생성형 AI 학습을 위한 멀티모달 데이터 품질특성 고찰 및 적용방안

박정하 TTA AI데이터품질팀 책임연구원

1. 머리말

데이터는 제조·생산, 금융, 의료, 교통, 교육 등 다양한 분야에서 생산·활용되고 있다. 특히, AI 시대에서 데이터는 의사결정, 예측모델, 자동화 시스템 등 다양한 영역에서 핵심 자산으로 작용하기에, 그 품질이 그 어느 때보다 중요해졌다.

데이터 품질이 높을수록 AI 모델이 올바른 학습을 하고, 신뢰할 수 있는 결과를 도출할 수 있다. 즉, AI를 통해 고품질의 결과를 생성하게 하려면 결국 좋은 데이터가 요구된다. AI 기술 발전과 활용이 확대됨에 따라, 데이터 품질은 성공적인 기술 적용에 필수적인 요소로 자리 잡고 있으며, 위와 같은 특성들을 종합적으로 관리하고 개선하는 것이 중요하다.

생성형 AI는 언어이해 뿐만 아니라 다양한 형태의 정보를 이해하고 처리하며 이를 바탕으로 새로운 콘텐츠를 생성하는 모델이다. 이러한 LMM(생성형 멀티모달 모델, Large MultiModal Model)을 학습하기 위해선 기초 모델(Foundation Model)을 기반으로 전이학습, 즉 파인튜닝을 통해 생성형 AI가 특정 임무 혹은 서비스에 적합하도록 성능을 향상시켜야 한다. 이때 파인튜닝 데이터 셋은 2개 이상의 다른 데이터 유형을 멀티모달 모델에 입력할 수 있는 형태로 구축하며, 생성형 AI가 특정 목적 임무를 정확하게 수행할 수 있도록 한다.

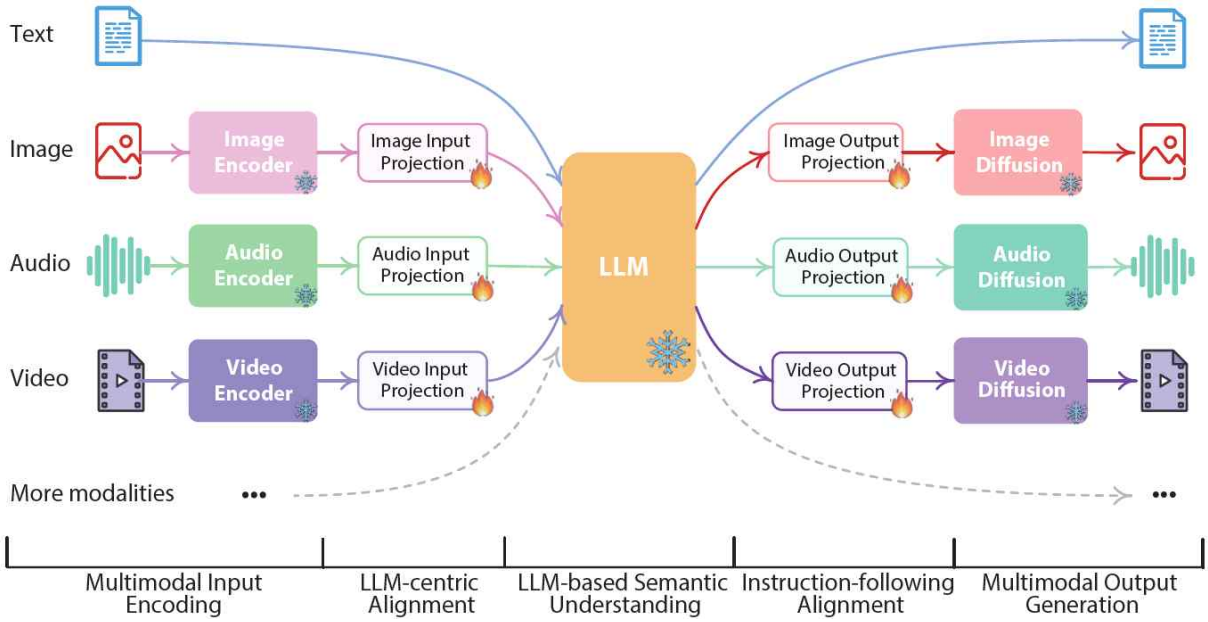
이번 원고에선 생성형 멀티모달 모델 파인튜닝을 위한 멀티모달 데이터 셋의 품질검증을 위해, 멀티모달 데이터를 평가하기에 적합한 품질특성을 제시하고 그 적용 방안에 대해서 알아보고자 한다.

2. 생성형 멀티모달 AI 모델 학습의 특성

생성형 멀티모달 AI 모델은 텍스트, 음성, 이미지, 비디오, 시계열 정보 등 다양한 유형의 데이터를 입력 후 결합한다. 이를 통해, AI 모델이 각 데이터간의 관계성을 학습하고 통합적으로 정보를 처리·이해함으로써 유니모달 모델의 생성 결과보다 더 풍부하고 통합된 결과를 생성한다. 이렇듯 생성형 멀티모달 AI는 텍스트·이미지·비디오·음성 생성 등 질적인 생성물을 출력하는데 여러 종류 데이터를 융합하고 학습할 수 있는 알고리즘으로 구현된다. 대표적인 멀티모달 AI 모델의 내부구조 사례[1]는 [그림 1]과 같다.

멀티모달 AI 모델의 내부 구조를 살펴보면, 학습 또는 튜닝(Tuning)을 위한 다양한 이기종 데이터들이 멀티모달 AI에 입력돼 처리된다. 이기종 데이터들을 각각의 양식, 모달리티(Modality)라고

1) 인간을 닮은 인공지능, 멀티모달 인공지능 기술 동향, IITP 주간기술동향, 2024.6.12

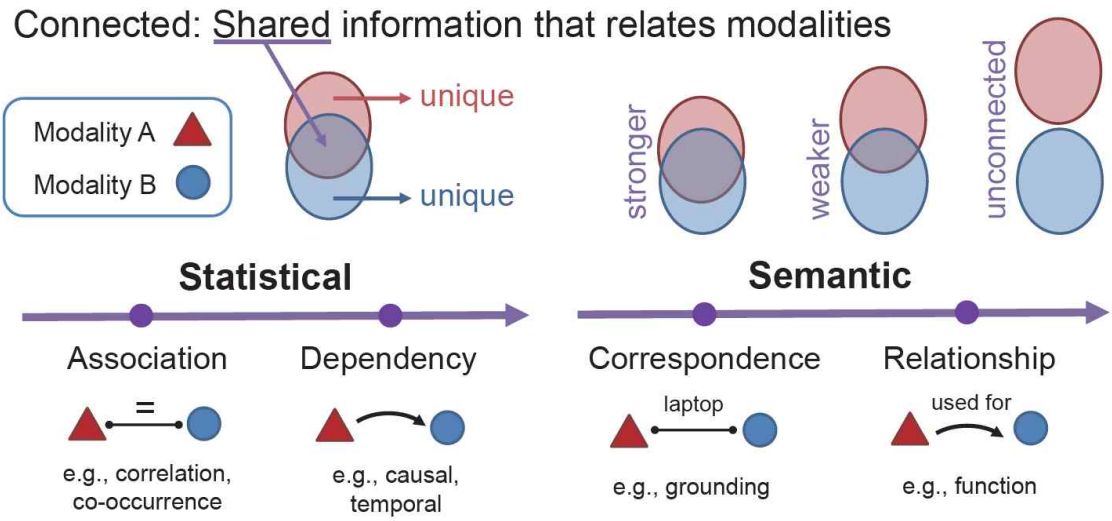


[그림 1] 멀티모달 AI 내부 구조 사례(NExT-GPT)

하는데, 멀티모달 AI에서 모달리티 연결은 모달리티가 어떻게 관련돼 있고 공통점을 공유하는지 명하는데 중요한 특성이다.

[그림 2]는 모달리티 간 연관된 정보가 통계적, 의미적 측면에서 공유되고 연결 및 고려되는 모습을 보여준다. 통계적 데이터 중심 관점에서의 연결은 다중모달 데이터 분포 패턴에서 식별된다. 반면, 의미적 접근방식은 모달리티가 고유한 정보를 공유하고 포함하는 방식에 대한 도메인 지식을 기반으로, 연결을 정의하고 있다.[2)]

따라서, 다양한 모달리티 정보 간 의미적 접근 방식을 기반으로, 멀티모달 AI 모델이 다양한 유형의 데이터를 입력 및 처리하는 메커니즘을 고려해, 멀티모달 데이터의 품질특성을 정의할 수 있다.



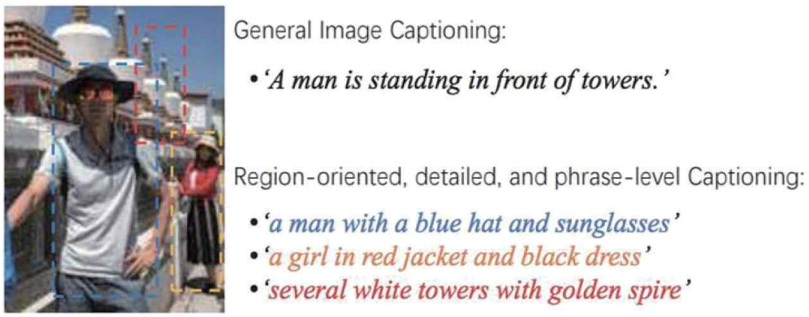
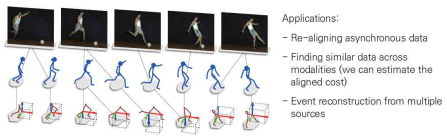
[그림 2] 모달리티 간 정보 연결

2) Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, arXiv:2209.03430v2, 20 Feb 2023

3. 멀티모달 데이터의 품질특성

이러한 생성형 멀티모달 AI 모델 학습의 특성을 고려해 멀티모달 데이터의 품질특성을 <표 1>과 같이 정의하고 멀티모달 데이터의 품질검증 시 품질특성으로 적용한다. 각 품질특성별 품질측정 지침은 다음과 같다.

<표 1> 멀티모달 데이터 품질특성

품질특성	지침
표현성 (Representation)	각 유니모달별 라벨링 정보와 실제 참값이 부합하는지를 확인하는 특성으로, 각 유니모달의 개체 식별이 가능한지를 판단한다. 유니모달의 참값 부합 여부에 대한 검증 기준은 TTA 단체표준 TTAK.KO-10.1419, TTAK.KO-10.1420의 학습임무(Task)별 품질검증 방법에 설명돼 있다.
일치(변환)성 (Translation)	멀티모달 간의 매핑이 정확한지를 판단한다. 이를 위해, 각 유니모달별 식별대상(Entity)이 같은 객체인지 등의 일치요소를 추출해 요소들 간 의미가 동일한지를 판단한다. 예를 들어, 언어와 이미지의 동일한 개념 간 대응관계를 찾기 위해, 원천데이터는 하나의 이미지 장면이나 컨텍스트를 가지는 텍스트 문장·문단, 발화 문장을 한 단위로 한다. 이미지 캡셔닝과 같이 캡션의 단어나 문구에 해당하는 이미지 영역을 찾아 매칭 정확성을 평가한다.[3]  [그림 3] 이미지와 텍스트(설명문) 간의 일치성
정렬(관계)성 (Alignment)	멀티모달 간 시간·공간적 요소에 종속되는 데이터의 품질특성으로, 둘 이상의 양식에서 인스턴스 하위 구성 요소 간 관계와 대응[그림 4][4]이 정확한지를 평가한다. 비디오 설명[그림 5][5]과 같이 비디오를 기반으로 한 대본이나 책의 챕터에 맞춰 시간적 정렬이 정확한지를 평가한다.  [그림 4] 인스턴스 하위 구성요소 (객체, 동작) 간의관계와 대응  C: a man is standing in a kitchen putting groceries away. He closes the cabinet when finished, walks over to a table and pulls out a chair and sits down. S: a man puts away his groceries and then sits at a kitchen table and stares out the window. [그림 5] 비디오 영상과 영상 프레임 이미지, 텍스트(묘사문(Caption)),설명문(Sentence) 간의 정렬(관계)성

3) MMML-Tutorial-ACL2017, CMU, 2017
4) ICML2023_Tutorial, CMU, 2023
5) Multimodal Transformer Networks for End-to-End Multimodal Transformer Networks for End-to-End, Hung Le etc, 2019

4. 멀티모달 데이터 학습업무별 품질특성 적용

4.1 멀티모달 데이터 구성 조합

생성형 멀티모달 AI 모델의 학습을 위해선 여러 유형의 멀티모달 데이터가 필요하다. 생성형 AI 학습을 위한 멀티모달 데이터를 데이터 관점에서 보면, 이미지, 텍스트, 오디오, 비디오, 수치형 데이터, 그래프로 구성된 복잡한 데이터 구조 등 다양한 유형의 조합으로 볼 수 있다. 생성형 멀티모달 모델의 대표적인 학습 업무별 멀티모달 데이터 구성 예시는 <표 2>와 같다.

<표 2> 대표 학습업무별 멀티모달 데이터 구성 조합

데이터유형 / 학습업무	텍스트	음성	이미지		비디오	수치형	멀티모달 데이터 구성
	text	speech (sound)	image(2D)	image(3D)	video	numerical	
시각적 음성·감정 인식	✔	✔			✔		• 비디오-음성 (전사텍스트) • 비디오-이미지(3D)-음성 (전사 텍스트)
시각적 음성 합성	✔	✔			✔		
이벤트 탐지(음성)	✔	✔			✔		
음성 및 모션 합성	✔	✔		✔	✔		
비디오 설명	✔				✔		• 비디오-텍스트
비디오 생성	✔				✔		
영상 검색	✔		✔		✔		• 이미지-음성 (전사 텍스트) • 이미지-텍스트
이미지 생성	✔	✔	✔	✔			
이미지 설명	✔		✔	✔			
이벤트 탐지(이미지)	✔		✔				
객체, 행동 분류	✔		✔				
시각적 질의응답	✔		✔				
감정인식	✔					✔	• 수치형-텍스트
이상 탐지	✔					✔	

4.2 멀티모달 데이터 학습업무별 품질특성 적용방안

4.2.1. 시각적 음성인식

시각적 음성인식은 비디오에서 입 모양과 같은 시각적 신호와 음성 신호를 분석하고, 말하는 사람이 무엇을 말하고 있는지 파악하는 것을 목표로 한다. 이는 청각 정보가 불분명하거나 소음이 있는 환경에서 유용하며, 시각적 정보를 통해 음성인식의 정확도를 높이는데 기여할 수 있다. 데이터 유형은 비디오 내 입 모양을 식별할 수 있는 어노테이션(Bounding Box, Keypoint 등), 그에 해당하는 음성, 전사 텍스트 등으로 이뤄져 있다. 또한, 다양한 발음, 얼굴 각도 등으로 이미지 데이터 셋의 다양성을 고려할 수 있다. 시각적 음성인식을 효과적으로 수행하기 위해선 음성, 비디오, 전사 텍스트 간의 일치성을 검증해야 하며, 음성과 이미지 프레임의 정렬성을 검증해야 한다.

4.2.2 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, 2D·3D 이미지, 수치형 데이터 등이 포함된다. 2D·3D 이미지에선 동일한 객체 일치성 및 비디오와 이미지 간 일치성 검증 작업이 필요하다. 비디오 데이터에선 객체가 시간에 따라 이동하는 경로를 추적, 수치형 데이터와 함께 객체

추적의 정렬성을 확인해야 한다.

4.2.3 이벤트 탐지 및 이상 탐지

이벤트 탐지 및 이상 탐지는 데이터 내 특정 사건이나 이상 상황을 탐지하고 식별하는 것을 목적으로 한다. 데이터 유형은 비디오, 이미지, 수치형 데이터, 음성, 텍스트 등 여러 가지 조합으로 구성돼 있다. 예를 들어, 기상 정보 데이터는 기상 정보 텍스트, 이미지, 센서 데이터 간 일치성 및 이미지와 센서 데이터 간 정렬성을 검증해야 한다. 또, 열화상 데이터에선 열화상 비디오에서 감지된 온도 변화와 텍스트 설명에 대한 일치성, 비디오와 온도의 변화 정렬성을 검증해 비디오에서 나타나는 시각적 변화와 실제 온도가 일관되게 나타나는지 확인할 필요가 있다.

4.2.4 비디오 및 이미지 설명

비디오 및 이미지 설명은 비디오나 이미지 내용을 정확하게 설명하거나 요약하는 것을 목표로 한다. 데이터 유형은 비디오, 이미지, 텍스트 등으로 이뤄져 있으며 비디오 설명의 경우, 비디오 데이터, 주요 객체, 주요 사건의 일치성을 검증하는 것이 필요하다. 또한, 비디오 데이터와 설명문 간 시간적 흐름이 정확한지에 대한 정렬성을 검증해야 한다. 이미지 설명의 경우, 이미지 데이터의 시각적 내용을 정확하게 반영하는지를 확인하기 위해 이미지 데이터와 설명문의 일치성을 검증한다. 이때 하나의 이미지 장면에 대한 설명 정보엔 시퀀스가 존재하지 않으므로 정렬성을 검증하기에 부적합하다.

4.2.5 시각적 질의응답

시각적 질의응답은 비디오나 이미지에 대한 자연어 질문에 답변을 생성하는 것이다. 이를 통해, 시각적 데이터의 이해·해석 능력을 향상시키는 것이 목표다. 데이터 유형으로는 비디오, 이미지, 텍스트 등이 있다. 이미지와 질의응답 텍스트 쌍으로 구성된 데이터는 이미지와 질의응답의 일치성을 검증한다. 비디오, 질의응답 텍스트, 번역 텍스트로 이뤄져 있는 데이터는 비디오 데이터와 질의응답의 일치성, 그리고 질의응답과 번역문의 일치성을 검증할 필요가 있다. 질의응답 구성에 비디오 순서를 고려한 경우엔 비디오 내용과 질의응답에 대한 정렬성 검증이 필요하다.

4.2.6 비디오 및 이미지 생성

비디오 및 이미지 생성은 AI 모델을 사용해 여러 가지 방식으로 데이터 증강, 콘텐츠 창작, 시뮬레이션 등을 목적으로 하는 시각적 데이터를 생성한다. 데이터 구성은 비디오, 2D와 3D이미지, 음성, 텍스트 등으로 이뤄져 있다. 2D와 3D 이미지의 객체 일치성 검증, 음성과 3D 이미지 입 모양 일치성 검증이 필요하다. 특히 스토리(Story)를 포함한 동작 데이터의 경우, 비디오 데이터와 동작의 일치성 및 비디오 데이터의 동작과 스토리 텍스트의 일치성을 검증하고, 동일한 화자 객체 확인을 위해 특정 시간 단위 프레임 기준으로 프레임별 동일 화자 여부를 확인하는 정렬성을 검증해야 한다.

3. 맺음말

생성형 AI는 멀티모달 AI가 활성화되면서 그 활용 영역이 확대되고 있다. 다양한 모달리티들 간 개별요소들이 상호작용하는 멀티모달 AI는 입력 데이터로부터 다양한 모달리티들의 피쳐(Feature)들을 생성하고, 이를 이용해 주어진 임무를 해결한다. 따라서, 멀티모달 AI 모델의 학습적 측면에 초점을 맞춰 멀티모달 데이터의 특성과 요소들을 추출하는 것이 중요하며, 모달리티 내 개체, 동작, 순서 간 관계를 설명할 수 있는 의미 있는 멀티모달 데이터의 품질특성을 정의할 필요가 있다.
생성형 AI 활용을 위한 멀티모달 AI의 목표는 다양한 모달리티 간 상호작용을 효과적으로 통합하고, 이를 통해 인간중심형 인터페이스 가치를 극대화하는 것이다. 특히 언어 모델에서 LLM(거대언어모델, Large Language Model) 기반 멀티모달 AI 시대로의 변화가 진행 중인데, 이때 다양한 영역에서 생성형 AI의 가치 실현이 구체화될 수 있어야 한다. 이를 위해, 향후엔 멀티모달 AI 모델의 여러 임무별로 구체화된 세부 품질특성들을 연구해, 최적의 성능을 위한 고품질 멀티모달 데이터 평가가 이뤄지길 바란다.

※ 본 연구는 과학기술정보통신부 초거대AI 생태계 조성사업(2024-0286, 2024년 초거대AI 확산 생태계 조성사업)에 의해서 수행됐음

[참고문헌]

[1] Multimodal Large Language Models: A Survey, arXiv:2311.13165v1, 2023.
[2] Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions, arXiv:2209.03430v2, 2023.
[3] Multimodal Transformer Networks for End-to-End Multimodal Transformer Networks for End-to-End, arXiv:1907.01166v1, 2019.
[4] 생성형 인공지능 학습을 위한 멀티모달 데이터의 품질검증 방법, TTAK.KO-10.1558, 2024.
[5] 인간을 닮은 인공지능, 멀티모달 인공지능 기술 동향, IITP 주간기술동향, 2024.
[6] MMML-Tutorial-ACL2017, CMU, 2017. https://www.youtube.com/playlist?list=PLTLz0-WCKX616TjsrgPr2wFzKF54y-ZKc
[7] ICML2023_Tutorial, CMU, 2023. https://cmu-multicomp-lab.github.io/mmml-tutorial/icml2023/

※ 출처: TTA 저널 제218호