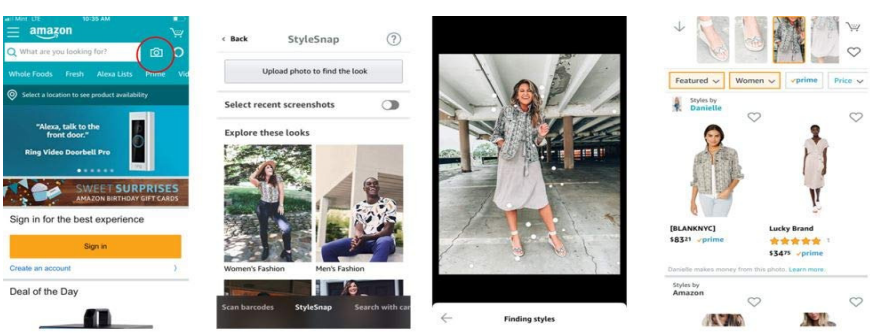
●○ 상품 이미지 및 고객 주문질의 응답 데이터 과제

상품 이미지 데이터

●○ 개요: 인공지능 학습용 이미지 데이터셋이란?

컴퓨터비전(CV, Computer Vision) 분야는 인간이 시각을 통해 수행할 수 있는 작업을 공학적으로 해석하여 기계로 하여금 동일한 처리를 할 수 있도록 자율적인 시스템을 구축하는 것을 목표로 하는 연구 분야이다. 최근 딥러닝(Deep learning)의 적용으로 성능의 비약적인 향상을 이루었으며 합성곱 신경망 (Convolutional Neural Network, 이하 CNN) 알고리즘의 사용이 대표적인 예이다. 이미지 인식 데이터셋은 CNN 모델을 학습하기 위한 데이터셋으로 이미지 분류(Image Classification) 기술을 적용하는데 활용할 수 있으며 ㈜롯데정보통신에서 구축하여, 10,000 종의 유통 상품에 대한 사진으로 구성되어 있다.

이미지 분류는 주어진 영상이 무엇인지를 분류하는 기술이며, 사람의 눈으로 보고있는 객체가 무엇인지에 대한 판단을 내리는 작업을 인공지능으로 구현하는 기술이며, 유통 분야에서는 상품 인식 및 자동 결제 처리 등에 활용하기 위해 활발한 연구 및 상용화를 진행하고 있다.이미지 분류 기술의 사례에 대해서는 아래를 참고할 수 있다.



[amazon stylesnap 사용 예시 (출처: amazon associate blog)]

StyleSnap:아마존의 옷 스타일 검색 서비스

미국의 IT 기업인 아마존(Amazon)은 자사 모바일앱에서 사진을 이용해 원하는 스타일의 의류를 검색할 수 있는 이미지 검색 기능을 제공한다. 소비자가 입력한 사진에 대해 컴퓨터 비전 처리를 통해 영상 분석을 수행하고, 유사한 스타일의 상품을 제시한다. 인공지능 기술을 사용한 이 서비스는 딥러닝 및 머신러닝을 통해 의류 아이템의 종류와 스타일에 대한 분류를 수행하여 상품을 추천한다.

그림1 | 이미지 분류의 사례

●○ 데이터셋의 구성

본 데이터셋은 일반적으로 이미지 분류 기술의 학습에 활용하는 이미지-정답 쌍의 상품 데이터셋 288만건과 상품 대표사진, 바코드, 영양정보를 담은 부가정보 데이터셋 6만건으로 구성되어 있다.

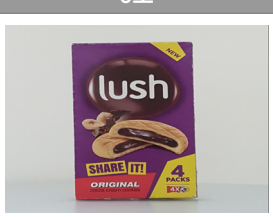
상품 데이터셋 288만건은 연구자가 연구를 진행하기에 충분한 양이며, 상용화 단계에서 강력한 사전학습모델을 만들 수 있는 양이다. 상품 대표사진, 바코드, 영양정보를 담은 부가정보 데이터셋 3만건은 사용자의 필요에 따라 상품 인식 결과에 대한 부가 정보 제공, 상품 가격 및 상품 바코드 검색 기능에 적용에 활용할 수 있다.

데이터 종류	포함 내용	제공 방식
상품 데이터셋	상품 이미지와 답(288만건)	JPG, XML 파일 (각 144만건)
부가정보 데이터셋	상품 대표사진, 바코드, 영양정보를 담은 부가 정보(6만 건)	JPG, XML 파일 (각 3만건)

●○ 데이터 원천 정보 요약

데이터 이름	상품이미지데이터
활용 분야	AI Hub 이용한 데이터 공개 통한 상품인식분야 AI 서비스 스타트업 기업 활용 무인 스토어, 물류창고, t-commerce 등 다양한 분야에서 활용
데이터 요약	원천데이터 : 국내 수퍼마켓, 편의점 등에서 판매되는 상품에 대해 Pitch 3개(0, 30, 60)와 Roll 20개(18, 36, 54, 중략..360)의 표준 촬영
데이터 출처	국내 수퍼마켓, 편의점의 판매 상품 대단위 마켓(대형마트, 수퍼마켓조합 물류센터 등)을 통한 구매 및 임대

| 이미지 샘플 : 단수/복수 상품 피치별 |

		0도	30도	60도
과자	단수			

		0도	30도	60도
	복수			
음료	단수			
	복수			
소스	단수			
	복수			

| 이미지 샘플 : 롤별 상품촬영(예시 : 0, 45, 60 90, 105, 150, 180도) |

복수							

| 이미지 샘플 : 부가정보 예시 |

상품대표사진	상품 영양정보	상품 바코드
		

●○ 데이터셋의 설계 기준과 분포

| 설계기준 |

- 무인결제대/단일/선반 상품의 인공지능 기반 인식기능의 구현을 위한 데이터셋 구축
  - 소상공인 매장 상품 유형과 비중 구성
  - 응용 서비스 모델에 맞는 상품 이미지 확보
    - 실제 생활에 매우 밀접한 수퍼마켓/편의점에서 현재 판매되는 상품
    - 특정 브랜드에 종속되지 않는 범용 상품

구분	상세 내용
편의점, 슈퍼 등 소상공인 상품 데이터	- 전국 편의점에서 판매되는 상품 이미지 데이터 - 전국 마트에서 판매되는 상품 이미지 데이터 - 전국 슈퍼마켓에서 판매되는 상품 이미지 데이터 ※ 편의점과 슈퍼마켓의 상품 분류가 상이하므로 편의점 분류 체계를 확장하며, 중복 상품 제거(10,000개 분류)
기 구축 상품 데이터	- 수행기관의 자체 데이터 구축 이미지 - 상품이미지 데이터 구축 후 증강용 활용

| 데이터 분포 |

데이터셋을 설계할 때 가장 중요하게 고려했던 점은 데이터 밸런스이다. 일반인들이 보편적으로 접할 수 있는 수퍼마켓, 편의점 등에서 판매되는 상품을 기준으로 분류기준을 만들었고, 해당 분류기준에 따라 골고루 데이터가 분포되도록 설계하여 학습 시 예상할 수 있는 데이터 편향성을 최소화하도록 했다.

- 각 상품별 원천 데이터 개수(10,022종이나, 상품 상황에 따라 10,000개 조정될 수 있음)

구분	데이터	Pitch(3개)+Roll(10개)	
		단수상품(1개)	복수상품(3개)
유제품	664	19,920	19,920
음료	1,048	31,440	31,440
주류	785	23,550	23,550
커피차	397	11,910	11,910
디저트	49	1,470	1,470
통조림/안주	771	23,130	23,130
상온HMR	426	12,780	12,780
소스	1,689	50,670	50,670
면류	381	11,430	11,430
과자	2,713	81,390	81,390
홈클린	1,099	32,970	32,970

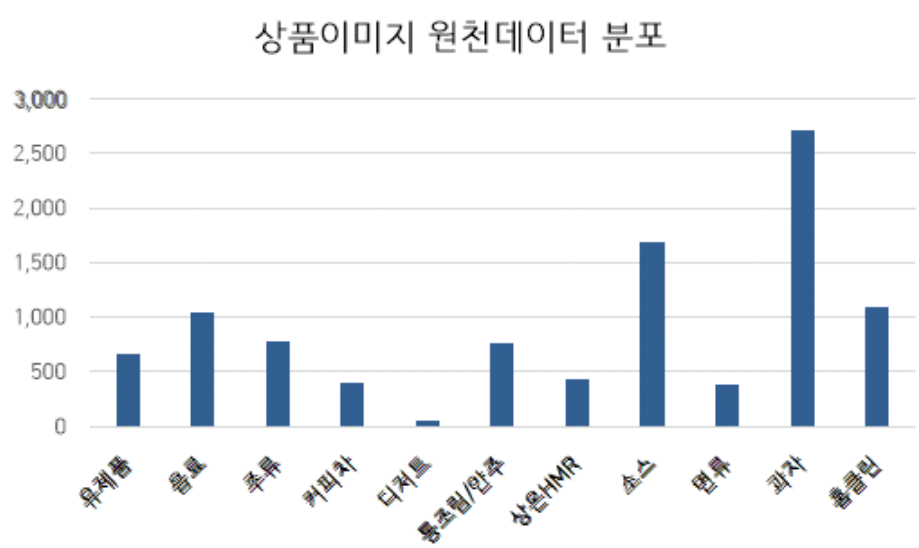


그림2 | 데이터셋 구성 개요

●○ 데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

※ PASCAL VOC object detection dataset

NO				항목명	설명
1				comp_cd	제조사 코드
	1-1			div_cd	분류 코드
		1-1-1		item_cd	상품코드
		1-1-2		item_no	상품번호
		1-1-3		div_l	대분류
		1-1-4		div_m	중분류
		1-1-5		div_s	소분류
		1-1-6		div_n	세분류
		1-1-7		comp_nm	제조사
		1-1-8		prod_nm	상품명
		1-1-9		vessel	용기
		1-1-10		volume	용량
		1-1-11		barcd	바코드번호
		1-1-12		width	가로
		1-1-13		length	세로
		1-1-14		height	높이
		1-1-15		img_prod_nm	상품명(이미지상)
		1-1-16		nutrition_info	영양기능정보
2				annotation	어노테이션 정보
	2-1			folder	이미지 디렉토리명
	2-2			filename	이미지 파일명
	2-3			path	이미지 파일 경로
	2-4			source	출처 정보
		2-4-1		database	DB명
	2-5			size	이미지 파일 크기
		2-5-1		width	너비 픽셀
		2-5-2		height	높이 픽셀
		2-5-3		depth	차원(RGB: 3)
	2-6			segmented	분할 여부
	2-7			object	라벨링 객체 정보
		2-7-1		name	상품코드
		2-7-2		pose	방향
		2-7-3		truncated	객체의 영역 초과
		2-7-4		difficult	난이도
		2-7-5		bndbox	바운딩 박스 좌표
			2-7-5-1	xmin	x최솟값
			2-7-5-2	ymin	y최솟값
			2-7-5-3	xmax	x최댓값
			2-7-5-4	ymax	y최댓값

●○ 데이터 예시

이 데이터는 설명 가능 데이터 기준이며, 표준 데이터셋, 정답 없는 데이터셋은 아래 예시에서 각각 clue, answers가 없는 구조를 가진다.

```
{
        "comp_cd" : {                    //제조사 코드
                "div_cd" : {               //분류코드
                        "item_cd"          : "", //상품코드
                        "item_no"          : "", //상품번호
                        "div_l"            : "", //대분류
                        "div_m"            : "", //중분류
                        "div_s"            : "", //소분류
                        "div_n"            : "", //세분류
                        "comp_nm"          : "", //제조사
                        "prod_nm"          : "", //상품명
                        "vessel"           : "", //용기
                        "volume"           : "", //용량
                        "barcd"            : "", //바코드번호
                        "width"            : "", //가로
                        "length"           : "", //세로
                        "height"           : "", //높이
                        "img_prod_nm"      : "", //상품명(이미지상)
                        "nutrition_info" : "", //영양기능정보
                        "sample"           : "", //샘플
                }
        }
}
```

●○ 데이터 구축 과정

인공지능 구현에 필요한 데이터를 확보하는 것은 매우 어렵다. 방대한 양의 데이터를 확보하는 것뿐만 아니라 데이터의 질적(Quality)인 측면까지 동시에3 고려되어야 하기 때문인데요. 데이터의 질(Quality)은 크게 두 가지 의미를 가져야 한다.

첫째, 다양한 상품이미지 데이터를 확보해야 한다. 정제되고 완벽한 상황만을 반영한 데이터가 많다고 해서 이를 기반으로 학습된 인공지능의 성능이 높아지는 것은 아니다.

이미지인식의 경우 완벽한 상태로 촬영된 데이터보다는 상품이 겹쳐져 있는 상태, 고객의 시선이나 진열대에 배치된 위치에 따른 다양한 각도의 사진 등의 데이터가 함께 학습 되어야 실제 제품이나 서비스로 출시 될 때 완성도를 높일 수 있다.

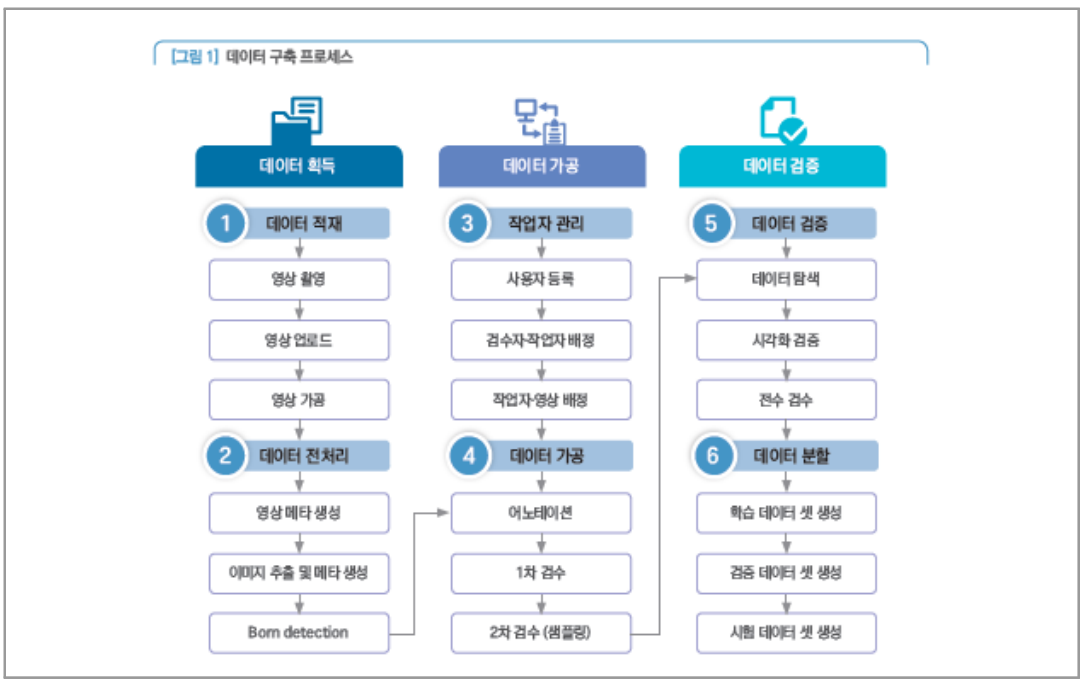


그림1 | 상품이미지 촬영 과정

둘째, 확보된 데이터가 기계 학습이 가능한 형태로 준비되어야 한다. 과거에는 데이터를 분석하는 주체가 인간이었지만 이제는 기계가 데이터를 직접 학습하고 분석한다. 인간에게는 단순하게 보이는 데이터라 할지라도 기계가 이해하기 위해서는 데이터의 전처리 과정이 필수적으로 요구된다.

예를 들어 상품이 진열되어 있을 때 상품의 경계선을 구분하는 것은 인간에게는 매우 쉽지만 동일한 이미지를 기계가 인간처럼 알아보기 위해서는 이미지 속의 각 상품별 경계 구분을 인간이 일일이 경계선으로 구분 짓고 해당 상품의 명칭을 이미지와 함께 기록해주어야 한다. 이미지 어노테이션(Annotation)이라 불리는 전처리 과정을 거쳐야 해당 이미지를 기계가 이해할 수 있게 되는 것이다.

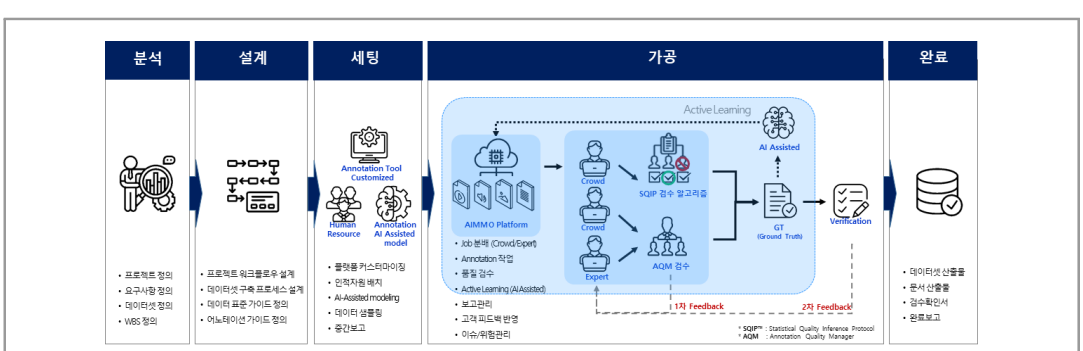


그림2 | 상품이미지 전처리 (어노테이션) 과정

●○ 검수와 품질 확보

대량의 데이터를 높은 품질로 생성하기 위하여, 크라우드 소싱 방식의 데이터 생성작업 품질을확보하기위한 검수 프로세스의 정립은 데이터셋 구축에 매우 중요한 의미를 갖는다. 철저한 검수를 통한 품질확보를 위해 3단계 검수 체계를 구축했는데, 가장 하위 레벨에는 크라우드 소싱 작업자들이 작업한 결과물을 상품이미지 촬영 가이드라인에서 제시한 형식에 맞는지 체크하는 검수자가 있었고, 이들이 검수한 결과물에 대해서 내용적으로 유효한지 검수하는 재검수자가 팀을 이뤄 활동했다. 이렇게 만들어진 데이터셋을 전체적으로 들여다보며 데이터셋의 밸런스나 가이드라인의 적절성을 제시해주는 관리자는 인공지능 데이터셋 처리 분야에서 다년간 지식과 경험을 축적한 산학연 데이터 품질관련 전문가로 구성된 품질관리위원회 인력으로 배치하여 최종적인 데이터셋의 품질을 담보했고 컨소시움사가 자체 개발한 저작도구를 사용해 품질 수준을 확보했다.

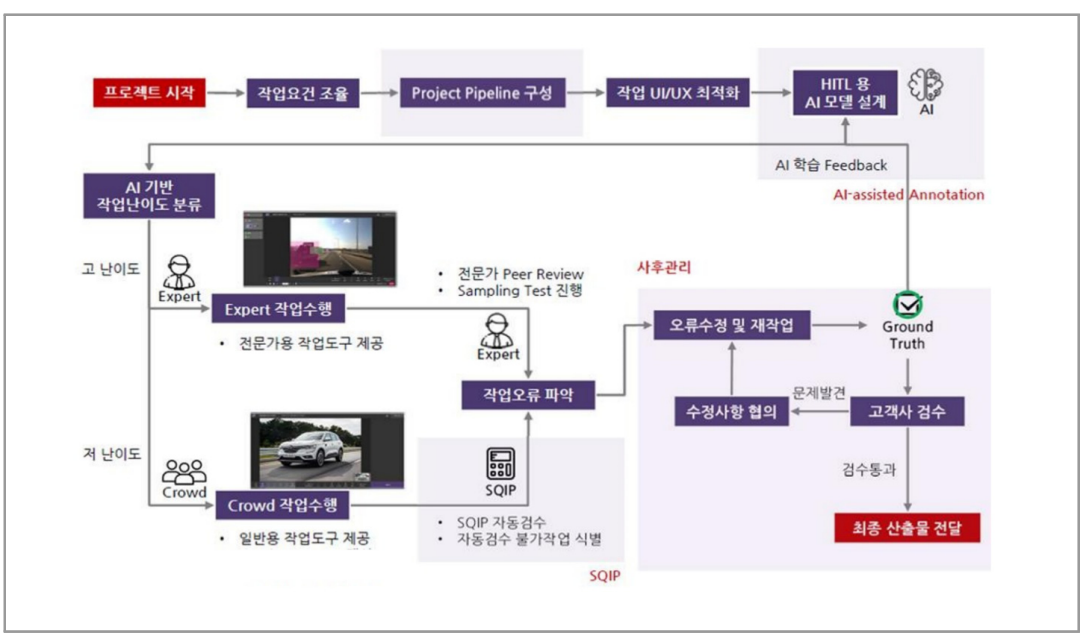


그림3 | 상품이미지 인식 품질 확보를 위한 검수 체계

●○ 데이터 구축 담당자

수행기관(컨소시움사) : ㈜롯데정보통신
(전화: 2626-4000, 이메일: aidata@lotte.net)