037 비전언어모델 / VLM

Vision-Language Model

이미지와 텍스트를 함께 이해하고 처리하는 AI 모델

- 시각 및 언어 정보를 결합해 이미지 해석, 설명 생성, 질의응답 등을 수행하는 멀티모달 AI 모델
- 화면·장면·문맥을 통합적으로 이해해 다양한 작업을 처리하는 구조

비전언어모델의 개념

비전언어모델(VLM)은 이미지와 텍스트를 동시에 입력받아 서로의 의미를 연결해 이해하도록 설계된 AI 모델을 의미합니다. 기존에는 이미지 분석과 언어 처리가 별도의 모델에서 이루어졌지만, VLM은 두 정보를 통합적으로 해석해 하나의 과제로 처리한다는 점에서 차별됩니다. 예를 들어 한 장의 사진에서 사물을 인식하는 것뿐 아니라, 사진 속 상황을 설명하거나 특정 부분에 대해 질문에 답하는 등 복합적 이해가 가능합니다. 이러한 모델은 시각·언어 정보를 결합해 더 자연스럽고 인간적인 방식으로 세계를 이해하려는 AI 발전 흐름 속에서 등장했으나, 최근에는 처음부터 다양한 모달리티(텍스트·이미지·음성 등)를 단일 모델에서 일관되게 처리하도록 설계된 거대 멀티모달 모델(LMM)로 발전하고 있습니다.

비전언어모델의 구성

VLM은 일반적으로 시각 정보를 처리하는 비전 인코더와 언어 정보를 처리하는 언어 모델을 결합해 작동합니다. 먼저 비전 인코더가 이미지에서 특징을 추출해 벡터 형태로 변환하고, 언어 모델은 이 벡터를 문맥 정보로 활용해 질문에 답하거나 설명을 생성합니다. 이 과정에서 두 정보가 서로 연결되도록 멀티모달 임베딩 공간이 활용되며, 이미지와 텍스트가 의미적으로 정렬될수록 모델의 이해 능력이 향상됩니다. 최근에는 LLM을 중심에 두고 이미지 인코더를 접목하는 구조가 주류가 되었으며, 이를 통해 언어 기반 지시를 시각적 판단과 결합해 더 복잡한 작업을 수행할 수 있게 되었습니다. 이러한 구조는 텍스트와 이미지가 서로의 부족한 부분을 보완하여 더 정교한 추론을 가능하게 합니다.

비전언어모델의 전망

비전언어모델은 고객지원 자동화, 시각 검색, 이미지 설명 생성 등 시각·언어 결합이 필요한 업무에서 핵심 기반 기술로서 활용되고 있습니다. 그러나 최근 AI 발전 흐름은 더 높은 통합성과 일반성을 갖춘 LMM 중심으로 이동하고 있습니다. LMM은 텍스트·이미지·음성·영상 등 다양한 정보를 단일 모델에서 일관되게 처리하며, 멀티모달 이해·추론·액션 수행까지 확장되는 경향을 보입니다. 이로 인해 전통적 VLM은 독립적 모델군이라기보다는 멀티모달 AI로 발전해가는 과정에서 등장한 전환기적 기술 단계로 평가받고 있습니다.