YOLO-ViT를 이용한 차량 외관 손상 탐지 및 분류 모델

하민수, 유철우*
명지대학교
ms054103@mju.ac.kr, *cwyou@mju.ac.kr

A YOLO-ViT-Based Model for Vehicle Exterior Damage Detection and Classification

Minsu Ha, Cheolwoo You*
Myongji Univ.

요 약

본 논문에서는 YOLOv8m과 ViT 모델을 이용하여 차량 외관 손상 탐지 및 분류 모델을 구현하고 성능을 분석한다. 손상 탐지는 YOLOv8m 모델을 통해 객체를 탐지하고, 손상 분류는 ViT 모델을 이용한다. 1단계에서는 YOLOv8m 모델을 이용하여 높은 재현율(Recall)을 목표로 하여 손상을 놓치는 경우를 줄이는 데 집중한다. 2단계에서는 ViT 모델을 이용하여 1단계에서 손상이라고 판단한 객체들을 더욱 정확하게 분류한다. AI Hub의 차량 파손 이미지 데이터셋을 이용하여 학습하였으며, 실험 결과들은 일관되고 명확한 기준을 가진 손상 탐지 및 분류 시스템이 구현되었음을 보여준다.

Ⅰ. 서론

최근 중고차와 렌터카 시장의 규모가 확대됨에 따라 차량 관리의 중요성이 커지고 있다. 그중에서도 차량의 가치와 가장 직관적으로 연결된 외관의 손상 확인과 판단은 매우 핵심적인 요소이다. 이를 관리자가 육안으로 검사하는 것은 많은 시간이 소요될 뿐만 아니라 검사자의 주관에 따라 판단 기준이 달라질 수 있다는 한계가 있다. 이에 본 논문에서는 객체 탐지에 특화된 YOLO 모델[1]과 이미지 분류 성능이 뛰어난 ViT 모델[2]을 결합한 2단계 손상 탐지 시스템을 제안한다. 그림 1과 같이, 1단계에서 의심 영역 검출하고 2단계의 전처리 된 이미지를 이용해 정밀 분류 단계를 통해 기존 육안 검사의 한계를 극복하고 진단 정확도를 높이고자 한다.

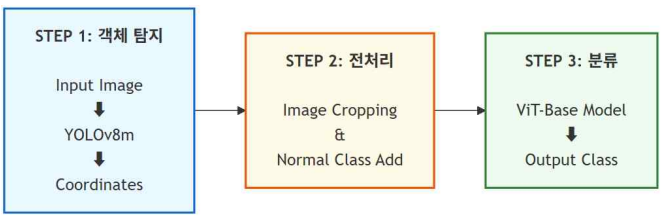

그림 1. YOLO-ViT 모델 구성도

Ⅱ. 본론

본 연구에서는 학습을 위해 AI Hub의 차량 파손 이미지 데이터셋을 활용했다. 해당 데이터셋은 스크래치(Scratched), 찌그러짐(Crushed), 파손(Breakage), 이격(Separated)의 4가지 클래스로 구성되어 있다.

객체 탐지 모델은 실시간 서비스 구현을 고려하여 빠른 속도를 장점으로 가지는 YOLOv8(You Only Look Once)을 선택했다. YOLOv8의 모델 크기와 이미지 해상도에 따른 비교 실험을 통해 세부적인 모델을 결정하였다. [표1]은 다양한 비교 실험 결과이며, Recall 값과 학습 소요 시간을 고려하여 최종적으로 YOLOv8m 모델에 640의 이미지 해상도를 선택하였다.

분류 모델의 후보로 CNN(ResNet, EfficientNet) 과 Transformer(ViT)이 존재한다. [표2]는 분류 모델 선택을 위해 동일한 조건에서 학습한 결과인데, 정확도가 가장 높았던 ViT를 분류 모델로 정해 연구를 진행했다.

표 1. YOLOv8의 모델크기와 이미지 해상도 성능지표

Model	Recall	Training Time (sec/epoch)
v8n, 640	0.3064	70
v8m, 640	0.3332	180
v8n, 1024	0.3169	140
v8m, 1024	0.3337	420

표 2. 분류 모델의 성능 지표

	ViT	ResNet-50	ResNet-101	ResNet-152	Efficient Net
Accuracy	0.6400	0.5625	0.6025	0.5000	0.6075

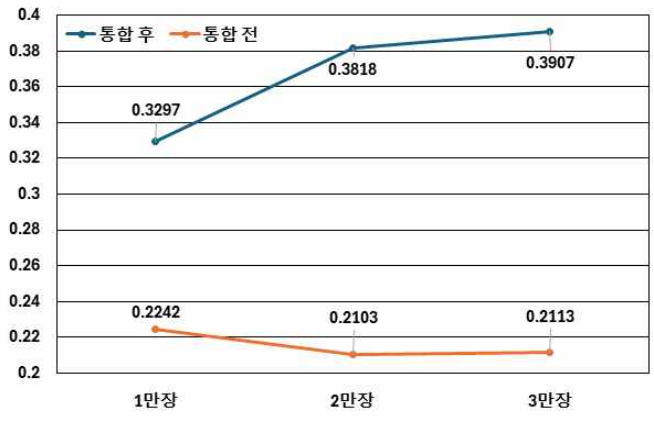

그림 2. 클래스 통합 전/후 Recall 그래프

첫 번째 단계인 YOLOv8m의 학습 목표는 높은 재현율(Recall)을 확보하는 것이다. 이를 위해 실제 손상을 놓치는 미탐(False Negative)을 줄이는 데 집중했다. [그림2]는 관련 실험 결과인데, 클래스가 나눠진 상태로 객체 탐지를 진행했을 때, 객체 탐지 성능 저하 현상이 발생했다. 이를 해결하기 위해 4개의 클래스를 단일 클래스인 '손상(Damage)'으로 통합

표 3. 클래스별 비율(단위 %)

Breakage	Crushed	Normal	Scratched	Separated
8.6	10	5.8	61.7	13.9

표 4. 클래스 불균형 해소 방식 별 성능 지표

	기본	Focal Loss	Resampling	Focal Loss + Resampling
Accuracy	0.7305	0.7465	0.5868	0.5589
Macro F1-score	0.5905	0.5670	0.4989	0.4595

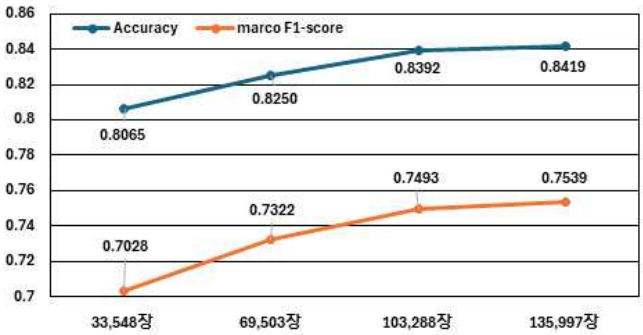

그림 3. 학습 데이터 별 성능 그래프

하여 학습을 진행했다. 이를 통해 분류를 배제한 객체 탐지 성능향상이 있었고, 그 결과 Recall 값이 약 1.5~2배 가까이 상승했다.

한편, YOLOv8m에서 나온 좌표를 이용하여 원본 이미지를 크롭한 데이터를 ViT의 Input 데이터로 사용하였다. 하지만, 이 데이터 중에는 먼지나 강한 빛 반사 등으로 인해 정상인데 손상으로 탐지한 데이터들이 존재했다. 이를 정상(Normal) 클래스로 추가하여 총 5개의 클래스로 학습을 진행했다. 이를 통해 손상의 분류뿐만 아니라 YOLOv8m의 오탐을 필터링하여 높은 정확도를 얻는 데 집중하였다.

[표3]을 보면 Scratched 클래스로 인해 데이터 불균형이 존재하는 것을 확인할 수 있다. 이를 해소하기 위해 Focal Loss 손실함수[3]와 Scratched의 수를 줄이는 Resampling 방식을 이용하여 비교 실험을 진행하였다. [표4]를 보면 기본 Fine-Tuning 방식의 성능이 가장 좋음을 확인할 수 있다. Focal Loss 손실함수를 사용하기엔 소수 클래스의 수가 학습에 충분했고, Resampling으로 인해 발생하는 정보 손실이 발생하여 성능이 하락한 것으로 분석된다.

데이터의 불균형으로 인해 정확도의 역설이 발생하는 상황을 방지하는 성능 지표가 필요했다. 따라서 모든 클래스를 동등한 비중으로 평균을 내는 Macro F1-score[4]를 성능 지표로 사용하여 모델의 성능을 정확하게 판단하였다. [그림 3]을 보면 학습 데이터의 양이 늘어날수록 성능이 향상하는 모습을 보였지만, 성능의 향상 폭이 점차 감소하여 추가로 학습 데이터의 양을 늘리더라도 성능이 특정 지점에 이르면 수렴할 것으로 판단하여 본 프로젝트는 학습 데이터 13만 장 지점에서 학습을 종료하였다.

[표5]를 보면 많은 수의 데이터가 존재하는 Scratched 클래스가 적은 수의 데이터가 존재하는 Breakage, Crushed, Normal 클래스보다 높은 결과를 얻은 것을 확인하였다.

이렇게 준비된 YOLOv8m와 ViT를 연결하여 최종적인 모델을 실행하면 손상으로 탐지한 영역에 bounding box로 시각화하며 해당 손상의 클래스 및 softmax 기반의 확률값을 함께 출력한다. [그림4]는 YOLO-ViT 모델의 시각화 결과이다.

표 5. 최종 성능 지표(135,997장 학습)

	Precision	Recall	F1 - Score
Breakage	0.7457	0.6204	0.6773
Crushed	0.6480	0.6773	0.6623
Normal	0.7675	0.6452	0.7011
Scratched	0.8976	0.9202	0.9088
Separated	0.8094	0.8307	0.8199

그림4. YOLO-ViT 모델 결과 예시

Ⅲ. 결론

본 연구에서는 YOLOv8m와 ViT 모델을 이용해 빠른 속도와 높은 정확도를 가진 차량 외관 손상 탐지 모델을 구현하였다. YOLOv8m의 Confidence Threshold를 낮게 설정하여 최대한 많은 의심 영역을 검출하고, 2단계 ViT에서 정밀하게 필터링하는 구조를 통해 전체 시스템의 정확도를 확보했다. 데이터의 수가 부족했던 Breakage, Crushed, Normal 클래스의 경우 추가적인 데이터를 확보하여 학습한다면 Scratched, Separated 클래스와 동등한 수준의 성능 향상이 기대된다. 본 연구에서 제안한 모델은 향후 렌터카 및 중고차 시장의 자동화된 차량 상태 점검 시스템 등에 효과적으로 활용될 수 있을 것이다.

ACKNOWLEDGMENT

본 과제(결과물)는 교육부와 경기도의 재원으로 지원을 받아 수행된 경기 지역혁신중심 대학지원사업(경기RISE사업)의 연구결과임(2025-RISE-09-A15). 또한, 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NO. RS-2024-00335012).

참 고 문 헌

[1] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[2] Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR.

[3] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[4] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.