JDFR Journal of Data Forensics Research
데이터포렌식연구, 2권 1호, 2025년 6월, 37-54면
Journal of Data Forensics Research, Volume 2, Issue 1, June 2025, Pages 37-54.
https://doi.org/10.12972/JDFR.2025.2.1.3
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

LLaVA-1.6 모델의 파인튜닝을 통한 오토바이 안전모 착용 여부 자동 판독 시스템 개발 연구

김태연[1], 김경종[2]

[1]경찰대학 학부생, [2]경찰대학 경찰학과 교수

교신저자: 김경종, leeyeongul@police.go.kr

요약

기존의 객체 검출 기반 자동화 연구들은 안전모 객체의 존재 유무는 판별할 수 있으나, 실제 '착용' 상태에 대한 맥락적 이해나 다양한 질의에 대한 유연한 응답 생성에는 어려움이 있었다. 본 연구는 이러한 한계를 극복하기 위해 최신의 대규모 언어-시각 모델(VLM)인 LLaVA-1.6을 활용하여, 오토바이 주행자 이미지에 대해 "이 오토바이 주행자가 안전모를 착용했습니까?"라는 자연어 질문을 입력받았을 때, "예, 착용했습니다." 또는 "아니요, 착용하지 않았습니다." 형태의 명확한 답변을 생성하는 자동 판독 시스템을 개발한다. 공개 데이터셋 기반 커스텀 질의응답 데이터셋을 구성하고, 파라미터 효율적 파인튜닝 기법인 QLoRA를 적용하여 LLaVA-1.6 모델을 특정 작업에 맞게 최적화하였다. 실험 결과, 제안 시스템은 순수 모델(휴리스틱 적용 시 약 88%) 대비 우수한 95% 정확도를 달성했으며, 특히 치명적 오류(False Negative)를 현저히 감소시켜 실제 단속 효용성을 높였다. 본 연구는 VLM의 특정 작업 최적화를 통해 교통안전 AI 활용성을 확장하고 자동 단속 시스템의 효율과 정확성 증진에 기여함을 시사한다.

주제어

LLaVA-1.6, VLM, QLoRA, 파인튜닝, 시각 질의 응답, 교통안전

Received: June 11, 2025
Revised: June 28, 2025
Accepted: June 30, 2025
Published: June 30, 2025

Original Article

A study on the development of an automated motorcycle helmet wearing detection system via fine-tuning of the LLaVA-1.6 model

Taeyeon Kim[1], Kyungjong Kim[2]

[1]Undergraduate Student, Korean National Police University, Republic of Korea
[2]Professor, Department of Police Science, Korean National Police University, Republic of Korea

Corresponding Author: Kyungjong Kim, leeyeongul@police.go.kr

ABSTRACT

Although existing object-detection studies can identify the 'presence' of a helmet, they often fail to contextually assess whether the helmet is actually being worn. To address this limitation, we developed a visual question answering (VQA) system based on LLaVA-1.6, a state-of-the-art vision-language model (VLM), to determine the helmet-wearing status of motorcyclists. A custom dataset of 1,555 image-question-answer pairs was constructed using public datasets, specifically designed to elicit clear "Yes" or "No" responses to a fixed question. We fine-tuned the model using Quantized Low-Rank Adaptation (QLoRA), a parameter-efficient technique that specializes the model by only updating a small set of adapter weights. Experimental results showed that the proposed system achieved 95% accuracy, significantly outperforming the vanilla model, which reached only 88% accuracy and required complex heuristic post-processing. Importantly, the system notably reduced critical false-negative errors—cases in which wearers were misidentified as non-wearers—highlighting its reliability for real-world enforcement scenarios. In conclusion, this study demonstrates that task-specific optimization of VLMs can enhance the applicability of AI in traffic safety, improving both the efficiency and accuracy of automated enforcement systems.

KEYWORDS

LLaVA-1.6, VLM, QLoRA, Fine Tuning, Visual Question Answering, Traffic Safety

I. 서론

최근 개인 이동성의 증대와 함께 이륜차 및 개인형 이동장치의 이용이 급증하면서 관련 교통사고 또한 중요한 사회적 문제로 부각되고 있다. 이러한 사고 발생 시, 운전자의 안전모 착용은 두부 손상을 치명적인 수준에서 심각한 수준 이하로 경감시켜 사망률 및 중상률을 낮추는 가장 효과적이고 기본적인 보호 수단으로 알려져 있다[1]. 이에 따라 도로교통법에서는 이륜차 및 PM 운전자의 안전모 착용을 의무화하고 있으나, 실제 현장에서의 착용률은 여전히 기대에 미치지 못하고 있으며, 특히 단속 인력의 부족, 시간적·공간적 제약 등으로 인해 기존의 인력 중심 수동 단속 방식은 그 실효성과 효율성 면에서 명확한 한계를 드러내고 있다[2].

이러한 문제를 해결하기 위한 노력의 일환으로, 인공지능(AI) 기술을 활용한 자동화된 안전모 착용 여부 감지 시스템에 대한 연구가 지속적으로 이루어져 왔다. 초기 연구들은 주로 You Only Look Once (YOLO)[3]와 같은 객체 검출(object detection) 알고리즘을 기반으로 안전모 객체 자체를 탐지하는 데 중점을 두었다. 이러한 시스템들은 특정 조건 하에서 높은 탐지율을 보일 수 있으나, 단순히 안전모 객체의 존재 유무를 넘어 실제 '착용' 상태, 즉 안전모가 운전자의 머리에 올바르게 장착되어 보호 기능을 수행할 수 있는 상태인지를 맥락적으로 판단하는 데에는 근본적인 어려움이 있다. 예를 들어, 안전모를 착용하지 않고 손에 들고 있거나 차량의 다른 부분에 걸쳐 놓은 경우에도 객체 탐지 모델은 이를 '안전모 존재'로 인식할 수 있어, 실제 규제 목적에 부합하는 정확한 판독에는 미흡한 결과를 초래할 수 있다[4].

최근 몇 년간 이미지와 텍스트 정보를 동시에 이해하고 처리할 수 있는 대규모 언어-시각 모델(Vision-Language Model, VLM)[5]이 자연어 처리 및 컴퓨터 비전 분야에서 괄목할 만한 발전을 이루었다. VLM은 방대한 양의 이미지-텍스트 쌍 데이터를 사전 학습하여, 이미지 캡셔닝, 시각적 질의응답(Visual Question Answering, VQA), 지시 따르기 등 복잡한 멀티모달 작업을 수행할 수 있는 능력을 갖추게 되었다. 교통 분야에서도 TrafficVLM[6]과 같이 VLM을 활용하여 교통 상황에 대한 전반적인 이해나 비디오 캡셔닝을 목표로 하는 연구가 등장하고 있으나, 본 연구에서와 같이 특정 객체(예: 오토바이 주행자)에 대해 세밀하고 구체적인 상태(예: 안전모의 올바른 착용 여부)를 자연어 질의응답 형태로 명확히 판독하도록 VLM을 특정 작업에 특화하여 파인튜닝하는 연구는 아직 초기 단계에 머물러 있으며, 그 필요성이 더욱 강조되고 있다.

따라서 본 연구는 최신의 고성능 VLM 중 하나인 LLaVA-1.6(llava-hf/llava-v1.6-mistral-7b-hf)[7]을 기반으로, 사전에 오토바이 주행자 영역이 특정되어 크롭된 이미지를 입력으로 받아, "이 오토바이 주행자가 안전모를 착용했습니까?"라는 자연어 질문에 대해 "예, 착용했습니다." 또는 "아니요, 착용하지 않았습니다."와 같이 명확하고 직관적인 답변을 생성하는 자동 판독 시스템 개발을 목표로 한다. 이를 위해, 공개 데이터셋으로부터 안전모 착용 여부가 명시된 주행자 이미지를 수집하고, 이를 질의응답 형식에 맞게 가공하여 커스텀 데이터셋을 구축하였다. 이후, 계산 효율성을 높이면서도 높은 성능을 유지할 수 있는 파라미터 효율적 미세조정 기법인 QLoRA(Quantized Low-Rank Adaptation)[8]를 적용하여 LLaVA-1.6 모델을 특정 작업에 최적화하였다. 본 연구를 통해 개발된 시스템은 미세조정 전 순수 LLaVA-1.6 모델이 개선된 답변 판독 로직을 적용했을 때 약 88%의 정확도를 보인 반면, QLoRA 미세조정을 거친 모델은 95%의 높은 정확도를 달성하며 안전모 착용 여부를 성공적으로 판독하였다. 이러한 결과는 제안된 시스템이 향후 교통안전 단속 업무의 효율화 및 자동화에 실질적으로 기여할 수 있는 높은 잠재력을 가지고 있음을 시사한다.

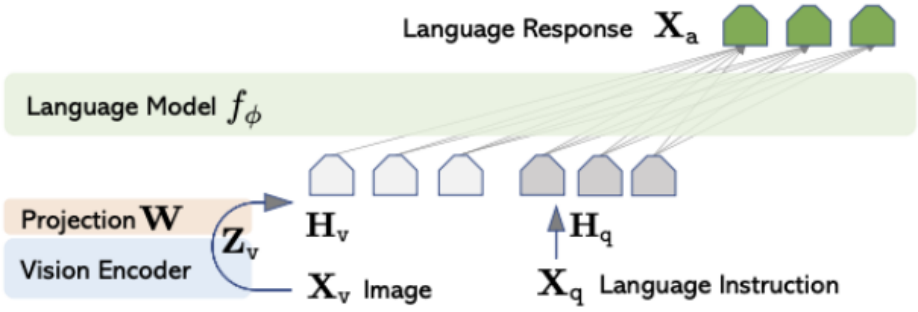
II. 관련 연구

2.1. 대규모 언어-시각 모델(VLM)의 발전 동향

대규모 언어-시각 모델(VLM)은 컴퓨터 비전과 자연어 처리 기술을 통합하여 시각적 데이터와 텍스트 데이터를 동시에 이해하고 생성하는 인공지능 모델이다. 초기 VLM 연구들은 주로 특정 작업에 국한된 성능을 보였으나[9], 트랜스포머 아키텍처[10]의 등장과 대규모 이미지-텍스트 쌍 데이터셋을 활용한 사전 학습 방식이 도입되면서 VLM의 성능은 비약적으로 발전하였다. CLIP[11]과 같은 모델들은 이미지와 텍스트를 동일한 임베딩 공간에 매핑하여 둘 사이의 의미론적 유사성을 학습함으로써 이미지 검색, 제로샷 이미지 분류 등에서 뛰어난 능력을 선보였다. 이후, VLM은 시각적 질의응답, 이미지 캡셔닝, 시각적 추론, 그리고 복잡한 지시 따르기 등 더욱 고도화된 멀티모달 작업으로 그 영역을 확장해왔다. 특히, Flamingo[12]와 같은 모델들은 사전 훈련된 대규모 언어 모델(LLM)의 강력한 추론 및 생성 능력을 시각 정보와 효과적으로 결합하는 다양한 아키텍처와 학습 전략을 제시하며 VLM 기술의 발전을 주도하고 있다. 이러한 모델들은 사용자와의 자연스러운 상호작용을 통해 이미지에 대한 깊이 있는 이해를 바탕으로 정보를 제공하거나 특정 작업을 수행할 수 있는 가능성을 열었다.

2.2. LLaVA 모델 아키텍처 및 발전

LLaVA (Large Language and Vision Assistant)[13]는 시각적 지시 따르기 능력에 중점을 둔 대표적인 오픈소스 VLM으로, 학계 및 산업계에서 광범위한 주목을 받고 있다. LLaVA의 핵심 아키텍처는 (1) 사전 훈련된 비전 인코더(vision encoder)를 통해 입력 이미지로부터 시각적 특징을 추출하고, (2) 이 시각적 특징을 LLM이 이해할 수 있는 텍스트 임베딩 공간으로 변환하기 위한 간단한 선형 프로젝션 레이어 또는 다층 퍼셉트론(MLP) 형태의 연결 모듈, 그리고 (3) 사전 훈련된 고성능 LLM으로 구성된다. LLaVA의 초기 버전은 GPT-4와 같은 고성능 모델을 활용하여 생성한 멀티모달 지시-따르기 데이터셋을 사용하여 엔드-투-엔드로 파인튜닝 되었으며, 이를 통해 뛰어난 대화 능력과 VQA 성능을 입증하였다.



<Figure 1> LLaVA network architecture [13]

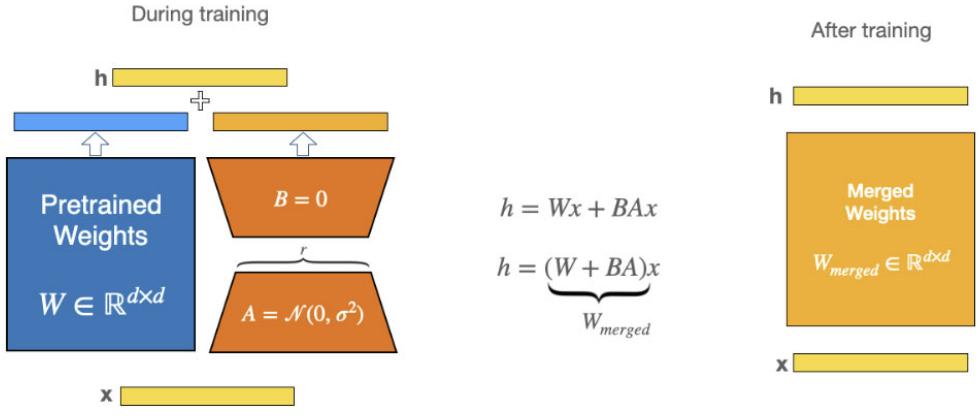
이후 LLaVA-1.5 [14]는 응답 품질 개선, 다양한 벤치마크에서의 일반화 성능 향상, 그리고 더 간단한 아키텍처 변경을 통해 효율성을 높였다. 본 연구에서 기반 모델로 사용한 LLaVA-1.6 (구체적으로 llava-hf/llava-v1.6-mistral-7b-hf 체크포인트)은 Mistral-7B[7]와 같은 최신 고성능 소형 LLM을 언어 백본으로 채택하고, 이미지 인코딩 과정에서 'AnyRes'와 유사한 동적 고해상도 처리 기법을 지원하여 다양한 종횡비의 이미지를 보다 효과적으로 처

리할 수 있도록 개선되었다[15]. 이를 통해 LLaVA-1.6은 이전 버전에 비해 향상된 시각 정보 이해 능력과 더욱 정교한 자연어 응답 생성 능력을 갖추게 되었으며, 특히 특정 작업에 대한 미세조정을 통해 그 성능을 극대화할 수 있는 잠재력을 지니고 있다.

2.3. 파라미터 효율적 파인튜닝(PEFT) 기법

LLM 및 VLM의 크기가 기하급수적으로 증가함에 따라, 모든 모델 파라미터를 업데이트하는 전체 파인튜닝 방식은 막대한 계산 자원(GPU 메모리, 학습 시간)을 필요로 하며, 이는 실제 응용 환경에서의 모델 배포 및 유지보수에 큰 부담으로 작용한다. 이러한 문제를 해결하기 위해 파라미터 효율적 미세조정(PEFT) 기법들이 활발히 연구되고 있다[16]. PEFT의 핵심 아이디어는 사전 훈련된 대규모 모델의 대부분 가중치는 동결시킨 채, 소수의 추가 파라미터 또는 특정 부분의 파라미터만을 학습하여 모델을 특정 작업에 적응시키는 것이다.

대표적인 PEFT 기법 중 하나인 LoRA(Low-Rank Adaptation)[17]는 트랜스포머 아키텍처 내의 어텐션 가중치 행렬과 같이 큰 행렬의 변화량이 저차원이라는 가정 하에, 원래의 가중치 행렬 W 옆에 두 개의 작은 저차원 행렬 A와 B의 곱(ΔW=BA)을 추가하여 이 A와 B만을 학습한다. 이를 통해 학습 가능한 파라미터 수를 크게 줄이면서도 전체 미세조정과 유사한 성능을 달성할 수 있다.



<Figure 2> LoRA adapter [17]

QLoRA(Quantized Low-Rank Adaptation)[8]는 LoRA의 효율성을 한층 더 끌어올린 기법이다. QLoRA는 (1) 사전 훈련된 기본 모델의 가중치를 4비트 NormalFloat(NF4)와 같은 정밀도로 양자화하여 메모리에 로드하고, (2) 이 양자화된 가중치 위에 LoRA 어댑터를 추가하여 학습을 진행한다. 또한, 이중 양자화(double quantization) 및 페이지드 옵티마이저(paged optimizer)와 같은 추가적인 메모리 절약 기법을 사용하여 매우 큰 모델도 단일 GPU 환경에서 미세조정할 수 있도록 지원한다. QLoRA는 특히 4비트 양자화를 사용함에도 불구하고 16비트 전체 미세조정 성능에 근접하는 결과를 보여주어, 본 연구와 같이 계산 자원이 제한적인 상황에서 대규모 VLM을 효과적으로 특정 작업에 특화시키는 데 매우 유용한 기법이다. 본 연구에서는 LLaVA-1.6 모델의 언어 모델 부분의 주요 선형 레이어(q_proj, k_proj, v_proj, o_proj 등)에 QLoRA를 적용하여 메모리 사용량을 최소화하면서도 안전모 착용 여부 판독이라는 특정 작업에 대한 성능을 극대화하고자 하였다.

2.4. 기존 안전모 착용 연구

딥러닝 기술의 발전과 함께, 합성곱 신경망(CNN)을 기반으로 하는 객체 탐지 모델들이 안전모 감지 연구의 주류를 이루게 되었다. YOLO[3]이나 Faster R-CNN[18]과 같은 딥러닝 객체 탐지 모델들은 대규모 데이터셋으로 학습될 경우 높은 정확도로 이미지나 비디오 프레임 내에서 안전모 객체의 위치를 바운딩 박스 형태로 식별할 수 있다. 실제로 다수의 연구에서 이러한 모델들을 활용하여 건설 현장 작업자나 이륜차 운전자의 안전모를 탐지하는 시스템을 개발하였다. GitHub 등에서도 “yolov3-Helmet-Detection”[19]과 같이 사전 훈련된 모델이나 관련 코드가 공개되어 있다.

<Figure 3> Yolov3-Helmet-Detection [19]

그러나 이러한 객체 탐지 기반 접근 방식은 다음과 같은 본질적인 한계를 지닌다. 첫째, 대부분의 시스템은 안전모 ‘객체’의 존재 유무만을 탐지할 뿐, 해당 안전모가 실제로 운전자의 머리에 ‘올바르게 착용된 상태’인지, 아니면 단순히 차량의 다른 곳에 놓여 있거나 손에 들고 있는 상태인지를 구분하기 어렵다. 이는 실제 안전 규정 준수 여부를 판단하는 데 있어 중요한 오류를 야기할 수 있다. 둘째, 탐지 결과가 주로 바운딩 박스와 클래스 레이블로 제한되어, “안전모를 착용했습니까?”와 같은 자연어 질의에 대한 직접적인 답변을 생성하거나, 착용 상태에 대한 부가적인 설명(예: “턱끈이 풀려있습니다.” 등)을 제공하는 데에는 한계가 있다.

2.5. 교통 및 안전 분야에서의 VLM 적용 연구

최근 VLM의 발전은 교통 및 공공안전 분야에서도 새로운 가능성을 열고 있다. 예를 들어, TrafficVLM[6]은 교통CCTV 비디오의 내용을 이해하고 설명하는 캡션을 생성하거나, 특정 이벤트에 대한 질문에 답변할 수 있는 제어 가능한 VLM을 제안하였다. 이러한 연구는 VLM이 복잡한 교통 상황에 대한 종합적인 이해를 바탕으로 유용한 정보를 추출할 수 있음을 보여준다.

<Figure 4> TrafficVLM [6]

그러나 현재까지 교통안전 분야에서 VLM을 적용한 연구들은 주로 거시적인 장면 이해나 이벤트 설명에 초점을 맞추고 있으며, 본 연구와 같이 특정 대상(오토바이 주행자)의 구체적이고 세밀한 상태(안전모 착용 여부)를 명확한 “예/아니요” 스타일의 질의응답 형식으로 판독하도록 VLM을 특화하여 파인튜닝한 사례는 상대적으로 드물다.

2.6. 본 연구의 차별성 및 독창성

앞서 살펴본 선행 연구들과 비교하여 본 연구는 다음과 같은 주요 차별성 및 독창성을 지닌다.

첫째, 고도의 작업 특화성 및 의미론적 상태 판독에 대한 성능을 입증하였다. 기존 안전모 감지 연구들이 주로 ‘안전모 객체’의 존재 유무를 탐지하는 데 그쳤다면, 본 연구는 LLaVA-1.6이라는 최신 VLM을 활용하여 ‘안전모 착용’이라는 행위 및 상태에 대한 의미론적 이해를 바탕으로 자연어 질문에 답변하도록 하였다. 이는 단순 객체 존재를 넘어 실제 안전 규정 준수 여부를 판단하는 데 더 직접적으로 기여할 수 있다.

둘째, 명확하고 직관적인 질의응답형 결과를 제공한다는 점이다. 본 연구에서 개발된 시스템은 “이 오토바이 주행자가 안전모를 착용했습니까?”라는 질문에 대해 “예, 착용했습니다.” 또는 “아니요, 착용하지 않았습니다.”와 같이 사용자가 이해하기 쉽고 명확한 형태의 답변을 생성한다. 이는 바운딩 박스나 단순 클래스 레이블을 제공하는 기존 객체 탐지 시스템에 비해 최종 사용자의 해석 용이성을 크게 높인다.

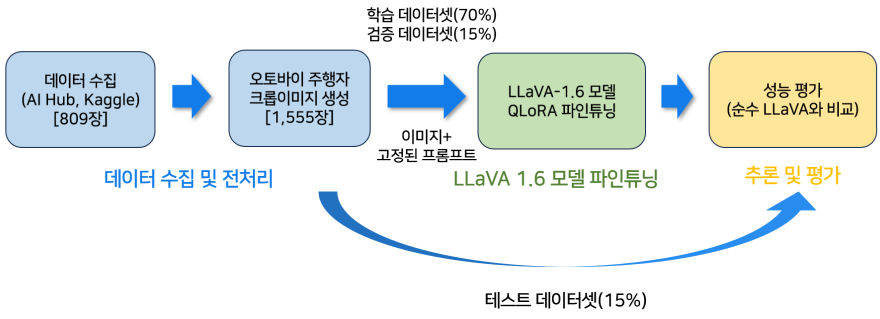
셋째, 최신 VLM과 효율적 미세조정 기법의 결합에 성공하였다. 본 연구는 Mistral-7B를 언어 백본으로 사용하는 강력한 LLaVA-1.6 모델을 기반으로 하며, QLoRA라는 파라미터 효율적 미세조정 기법을 적용하여 제한된 자원 하에서도 특정 작업에 대한 높은 성능(정확도 95%)을 달성하였다. 이는 대규모 VLM을 실제 특정 응용 분야에 맞게 최적화하는 효과적인 방법론을 제시한다.

이러한 차별점들을 통해 본 연구는 기존 안전모 감지 시스템의 한계를 극복하고, VLM을 활용한 보다 지능적이고 실용적인 교통안전 솔루션 개발에 기여하고자 한다.

III. 제안 시스템 설계 및 방법론

3.1. 시스템 전체 구조

제안하는 오토바이 안전모 착용 여부 자동 판독 시스템은 크게 세 단계로 구성된다: (1) 데이터 수집 및 전처리, (2) LLaVA-1.6 모델 미세조정, (3) 추론 및 평가 단계이다.



<Figure 5> Automated motorcycle helmet detection framework

첫째로 데이터 수집 및 전처리 단계에서는, 다양한 출처(AI Hub, Kaggle 등)로부터 오토바이 주행자 및 안전모 관련 이미지와 주석(어노테이션) 데이터를 수집한다. 수집된 데이터에서 주행자 영역을 크롭하고, 이를 고정된 자연어 질문(“이 오토바이 주행자가 안전모를 착용했습니까?”) 및 정형화된 답변(“예, 착용했습니다.” / “아니요, 착용하지 않았습니다.”)과 쌍을 이루는 JSON 형식의 학습용 데이터셋을 구축한다. 또한, 이 데이터셋의 일부를 무작위로 샘플링하여 평가용 테스트 메타데이터를 생성한다.

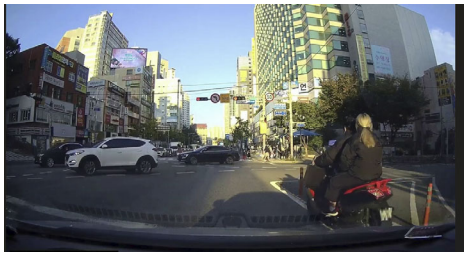
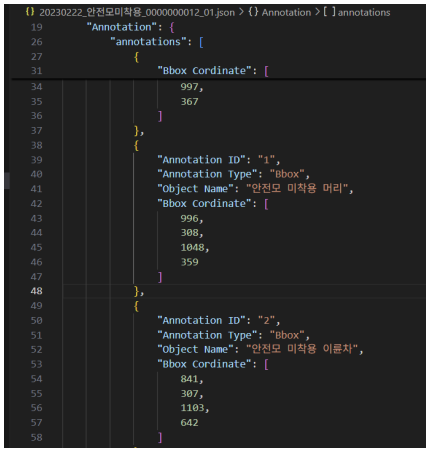
둘째로 LLaVA-1.6 모델 파인튜닝 단계에서는, 구축된 학습용 데이터셋을 사용하여 LLaVA-1.6 모델을 QLoRA 기법과 PyTorch Lightning 프레임워크를 통해 미세조정한다. 이 과정에서 모델은 특정 질문 형식에 대해 정해진 스타일의 답변을 생성하도록 학습된다.

셋째로 추론 및 평가 단계에서는, 미세조정된 모델 또는 순수(vanilla) LLaVA-1.6 모델에 테스트용 이미지를 입력하여 안전모 착용 여부에 대한 답변을 생성하게 한다. 생성된 답변과 실제 레이블을 비교하여 정확도, 정밀도, 재현율, F1-점수 등의 지표로 모델 성능을 정량적으로 평가한다.

3.2. 데이터셋 구축

AI Hub의 ‘교통법규 위반 상황 데이터’[20]와 Kaggle의 ‘Helmet Detection’[21] 공개 데이터셋에서 총 809장의 원본 이미지를 초기 데이터로 사용하였다. 이들 이미지는 안전모 착용 정보가 JSON 또는 XML 형식으로 어노테이션 정보가 함께 제공되어 있으며, 그 형식은 Table 1과 같다. 전처리 스크립트를 작성하여 주행자의 머리 또는 상반신 영역을 자동으로 크롭하여 총 1,555개의 크롭 이미지를 생성하였다. 각 크롭 이미지에 대해서는 “이 오토바이 주행자가 안전모를 착용했나요?”라는 고정된 질문과, 원본 어노테이션의 클래스 레이블(예: “안전모 미착용 머리”, “안전모 미착용 이륜차” 등)을 분석하여 “예, 착용했습니다.” 또는 “아니요, 착용하지 않았습니다.” 형태의 정형화된 답변을 생성하였다. 이렇게 생성된 (크롭 이미지 경로, 고정 질문, 정형화된 답변) 쌍들은 JSON 형식의 학습용 메타데이터 파일에 리스트 형태로 저장되어 모델 학습에 사용되었다. 이후 모델의 일반화 성능을 객관적으로 평가하기 위해, 스크립트를 작성하여 데이터 파일의 전체 데이터 중 15%를 무작위로 샘플링하여 별도의 테스트용 메타데이터 파일을 생성하였다.

<Table 1> Example dataset of AI Hub ‘Traffic Violation’

Image	Annotation
	

3.3. LLaVA-1.6 모델 파인튜닝

본 연구에서는 llava-hf/llava-v1.6-mistral-7b-hf 모델을 기반으로, 특정 작업인 오토바이 안전모 착용 여부 판독에 최적화하기 위해 미세조정을 수행하였다. 이 과정은 크게 QLoRA 설정을 통해 모델을 준비하는 단계와 PyTorch Lightning 프레임워크를 사용하여 실제 학습을 진행하는 단계로 나눌 수 있다.

3.3.1. 파인튜닝을 위한 QLoRA 설정

Table 2에 나타난 바와 같이, 먼저 BitsAndBytesConfig를 통해 4비트 NF4 양자화 설정을 정의하고, 이 설정을 적용하여 llava-hf/llava-v1.6-mistral-7b-hf 기본 모델을 로드한다. 이때 모델 가중치는 4비트로 로드되지만, 실제 연산은 torch.float16 정밀도로 수행되어 메모리 효율성과 연산 안정성을 동시에 확보한다. 다음으로, LoraConfig를 사용하여 LoRA 어댑터를 추가할 대상 모듈(본 연구에서는 언어 모델의 q_proj, k_proj, v_proj, o_proj 등 주요 선형 레이어), LoRA 랭크(r=8), 스케일링 인자 알파(alpha=16) 등을 지정한다. 마지막으로 peft 라이브러리의 prepare_model_for_kbit_training 함수와 get_peft_model 함수를 순차적으로 호출하여 최종적으로 QLoRA 설정이 완료된, 학습 준비된 LLaVA-1.6 모델을 생성한다. 이 과정을 통해 전체 모델 파라미터 중 약 0.29%에 해당하는 소수의 파라미터만이 학습 대상으로 지정된다.

<Table 2> Procedure of preparing LLava-1.6 model based on QLoRA

단계	설명
입력:	BASE_MODEL_ID (사전 훈련된 LLaVA 모델 경로), LORA_RANK, LORA_ALPHA, LORA_TARGET_MODULES (LoRA 적용 대상 레이어)
출력:	QLoRA 설정이 완료된 LLaVA 모델
1. 4비트 양자화 설정 정의	BitsAndBytesConfig를 사용하여 NF4 타입 4비트 양자화 및 float16 연산 정밀도 설정 quantization_config = BitsAndBytesConfig(load_in_4bit=True, bnb_4bit_quant_type="nf4", bnb_4bit_compute_dtype=torch.float16)
2. 기본 LLaVA 모델 로드 (양자화 적용)	AutoModelForCausalLM.from_pretrained를 사용하여 BASE_MODEL_ID로부터 모델을 로드하며, 1단계에서 정의한 quantization_config와 torch_dtype=torch.float16 적용.
3. LoRA 어댑터 설정 정의	LoraConfig를 사용하여 LoRA 랭크 (r), 알파 (lora_alpha), 대상 모듈 (target_modules), 드롭아웃, 편향(bias) 등을 지정. (예: task_type="CAUSAL_LM")
4. PEFT 라이브러리를 통한 QLoRA 모델 준비 및 적용	prepare_model_for_kbit_training으로 모델을 k-bit 학습에 맞게 준비. get_peft_model을 사용하여 LoRA 설정을 모델에 최종 적용.

3.3.2. Pytorch Lightning 프레임워크를 활용한 학습 진행

Table 3은 PyTorch Lightning의 LightningModule을 활용한 학습 절차를 보여준다. CustomLLaVAPLModule 클래스 내에 모델(Algorithm 1에서 준비된 QLoRA 모델), 데이터 처리 로직, 학습 스텝, 검증 스텝, 그리고 옵티마이저 구성을 정의한다.

<Table 3> Procedure of fine-tuning LLaVA-1.6 model based on Pytorch Lightning

구성 요소	설명
__init__ 메서드	QLoRA가 적용된 모델, 토큰화 및 전처리를 위한 프로세서, 학습률(learning_rate), 학습용 프롬프트 템플릿(prompt_template_train) 등을 초기화 매개변수로 받아 멤버 변수로 저장.
training_step 메서드	배치(batch) 단위 학습 데이터에 대해 다음을 수행: 1. 프로세서를 사용해 입력 데이터 전처리 (예: 이미지 로드, 프롬프트 구성 및 토큰화, 어텐션 마스크 및 레이블 생성). 학습 프롬프트는 "USER: <image>₩n{질문}₩nASSISTANT: {정답}</s>" 형식 사용. 2. 전처리된 입력을 모델에 전달하여 순전파(forward pass) 수행 및 손실(loss) 계산. 3. 계산된 손실 값을 반환 (필요시 self.log를 통해 로깅).
configure_optimizers 메서드	torch.optim.AdamW와 같은 옵티마이저를 모델의 학습 가능한 파라미터(QLoRA 적용 시 LoRA 어댑터 파라미터) 및 학습률을 사용하여 설정하고 반환.

_prepare_inputs_for_step는 데이터로더로부터 받은 원시 데이터를 AutoProcessor를 사용하여 모델이 요구하는 최종 입력 텐서(이미지 텐서, 토큰화된 프롬프트, 마스킹된 레이블 등)로 변환하는 역할을 담당한다. 학습 시 사용되는 프롬프트는 “USER: <image>\n{질문}\nASSISTANT: {정답}</s>” 형식이다.

training_step에서는 준비된 입력을 모델에 통과시켜 손실을 계산하고, 이 손실 값은 PyTorch Lightning에 의해 자동으로 역전파 및 옵티마이저 스텝(AdamW, 학습률 1e-5)으로 이어져 LoRA 어댑터의 가중치를 업데이트한다.

L.Trainer[22]는 이렇게 정의된 LightningModule과 데이터로더(내부적으로 HelmetDataset과 collate_fn 사용)를 받아, 지정된 하이퍼파라미터(총 3 에포크, 실질 배치 크기 16을 위한 그래디언트 누적 스텝 16, 혼합 정밀도 학습, 그래디언트 클리핑 등)에 따라 전체 학습 과정을 관리하고 실행한다. 학습이 완료되면, 최적화된 LoRA 어댑터의 가중치와 추론에 필요한 프로세서 설정이 최종적으로 저장된다. 이처럼 PyTorch Lightning은 학습 파이프라인의 복잡성을 줄이고 실험의 재현성과 확장성을 높이는 데 기여한다.

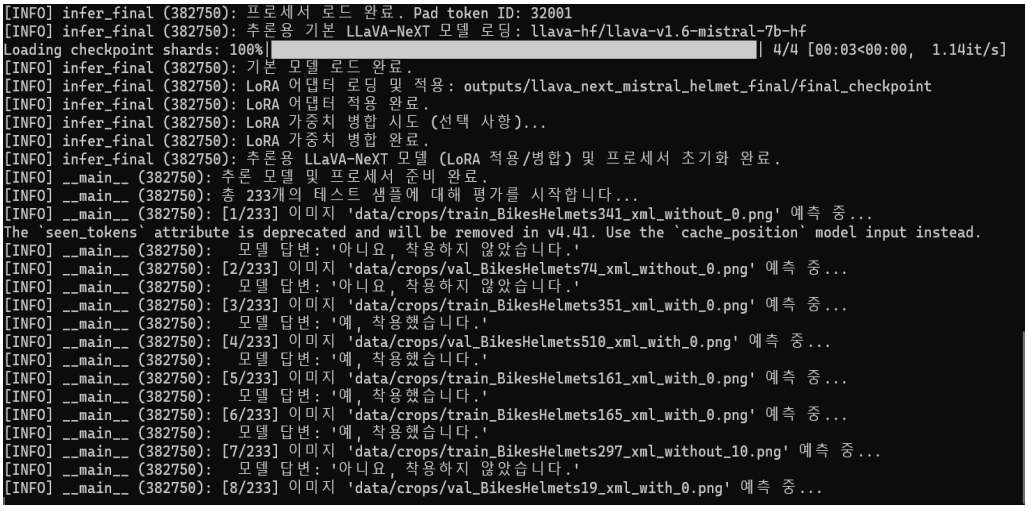
IV. 실험 결과

개발된 시스템의 객관적인 성능을 평가하기 위해, 미세조정을 거친 LLaVA-1.6 모델과 미세조정 전의 순수 LLaVA-1.6 모델 각각에 대해 3.2절에서 구축한 테스트 데이터셋을 사용하여 추론을 수행하고 그 결과를 분석하였다.

4.1. 미세조정된 LLaVA-1.6 모델의 추론 및 결과 해석

미세조정된 LLaVA-1.6 모델의 안전모 착용 여부 판독 성능 평가는 다음과 같은 절차로 진행되었다.

먼저, 3.3절에서 학습시킨 LoRA 어댑터 가중치를 불러와 기본 LLaVA-1.6 모델에 적용하였다. 이 과정을 통해 미세조정으로 학습된 특정 작업 수행 능력이 주입된 최종 추론용 모델이 준비된다. 추론 시에는 불필요한 그래디언트 계산을 비활성화하기 위해 모델을 평가 모드로 설정하였다.



<Figure 6> Testing fine-tuned LLaVA 1.6

테스트 데이터셋의 각 이미지에 대해, “이 오토바이 주행자가 안전모를 착용했습니까?”라는 표준화된 자연어 질문을 함께 모델에 제시하였다. 이 이미지와 텍스트 질문은 LLaVA-1.6 (Mistral-7B) 모델의 고유한 입력 형식인 “<s>[INST] <image>\n{질문} [/INST]” 프롬프트 템플릿에 맞춰 구성되었다. 여기서 <image> 부분은 AutoProcessor에 의해 실제 이미지의 시각적 특징으로 대체된다. 이렇게 구성된 입력은 모델의 generate 함수를 통해 처리되어, 안전모 착용 여부에 대한 텍스트 답변을 생성한다. 답변 생성 시 최대 토큰 길이를 32토큰으로 제한하여 간결하고 명확한 답변을 유도하였다.

미세조정된 모델은 학습 데이터에서 정의된 대로 “예, 착용했습니다.” 또는 “아니요, 착용하지 않았습니다.”와 유사한 스타일의 답변을 생성하도록 학습되었다. 따라서, 모델이 생성한 텍스트 답변에서 핵심 키워드인 “예”의 포함 여부를 기준으로 판독 결과를 이진화하였다. 즉, 답변에 “예”가 포함되어 있으면 해당 주행자가 안전모를 “착용(1)”한 것으로, 그렇지 않으면 “미착용(0)”으로 분류하였다. 이 이진 분류 결과가 모델의 최종 예측값으로 사용되어 테스트 데이터셋의 실제 정답 레이블(“yes”/“no”)과 비교된다.

4.2. 순수 LLaVA-1.6 모델의 추론 및 결과 해석

미세조정의 효과를 명확히 비교하기 위해, 어떠한 추가 학습도 거치지 않은 순수 LLaVA-1.6 모델에 대해서도 동일한 테스트 데이터셋과 질문을 사용하여 평가를 진행하였다. 순수 LLaVA-1.6 모델을 로드한 후, 미세조정된 모델과 동일한 방식으로 각 테스트 이미지와 표준 질문을 입력으로 제공하여 텍스트 답변을 생성하도록 하였다.

순수 모델은 특정 작업이나 답변 형식에 대해 사전 학습된 바 없으므로, 생성하는 답변이 매우 다양하고 때로는 모호하거나 질문의 의도와 직접적으로 관련 없는 내용을 포함할 수 있다. 예를 들어, “이미지에는 오토바이를 탄 사람이 보입니다.”와 같이 단순히 이미지를 묘사하거나, “안전모 착용 여부는 명확하지 않습니다.”처럼 판단을 유보하는 답변을 생성할 수 있다. 이러한 자유 형식의 답변을 미세조정된 모델의 “예/아니요” 스타일 답변과 공정하게 비교하기 위해서는 일관된 기준에 따라 그 의미를 해석하고 분류하는 과정이 필수적이다. 따라서 순수 모델이 생성한 다양한 답변들을 “착용”, “미착용”, “확인 불가”의 세 가지 범주로 일관되게 분류하기 위해, 키워드 기반의 휴리스틱 규칙을 적용하였다. 이 규칙은 다음과 같은 분류 기준을 따른다.

<Table 4> Rule-based labeling for answer texts from Vanilla LLaVA-1.6

판독 라벨링	답변 텍스트
긍정(Positive)	예, 착용했습니다, 쓰고 있습니다 등
부정(Negative)	아니요, 미착용, 착용하지 않았습니다, 없습니다 등
확인 불가	흐릿, 모호, 불분명, 확인하기 어렵습니다 등의 키워드가 발견된 경우 + '긍정', '부정'에 해당하는 키워드가 발견되지 않은 경우

부정 판독: 답변 텍스트에 "아니요", "미착용", "착용하지 않았습니다", "없습니다" 등과 같이 사전에 정의된 '부정 키워드' 목록에 포함된 단어가 하나라도 나타나면, 해당 답변은 "미착용"으로 분류된다.

긍정 판독: 부정 키워드가 발견되지 않은 상태에서, "예", "착용했습니다", "쓰고 있습니다" 등과 같이 사전에 정의된 '긍정 키워드' 목록에 포함된 단어가 발견되면, 해당 답변은 "착용"으로 분류된다. 단, 이때 "확인하기 어렵습니다", "불분명합니다", "흐릿합니다" 등 '불확실/모호성 키워드'가 동시에 발견될 경우에는 "긍정"으로 분류하지 않고 아래의 "확인 불가"로 처리한다.

확인 불가 판독: 위의 두 조건에 해당하지 않거나, 명시적으로 '불확실/모호성 키워드'가 발견된 경우, 해당 답변은 "확인 불가"로 분류된다. 이러한 분류 로직을 통해 순수 모델의 다양한 답변을 정량적 평가가 가능한 형태로 변환하였다. "확인 불가"로 분류된 답변은 전체 정확도 계산 시 제외하거나, 특정 시나리오에서는 "미착용"으로 간주하여 분석에 포함될 수 있다.

4.3. 평가 지표

위의 절차를 통해 얻어진 각 모델(미세조정된 모델, 순수 모델)의 이진 예측 결과("착용" 또는 "미착용")와 테스트 데이터셋의 실제 정답 레이블을 비교하여 성능을 정량적으로 평가하였다. 주요 평가지표로는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), 그리고 F1-점수(F1-score)가 사용되었다. 이 지표들은 sklearn.metrics 라이브러리의 classification_report 및 confusion_matrix 함수를 통해 계산되었다. 특히, 오차 행렬 분석을 통해 참 긍정(True Positive, TP), 거짓 긍정(False Positive, FP), 참 부정(True Negative, TN), 거짓 부정(False Negative, FN)의 구체적인 수를 파악하고, 이를 바탕으로 모델이 '착용' 클래스와 '미착용' 클래스 각각에 대해 어떤 판독 특성을 보이는지, 그리고 어떤 유형의 오류를 주로 범하는지를 심층적으로 분석하였다.

4.4. 실험 결과 및 분석

본 연구의 모든 실험은 Python 3.9 버전 기반의 Conda 가상환경에서 수행되었다. 주요 소프트웨어 라이브러리 및 버전은 PyTorch 2.1.0 (CUDA 12.1 지원), transformers 4.38.2, peft 0.11.1, bitsandbytes 0.43.1, pytorch-lightning 2.3.0을 사용하였다. 모델 학습 및 추론에는 NVIDIA A100 (80GB VRAM) GPU 1대가 활용되었다.

데이터셋은 III장에서 기술한 바와 같이 AI Hub의 '교통법규 위반 상황 데이터'와 Kaggle의 'Helmet Detection' 데이터셋으로부터 총 1,555개의 크롭 이미지를 생성하여 학습 및 검증에 사용하였고, 이 중 15%에 해당하는 233개 이미지를 무작위로 추출하여 테스트 데이터셋으로 사용하였다. 평가는 이 233개의 테스트 샘플에 대해 이루어졌다.

4.4.1. 정량적 성능 비교 분석

미세조정을 거치지 않은 순수 LLaVA-1.6 모델(Vanilla LLaVA-1.6)과 본 연구에서 제안한 QLoRA 기법으로 미세조정한 LLaVA-1.6 모델(Fine-tuned LLaVA-1.6)의 안전모 착용 여부 판독 성능을 비교하였다. 순수 모델의 경우, 자유 형식의 답변을 생성하므로 Table 4에서 기술한 답변 분류 로직을 적용하여 “긍정”, “부정”, “확인 불가”로 분류하였으며, “확인 불가”로 판별된 답변(테스트셋 233개 중 80건, 약 34.3%)을 제외한 나머지 샘플에 대해 이진 분류 성능을 측정하거나, 보다 보수적인 관점에서 “확인 불가” 답변을 “미착용”으로 간주하여 정확도를 계산하였다. 사용자 소논문 초안에서 언급된 순수 모델의 약 88% 정확도는 이러한 보수적 기준 또는 답변 형식 제한 프롬프팅을 적용했을 때의 수치이다.

<Table 5> Helmet wearing detection performance metrics by model

평가 지표	순수 LLaVA-1.6	파인튜닝 LLaVA-1.6
전체 정확도	88%	95%
미착용 F1	0.86	0.94
착용 F1	0.90	0.96
미착용 Precision	0.77	0.96
미착용 Recall	0.97	0.91

Table 5에서 볼 수 있듯이, QLoRA를 통해 미세조정된 LLaVA-1.6 모델은 전체 정확도에서 약 95.3%를 달성하여, 순수 모델의 약 88% 대비 7%p 이상 향상된 성능을 보였다. 이는 특정 작업에 대한 미세조정의 효과를 명확히 보여준다. F1-점수 또한 ‘미착용’ 클래스에서 0.86에서 0.936으로, ‘착용’ 클래스에서 0.90에서 0.963으로 모두 유의미하게 상승하여, 두 클래스에 대한 전반적인 판독 능력이 균형 있게 개선되었음을 나타낸다. 특히, ‘미착용’ 클래스에 대한 정밀도(Precision)가 0.77에서 0.964로 크게 향상된 점이 주목할 만하다. 이는 모델이 ‘미착용’으로 예측했을 때 실제 미착용일 확률이 높아졌음을 의미하며, 오판으로 인한 불필요한 확인 작업을 줄이는 데 기여할 수 있다. 반면, ‘미착용’ 클래스의 재현율(Recall)은 소폭 감소하였는데, 이는 파인튜닝된 모델이 실제 미착용 사례 중 일부를 착용으로 잘못 판단하는 경우가 순수 모델보다 다소 늘어났음을 시사할 수 있으나, 정밀도의 큰 향상과 전체적인 F1-점수 상승을 고려할 때 모델의 전반적인 판독 신뢰도는 향상된 것으로 판단된다.

파인튜닝된 모델의 구체적인 오분류 양상을 파악하기 위해 Table 6과 같이 오차 행렬(Confusion Matrix)을 분석하였다. 테스트 샘플 총 233개에 대한 결과이다. 총 233개의 테스트 샘플 중 모델은 222개(TN 80 + TP 142)를 정확히 예측하여 95.28% ((80+142)/233)의 전체 정확도를 보였다.

<Table 6> Detailed prediction results analysis for the fine-tuned model

평가 지표	Predicted No	Predicted Yes
Actual No	80 (TN)	8 (FP)
Actual Yes	3 (FN)	142 (TP)

안전 규정 준수 판독에서 특히 중요한 오류는 실제 안전모를 착용했음에도 불구하고 미착용으로 잘못 판단하는 False Negative (FN) 오류이다. 파인튜닝된 모델의 경우, 이러한 FN 오류

는 단 3건에 불과하여 '착용' 클래스에 대한 재현율이 97.9% (142/145)로 매우 높게 나타났다. 이는 사용자 소논문 초안에서 언급된 순수 모델의 FN 25건(개선된 판독 로직 적용 및 특정 해석 기준을 적용한 경우)과 비교했을 때 획기적인 감소이며, 실제 착용자를 놓치지 않는 능력이 크게 향상되었음을 의미한다.

'미착용' 클래스에 대한 재현율은 90.9% (80/88)였으며, 이는 실제 미착용자 88명 중 80명을 정확히 '미착용'으로 판독했음을 나타낸다. False Positive (FP) 오류, 즉 미착용자를 착용으로 잘못 판단한 경우는 8건으로 나타났다. '미착용' 클래스에 대한 정밀도는 96.4% (80/83)로, 모델이 '미착용'이라고 예측한 경우 중 실제 미착용일 확률이 매우 높음을 알 수 있다. '착용' 클래스에 대한 정밀도는 94.7% (142/150)였다.

이러한 정량적 분석 결과는 QLoRA 기반 미세조정이 LLaVA-1.6 모델의 안전모 착용 여부 판독 성능을 전반적으로 향상시켰으며, 특히 실제 단속 업무에서 중요한 오탐(FP) 및 미탐(FN) 비율을 효과적으로 관리할 수 있는 수준으로 개선했음을 보여준다.

4.4.2. 정성적 성능 비교 분석

순수 LLaVA-1.6 모델의 답변 특성을 살펴보면, 파인튜닝을 거치지 않은 순수 LLaVA-1.6 모델은 "이 오토바이 주행자가 안전모를 착용했습니까?"라는 동일한 질문에 대해 일관되지 않고 다양한 형태의 답변을 생성하는 경향을 보였다. 테스트 샘플 233개 중 약 34.3%에 해당하는 80건의 답변에서 "이미지가 흐릿하여 확인하기 어렵습니다.", "제공된 이미지로는 안전모 착용 여부를 명확히 판단할 수 없습니다.", "정보가 부족하여 답변할 수 없습니다."와 같이 명시적으로 "확인 불가" 또는 모호함을 표현하였다. 이는 모델이 특정 작업에 대한 전문 지식이나 명확한 답변 형식에 대한 학습이 이루어지지 않았을 때 나타나는 전형적인 특성이다. 답변 형식을 "예 또는 아니오로만 답해주세요"와 같이 프롬프트를 통해 제한하더라도, 이러한 모호한 답변의 비율이 크게 줄어들지 않거나, "예." 또는 "아니요."로 답변하더라도 그 근거가 부족해 보이는 경우가 많았다.

<Table 7> Model response analysis for identical input images: Not fine-tuned vs. fine-tuned

파인튜닝되지 않은 LLaVA-1.6 모델의 응답
The `seen_tokens` attribute is deprecated and will be removed in v4.41. Use the `cache_position` model input instead. [INFO] __main__ (486706): 모델이 'data/crops/train_BikesHelmets341_xml_without_0.png'에 대해 '확인 불가'로 답변 (원본 답변: '이 사진에서 주행자가 안전모를 착용하고 있는지 확인하기 어렵습니다. 주행자의 얼굴이 이미지에서 잘 보이지 않으며, 안전모가 얼굴 앞에 있는 것은 눈에 보이지 않습니다. 따라서 이 사진에서 주행자가 안'). 평가지표 계산에서는 제외하거나 특정 값으로 편향시킬 수 있음. 현재는 제외. [INFO] __main__ (486706): 모델이 'data/crops/val_BikesHelmets74_xml_without_0.png'에 대해 '확인 불가'로 답변 (원본 답변: '이 사진에서 주행자가 안전모를 착용하고 있는지 확인하기 어렵습니다. 주행자의 얼굴이 이미지에서 잘 보이지 않으며, 안전모가 눈앞에 있는 것은 주행자가 안전모를 착용하고 있는 것을 의미할 수 있지만, 이'). 평가지표 계산에서는 제외하거나 특정 값으로 편향시킬 수 있음. 현재는 제외.
파인튜닝된 LLaVA-1.6 모델의 응답
The `seen_tokens` attribute is deprecated and will be removed in v4.41. Use the `cache_position` model input instead. [INFO] __main__ (486823): 모델 답변: '아니요, 착용하지 않았습니다.' [INFO] __main__ (486823): [2/233] 이미지 'data/crops/val_BikesHelmets74_xml_without_0.png' 예측 중... [INFO] __main__ (486823): 모델 답변: '아니요, 착용하지 않았습니다.'

반면, 본 연구에서 제안한 QLoRA 기법을 사용하여 특정 작업에 맞게 파인튜닝된 LLaVA-1.6 모델은 순수 모델과 확연히 다른 답변 양상을 보였다. 파인튜닝된 모델은 테스트셋 전체에 걸쳐 "확인하기 어려움"과 같은 모호하거나 회피적인 답변을 거의 생성하지 않았으며, 학습 데이터에서 정의한 대로 "예, 착용했습니다." 또는 "아니요, 착용하지 않았습니다."의 두

가지 형태로 매우 일관되고 명확한 답변을 안정적으로 생성하였다. 이는 미세조정을 통해 모델이 특정 질문의 의도를 정확히 파악하고, 요구되는 답변 형식에 맞춰 응답하도록 효과적으로 학습되었음을 시사한다.

4.5. 파인튜닝 효과 고찰

본 연구의 실험 결과는 LLaVA-1.6 모델에 대한 특정 작업 중심의 QLoRA 미세조정이 안전모 착용 여부 판독 성능에 미치는 긍정적이고 결정적인 영향을 명확히 보여준다. 파인튜닝의 효과는 다음과 같은 측면에서 두드러지게 나타났다.

첫째로, 작업 특화성 및 정확도 향상이다. 순수 LLaVA-1.6 모델은 광범위한 일반 지식을 학습했지만, '오토바이 주행자의 안전모 착용 여부 판독'이라는 매우 구체적이고 세밀한 작업에 대해서는 최적의 성능을 발휘하지 못했다. 반면, 특정 작업 데이터셋(크롭된 주행자 이미지와 "착용/미착용" 답변 쌍)으로 미세조정함으로써, 모델은 안전모 착용과 관련된 주요 시각적 특징(예: 안전모의 형태, 머리와의 결합 상태, 턱끈 유무 등)을 보다 정교하게 학습하고 이를 질문과 연관 지어 정확한 판단을 내릴 수 있게 되었다. 이는 전체 정확도가 88%에서 95.3%로 향상된 결과로 입증된다.

둘째로, 답변의 일관성 및 명확성을 확보했다는 점이다. 파인튜닝 전 모델은 다양한 형태로 답변하거나 판단을 유보하는 경향이 강했던 반면, 파인튜닝 후 모델은 학습된 답변 형식("예, 착용했습니다." / "아니요, 착용하지 않았습니다.")을 일관되게 따르며 명확한 답변을 생성하였다. 이는 모델의 답변 불확실성을 크게 줄이고, 후처리 과정 없이 결과를 바로 활용할 수 있게 만들어 시스템의 실용성을 높인다.

셋째로, 치명적 오류 감소 및 신뢰도 증진의 효과를 달성하였다. 특히 안전 규정 관련 시스템에서 중요한 False Negative(실제 착용자를 미착용으로 오판) 오류가 파인튜닝을 통해 획기적으로 감소(순수 모델 대비 약 88% 감소, 25건→3건)하였고, 이로 인해 '착용' 클래스에 대한 재현율이 97.9%에 도달하였다. 이는 시스템의 판독 결과를 신뢰하고 실제 단속 업무에 참고 자료로 활용할 수 있는 가능성을 높인다.

넷째로, QLoRA를 통한 효율적 학습을 수행하였으며 이러한 성능 개선이 전체 모델 파라미터의 극히 일부(약 0.29%)만을 학습하는 QLoRA 기법을 통해 달성되었다는 점은 매우 고무적이다. 이는 대규모 VLM을 특정 산업 현장이나 제한된 자원 환경에 적용하고자 할 때, QLoRA가 효과적이고 실용적인 미세조정 전략이 될 수 있음을 시사한다.

V. 결론

본 연구의 가장 큰 의의는 특정 교통안전 규정(안전모 착용) 준수 여부를 판독하는 세밀하고 특화된 작업에 VLM을 효과적으로 적용하고, 그 실용적 가능성을 입증했다는 점이다. 미세조정된 LLaVA-1.6 모델이 달성한 95.28%의 높은 정확도는 AI 기술이 교통 단속 업무의 효율성을 증대시키고, 수동 단속의 인력 및 시간적 한계를 극복하는 데 기여할 수 있음을 시사한다. 이는 궁극적으로 안전모 착용률을 높여 교통 사고 시 중상 및 사망자 감소에 긍정적인 영향을 미칠 수 있을 것으로 기대된다.

또한, 본 연구는 기존의 객체 탐지 기반 자동 감지 시스템과 차별화되는 VLM 접근 방식의 장점을 명확히 보여주었다. 단순히 '안전모'라는 객체의 존재 유무를 탐지하는 것을 넘어, "이 오토바이 주행자가 안전모를 착용했습니까?"라는 자연어 질문에 대해 "예, 착용했습니다." 또는

“아니요, 착용하지 않았습니다.”와 같이 맥락을 이해하고 명확한 답변을 생성하는 능력은 VLM만이 제공할 수 있는 독특한 가치이다. 이러한 질의응답 방식은 단속 결과의 직관성을 높이고, 향후 다양한 형태의 세부 질의(예: “안전모의 종류는 무엇입니까?”, “턱끈이 올바르게 착용되었습니까?”)로 확장될 수 있는 유연성을 제공한다.

마지막으로, QLoRA라는 파라미터 효율적 미세조정 기법을 대규모 VLM에 성공적으로 적용하여 자원 효율성과 높은 성능을 동시에 달성했다는 점도 중요한 의의를 갖는다. 이는 복잡하고 큰 모델을 특정 도메인이나 응용 분야에 맞게 최적화하고자 할 때, 계산 자원의 제약 없이도 효과적인 맞춤형 AI 솔루션 개발이 가능함을 보여주는 사례이다.

그러나 본 연구는 다음과 같은 몇 가지 한계점 또한 내포하고 있다. 첫째, 입력 이미지의 의존성이다. 현재 시스템은 사전에 오토바이 주행자 영역이 크롭된 이미지를 입력으로 가정한다. 따라서 실제 도로 환경에 적용되기 위해서는 고성능의 실시간 주행자 탐지 및 크롭 모듈이 선행되어야 하며, 이 선행 모듈의 성능이 전체 파이프라인의 성공을 좌우하는 병목 현상을 일으킬 수 있다.

둘째, 데이터셋의 규모 및 다양성 한계이다. 본 연구에서 사용된 데이터셋은 공개된 데이터를 기반으로 구축되었으나, 실제 도로 환경에서 마주할 수 있는 극한의 조명 변화(역광, 야간 등), 악천후(우천, 안개 등), 심한 가려짐(occlusion), 매우 다양한 종류와 형태의 안전모, 혹은 비정형적인 착용 상태 등에 대한 모델의 강인성은 추가적인 검증과 개선이 필요하다.

셋째, 판단 범위의 제한성이다. “예/아니요”라는 명확한 답변은 장점이지만, “턱끈이 올바르게 조여졌는가?”와 같이 더 복합적인 안전 상태를 판단하지 못하는 한계로 작용한다. 향후 연구에서는 이를 확장하여, 멀티 라벨 분류나 구조화된 답변 생성을 통해 더 세밀한 수준의 안전 규정 준수 여부를 판독하는 방향으로 나아갈 수 있을 것이다.

이와 같은 한계에도 불구하고, 본 연구는 대규모 비전-언어 모델(VLM)에 QLoRA 기반의 파인튜닝 기법을 적용하여 실제 교통안전 단속 업무에 활용 가능한 수준의 정확도와 실용성을 달성했다는 점에서 중요한 시사점을 제공한다. 향후에는 보다 다양한 실환경 데이터를 반영하고, 주행자 검출 및 전처리 모듈과의 통합을 통해 전체 시스템의 완성도를 높이는 방향으로 연구를 확장할 수 있을 것이다. 본 연구는 VLM 기반 교통안전 단속 기술의 가능성을 제시한 첫걸음으로서, 향후 유사한 과제들에 대한 기술적 발전과 정책적 논의에 기초 자료로 활용될 수 있기를 기대한다.

참고문헌 (References)

[1] WHO (World Health Organization). 2006. Helmets: a road safety manual for decision-makers and practitioners. WHO.

[2] WHO (World Health Organization). 2018. Global status report on road safety 2018. pp. 24-29. WHO.

[3] Redmon J, Divvala S, Girshick R, et al. 2016. You only look once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[4] Bergstrom T, Shi H. 2020. Human-object interaction detection: A quick survey and examination of methods. HuMA'20: Proceedings of the 1st International Workshop on Human-centric Multimedia Analysis, pp. 63-71. https://doi.org/10.1145/3422852.3423481

[5] Li Z, Wu X, Du H, et al. 2025. A survey of state of the art large vision language models: alignment, benchmark, evaluations and challenges. arXiv:2501.02189. https://doi.org/10.48550/arXiv.2501.02189

[6] Dinh QM, Ho MK, Dang AQ, et al. 2024. TrafficVLM: A controllable visual language model for traffic video captioning. arXiv:2404.09275. https://doi.org/10.48550/arXiv.2404.09275

[7] Jiang AQ, Sablayrolles A, Mensch A, et al. 2023. Mistral 7B. arXiv:2310.06825. https://doi.org/10.48550/arXiv.2310.06825

[8] Dettmers T, Pagnoni A, Holtzman A, et al. 2023. QLoRA: Efficient finetuning of quantized LLMs. arXiv:2305.14314. https://doi.org/10.48550/arXiv.2305.14314

[9] Li Z, Wu X, Du H, et al. 2025. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. arXiv:2501.02189. https://doi.org/10.48550/arXiv.2501.02189

[10] Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, pp. 5998-6008.

[11] Radford A, Kim JW, Hallacy C, et al. 2021. Learning transferable visual models from natural language supervision. 38th International Conference on Machine Learning (ICML), 139, pp. 8748-8763.

[12] Alayrac JB, Donahue J, Luc P, et al. 2022. Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems 35 (NeurIPS 2022), pp. 23716-23736.

[13] Liu H, Li C, Wu Q, et al. 2023. Visual instruction tuning. Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

[14] Liu H, Li C, Li Y, et al. 2024. Improved baselines with visual instruction tuning. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26296-26306.

[15] Liu H, Li C, Li Y, et al. 2023. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Blog. Available at: https://llava-vl.github.io/blog/2024-01-30-llava-next/ accessed on 2024. 01. 30

[16] Lei S, Hua Y, Zhihao S. 2025. Revisiting fine-tuning: A survey of parameter-efficient techniques for large AI models. Preprints 2025040743.v1. https://www.preprints.org/manuscript/202504.0743/v1

[17] Hu E, Shen Y, Wallis P, et al. 2022. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

[18] Ren S, He K, Girshick R, et al. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems 28 (NIPS 2015), pp. 91-99.

[19] BlacKHat. 2018. yolov3-Helmet-Detection. GitHub. Available at: https://github.com/BlcaKHat/yolov3-Helmet-Detection accessed on 2025. 6. 10.

[20] AI-Hub. n.d. Traffic violation incident data [교통법규 위반 상황 데이터]. Korean National Police

Agency, Seoul, Korea. Available at:
https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71555 accessed on 2025. 6. 10.

[21] Larxel. 2020. Helmet detection. Kaggle. Available at:
https://www.kaggle.com/datasets/andrewmvd/helmet-detection accessed on 2025. 6. 10.

[22] Lightning AI. 2019. PyTorch Lightning. GitHub. Available:
https://github.com/Lightning-AI/pytorch-lightning accessed on 2025. 6. 10.