026 멀티모달

Multimodal

다른 형태의 데이터를 연계해 의미를 통합적으로 이해하는 기술

- 텍스트·이미지·음성 등 다양한 입력 정보를 동시에 처리해, 단일 데이터만으로는 파악하기 어려운 의미를 종합적으로 이해하고 연관짓는 기술
- 서로 다른 정보가 연결되며, 상황과 맥락을 인간처럼 종합적으로 이해하도록 돕는 핵심 기반

멀티모달의 개념

멀티모달은 '여러 형태(modality)의 데이터가 결합된다'는 뜻으로, 한 가지 형태의 정보만 처리하던 기존 방식에서 벗어나 텍스트, 이미지, 음성, 영상 등 다양한 형태의 데이터를 함께 이해하고 연관짓는 기술을 의미합니다. 사람은 시각, 청각, 언어 등 여러 감각을 동시에 사용해 상황을 인식합니다. 멀티모달 AI의 경우에도 여러 데이터의 상호 관계를 학습해 보다 풍부한 맥락을 이해할 수 있도록 설계됩니다. 예를 들어, "웃는 사람"이라는 문장을 보고 실제 웃고 있는 얼굴 이미지를 연관지을 수 있고, 음성·영상 데이터를 함께 분석해 감정이나 의도를 추론할 수도 있습니다.

멀티모달의 활용

멀티모달 기술은 생성형 AI, 자율주행, 의료 진단, 감정 인식, 로봇 비전 등 다양한 분야에서 활용됩니다. 예를 들어, 텍스트 설명으로 이미지를 생성하는 모델, 이미지와 텍스트를 동시에 이해하는 검색 시스템, 영상과 음성을 결합한 감정 분석 모델 등이 있습니다. 이러한 AI는 복합적인 맥락을 고려해 더 정확한 판단과 표현을 수행하며, AI가 인간의 감각과 사고 방식을 닮아가는 과정으로 평가됩니다. 다만 데이터 결합 과정에서 편향이 증폭되거나, 특정 모달이 과도하게 영향을 미치는 문제가 발생할 수 있습니다. 따라서 멀티모달 AI의 발전은 단순한 성능 향상을 넘어, 정보 간 균형과 의미의 정합성을 확보하는 방향으로 이어지고 있습니다.

관련 용어

멀티모달 거대 언어모델 / MLLM (Multimodal Large Language Model)

기존의 텍스트 기반 LLM을 확장해 이미지·음성·영상 등 다양한 형태의 데이터를 함께 이해하고 생성하는 AI 모델입니다. 언어 모델의 언어적 추론 능력에 시각·청각 정보 처리를 결합해, 복합적 맥락을 통합적으로 해석하고 다중 입력 간 의미를 정렬합니다. 예를 들어 이미지 속 장면을 설명하거나, 사용자의 음성 질문에 시각 정보를 결합해 답변하는 등 언어와 감각 정보를 동시에 활용하는 통합형 AI를 구현합니다. GPT-4o, Gemini 1. 5, Claude 3 등은 대표적 MLLM으로, 인간의 감각 인식 구조를 모방해 AI의 이해력과 표현력을 한 단계 확장한 차세대 모델로 평가됩니다.