

2022년 ‘인공지능 학습용 데이터 구축 사업’ 과제 발굴을 위한 공공부문 수요조사 추진 계획(안)

한국지능정보사회진흥원, 2021.9.14.(화)

- ◆ 디지털 뉴딜 성과 조기 달성과 인공지능 산업 활성화를 위해
‘22년 인공지능 학습용 데이터 360종 구축 예정

□ 사업개요

- 원천데이터(이미지, 동영상, 텍스트, 음성 등)를 수집하여 기계학습(머신러닝) 개발에 필요한 형태로 구축·가공하여 개방하는 사업
 - ‘AI 허브(aihub.or.kr)’를 통해 개방 중이며, ‘21년 과제(190종) 추진 중

□ 수요조사

- (조사내용) 신규로 구축하고자 하는 인공지능 학습용 데이터
 - * 인공지능 활용 영역, 활용 목적, 필요한 데이터명, 보유데이터 현황, 학습용 데이터 형식, 세부 형식, 필수 데이터 수량, 데이터 수집 방법, 데이터 라벨링 방법
- (조사대상) AI 대학원
- (조사기간) ‘21.9.16.(목) ~ 9.30.(목)
- (조사기관) 한국지능정보사회진흥원
- (결과활용) 인공지능 학습용 데이터 신규과제 선정의 기초 자료로 활용

□ 제출 및 문의처

- (제출처) 네이버폼(<http://naver.me/GMRrVpOt>)링크를 통해 설문 작성
- (문의처) NIA 인공지능데이터전략팀(02-6747-2169, 2166, 2174, 2167),
airoadmap@nia.or.kr

◆ AI 학습용 데이터는 현실세계의 정형·비정형 데이터에 각종 지식과 정보를 라벨링하여 인공지능이 이해할 수 있는 형태로 가공

① (데이터 설계) 기계학습이 가능한 데이터의 형식과 구조*를 모델링하고 데이터의 구축 공정을 개발하여 데이터 작업자에게 제공

* 데이터 크기 및 길이, 파일 형식, 코딩문법 체계, 정보 라벨링 범위 및 방법 등

② (데이터 수집) 저작권, 개인정보 등 법적문제가 없는 원천 데이터를 온·오프라인을 통해 수집·제작·축적

※ 자율주행AI를 위한 도로주행영상을 데이터수집 장비 등 통해 직접 촬영하여 획득

③ (데이터 가공) 원천 데이터에 AI가 인지하고 판단해야할 각종 정보를 사람이 라벨링(태깅, 어노테이션)함으로써 데이터셋으로 가공

④ (데이터 확장) 고성능·다기능 AI 구현을 위해 복잡하고 고차원적인 정보*를 개발자가 추가로 라벨링하여 데이터 확장

* Level 4이상의 자율주행기술의 구현을 위해 영상에 나타난 사람(객체)의 시선 및 동선, 표지판(객체)이 담고 있는 정보(지명, 위치, 방향) 등 고차원 정보를 추가

⑤ (데이터 검증) AI에게 학습시킬 데이터의 다양성, 정확성, 유효성 등을 품질검증 전문기관에서 AI 학습을 위한 데이터셋의 적합성을 평가

< 자율주행차 학습 데이터 가공(예시) >

원천데이터 제작	자율주행용 정보 라벨링	AI 데이터셋 제공
<p><데이터 수집 차량></p>  <p><주행영상 획득></p> 	<p><주행영상내 객체정보 라벨링 ></p>  <p><전용 라벨링 툴 활용></p> 	<p><원천데이터, 가공데이터></p>  <p><라벨링 정보 코드파일></p> <pre> 1 [2 { 3 "filename": "IMG_20181022_173137_640.jpg", 4 "size": 46593, 5 "regions": [6 { 7 "name": "polygon", 8 "all_points_x": [좌표값 X], 9 "all_points_y": [좌표값 Y], 10 }, 11], 12 }, 13], 14 "region_attributes": { 15 "annotation_id": "승용차/SEDAN", 16 "annotation_en": "다목적차량/SUV", 17 "annotation_cn": "버스/BUS", 18 "annotation_jp": "화물차/CARGO TRUCK" 19 } 20] </pre>