

# InstructBLIP review



InstructBLIP은 이미지와 선택적 텍스트 프롬프트가 주어지면 텍스트를 생성할 수 있는 시각-언어 모델임.

이 모델은 세 가지 주요 구성 요소로 이루어져 있음: Visual Encoder, Q-Former(from BLIP-2) and Large Language Model

Visual Encoder는 이미지에서 시각적 특성을 추출하는 역할을 함. Q-Former는 주어진 Instruction에 맞게 정보적인 특성을 추출하는 새로운 모듈임. 언어 모델은 시각적 및 텍스트 특성을 바탕으로 텍스트를 생성하는 responsibility가 있음.

모델에는 선택적으로 input\_ids를 Text prompts로 전달할 수 있어, LLM이 프롬프트를 이어서 생성하게 할 수 있음.

**InstructBLIP은 400M 이미지-텍스트 쌍으로 훈련된 대규모 시각-언어 모델인 BLIP-2 모델을 기반으로 함. BLIP-2 모델은 이미지 캡셔닝, 시각적 질문 답변, 시각적 추론, 이미지 생성 등 다양한 시각-언어 작업에서 인상적인 결과를 보여줌.**

그러나, BLIP-2 모델은 각 downstream task에 대해 특정한 미세조정이나 적응이 필요하여 유연성이나 일반화가 부족함.

또한, BLIP-2 모델은 입력에 따라 관련 없거나 모순된 텍스트를 생성할 수 있어 robustness나 consistency가 떨어짐.

**이러한 한계를 극복하기 위해, InstructBLIP은 다양한 작업과 입력에 적응할 수 있도록 Instruction Tuning을 사용함.**

**Instruction Tuning**은 모델이 무엇을 해야 하는지 또는 생성해야 하는지를 지정하는 자연어 지시를 따를 수 있게 하는 기술임.

예를 들어, "이미지의 요약물 작성하라"는 지시가 주어진다면, InstructBLIP은 이미지 내용의 간결하고 정보적인 요약물 생성할 수 있어야 함. 마찬가지로, "이미지에 대한 질문에 답하라"는 지시가 주어진다면, InstructBLIP은 이미지와 질문에 기반하여 올바른 답변을 생성할 수 있어야 함.

지시 튜닝을 통해, InstructBLIP은 특정 작업에 대한 미세조정이나 적응 없이도 다양한 시각-언어 작업을 수행할 수 있게 됨. 또한, 입력과 지시에 기반하여 더 관련성 있고 일관된 텍스트를 생성할 수 있게 됨.

---

InstructBLIP은 이미지, Instruction, 그리고 Selective Text Prompt 세 가지 입력을 받아 작동함. 이미지는 Visual Encoder에 의해 처리되어 이미지에서 시각적 특성을 추출함. 지시문과 선택적 텍스트 프롬프트는 LLM에 의해 처리되어 텍스트 특성으로 인코딩됨.

시각적 및 텍스트 특성은 그 다음 Q-Former에 공급됨. Q-Former는 주어진 지시에 맞춤형 정보적 특성을 추출하는 새로운 모듈임. Q-Former는 지시에 기반하여 시각적 및 텍스트 특성의 다른 부분에 선택적으로 주목하기 위해 Attention Mechanism을 사용함.

Q-Former는 지시에 기반한 모델이 생성해야 할 내용을 나타내는 쿼리 특성을 출력함. 쿼리 특성은 다시 언어 모델에 공급되어 텍스트로 Decode됨.

언어 모델은 Auto-regressive으로 텍스트를 생성함,  
즉 이전 단어와 쿼리 특성을 기반으로 한 번에 한 단어씩 생성함. 언어 모델은 특별한 텍스트  
종료 토큰에 도달하거나 최대 길이 한도에 도달할 때까지 텍스트 생성을 멈춤.

InstructBLIP의 출력은 이미지와 선택적 텍스트 프롬프트에 대한 지시를 따라 생성된 텍스트로 반환됨.

---

InstructBLIP의 장점.

InstructBLIP은 다른 시각-언어 모델에 비해 여러 가지 장점이 있음:

**Flexibility:** InstructBLIP은 자연어 지시를 따라 다양한 시각-언어 작업을 수행할 수 있음.

이는 InstructBLIP이 각 하류 작업에 대해 특정한 미세 조정이나 적응이 필요하지 않음을 의미함. 또한, InstructBLIP은 다양한 지시를 따라 다양한 입력과 출력을 처리할 수 있음.

**Generality:** InstructBLIP은 어떠한 특정 작업 미세 조정이나 적응 없이도 다양한 시각-언어 작업에 대해 최신 성능을 달성할 수 있음. InstructBLIP은 기존의 시각-언어 데이터셋이나 벤치마크에 포함되지 않은 작업들, 예를 들어 이미지 생성, 이미지 편집, 이미지 요약 등을 수행할 수도 있음.

**Robustness:** InstructBLIP은 입력과 지시에 기반하여 더 관련성 있고 일관된 텍스트를 생성할 수 있음. 예를 들어, 이미지가 흐릿하거나 자른 것이거나 부분적으로 가려져 있더라도 InstructBLIP은 지시를 따라 텍스트를 생성할 수 있음.

**Explainability:** InstructBLIP은 Q-Former의 Attention weight를 보여주면서 텍스트 생성에 대한 설명을 제공할 수 있음. 주의 가중치는 주어진 지시에 가장 관련성 있는 시각적 및 텍스트 특성의 어떤 부분인지를 나타냄. 이는 사용자가 InstructBLIP이 어떻게 작동하고 무엇을 생성하는지 이해하는 데 도움을 줄 수 있음.