% |
| | 이미지캡션 어절 수 | 최소 수량 | 5 어절 |
| 구문 정확성 | 구조 정확성 | 정확도 | 100% |
| | 형식 정확성 | 정확도 | 100% |
| 의미 정확성 | 3D 모델링 정확성 | 정확도 | 93% |
| | 이미지캡션 적정성 | 정확도 | 95.56% |
| 유효성 | 3D Generation | CLIP SCORE | 32.25 |

두 번째 적용 사례는 영상 교육자료 콘텐츠에서 문제 생성 학습을 위한 데이터, 즉 비디오-캡션 데이터이다. 해당 데이터는 비디오 영상-스크립트-문제(문항-답) 세트로 구성되어 있으므로 각 모달리티에 맞는 검증 항목을 설정해야 한다. 다양성은 비디오 주제 분포, 영상 길이, 스크립트 및 문제의 어절 수, 문제 난이도 등을 포함하며, 이를 통해 데이터의 균형을 평가한다. 단일 모달리티의 표현성은 비디오 영상 자체 오류 여부와 스크립트, 문제 텍스트의 문법적 오류 여부를 확인하고, 의미적 일치성은 영상에서 전사한 스크립트 내용이 영상과 일치하는지, 문제(문항-답안 쌍)가 영상-스크립트 내용과 일치하는지 등의 검사 기준으로 영상-스크립트-문제 쌍 내용의 상호 일치성을 검토한다. 시간적 정렬성은 각 모달리티의 시간적 순서가 맞는지 확인하며, 마지막으로 유효성 모델 검증은 동영상 콘텐츠 기반 문제 생성 AI task를 검증하기 위해 GPT나 BERT 계열 모델을 활용할 수 있다. 비디오 동영상에서 전사를 통해 변환 추출한 스크립트 텍스트에서 핵심 정보를 식별하고 문제를 생성하는 AI 모델의 성능을 평가할 수 있다.

## Ⅲ. 결론

본 논문에서는 비전-언어 멀티모달 모델 학습용 데이터의 품질검증 방법과 그 적용 사례를 통해 다양한 검증 기준을 살펴보았다. 하지만 데이터의 특성이나 구축 목적에 따라 검사 기준이 달라지는 검증 기준의 주관성과 여러 모달리티 간의 상호작용 때문에 단일 모달리티 데이터의 품질이 모델의 성능에 미치는 영향을 명확히 분석하기 어렵다는 점 등의 한계점이 존재한다[6]. 따라서 앞으로의 연구에서는 이러한 한계를 극복하기 위해 검사 기준의 주관성을 배제할 수 있는 평가 기준의 표준화 방법론의 발전과 멀티 모달리티 및 단일 모달리티 데이터가 모델에 미치는 영향에 대한 연구가 요구된다. 또한 비전-언어 멀티모달 모델의 학습용 데이터 품질 검증 방법을 기반으로 다양한 형태로 결합된 멀티모달 데이터의 품질 검증 방안을 모색하고자 한다.

## 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] Huang, Dawei, et al. "From Large Language Models to Large Multimodal Models: A Literature Review." Applied Sciences 14.12 (2024): 5068.

[2] Zhang, Duzhen, et al. "Mm-llms: Recent advances in multimodal large language models." arXiv preprint arXiv:2401.13601 (2024).

[3] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회. "인공지능 학습용 데이터 품질관리 가이드라인 v3.1 - 제 1권 품질관리 가이드라인" 2024.

[4] Alikhani, Malihe, Baber Khalid, and Matthew Stone. "Image - text coherence and its implications for multimodal AI." Frontiers in Artificial Intelligence 6 (2023): 1048874.

[5] Zhang, Duzhen, et al. "Mm-llms: Recent advances in multimodal large language models." arXiv preprint arXiv:2401.13601 (2024).

[6] Liang, Paul Pu, Amir Zadeh, and Louis-Philippe Morency. "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions." ACM Computing Surveys 56.10 (2024): 1-42.