LLaVA: Visual Instruction Tuning

LLaVA 연구는 멀티모달 모델이 단순히 이미지나 텍스트를 각각 처리하는 수준을 넘어서, 두 정보를 자연스럽게 통합해 지시를 이해하고 수행하는 assistant 로 발전하고 있다는 흐름 속에서 등장했다. 최근 대형 언어 모델들은 In-context Learning 과 Instruction Tuning 을 기반으로 높은 수준의 언어적 추론 능력을 보여주고 있지만, 이렇게 지시를 따르는 능력이 시각적 정보로까지 확장된 사례는 많지 않았다. 따라서 연구자들은 텍스트뿐 아니라 이미지를 함께 받아들여 인간 의도에 맞게 응답하는 모델, 즉 실제 멀티모달 assistant 가 필요하다고 보았다. 이 배경이 바로 LLaVA 가 제시하는 Visual Instruction Tuning 의 출발점이다.

기존 비전-언어 모델들은 비교적 한정적인 태스크를 위해 설계되거나, 여러 모델을 외부 도구로 조합해서 작동하는 경우가 많았다. 이런 방식은 모델의 능력을 end-to-end 로 확장하는 데 어려움이 있고, '범용 assistant'라는 목표와는 거리가 멀었다. 특히 멀티모달 instruction 데이터를 구축하는 것이 어려웠기 때문에, 모델이 이미지 기반의 지시를 자연스럽게 따르는 능력을 직접적으로 학습하지 못했다. LLaVA 는 이러한 문제를 해결하기 위해 언어 모델이 이미지에 기반한 지시를 수행하도록 만드는 새로운 학습 전략과 데이터 생성 파이프라인을 제안한다. 즉, Vision–Language 분야에 instruction tuning 개념을 본격적으로 적용하려는 시도이며, 이를 통해 보다 자연스럽게 대화하고 추론하는 멀티모달 모델을 생성하는 것이 목표다.

LLaVA 의 방법론에서 가장 중요한 구성요소는 GPT-assisted Visual Instruction Data Generation 이다. 기존의 이미지–텍스트 데이터셋은 단순 캡션 중심이기 때문에 '지시를 따르는 방식의 데이터'가 아니다. 이를 보완하기 위해 연구팀은 ChatGPT 와 GPT-4 를 활용해 기존 이미지 캡션과 bounding box 정보만을 입력으로 제공하고, 이를 기반으로 instruction-following 형식의 고품질 데이터를 자동 생성한다. 이 방식은 실제 이미지를 GPT 에 직접 보여주지 않더라도, 이미지의 구조적 정보를 설명하는 캡션과 위치 정보를 이용해 대화형, 상세 묘사형, 복잡한 추론형의 세 가지 유형 데이터를 얻을 수 있다는 점에서 효율적이다. 결국 사람이 직접 annotation 하지 않고도 방대한 시각적 instruction 데이터를 확보할 수 있었고, 이것이 모델 정렬 과정의 핵심이 된다.

모델 아키텍처는 단순하지만 효과적이다. LLaVA 는 CLIP 의 vision encoder 를 사용해 이미지 임베딩을 뽑고, 이를 projection layer 를 통해 Vicuna 기반 LLM 의 입력 공간과 동일한 차원으로 정렬한다. 이후 이미지 임베딩과 텍스트 토큰을 함께 언어 모델에 입력하면, 모델은 이미지 내용을 이해한 상태에서 질의에 대한 답변을 생성할 수 있게 된다. 학습은 두 단계로 진행되는데, 1 단계에서는 이미지 임베딩을 LLM 의 토큰 공간에 안정적으로 정렬시키는 데 집중한다. 이 단계에서 vision encoder 와 LLM 은 모두 고정하고 projection layer 만 학습한다. 반면 2 단계에서는 multi-turn 대화와 상세 묘사, 복잡한 추론 데이터를 활용해 모델을 end-to-end 로 미세 조정하며, 이 과정에서 LLM 과 projection layer 가 함께 학습된다. 이러한 two-stage 전략 덕분에 모델은 시각–언어적 이해를 동시에 강화하면서도 기존 LLM 의 능력을 해치지 않는 방식으로 alignment 가 가능해진다.

실험 결과에서도 LLaVA 의 장점이 뚜렷하게 나타난다. Multimodal Chatbot 평가에서 LLaVA 는 GPT-4 technical prompts 를 기반으로 한 대화 시나리오에서 다른 비전–언어 모델보다 더 나은 지시 수행 능력을 보였다. 특히 복잡한 reasoning 상황에서 모델답게 구조화된 설명을 제공하며, 실제 assistant 처럼 자연스럽게 응답하는 점이 인상적이다. 또한 COCO 실험에서는 모든 데이터 유형을 조합해 학습했을 때 가장 높은 점수를 기록했으며, 복합 추론 유형에서는 GPT-4 와 매우 근접한 성능을 보였다. LLaVA-Bench 에서도 다양한 도메인 이미지에 대해 높은 일반화 성능을 보여 멀티모달 대화 능력이 실제 상황에서도 강하다는 점을 확인할 수 있었다. ScienceQA 실험에서는 단독 모델 기준으로 기존 SoTA 모델에 근접한 90.92% 정확도를 기록했고, GPT-4 를 judge 로 활용하는 앙상블 방식에서는 92.53%의 정확도로 새로운 SoTA 를 달성했다. 이는 LLaVA 가 단순한 이미지 이해를 넘어서 과학적 reasoning 까지 수행할 수 있는 강력한 잠재력을 갖고 있음을 시사한다.

결론적으로 LLaVA 는 멀티모달 instruction-following 모델의 새로운 기준을 제시했다. 모델 자체는 비교적 단순한 구조임에도 불구하고, GPT 를 활용한 데이터 생성 파이프라인과 two-stage 학습 전략을 통해 시각적 지시 수행 능력을 성공적으로 확보했다. 특히 멀티모달 instruction 데이터를 자동 구축하는 방법을 정립했다는 점, CLIP 과 Vicuna 를 결합해 효과적으로 alignment 를 수행했다는 점, 그리고 전체 과정과 모델을 오픈소스로 공개해 연구 생태계를 확장했다는 점에서 중요한 의미를 갖는다. LLaVA 는 이후 멀티모달 assistant 연구의 방향성을 제시한 모델로, 이미지/텍스트/추론을 통합한 실제 사용자 친화적 멀티모달 시스템이 어떻게 구축될 수 있는지를 잘 보여준다