과학기술정보통신부
Ministry of Science and ICT
www.msit.go.kr

보도자료

혁신이 혁신을 낳는 과학기술 생태계를 만들겠습니다

힘내라 대한민국

보도일시	2020. 6. 22.(월) 조간(온라인 6. 21.(일) 12시)부터 보도해 주시기 바랍니다.		
배포일시	2020. 6. 19.(금)	담당부서	빅데이터진흥과
담당과장	양기성(044-202-6290)	담 당 자	박재수 사무관(044-202-6293)

과기정통부, 디지털 뉴딜 핵심, 「데이터 댐」 구축에 나서다

- 2020년도 인공지능(AI) 학습용 데이터 20종 선정(390억원) -
- 올해내 추경예산을 통한 총 150종(2,925억원, 정부안) 추가 구축-

□ **과학기술정보통신부**(장관 최기영, 이하 '과기정통부')는 6월 22일, 인공지능(AI) 개발에 필수적인 양질의 **데이터를 대규모로 구축·개방**하는 **"AI 학습용 데이터 구축 사업"의 20개 과제를 확정**하였다고 밝혔다.

< AI 학습용 데이터 구축사업 현황 >

○ '17년부터 시작된 AI 학습용 데이터 구축 사업은 한-영 번역 말뭉치, 한국어 음성, 이상 행동 영상 등 텍스트, 이미지, 영상 분야의 인공지능 서비스 개발을 위한 **기계학습용 데이터 21종, 4,650만건**을 **구축**하여 AI통합지원 플랫폼인 **AI허브**(www.aihub.or.kr)**를 통해 공개** 중이다.

< AI 학습용 데이터 구축·개방 현황('17~'19) >

□ 올해는 작년보다 예산 및 과제가 **2배 늘어난 20개 과제**(10개 지정공모, 10개 자유공모)를 **390억 원 규모**('19년 195억 원)로 추진하였으며, 총 92개 사업자가 참여하여 **평균 4.6대 1의 경쟁률**을 보였다.

1. 선정 분야

□ **(국가적·산업적 필요성)** ❶사람의 감성과 문맥을 이해할 수 있는 **자연어 처리분야**, ❷자동차, 드론 등 **자율주행기술 분야**, ❸음성, 시각, 언어 등 **융합 분야** 등 국가적으로 산업적으로 **활용가치가 높고 데이터 확보 필요성이 시급**한 과제를 선정하였다.

□ **(국민편의 향상)** 이와 함께, ❶질병진단(치매, 구강질병 등), 운동 등 **헬스케어 분야,** ❷사람의 얼굴을 악의적으로 변조한 **딥페이크 방지 기술 분야,** ❸**장애인의 삶을 향상 시킬 수 있는 분야** 등 **국민생활을 윤택**하게 하고 **사회적 문제를 해결**할 수 있는 과제를 선정하였다.

< '20년 AI 학습용 데이터 구축 과제 선정결과 >

지정공모(10개)			자유공모(10개)	
	과제명	수행기관	과제명	수행기관
1	대용량 동영상 콘텐츠 AI데이터	KDX	질병진단 이미지 AI데이터	국립암센터
2	자율주행드론 비행 영상 AI데이터	울산대학교 산학협력단	도로환경 파노라마 이미지 AI데이터	올포랜드
3	시각정보 기반 질의응답 AI데이터	유클리드소프트	피트니스 자세 이미지 AI데이터	슬릭코퍼레이션
4	수어 영상 AI데이터	테스트웍스	K-Fashion 이미지 AI데이터	오피니언라이브
5	한국인 대화음성 AI데이터	솔루게이트	한국인 재식별 이미지 AI데이터	한국과학기술연구원
6	딥페이크 방지영상 AI데이터	머니브레인	도로주행영상 AI데이터	티큐에스코리아
7	랜드마크 이미지 AI데이터	피씨엔	치매진단 뇌파영상 AI데이터	디노플러
8	사람 인체·자세 3D AI데이터	스위트케이	감성 대화 말뭉치 AI데이터	미디어젠
9	문서요약 텍스트 AI데이터	비플라이소프트	위성영상 객체판독 이미지 AI데이터	한국항공우주연구원
10	전문분야 한영 말뭉치 AI데이터	플리토	구강악 2D·3D 이미지 AI데이터	헬스허브

2. 크라우드소싱 방식 도입으로 일자리 창출

□ AI 학습용 데이터를 수집하고 가공하는데 **많은 인력이 필요**하다.

○ 이에, 미국, 중국 등 해외에서는 누구나 참여하여 데이터를 가공할 수 있는 **크라우드소싱 방식***을 **경쟁적으로 도입**하여 **인력 부족 문제를 해결**하면서 **양질의 데이터 확보**와 **일자리를 창출**하고 있다.

* 크라우드소싱 : 언제 어디서든 누구나 데이터 수집 및 가공에 참여하는 방식

□ 과기정통부는 작년에 **2개 과제**를 **크라우드소싱 방식으로 추진**한데 이어, 올해는 **모든 과제**에 전면적으로 **크라우드 소싱 방식**을 **적용**하여 **AI 학습용 데이터를 구축**함으로써 **일자리를 만들어갈 예정**이다.

○ 특히, **청년**과 **취업준비생**, **경력단절여성**, **장애인** 등에게 **많은 일자리**가 **제공**되고, **데이터 가공 전문성**을 **쌓을 수 있는 기회**도 될 것으로 기대된다.

* '19년 AI 학습용 데이터 구축사업 분석결과, 일반적 데이터가공은 10억당 38.1명, 크라우드소싱 방식은 10억당 200명 일자리 창출 효과 발생

3. 추경 예산을 통해 AI 학습용 데이터 확대 구축

□ 아울러, **과기정통부**는 **'20년 추경예산**을 통해 코로나19 이후 경기 침체를 극복하기 위해 **일자리 창출효과가 크고 AI모델 개발에 필수**적인 **AI 학습용 데이터 구축 사업을 확대 추진**할 계획이다.

○ **총 150개 종류의 AI 학습용 데이터를 구축**하고 **AI 통합 지원 플랫폼인 AI 허브**(www.aihub.or.kr)**를 통해 무료 개방할 계획**이다.

□ 과기정통부는 중소기업과 스타트업 등은 **비용부담과 인력부족** 때문에 필요한 AI 학습용 데이터를 직접 구축하는데 어려움 많아서, **양질의 데이터 확보에 대한 수요가 크다**고 밝히며,

○ 시장에서 필요한 양질의 AI 학습용 데이터를 많이 확보하는 것이 **단기간에 우리나라 AI 경쟁력**을 **끌어 올릴 수 있는 해결책**이 될 수 있다고 강조하였다.

○ 과기정통부는 이번 추경을 통해 **AI 학습용 데이터 구축을 대규모로 확대·구축**하여 **데이터 댐**에 모으고, **다양한 AI 기술연구, 상용화 서비스 개발에 활용**할 수 있도록 지원하겠다고 하며,

- 이를 통해, **양질의 일자리 창출**과 **경제성장의 새로운 원동력을 확보**할 수 있을 것으로 기대한다고 밝혔다.

【 데이터 댐 구상 예시, 추경 정부안 기준 】

붙임 : ’20년 AI 학습용 데이터 과제선정 및 주요 내용(요약)

이 자료에 대하여 더욱 자세한 내용을 원하시면
과학기술정보통신부 박재수 사무관(☎ 044-202-6293)에게 연락주시기 바랍니다.

붙 임 20년 과제 선정결과 및 주요 내용

□ 지정과제(10개)

과제명	주관/참여기관	주요 내용
대용량 동영상 콘텐츠	**KDX 한국데이터거래소**, 씨이랩, 매경닷컴, 에버영피플, 서울대학교 산학협력단, 씨드롭, 상상우리, 에스이앤티, 베어버터, 디앤디클라우드	o 대용량 동영상 내 객체 탐지, 상황 이해, 행동 분석을 위한 대용량 동영상 AI 데이터 구축 - 원천 데이터 30종류 1,630시간 이상 확보, 객체 행동 카테고리 분류 7,500개 이상 구성, 바운딩 박스 700만개 이상 구성, 학습 데이터 500시간 이상 확보
자율주행 드론 비행 영상	**울산대학교 산학협력단,** 경북대학교 산학협력단, 서홍테크, 에이테크, 엠엠피, 휴먼드론개발, 유시스, 단트넷, 울산정보산업진흥원	o 관광지, 도심지, 산림지 4K, 25FPS 360시간 및 LiDAR 영상데이터 20시간 구축 o 별도 품질 관리 지표 및 방안 마련하여 데이터 품질 관리 실시
시각정보 기반 질의응답	**유클리드소프트,** 한국원자력연구원, 국립공주대학교, 터치스톤	o 생활 이미지와 이미지에 대한 질문을 입력받아 질문에 대한 답을 생성하는 AI데이터 구축(이미지 135만장, 한국어 질의응답 750만쌍)
수어 영상	**테스트웍스,** 이큐포올, 한국농아인협회, 카이스트, 나사렛대학교	o 청각 및 언어장애를 가진 사람들이 사용하는 수어를 영상 기반으로 인식하여 의사를 전달할 수 있도록 AI 기술 및 응용서비스 개발에 필요한 수어 영상 학습 데이터 구축
한국인 대화음성	**솔루게이트,** 타임소프트, 코난테크놀로지	o 한국인의 일상 대화를 인식하고 음성을 문자로 실시간 변환하는 AI 기술 개발을 위한 대화 음성 데이터셋 구축 (원본 음성 데이터 4,000시간 이상, 음성을 문자로 변환한 텍스트 데이터 400만 문장)
딥페이크 방지영상	**머니브레인,** 크라우드웍스, 서울대학교	o GAN(적대적 생성 신경망) 기반의 다양한 변형 알고리즘을 통해 생성된 변조 영상을 탐지하는 AI기술 개발에 필요한 원본 및 변조 영상 데이터 구축 및 응용서비스 개발
랜드마크 이미지	**피씨엔,** 크라우드웍스, 데이콘	o 인공지능 기반의 시각지능 기술 및 서비스 개발에 활용하기 위한 국내 특성이 반영된 국내 도심 민간건물, 공공기관, 관광명소, 편의시설 등 국내 도시별 주요 랜드마크 이미지 데이터 구축
사람 인체자세 3D	**스위트케이,** 서울대학교, 한국디자인진흥원, 모션테크놀로지	o 2D인체 영상을 3D모델로 변환할 때, 자세(pose)와 형태(shape)를 추론하여 커머스, 스포츠 및 AR·VR 서비스를 개발하기 위한 2D-3D 인체 데이터셋 구축
문서요약 텍스트	**비플라이소프트,** 위고, 테스트웍스, 고려대학교, 에이아이닷엠	o AI가 텍스트를 이해하고 핵심 내용을 요약적으로 전달하기 위해 AI SW가 해당 텍스트의 주요 내용이 무엇인지를 이해할 수 있는 형태로 가공된, 다양한 유형의 대규모 요약 텍스트 데이터 구축
전문분야 한영 말뭉치	**플리토,** 솔트룩스파트너스, 에버트란	o 한영 병렬 번역 말뭉치 155만건 구축 o 대법원 판례(인공지능 판례 번역), 의료/보건(코로나19 등 pandemic 관련 공문) 등 전문분야별 한영 말뭉치 구축

□ 자유과제(10개)

과제명	주관/참여기관	주요 내용
질병진단(암조직, 부비동) 이미지 AI데이터	**국립암센터,** 건양대학교병원, 인피니트헬스케어, 딥노이드, 마인즈앤컴퍼니, 유비즈정보기술, 오엠인터랙티브, 딥네츄럴, 네이버비즈니스플랫폼	o 유방암 및 부비동 질환의 진단을 위한 의료 영상 이미지 AI데이터 구축
도로환경 파노라마 이미지 AI데이터	**올포랜드,** 스티리스, 지디에스컨설팅그룹, 에스이앤티, 가천대학교 산학협력단	o 영상데이터 수집 후 가공을 통한 자율주행용 이미지 AI데이터 구축
피트니스 자세 이미지 AI데이터	**슬릭코퍼레이션,** 데이터연구소, 서울대학교 산학협력단, 위힐드	o 피트니스 자세 평가/피드백 AI Application을 개발하고자 하는 기관들이 사용할 AI데이터셋 및 관련 모델/응용서비스를 구축
K-Fashion 이미지 AI데이터	**오피니언라이브,** 웨얼리, 에이아이닷엠, 이화여자대학교 산학협력단, 한국패션산업연구원	o 구매 또는 직접 촬영하여 저작권 문제가 해결된 패션 이미지의 패션 요소 정보를 어노테이션한 이미지 100만장 이상 구축
한국인 재식별 이미지 AI데이터	**한국과학기술연구원,** 휴먼아이씨티, SQI소프트	o 대한민국의 실내/외 구축된 공공 CCTV 환경을 고려한 한국인(1,000명) 재식별 데이터셋 구축
도로주행영상 AI데이터	**티큐에스코리아,** 지어소프트, 와토시스, 한국자동차연구원	o 70건 이상 실도로 주행 데이터 Use-Case 기반 총 175TB 상당의 자율주행 원천데이터 수집, 총 85만 5천 프레임 구축
치매진단 뇌파영상 AI데이터	**디노플러스,** 엔브레인, 삼성서울병원	o 치매 및 난청 Active 영상·이미지 데이터(PACS) 및 임상전문의 진단정보 AI데이터 구축
감성 대화 말뭉치 AI데이터	**미디어젠**	o 우울증 등 심리 장애로 인한 사회문제 해결을 위해 감성대화 코퍼스 데이터 구축
위성영상 객체판독 이미지 AI데이터	**한국항공우주연구원,** 에스아이아이에스, 에스아이에이, 슈퍼브에이아이	o 국내 위성 영상 활용 산업의 발전을 위해 아리랑 위성영상을 이용한 범용 위성정보 데이터 구축
구강악 2D·3D 이미지 AI데이터	**헬스허브,** 서울대학교 치과병원, 사회적 협동조합 굿임팩트	o 치아 및 치주질환 진단과 치료계획 수립을 위한 파노라마 영상과 CBCT (Cone Beam Computed Tomography) 영상 데이터 구축