

●○ 교통안전 과제

자동차 차종/연식/번호판 인식용 데이터



●○ 개요: 차량 인식 데이터셋이란?



그림1 | 차량 인식 데이터셋 구축 개요

차량 인식 데이터셋은 동영상이나 이미지에서 차량의 차종, 연식(세대별), 번호판 식별 등 차량 인식용 인공지능 개발에 활용할 수 있는 학습 데이터셋이다. 2020년 답핑소스(주)에서 구축했으며, 실제 500시간 이상의 CCTV 동영상에서 추출한 차량 50만장, 번호판 10만장의 이미지로 구성되어 있다.

차량 인식 기술은 차량 도난 등의 보안, 관제 등을 위한 차량 식별 자동화나 주차장과 주유소 등에서의 입출차 자동화에 적용할 수 있는 기술이다. 특히, 이 데이터셋의 경우 통제된 환경이 아닌 실제 CCTV 영상 기반의 이미지를 학습데이터로 구축하였기에, 실제 상용화에서 더 높은 효율을 가져올 것으로 기대된다.

차량 인식용 학습데이터의 어노테이션은 차량 이미지에서의 차량과 번호판의 바운딩박스 및 차량 특성 정보(메이커 / 모델 / 연식 / 색상) 와 번호판 이미지에서의 번호판 텍스트 정보이다. 이번에 구축한 데이터셋의 경우, 대한민국 국토부에 등록된 차종 중 XX & 이상을 포함하는 것을 목표로 하였다. 이를 통해 영상에서 차량과 번호판의 위치를 특정하고, 차량의 특성, 번호판의 글씨

및 숫자를 확인하는 인공지능 학습에 사용 가능하다.

●○ 데이터셋의 구성

본 데이터셋은 차량 인식용 인공지능 딥러닝 모델 학습에 활용하는 이미지와 이미지의 특성값의 데이터셋 50만건과 번호판 이미지와 해당 텍스트로 구성된 데이터셋 10만 건으로 구성되어 있다.

본 데이터셋은 연구자가 일반적으로 연구를 진행하기에 충분한 양이며, 상용화 레벨에서는 강력한 사전학습모델을 만들 수 있는 양이다.

표1 | 데이터셋 구성

데이터 종류	포함 내용	제공 방식
차량 인식 데이터셋	크롭된 차량이미지와 차종 라벨링 정보가 포함된 JSON파일	차량 크롭이미지+JSON 포맷 파일
번호판 데이터셋	번호판 크롭 이미지와 번호판OCR 라벨링 정보가 포함된 JSON파일	번호판 크롭이미지+JSON 포맷 파일

●○ 데이터셋의 설계 기준과 분포

본 데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터의 밸런스이다. 대한민국 국토부에 등록된 DB 와 한국자동차산업협회 DB 를 참조하여, 인식하고자 하는 차량의 모델을 특정하였다. 현실적으로 실제 CCTV 에서 포착하기 힘든 차종(예: 람보르기니 미우라, 닛지 바이퍼 등)의 경우 학습 데이터셋에서 제외하였다.

실제 환경 (Wild Environment) 에서 수집하기에 발생하는 편향을 방지하기 위해, 수집 후 차종 분포를 확인하여 실제 등록 및 판매 분포를 따르는지 검증하였다.

Class	모집단 분포	구축 데이터 분포	비고																					
색상	<div><div>1 WHITE 28%</div><div>2 BLACK 23%</div><div>3 SILVER 23%</div><div>4 GRAY 9%</div><div>5 PEARL 7%</div><div>6 BLUE 3%</div><div>7 RED 2%</div><div>8 ETC 5%</div></div>	<div>* 데이터셋 구축 후 작성</div>																						
메이커	<div><div>국내 자동차 시장 점유율 변화</div><div><table border="1"><thead><tr><th>Year</th><th>Hyundai</th><th>Kia</th><th>GM</th><th>Ford</th><th>Renault</th><th>Others</th></tr></thead><tbody><tr><td>2017</td><td>35.5%</td><td>3.2%</td><td>29.8%</td><td>6.0%</td><td>5.8%</td><td>13.1%</td></tr><tr><td>2024</td><td>36.0%</td><td>3.6%</td><td>29.8%</td><td>5.7%</td><td>5.4%</td><td>13.8%</td></tr></tbody></table></div></div>	Year	Hyundai	Kia	GM	Ford	Renault	Others	2017	35.5%	3.2%	29.8%	6.0%	5.8%	13.1%	2024	36.0%	3.6%	29.8%	5.7%	5.4%	13.8%	<div>* 데이터셋 구축 후 작성</div>	<div>모델 별 세부 수치는 별도 기재</div>
Year	Hyundai	Kia	GM	Ford	Renault	Others																		
2017	35.5%	3.2%	29.8%	6.0%	5.8%	13.1%																		
2024	36.0%	3.6%	29.8%	5.7%	5.4%	13.8%																		
모델/연식수	<div>n 개</div> <div>브랜드별 세부 통계 별도 기재</div>	<div>* 데이터셋 구축 후 작성</div>																						

●○ 데이터 구조 및 예시

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

표2 | 데이터 구조

분류		차량 데이터셋	번호판 데이터셋
내용		크롭 이미지 차량에 대한 차종/연식 정보 JSON파일	번호판 이미지에 대한 번호판 값 JSON파일
수량		50만장	10만장
항목			
이미지 파일 경로	image_path	Y	Y
고유 키	id	Y	Y
차량 제조사	brand	Y	-
차량 색상	color	Y	-
차량 모델 종류	model	Y	-
차량 모델 연식	year	Y	-
번호판값	value	-	Y

●○ 데이터 예시

```
"imagePath":"SUV/현대자동차/SUV_싼타페-2.jpg"
"attributes":{
  "brand":"현대자동차"
  "color":"흰색"
  "model":"SUV_싼타페"
  "year":"2010-2012"
}
```

●○ 데이터 구축 과정

학습데이터 가공을 담당하고 있는 기관의 클라우드소싱 플랫폼을 통해 크라우드워커를 모집하고, 크라우드워커에게 데이터 어노테이션 환경을 갖추도록 해주는 플랫폼인 Nachos를 통해 데이터 어노테이션 및 데이터셋 구축을 수행하였다.

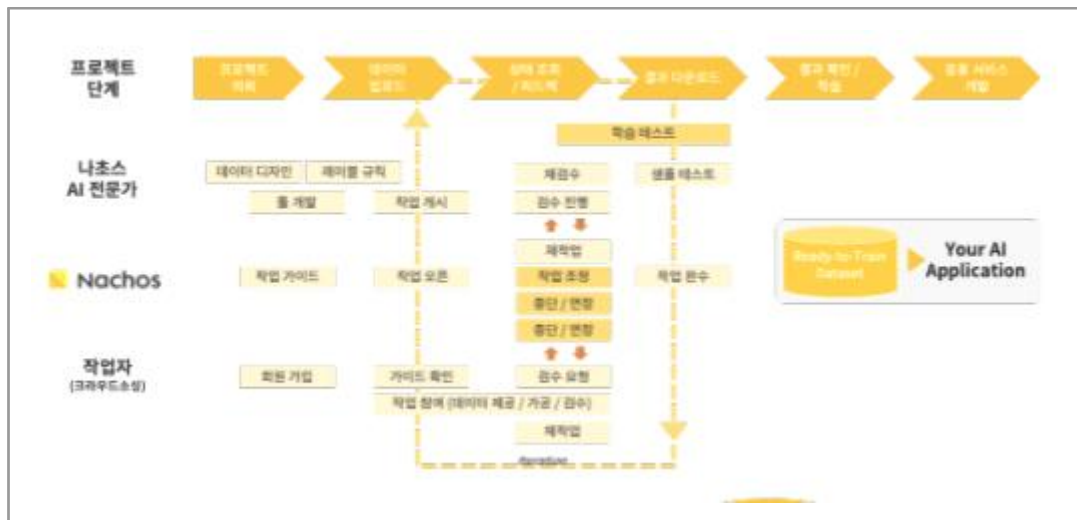


그림4 | 데이터 구축 로세스

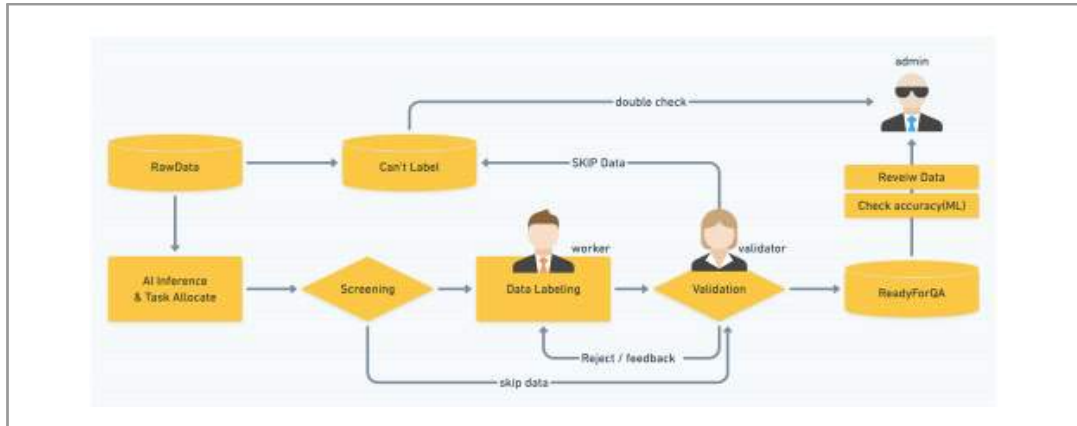


그림5 | 데이터 구축 흐름도

●○ 검수와 품질 확보

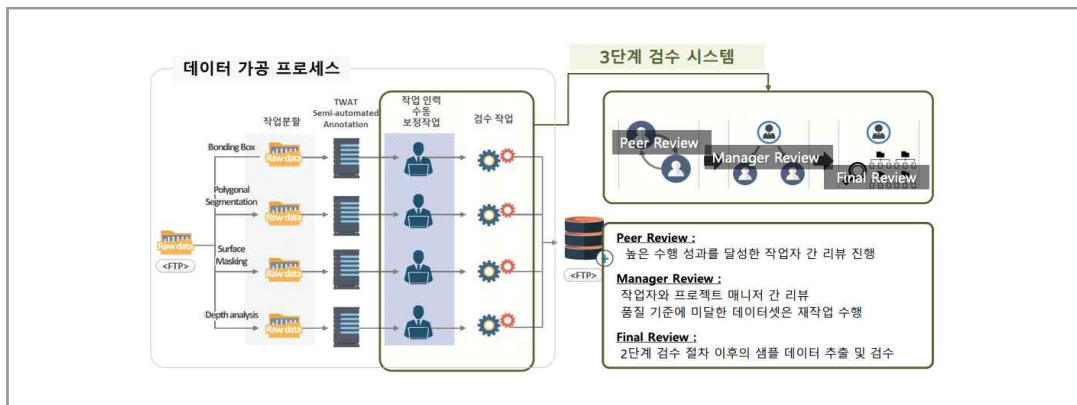





그림6 | 데이터 검수 프로세스

대량의 데이터를 생성하기 위하여 클라우드 소싱이 필요하다. 또, 그 과정에서 레이블 작업의 신뢰도를 높이기 위한 검수 프로세스의 정립이 필수적이다. 데이터셋 구축에서 매우 중요하다.

이 데이터셋에서는 라벨링 과정을 바운딩박스, 차량 특성 찾기, OCR 등으로 세분화하여, 과정 단계마다 이전 단계 데이터에 대한 검수를 진행하였다. 기존 미달 작업의 경우에는 재작업을 진행하였다.

각 단계별로 클라우드 워커들이 작업한 결과물을 가이드라인에서 제시한 형식에 맞는지(중복, 형식 규정 등) 체크하는 검수자가 있었고, 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성을 제시해주는 관리자는 딥핑소스의 직원으로 학습 경험이 풍부한 인력을 배치하여 최종적인 데이터셋의 품질을 담보할 수 있었다.

표3 | 데이터 검수 과정별 설명

항목	설명	예시
바운딩 박스 검수 및 재작업	<ul style="list-style-type: none"> 차량이미지 크롭이 올바르게 되었는지에 대해 전수 검수 진행 	<p>예) 아래와 같이 차량 일부가 잘린 경우 재작업 진행</p> 
메이커/모델/연식/색상 검수 및 재작업	<ul style="list-style-type: none"> 메이커/모델/연식/색상 전수 검수 	<p>예)</p> 
번호판 OCR 검수 및 재작업	<ul style="list-style-type: none"> 번호판 전사 결과 전수 검수 	<p>예)</p> 
데이터셋 통계 기반 검증	<ul style="list-style-type: none"> 클래시피케이션 결과 검증 	예)
학습 테스트 및 검수	<ul style="list-style-type: none"> 실제 학습 테스트 및 검수 	예)

●○ 데이터 구축 담당자

수행기관 : 딥핑소스(주) 이메일: contact@deepingsource.io