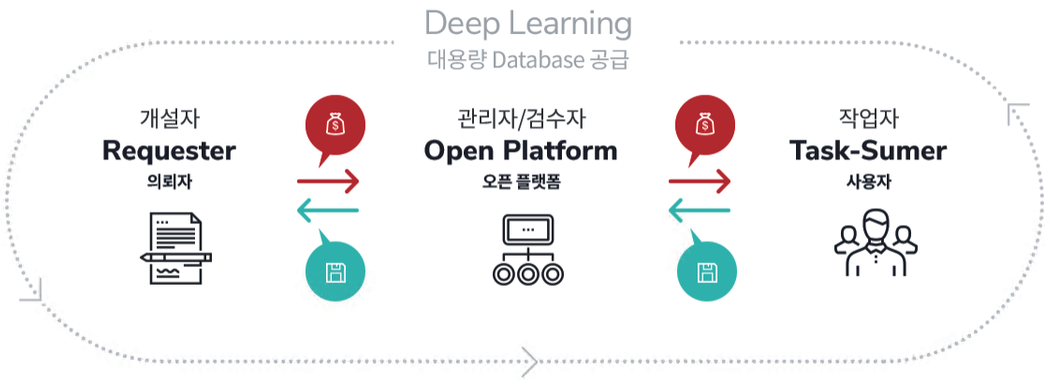
의미 있는 데이터를 모아

딥러닝 학습데이터 플랫폼 크라우드웍스

전문가들은 인공지능(AI)이 세상을 바꿀 것이라고 말한다. 그런데 AI는 대부분 데이터를 기반으로 작동한다. 의미 있는 데이터가 많아야 AI가 제대로 기능한다는 얘기다. 크라우드웍스는 딥러닝 학습데이터 플랫폼으로 기업이 필요로 하는 데이터 수집을 돕고 있다.

글 김태환 머니투데이방송 기자 kimthin@mtn.co.kr



인공지능(AI)의 역할이 확대되고 있다. AI는 사람이 하기 어려운 일을 대신하거나 단순한 작업을 대체해 사람이 더 창의적인 일에 몰입할 수 있게 돕는다. 그런데 이런 AI가 제대로 성능을 발휘하려면 데이터의 질이 좋아야 한다. AI는 데이터를 기반으로 학습하고 연산하기 때문에 데이터가 풍성하고 확실할수록 결과값도 정확하게 나온다.

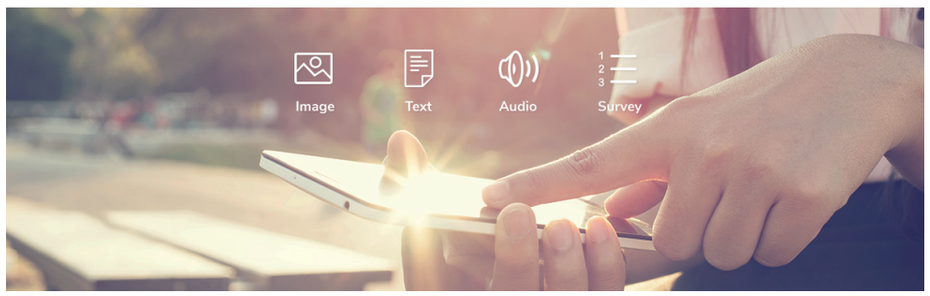
AI는 사람의 사고방식을 차용해 연산을 한다. 컴퓨터가 사람처럼 스스로 정보를 확인하고 공부하도록 학습시키는데, 이런 학습을 머신러닝이라고 부른다. 사람이 특정 데이터를 주면, 이를 학습하게 만든다. 바둑AI 알파고도 초창기 개발 단계에서 머신러닝을 진행했다.

딥러닝은 머신러닝에서 한 발짝 더 나아간 개념이다. 딥러닝 단계에 들어서면 사람의 도움 없이 AI가 스스로 데이터를 학습하고 분석한다. 예를 들어 강아지가 어떻게 생겼는지 학습할 때, 기존에는 사람이 강아지 사진만을 주면서 특징을 익히도록 만들었다. 하지만 딥러닝 단계에서는 AI가 수십만 장의 다양한 사진을 스스로 보고 판별한다.

한국 최초의 데이터 크라우드 소싱 플랫폼

AI가 딥러닝을 효율적으로 수행하려면 제공하는 데이터의 정확도가 높아야 한다. 예를 들어 강아지 사진을 학습한다면 정확한 강아지 사진이 많을수록 학습 효과가 높다. 오답도 미리 오답과 아닌 것을 확실하게 분류해 놓으면 효율이 더 높아진다. 오답을 분별하는 속도가 줄어 연산 속도가 빨라진다. 즉 가공된 비정형 데이터가 많을수록 유용하다.

기존에는 기업이 단기 아르바이트 직원을 직접 고용해 이 같은 분류 작업을 진행했다. 그런데 이렇게 하니 관리 문제가 발생했다. 주로 대학생들이 작업에 많이 참여했는데, 방학 때 업무를 수행하다 숙련됐을 즈음인 3개월이 지나면 모두 학교로 돌아가 버렸다. 교육하는 데도 비용이 드는데, 다시 새로운 인력을 채용해야 하는 곤란한 상황이 반복됐다.



크라우드웍스는 이 같은 관리 비용과 성과 관리에 대한 해결책으로 플랫폼을 도입했다. 데이터가 필요한 회사와 원하는 데이터를 제공할 수 있는 사람들을 연결해주는 시스템이다. 예를 들어 개와 고양이를 구분하는 AI를 만드는 A회사가 개 사진 데이터를 원하면, 크라우드웍스 홈페이지에서 개 사진을 모아줄 수 있는 작업자를 모집한다. 회원들은 1장당 일정 수준의 포인트를 지급받으며 사진을 업로드한다.

이런 형태를 데이터 크라우드 소싱 플랫폼이라고 한다. 세계 최대 IT업체 중 하나인 아마존이 '메커니컬 터크(Mechanical Turk)'라는 이름으로 먼저 시작한 서비스다. 한국에서는 크라우드웍스가 최초다.

크라우드 소싱 방식은 동물 사진처럼 단순한 데이터 수집뿐만 아니라 고도화된 데이터를 모을 수 있다는 장점이 있다. 크라우드웍스에서 진행한 '아기 울음소리 수집'을 예로 들 수 있다. 초보 부부들은 아기가 울 때 왜 우는지 모르는 경우가 많다.

크라우드웍스는 태어난 지 6개월 미만의 아기가 있는 데이터 수집자를 모았다. 이들에게 소리를 수집할 수 있는 장치를 지급해 아기가 우는 소리를 녹음하고, 왜 울었는지를 태깅(tagging)하도록 했다. 이렇게 수집한 소리가 축적될수록 아기 울음소리를 판별하는 AI 정확도가 높아졌다.

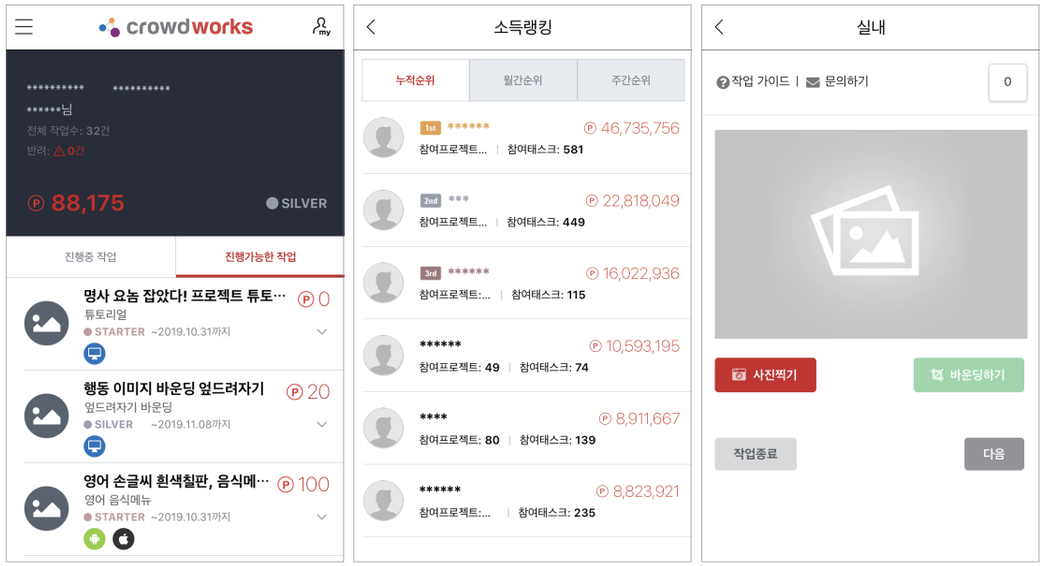
국내 AI 정확도 70%

채팅이나 전화 상담에서 음란한 메시지나 이야기를 차단하는 AI를 개발할 때도 크라우드웍스의 데이터 수집 플랫폼이 동원됐다. 성희롱은 직접적으로 음란한 단어를 이야기하기도 하지만 단어나 상황을 비틀며 나타나기도 한다. 크라우드웍스에서는 이를 위해 다양한 음란 메시지를 수집했고, AI를 만드는 데 기여할 수 있었다.

AI 데이터에서 가장 많은 비중을 차지하는 분야는 텍스트다. QA셋 수집, 요약문 생성, 형태소 분석과 같은 다양한 프로젝트를 진행한다.

작업한 데이터들은 모두 '검수'를 거친다. 검수를 통과하지 못한 작업물에는 리워드를 지급하지 않는다. 이 과정을 통해 데이터 신뢰도를 99%까지 높였다고 크라우드웍스 측은 설명했다. 반면 아마존 메커니컬 터크는 검수 과정이 없다. 이처럼 사람이 데이터를 정제하는 작업이 필요한 이유는 AI의 정확도가 아직은 다소 떨어지기 때문이다.

크라우드웍스 손유이 매니저는 "현재 국내 AI는 태깅과 라벨링 업무를 할 수는 있지만 정확도가 70% 수준이다. 결국 사람이 다시 봐야 하는 상황"이라며 "국내 AI 데이터는 예전에는 신생아 수준이었다면 지금은 중고생 수준까지 올라온 상태"라고 설명했다.



크라우드웍스 플랫폼 앱 화면.

그는 "특히 AI 수준이 낮을수록 데이터 품질이 높아야 한다"면서 "안타깝게도 미국이나 유럽의 AI 기술력에 비해 한국은 수준이 낮아 데이터 질을 극단적으로 끌어올려야 비슷한 수준의 서비스를 개발할 수 있다"고 지적했다.

이처럼 데이터에 대한 품질을 높이기 위해서는 상황에 따라 딱 맞는 인력을 그때마다 모집해야 한다. 예를 들어 CT 영상을 보고 종양을 찾는 AI 서비스를 만들려면 해당 영상을 보고 분석할 수 있는 의사들을 모집해야 한다. 이런 경우 크라우드웍스에서 따로 해당 직군의 사람을 모집하기도 한다. 보상금도 일반 과제보다 더 많이 지급된다.

AI와 콘텐츠의 만남

크라우드웍스는 최근 인공지능연구원과 함께 한국콘텐츠진흥원(이하 콘진원)의 지원을 받아 '지능형 캐릭터 저작 및 서비스 모델 개발' 사업을 진행하고 있다. 이 사업은 AI가 앞에 있는 사람의 상황을 판단하고, 이에 알맞는 감정을 표현하는 '지능형 아바타'를 개발하는 것이 목표다. 예를 들어 가정용 AI 로봇을 만드는데 앞에서 아이가 울고 있다면 어떤 상황인지 파악하고 소통하고 공감해야 한다. 이때 로봇이 생글생글 웃으면 오히려 역효과를 낼 수 있다.

지능형아바타는 사람의 감정과 상황을 파악해 어떤 상태인지를 '입력'하고, AI가 어떤 행동을 취해야 할지 '선택', 그리고 이를 밖으로 내보내는 '출력'까지 3단계 과정을 처리한다. 이 모든 과정을 AI로 구현하려면 모아야 할 데이터가 굉장히 많아진다. 우선 얼굴인식을 하기 위해서는, 나이와 성별 등 다양한 조건에 따라 개별적으로 수집해야 한다. 말을 할 때 함께 나오는 제스처도 분석해야 한다. 제스처를 통해 사람들의 정확한 감정 상태를 확인할 수 있다. 또 AI가 말을 하면서 제스처를 사용하면 더욱 사람처럼 표현할 수도 있다.

지능형아바타는 사람과 상호작용을 할 수 있어 콘텐츠 분야에서도 유용하게 활용할 수 있을 것으로 기대되고 있다. 아바타를 활용해 더 자연스러운 AI 아이돌 그룹을 만들거나 사이버 연예인을 만들 수도 있다.

크라우드웍스 김지선 사업개발팀 팀장은 "이번 콘진원 연구개발(R&D) 사업을 통해 새로운 경험을 할 수 있어 긍정적"이라고 설명했다. 이번 프로젝트는 2000명의 데이터 제공자를 통해 총 2만 건의 안면데이터를 모으는 대규모 사업이다. 크라우드웍스는 지금까지 해보지 못했던 대규모 사업을 해 볼 수 있다는 점에서 AI 기술 개발에 큰 도움이 될 것으로 기대하고 있다. Ⓝ

"세상에 존재하지 않는 이미지를 만들다"

김지선 크라우드웍스 사업개발팀 팀장

크라우드웍스 플랫폼 참여도는 어떤가.

작업자로 3만 여명이 투입되고 있다. 프로젝트를 오픈하고 24시간 이내에 끝나는 비교적 간단한 프로젝트는 한 프로젝트에 1,000명이 순식간에 들어오기도 한다. 프로젝트 수는 2019년 10월 기준 560개가 등록됐으며, 데이터 고객은 80여 회사다.

중복 데이터 관리는 어떻게 하나.

사진 같은 경우 앨범에서 불러온 사진을 업로드하지 못하게 막아 놨다. 직접 찍을 수밖에 없도록 만들어 사진 중복을 최소화하려는 조치다. 또 한 작업자의 수집물을 한 화면에서 검수하기 때문에 중복 데이터를 쉽게 걸러낼 수 있다. 이런 식으로 작업자의 부주의나 실수를 막을 수 있는 기능을 늘 고민하고 있다. 작업자와 검수자를 관리하기 위한 특허를 34개 출원해 현재 5건이 등록된 상태다.

어떤 회사들이 데이터를 원하는가.

AI를 다루는 거의 모든 기업과 학교가 크라우드웍스의 고객이다. 네이버와 카카오 같은 IT업체를 비롯해 SK텔레콤과 KT 같은 이동통신사, 카드사인 현대카드와 함께 했다. 서울대와 KAIST 같은 학계와도 다양한 데이터 프로젝트를 진행하고 있다.

콘진원과의 협업은 어떤 의미인가.

기업들과 업무를 진행할 때 우리의 영역은 늘 '데이터 전처리'까지다. 우리가 생성한 데이터가 어떻게 활용되는지 직접적으로 알 수 있는 기회가 적다. 이번 콘진원의 R&D 사업은 인공지능연구원이라는 주관기업과 함께 진행하기 때문에 우리가 만든 데이터가 어떻게 활용되는지 제대로 알 수 있어 많은 도움이 된다. 새로운 동기부여를 받았다.

콘텐츠 분야에 적용될 수 있는 AI 기술이 있다면.

생성적 적대 신경망(GAN)이라는 기술이 있다. 기존에 있는 데이터를 합쳐 새로운 것을 창조해낸다. '말하는 모나리자' 같이 세상에 존재하지 않는 이미지를 만드는 기술이다. 이런 기술을 활용하면 기존에 없던 작품을 만들 수 있다.

AI를 활성화하는데 필요한 정책 지원을 꼽는다면.

딥러닝 학습용 데이터에 대한 규제를 완화해야 한다. 현재 한국은 개인정보보호법이 강력해 데이터 활용이 매우 제한적이다. 반면 일본에서는 2017년에 저작권법을 개정해 연구 용도로 데이터를 자유롭게 활용할 수 있다. 데이터를 자유롭게 활용할 수 있어야 한국 AI 연구가 더 활성화되고 발전할 수 있다.