

방송영상 데이터 117만 건, 국민에 개방

- 인공지능 학습용 데이터 누구나 활용토록 공개...다양한 인공지능 모델 연구 기대 -

방송영상 이미지와 배경 및 인물, 한글 서체와 자막 등 우리나라 정서와 문화적 요소를 반영한 방송영상 데이터 117만여 건이 국민에 개방된다.

방송미디어통신위원회(위원장 김종철)와 한국전파진흥협회(협회장 홍범식)는 방송영상 인공지능 학습용 데이터 일부를 교육·연구용으로 국민 누구나 활용할 수 있도록 한국지능정보사회진흥원(NIA)의 ‘인공지능 허브 안심 구역(AI-Hub Safe Zone)’을 통해 정식 개방한다고 5일 밝혔다.

‘인공지능 허브 안심 구역(AI-Hub Safe Zone)’은 관련 개발자와 국민이 제공된 데이터를 안전하게 사용할 수 있도록 제공된 보안 통제 환경으로, 사전에 안심 구역에 대한 이용 신청 및 승인 절차를 거쳐 누구나 무료로 이용할 수 있다.

그 간 최적의 인공지능 학습용 데이터로 평가받는 방송영상에 대해 많은 수요가 있었지만 지식재산권, 초상권, 인접권 등의 문제로 이용이 어려웠는데, 이번 데이터 공개를 통해 개발자들이 걱정 없이 고품질의 영상 데이터를 활용해 다양한 인공지능 모델을 연구할 수 있게 됐다.

개방되는 데이터는 인공지능 허브 안심 구역 누리집(<https://www.aihub.or.kr/>)을 통해 확인 가능하며 ▲영상·이미지 이해 및 설명 ▲배경 및 인물, 객체 생성 ▲제작 환경 고도화 분야에서 배경, 인물, 객체, 서체, 자막 등 10개 유형으로 총 117만여 건이 제공된다.

천지현 방송미디어진흥국장은 “방송영상은 한국적인 문화와 정서, 고도의 제작 기술과 전문성이 집약된 인공지능 산업의 핵심 자산”이라며 “앞으로도 다양한 산업 분야에서 방송영상이 활용될 수 있도록 고품질의 인공지능 학습데이터를 지속적으로 구축, 공개할 계획이다”라고 말했다.

붙임. 방송영상 데이터 개방 내역

담당 부서	방송미디어진흥국 디지털방송미디어정책과	책임자	과 장	함형철	02-2110-1620
		담당자	사무관	성창원	02-2110-1635
관련 기관	한국전파진흥협회 방송미디어진흥본부	책임자	본부장	곽기훈	(02-317-6110)
		담당자	팀 장	김지수	(02-317-6197)

구 분	데이터셋명	수량	세부 내용
영상·이미지 이해 및 설명	방송영상 이미지 캡셔닝 데이터	294,906	이미지 기반 한국 지역적 특색(문화유산, 자연 경관 등) 을 반영한 캡셔닝 데이터셋
	방송영상 비디오 캡셔닝 데이터	217,587	영상 기반 한국 지역적 특색(문화유산, 자연 경관 등)을 반영한 캡셔닝 데이터셋
	K-콘텐츠 영상 이해 데이터	37,940	K-콘텐츠 영상 내 객체·행동·장면·VQA 등 유형별 다양한 요소를 상세 설명한 데이터셋
배경 및 인물 객체 생성	배경영상 상세 설명·생성 데이터	22,953	한국 고유의 자연, 도시, 건축 양식 등 다양한 배경을 상세 묘사한 데이터셋
	한국적 배경 및 객체 생성 데이터	128,689	한국 고유의 자연, 도시 등 다양한 환경 이미지에 상세 설명을 결합한 데이터셋
	한국적 감정 인물 생성 데이터	96,991	K-콘텐츠의 컨텍스트 기반 한국인의 표정과 감정이 담긴 데이터셋
	K-콘텐츠 인물 페르소나 생성 데이터	95,446	방송에 등장하는 인물의 성격, 말투, 관계 등 정체성 정보를 상세 묘사한 데이터셋
제작 환경 고도화	K-콘텐츠 VLM 데이터	43,604	영상의 객체, 장면, 행동뿐만 아니라 맥락적 의미까지 반영한 멀티모달 데이터셋
	K-콘텐츠 서체 생성 데이터	151,334	방송영상 내에 노출된 한글 이미지 특징을 상세 묘사한 데이터셋
	K-콘텐츠 자막 생성 데이터	86,350	방송 자막의 위치, 색상 등 시각적 속성과 표현 방식을 상세히 묘사한 데이터셋
합 계		1,175,800	