멀티모달 입력 기반 생성형 AI 학습을 위한 비디오-텍스트 데이터 품질검증 방법 및 적용 사례

박경은[1], 이상복[1], 조연제[2], 유호진[2]
[1]한국정보통신기술협회 AI데이터품질팀
[2]한국지능정보사회진흥원 AI데이터사업팀
kepark@tta.or.kr, jangpo@tta.or.kr, yjyjyjcho@nia.or.kr, yuhj@nia.or.kr

Video-Text Data Quality Verification Method and Case Study for Training Multimodal Input-based Generative AI

Kyung-Eun Park[1], Sang-Bok Lee[1], Yeon-Je Cho[2], Ho-jin Yu[2]
[1]AI Data Quality Team, Telecommunications Technology Association
[2]AI Data Project Team, National Information Society Agency

요 약

본 논문은 멀티모달 입력 기반의 생성형 AI 학습을 위한 비디오-텍스트 데이터의 품질을 평가하는 검증 방법을 제안한다. 비디오-텍스트 데이터에 적용할 수 있는 품질특성을 구분하고, 품질특성별 검증 방법 및 예시를 설명한다. 그리고 해당 품질검증 방법을 K-콘텐츠 특화 생성형 AI 솔루션을 위해 구축된 비디오-텍스트 데이터에 적용한 품질검증 사례를 소개한다.

1. 서론

최근 생성형 AI 기술이 발전함에 따라 비디오-텍스트 데이터와 같은 멀티모달(Multimodal) 입력을 활용하는 모델이 주목받고 있으며, 기존의 텍스트 기반 생성형 AI에서 벗어나 비디오와 같은 복합적인 데이터를 활용하는 방식으로 확장되면서 더욱 정교한 콘텐츠 생성이 가능해졌다[1]. 이러한 멀티모달 AI 모델이 높은 성능을 발휘하기 위해서는 학습 데이터의 품질이 중요한 요소로 작용한다. 특히, 비디오-텍스트 데이터는 영상과 캡션 간의 정합성(Consistency), 문장의 다양성 등이 보장될 때 모델의 성능이 향상될 수 있다[2]. 즉, 멀티모달 AI의 발전과 함께 학습 데이터의 품질 수준을 확보하기 위한 체계적인 검증 방법이 요구된다. 따라서 본 연구에서는 비디오-텍스트 데이터의 품질을 다각적으로 검증하는 방법을 제안하고, 실제 데이터에 적용하여 그 유효성을 평가한다.

2. 본론

2.1 품질검증 대상 정의

본 논문에서 품질검증 대상 데이터는 비디오와 해당 비디오를 설명하는 캡션으로 구성된 비디오-텍스트 데이터로 정의한다. 비디오-텍스트 데이터는 비디오를 입력하여 상세한 설명문을 생성하거나, 설명문을 입력하여 대응되는 비디오를 생성하는 멀티모달 생성형 AI 모델의 학습에 사용되는 데이터다.

2.2 품질검증 방법

2.2.1 의미적 정확성

멀티모달 입력 기반 생성형 AI의 학습을 위한 비디오-텍스트 데이터는 비디오와 텍스트 간의 의미적(Semantic) 정합성 및 묘사 수준이 학습 성능에 중요한 영향을 미친다[3]. 따라서 모달리티 간 의미적 요소가 일치하는지 판단하여 묘사 수준을 검증하는 멀티모달 간 의미적 정합성을 중심으로, 표1과 같이 의미적 정확성 세부 품질특성을 구분한 후 품질 수준을 검증할 수 있다.

<표 1> 의미적 정확성 세부 품질특성

구분	품질특성	설명	
유니모달	이해가능성	유니모달 각각 이해할 수 있는 형태로 표현하고 있는지 확인	
		예시	비디오 재생 여부, 객체 식별 여부, 텍스트 문법 등
멀티모달	내용정합성	모달리티 간 내용적 요소가 일치하는지 확인	

구분	품질 특성	설명	
멀티 모달	내용 정합성	예시	비디오-캡션 간 묘사 내용 일치 여부 확인 등
	시간 정합성	모달리티 간 시간적 요소가 일치하는지 확인	
		예시	비디오-캡션 간 Timestamp 일치 여부 확인 등
	시간 정렬성	모달리티 간 시간적 요소의 정렬이 올바른지 확인	
		예시	비디오-캡션 정렬 순서 확인 등

먼저, 유니모달(Unimodal)인 비디오와 텍스트가 각각 이해 및 식별할 수 있는 형태로 표현하고 있는지 확인한다[4]. 그리고 각 모달리티의 의미적 요소를 추출한 후 서로 비교하여 정합성을 확인한다. 이때 의미적 요소는 묘사 내용 및 시간적 요소 등이 해당한다. 만약 비디오의 구간 순서에 따라 캡션이 작성된 다중캡션(Multiple Captions) 형태라면, 모달리티 간 시간적 요소의 정렬(Alignment)이 올바른지 확인한다.

2.2.2 다양성

다양성은 비디오-텍스트 데이터가 특정 분야에 편중되지 않고 다양한 도메인과 장르를 포함하는지 검증한다[5]. 또한, 캡션의 글자 수 분포를 분석하여 AI 학습에 충분한 정보량을 제공할 수 있는지 평가한다. 이를 위해 캡션 길이 및 평균을 측정하여, 모델 학습에 적절한 문장 표현의 깊이를 갖추었는지 확인한다. 추가로, 캡션 간 중복성 또는 유사성 분포를 분석하여 문장 표현의 다양성이 충분히 확보되었는지를 검증한다.

2.2.3 구문적 정확성

구문적 정확성은 비디오-텍스트 데이터가 데이터 정의서에서 규정한 구조와 형식을 일관되게 준수하는지를 검증하는 품질특성이다. 이를 위해 데이터 내 필수 항목이 누락되지 않았는지 확인하고, 유효하지 않은 값이 포함되어 있는지 검출하며, 각 항목이 정의된 데이터 타입과 일치하는지 점검한다. 또한, 캡션의 문장 형식, 패턴 구조 등이 일관되게 유지되는지를 평가하여 데이터의 정형성을 확인한다.

2.2.4 유효성

유효성은 구축된 비디오-텍스트 데이터가 AI Task에 따른 AI 모델 학습에 적절한지를 검증하는 품질특성이다. 이를 위해 실제 AI 모델을 활용하여 데이터셋을 학습하거나 추론한 후, 성능 평가지표를 분석하여 데이터의 유효성을 평가한다. 대표적인 평가지표로 BLEU, METEOR, ROUGE 등이 활용되며, BLEU는 n-그램 기반 정밀도를, METEOR는 어휘적 변형을 고려한 유사도를, ROUGE는 캡션의 요약 성능을 측정하는 데 활용된다[6].

2.3 적용 사례

2.2절에서 언급한 품질검증 방법 중 일부를 K-콘텐츠 특화 생성형 AI 솔루션을 위해 구축된 비디오-텍스트 데이터에 적용해 보았다. 해당 데이터의 AI Task는 비디오 상세 설명문 생성이며, 한국 고유 배경을 촬영한 비디오와 500자 이상의 상세 설명문이 쌍(Pair)으로 구성되어 있다. 해당 비디오-텍스트 데이터에 품질검증 방법을 적용한 사례는 표 2와 같으며, 품질특성별 세부 항목명 및 측정 지표를 기술하였다.

<표 2> 비디오-텍스트 데이터 품질검증 방법 적용 사례

품질특성	항목명	측정 지표
의미적 정확성	영상-캡션 내용 일치성	정확도
	배경 분류 태깅 정확성	정확도
다양성	배경별 연도 분포	구성비
	배경별 영상 길이 분포	구성비
	캡션 어절 수	수량
	캡션 중복성	구성비
구문적 정확성	구조 정확성	정확도
	형식 정확성	정확도
유효성	영상 상세 설명문 생성 성능	METEOR

첫째, 의미적 정확성 중 '영상-캡션 내용 일치성' 항목을 통해 멀티모달 간 묘사 내용이 일치하는지 검증하였다. 이때 객체, 계절 및 날씨 등 영상 묘사 내용이 캡션의 내용과 일치하는지 의미적 정합성을 확인하였다. 또한, 유니모달인 비디오 재생에 오류가 없는지 확인하고, 텍스트에 문장 잘림, 조사/전치사/복수 접미사 및 시제 등 문법적 오류가 없는지 확인하였다. 그리고 '배경 분류 태깅 정확성' 항목은 영상에 어노테이션(Annotation)된 배경 분류 정보가 실제 영상에 나타난 주요 주제와 의미적으로 일치하는지를 확인하였다. 단, 해당 데이터는 Timestamp 및 시간 순서 등 시간적 요소가 존재하지 않는 단일

영상의 단일 캡션 형태로, 시간 정합성 및 정렬성은 검증 대상에서 제외하였다.

둘째, 다양성 중 '배경별 연도 분포' 항목을 통해 특정 분야에 편향되지 않았는지 확인하였고, '배경별 영상 길이 분포' 항목을 통해 배경별로 영상 길이가 충분한지 확인하였다. '캡션 어절 수' 항목을 통해 AI 학습에 충분한 정보량을 제공하고 있는지 확인하였으며, '캡션 중복성' 항목을 통해 문장 표현의 다양성이 확보되었는지 확인하였다.

셋째, 구문적 정확성 중 '구조 정확성' 항목을 통해 데이터 정의서의 구조를 준수하는지 확인하였고, '형식 정확성' 항목을 통해 데이터 타입, 유효값 등이 올바르게 구축되었는지 확인하였다.

마지막 유효성은 '영상 상세 설명문 생성 성능' 항목에서 METEOR 지표를 활용하여 검증하였다. METEOR는 의미적으로 유사한 단어와 형태소 변형을 고려하여 생성된 캡션과 참조 캡션 간의 정합성을 평가하는 지표로, 단순한 n-그램 정밀도를 측정하는 BLEU보다 문장 유사도를 더 정밀하게 반영할 수 있다[7]. 이를 통해 AI 모델이 비디오 상세 설명문을 적절하게 생성하는지 확인하였다.

3. 결론

본 논문은 멀티모달 입력 기반 생성형 AI 학습을 위한 비디오-텍스트 데이터의 품질검증 방법을 제안하고, 이를 실제 데이터에 적용하여 품질을 정량적으로 평가하였다. 본 논문의 품질검증 대상은 비디오-텍스트 데이터로 한정되어 있어, 다른 멀티모달 데이터 유형(예: 오디오-텍스트, 이미지-텍스트)에는 직접 적용하기 어렵고 일부 변형 후 적용해야 한다는 한계가 존재한다. 또한, 제안한 품질검증 방법이 데이터의 특정 품질 요소를 평가할 수는 있으나, 비디오-텍스트 데이터의 전체적인 품질 수준을 완벽히 대변하는 것은 아니므로 추가적인 보완 연구가 필요하다. 향후 다양한 멀티모달 데이터 유형에 적용할 수 있는 확장된 품질검증 방법을 모색하고, 데이터 품질을 더욱 정밀하게 검증하기 위한 심층적인 평가 기법을 연구하고자 한다.

사사문구

본 연구는 과학기술정보통신부 초거대AI 확산 생태계 조성 사업(2600-2604-301, 2025년 초거대AI 확산 생태계 조성 사업)에 의해서 수행되었습니다.

참고문헌

[1] L. Zhi, E. Lee, Y. Kim, "Analysis of Research Trends in Deep Learning-Based Video Captioning," KIPS Transactions on Software and Data Engineering, vol. 13, no. 1, pp. 35-49, 2024.

[2] 이동훈, 허찬, 박혜영 and 박상효. "텍스트-비디오 검색 모델에서의 캡션을 활용한 비디오 특성 대체 방안 연구" 대한임베디드공학회논문지 17, no.6, 2022.

[3] Yaya Shi, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. "Learning Video-Text Aligned Representations for Video Captioning," ACM Trans, Multimedia Comput. Commun, Appl. 19, 2, Article 63, March 2023.

[4] 김아름 외, "비전-언어 멀티모달 모델 학습용 데이터 품질검증 방법 및 적용 사례," 한국통신학회 인공지능 학술대회, 2024.

[5] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회. "인공지능학습용데이터품질관리 가이드라인v3.1-제1권품질관리 가이드라인" 2024.

[6] Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. "Curious Case of Language Generation Evaluation Metrics: A Cautionary Tale.," In Proceedings of the 28th International Conference on Computational Linguistics, pages 2322 - 2328, Barcelona, Spain (Online). International Committee on Computational Linguistics. 2020.

[7] Satanjeev Banerjee and Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65 - 72, Ann Arbor, Michigan. Association for Computational Linguistics, 2005.