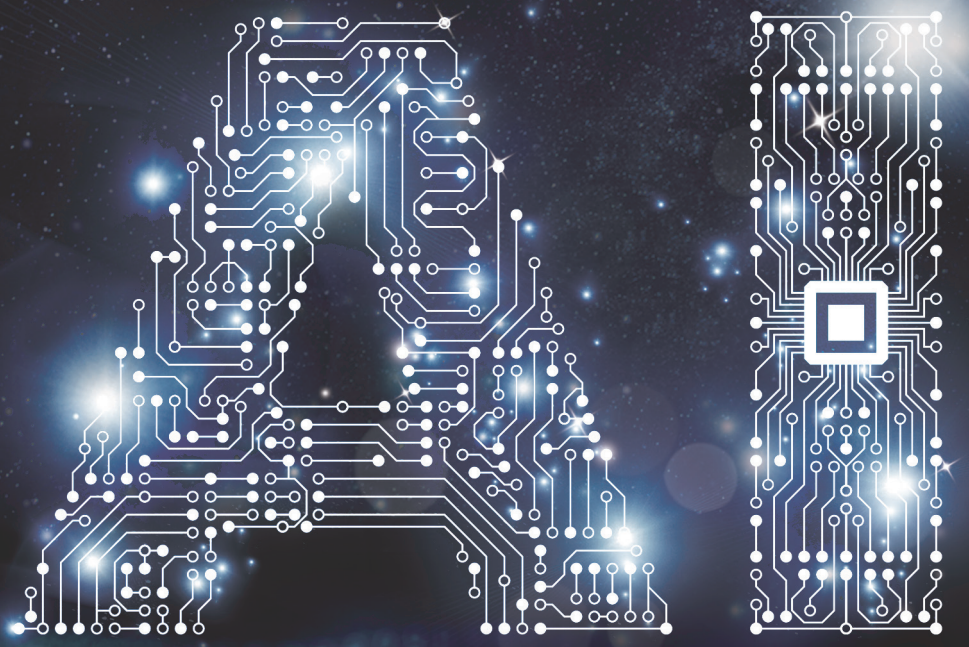
AI INSIGHT REPORT

VOL.03
2020. 10

AI 학습용 데이터
크라우드소싱 현황과 시사점

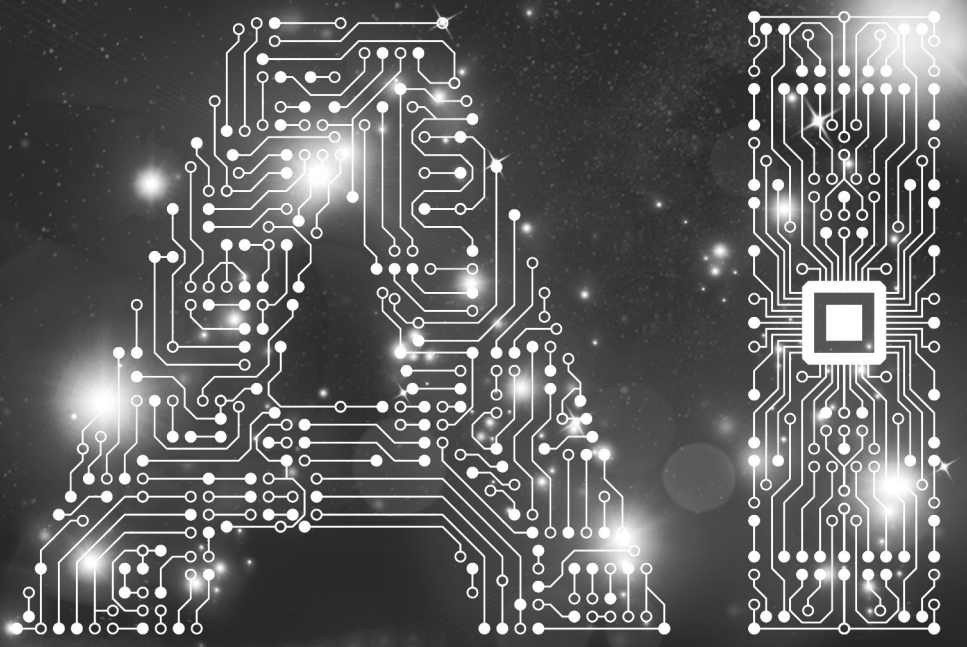


NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

AI INSIGHT REPORT

VOL.03
2020. 10

AI 학습용 데이터
크라우드소싱 현황과 시사점



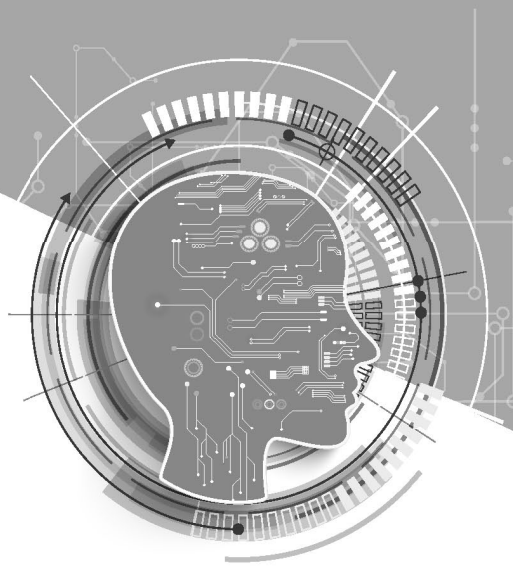
NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

AI INSIGHT REPORT

AI 학습용 데이터 크라우드소싱 현황과 시사점

- 해외 기업을 중심으로 -

제 3 호 (2020.10.)



Contents

NIA 한국정보화진흥원

AI Insight Report

- **"AI INSIGHT REPORT"**는 급변하는 인공지능산업의 기술, 서비스, 정책에 대한 시의성 있는 정보 제공과 심도 있는 분석을 위해 한국정보화진흥원에서 기획·발간하는 보고서입니다.
- 한국정보화진흥원의 승인 없이 본 보고서의 무단전제와 복제를 금하며, 인용 시에는 반드시 "한국정보화진흥원, 「AI INSIGHT REPORT」" 임을 밝혀주시기 바랍니다.
- 본 보고서의 내용은 한국정보화진흥원(NIA)의 공식 견해와 다를 수 있습니다.

▶ 작 성 AI데이터추진단 AI데이터기획팀
홍효진 수석(hhyoj@nia.or.kr)

▶ 기 획 박정은 단장, 신다울 팀장

▶ 발행인 문용식

▶ 보고서 온라인 서비스 www.nia.or.kr, www.aihub.or.kr
https://ko-kr.facebook.com/kict.bigdata

요 약 summary

◇ 크라우드소싱 개요

○ 전 세계적인 인공지능산업 성장과 데이터 가공 수요 증가에 따라, 대량의 데이터를 효율적으로 수집·가공하는 크라우드소싱 확대

* "크라우드소싱"은 대중(crowd)이 온라인 오픈 플랫폼을 통해 고객에게 의뢰 받은 데이터 수집·가공 작업을 수행하는 것을 의미

○ 크라우드소싱은 기계학습에 필요한 대량의 이미지, 텍스트, 오디오 데이터 구축에 주로 활용되며, 의료, 통·번역 등 다양한 산업에 적용

○ 글로벌 데이터 기업들은 주로 비용절감을 위해 크라우드소싱을 활용하지만, 저소득·저개발 국가의 일자리 창출에도 기여

◇ 주요 크라우드소싱 기업 현황

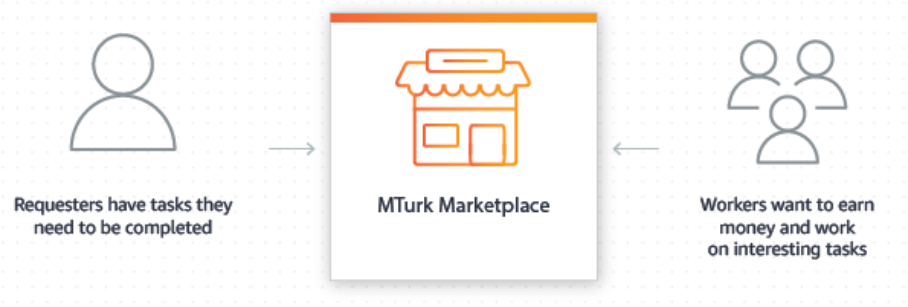
기업명	적용 분야	인력
Amazon Mechanical Turk (미국)	· 데이터 수집, 주석 달기 등 간단한 반복 작업	· 190개국 50만명의 크라우드소싱 인력 보유
CloudFactory (미국)	· 클라우드 플랫폼을 활용한 이미지 주석 처리	· 영국, 미국, 케냐, 네팔 등 5천명 이상의 인력 보유
Appen Limited. (호주)	· 컴퓨터 비전, 자연어 처리, Speech Data Transcription	· 전 세계 130개 이상 국가의 40만명 이상의 인력을 보유
Figure Eight (미국)	· 자연어 처리, 컴퓨터 비전, 데이터 증강 및 분류 * Appen에 합병('19)	· 외주 파트너를 별도로 두고 특정 언어 사용 가능 인력 또는 특정 국적의 인력 활용 · 1백만명 이상의 인력 보유
Samasource (미국)	· 컴퓨터비전, 자연어 처리 학습 및 검증과 관련된 작업을 주로 수행	· 디지털 경제(데이터 분야)를 통한 저소득층의 소득 창출을 위해 케냐, 우간다, 인도, 아이티 국적의 인력 11,480명 활용
Clickworker (독일)	· 모바일 크라우드소싱 플랫폼을 활용	· 전 세계 136개국 190만명 보유
MBH (중국)	· 영상과 이미지 가공 작업을 수행하며, 실시간으로 빠르게 작업	· 중국 내에서 경제적으로 낙후된 지역 중심으로 30만명의 인력을 소싱

Ⅰ 크라우드소싱 개요

□ 전 세계적인 인공지능산업 성장과 데이터 가공 수요 증가에 따라, 대량의 데이터를 효율적으로 수집·가공하는 크라우드소싱 확대

○ "크라우드소싱"은 대중(crowd)이 온라인 오픈 플랫폼을 통해 고객에게 의뢰 받은 데이터 수집·가공 작업을 수행하는 것을 의미

- 크라우드소싱 플랫폼 제공기업은 작업자를 모집하고, 작업(task)을 정의하며, 생산된 데이터를 검수·관리하며, 작업자에게 임금을 지급
- 크라우드소싱의 원조격인 Amazon Mechanical Turk는 라벨링 작업 플랫폼으로, 데이터 구축 과정을 최대한 세분화하여 단순 업무에 적용



< Amazon Mechanical Turk 크라우드소싱 과정 >

출처 : www.mturk.com

○ 전 세계 크라우드소싱 시장 규모는 '18년 95억달러에서 '27년 1,548억 달러로, 연평균 36.5%의 성장률을 보일 것으로 예상[1)]

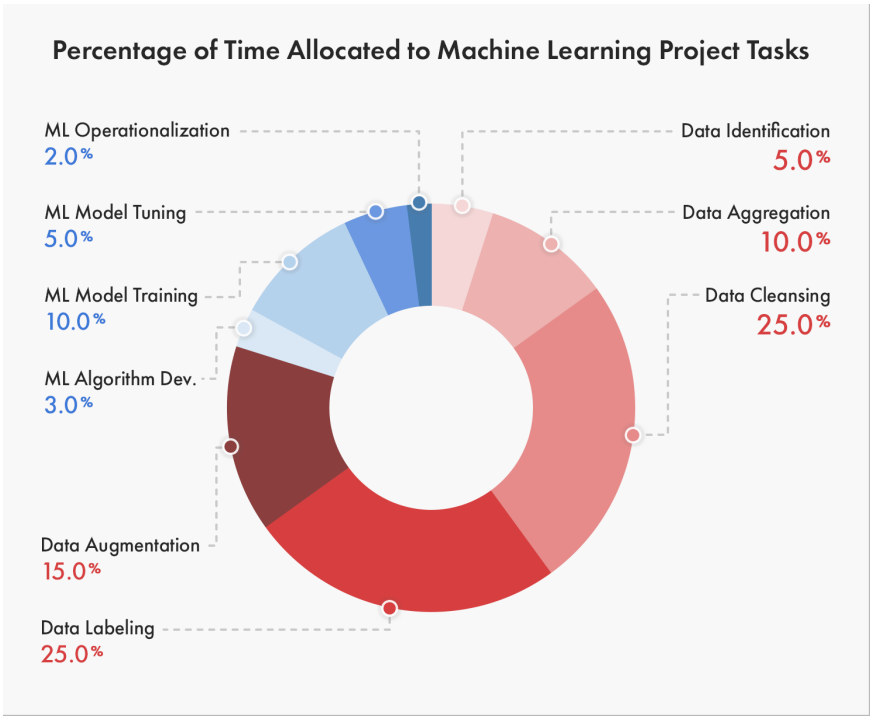
[참고1] 글로벌 데이터 가공 시장

- 글로벌 데이터 가공(AI Training Data) 시장은 '18년 3억천달러에서 '25년 16억1천달러로, 연평균 26.6%씩 성장할 것으로 전망[2)]
  - '18년, 지역별로는 북미 38.0%, 유럽 26.4%, 아시아 28.9%

1) Absolute Market Insights(2020.1), 'Crowdsourcing Market 2019-2027'
2) Grand View Research(2019), 'Data Annotation Tools Market Analysis'

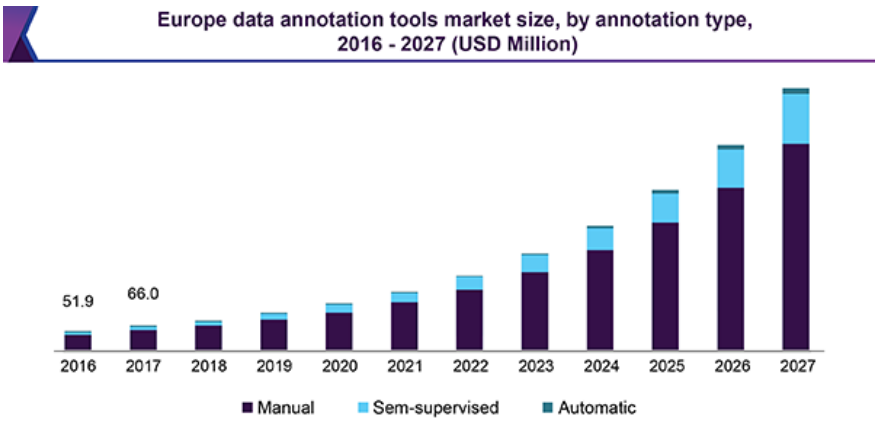
[참고2] 기계학습을 위한 데이터 작업

◇ 통상 기계학습에 소요되는 시간의 약 80%가 데이터 관련 작업이 차지하면서, 단순 업무를 크라우드소싱으로 대체



출처 : Cognilytica & CloudFactory

▸ 글로벌 데이터 시장은 사람이 데이터를 가공(Manual)하는 시장이 가장 크며, 반자동(Semi-supervised, 사람+시스템)과 AI자동가공(Automatic)은 미미



출처 : Grand View Research(2019)

□ 크라우드소싱은 기계학습에 필요한 대량의 이미지, 텍스트, 오디오 데이터 구축에 주로 활용되며, 의료, 통번역 등 다양한 산업에 적용

○ 아직은 Amazon Mechanical Turk처럼 데이터 구축 과정 중 단순 업무를 크라우드소싱하는 마이크로 테스킹 방식이 다수

- 크라우드소싱 인력은 단순 데이터 분류(수집)부터 이미지 주석 처리, 번역, AI모델 피드백, 자율주행 영상 데이터 라벨링 등을 수행
- 작업자 기초 훈련, 자격시험, 작업 난이도에 따른 임금 책정, 독립된 검사팀 등을 운영하여 품질 이슈 발생을 방지

○ 원천 데이터 수집과 단순 가공부터 일정 수준의 AI기술과 도메인 지식이 필요한 작업에 이르기까지, 크라우드소싱의 스펙이 광범위해지는 추세

- 의료, 법률, 심지어 AI전문가들도 크라우드소싱에 참여하는 사례도 등장

□ 글로벌 데이터 기업들은 주로 비용절감을 위해 크라우드소싱을 활용하지만, 저소득·저개발 국가의 일자리 창출에도 기여

○ 저소득·저개발 국가의 인력이나 사회적 약자를 크라우드소싱 인력으로 활용하는 데이터 가공 사회적 기업(美 Samasource)도 등장

- 미국, 독일, 호주, 중국의 데이터 기업들은 케냐, 우간다, 네팔, 인도, 아이티, 혹은 자국 내 낙후된 지역의 인력을 크라우드소싱에 활용

< 주요 기업의 크라우드소싱 인력 현황 >

기업명	국가(HQ)	크라우드소싱 인력
Samasource	미국	디지털 경제(데이터 분야)를 통한 저소득층의 소득 창출을 위해 케냐, 우간다, 인도, 아이티 국적의 인력 11,480명
CloudFactory		케냐, 네팔 등 저소득 국가의 인력 5천명 이상
Appen Limited.	호주	전 세계 130개 이상 국가의 180개의 외국어 작업이 가능한 40만명 이상
Clickworker	독일	전 세계 136개국 190만명(미국 35%, 독일 15%, 유럽 25%, 기타(캐나다, 호주, 남미 등) 25%
MBH	중국	중국 내에서 경제적으로 낙후된 지역 중심으로 30만명의 인력을 소싱

출처 : 각 사 홈페이지

Ⅱ 주요 크라우드소싱 기업 현황

1. Amazon Mechanical Turk, Inc. (미국)

□ 크라우드소싱 적용 분야

○ 데이터 수집, 주석 달기 등 간단한 수준의 반복 작업에 특화

○ HITL(Human-in-the-loop) 작업을 통해 고객의 머신러닝 모델에 대해 사람이 반복적으로 피드백을 주어 모델을 고도화시키는 작업을 지원

< AMT 적용 사례 >

고객사	주요 내용
WikiHow	방대한 양의 질의응답 데이터를 유관한 질문들로 분류
Baidu	AMT 플랫폼을 활용하여 음성 합성 데이터 테스트(다수의 작업자들로부터 데이터 테스트 피드백 수집)를 신속하게 진행
Pinterest	이미지 및 상품 분류 작업 수행
F&B 산업군	잠재소비자 수요 이해하기 위한 웹사이트 데이터 수집

출처 : 각 사 홈페이지

□ 크라우드소싱 인력 운영

○ 190개국의 50만명의 크라우드소싱 인력을 보유

○ 기본적으로 인터넷 작업을 할 수 있는 사람이면 누구나 작업 가능

○ 태스크 당 임금을 지급하는 모델을 통해 신속하게 인력 소싱 가능

○ AWS Software Development Kit를 통해 작업 의뢰자의 API를 지원하여, 프로그래밍 언어*를 선택해 최적의 맞춤 환경에서 작업 가능

* 안드로이드, 자바스크립트, iOS, 자바, NET, Node.js, PHP, 루비, 파이썬 등

□ 수수료 정책

○ 작업 의뢰자(고객社)가 작업별 임금을 설정하고, Amazon은 해당 임금의 20%의 수수료를 부과

○ 원하는 작업자(Premium Qualifications)를 모집하기 위해서는 작업자에게 추가 보상(작업료의 5%)을 지불

○ 특정 조건에 부합하는 인력을 활용할 경우, 132개의 조건(예. 성별, 흡연 여부, 정치 성향 등)별로 상이한 추가 수수료 부과

2. CloudFactory Limited. (미국)

□ 크라우드소싱 적용 분야

○ 클라우드 플랫폼*을 기반으로 하여, 데이터 분류 및 이미지 주석 달기, 웹 서치 등에 크라우드소싱을 주로 활용

* 코로나 사태에도 더 신속하게 서비스를 제공할 수 있다는 점을 홍보

○ 의료 이미지와 같은 특수 데이터에 주석을 다는 작업의 경우, 'train the trainer' 접근 방식을 통해 작업 의뢰자와 작업자가 피드백을 주고 받으며 데이터 작업 능력을 제고

□ 크라우드소싱 인력 운영

○ 영국, 미국, 케냐, 네팔에서 1,820명의 핵심 팀원과 5천명 이상의 크라우드소싱 인력을 보유 중이며, 주로 저소득 국가의 인력을 소싱

○ 구독제를 기반으로 요금이 부과되고, 구독 시간을 다양하게 설정 가능

- 지속적인 작업의 경우 월 최소 300시간을, 일회성 작업의 경우 월 최소 600시간 작업 가능

3. Appen Limited. (호주)

□ 크라우드소싱 적용 분야

○ 글로벌 기술 기업을 고객사로 두고 있으며, 데이터 분류, 주석 달기, 번역, 전사(Speech data Transcription)에 특화

○ 미국의 크라우드소싱 기업인 Figure Eight을 3억 달러에 인수('19년 3월)

□ 크라우드소싱 인력 운영

○ 전 세계 130개 이상 국가의 180개의 외국어 작업이 가능한 40만명 이상의 인력을 보유하여, 인력 구성이 다양하고 규모가 큰 것이 특징

- 호주, 미국, 영국, 필리핀, 중국에 오피스를 두고, 크라우드소싱으로 작업된 데이터를 검수하고, 고도화된 작업 수행 인력 600명을 채용

○ 업무에 참여하려면 자격시험을 통과해야 하고, 응시기회는 2회로 제한

○ 프로젝트 매니저가 의뢰자의 타임라인을 설정하고, 요구사항을 상세화하며, 임금은 작업 시간에 비례하여 책정

○ 크라우드소싱 윤리 원칙을 제정 : ①Fair Pay, ②Inclusion, ③ Crowd Voice, ④Privacy and Confidentiality, ⑤Communication, ⑥Well-being

4. Figure Eight Inc. (미국)

□ 크라우드소싱 적용 분야

○ '07년에 설립되어, '19년에 Appen Limited.에 인수

○ 텍스트, 자연어처리, 컴퓨터 비전, 데이터 증강, 데이터 분류에 중점

○ 일반적인 데이터 수집 등 기초 작업보다는 HITL(Human-in-the-loop)에 집중해 머신러닝 모델을 고도화시키기 업무에 주력

< Figure Eight 적용 사례 >

고객사	주요 내용
Adobe Stock	방대한 양의 이미지 내 영역 분리 작업 수행으로, 어도비 스톡의 사용자 검색 기반의 이미지 노출 머신러닝 모델을 고도화
eBay	상품분류 머신러닝 모델 디자인 초기부터 협력하고, 해당 모델을 기반으로 데이터가 얼마나 정확하게 리턴되는지 반복적인 피드백 제공

□ 크라우드소싱 인력 운영

○ 외주 파트너를 별도로 두고, 이들 인력에 대해 힌디어, 러시아어 사용자 또는 일본인, 독일인과 같은 국적 조건 설정 가능

○ 의뢰인은 데이터 보안 유지를 위해 기밀유지협약 하에 특수하게 진행하는 작업 옵션도 선택 가능

5. Samasource (미국)

□ 크라우드소싱 적용 분야

○ 컴퓨터비전, 자연어 처리 학습 및 검증과 관련된 작업을 주로 수행

○ 스타트업부터 Fortune 50대 기업에 이르는 다양한 규모의 고객社 확보

○ Vulcan(AI를 활용한 환경 보존 연구단체)과 파트너십을 맺고, 야생동물 사진 60만장에 대한 라벨링 작업을 하고 머신러닝 모델 구축에 참여

□ 크라우드소싱 인력 운영

○ 대표적인 임팩트 소싱(Impact Sourcing)* 기업으로, 케냐, 우간다, 인도, 아이티의 빈민 지역 주민 11,480명을 크라우드워커로 활용

* 기부나 후원 등 직접 원조 대신 안정적인 일자리를 제공함으로써, 자본주의 시스템 내에서 저소득 국가의 시민들이 스스로 경제적 자립 기반을 마련하고 인간 존엄을 회복하게 하려는 비즈니스 모델을 일컫는 용어[3)]

○ 소량의 쉬운 작업, 다소 복잡한 작업, 대량의 복잡한 작업(기업용)으로 가격 모델을 분류하여 요금을 부과

○ 작업자들은 월급 형태로 임금을 받고, 성과에 따라 인센티브도 수령

3) 한국일보(2020.3.9.), "레일라 자라, 공정·평등하게 …빈민국 노동자 모아 글로벌 IT기업 일구다"

6. Scale AI (미국)

□ 크라우드소싱 적용 분야

○ 자율주행차량의 사물인식모델에 필요한 데이터 구축에 강점이 있다고 피력하고 있으며, Lyft, NuScenes, Hesai와 구축한 오픈 데이터 3개 공개

○ 컴퓨터 비전과 자연어 처리를 위한 작업 외에도 최근에는 드론, 로봇에 활용되는 머신러닝 모델 고도화에 필요한 데이터 작업을 위주로 수행

□ 크라우드소싱 인력 운영

○ 작업 시간당 임금을 책정하며, 월별 최대 작업 수는 500개로 제한

○ 영업 상담 데이터 분류, 사물에 주석 처리하는 라벨링 작업을 수행

○ 수작업 → 통계 신뢰도 검증 → 머신러닝 모델 검증 등의 단계를 거쳐, 작업한 데이터의 정확성을 제고

7. Mighty AI (미국)

□ 크라우드소싱 적용 분야

○ '14년에 설립되어, '19년에 Uber에 인수

○ 자율주행차량 운행 데이터(사물추적, 의도분류, 도로·레인 등의 교통지표 주석 달기 등)와 관련된 모든 작업을 전문적으로 지원

○ 이커머스, 헬스케어 분야에도 크라우드소싱을 적용

□ 크라우드소싱 인력 운영

○ 차량 센서 데이터 라벨링을 전문으로 하는 인력 풀 40만명 이상 구축

8. Clickworker GmbH (독일)

□ 크라우드소싱 적용 분야

○ 이커머스, 패션업, 언론, 지식전달 서비스 영역에서 솔루션을 제공

○ 모바일 크라우드소싱 플랫폼을 통해 식당 메뉴, 가게나 건물의 전경 등에 대한 데이터 수집·가공

□ 크라우드소싱 인력 운영

○ 전 세계 136개국, 190만명의 대규모 크라우드 소싱 인력*을 보유

* USA 35%, 독일 15%, 유럽 25%, 이외(캐나다, 호주, 남미 등) 25%

○ 크라우드소싱 작업자들은 온라인 테스트 및 훈련을 거쳐 자격을 검증받아야 하며, 급수에 따른 임금 체계를 구축

9. MBH (중국)

□ 크라우드소싱 적용 분야

○ 의료, 안전, 도시 등 다양한 분야의 영상과 이미지 가공 작업을 수행하며, 실시간으로 빠르게 작업하는 것이 특징

* '17년에는 안면 인식 데이터를, '20년에는 AI진단 모델용 의학 데이터 작업을 주로 수행

○ 실제로 TikTok에 업로드된 영상 중 포르노로 구분될만한 것을 빠르게 판단하고 분류해, 거의 1초 내로 TikTok에 결과를 전달

□ 크라우드소싱 인력 운영

○ 중국 내에서 경제적으로 낙후된 지역 중심으로 30만명의 인력을 소싱

○ 작업자들은 하루에 6시간씩 안면 데이터, 의학 촬영본, 도시 촬영 이미지에 대한 태깅 작업을 수행

○ 아마존이 고객에게 상품을 추천할 때 사용하는 머신러닝 시스템과 동일한 것을 채택해, 작업자들에게 작업분을 효율적으로 빠르게 배정

10. Lionbridge AI (미국)

□ 크라우드소싱 적용 분야

○ 번역 작업에 특화되어 있으며, 언어 데이터 주석 처리, 언어처리 모델에 필요한 어휘집 개발, 언어 규칙 개발에 필요한 머신러닝 모델 컨설팅 서비스를 제공

○ 인공지능 비서의 14개 언어 지원이 가능한 음성인식 모델을 고도화

□ 크라우드소싱 인력 운영

○ 1백만명의 준전문가를 통해 300개 이상의 언어 텍스트 데이터(요약)와 손글씨와 같은 이미지 데이터를 수집

○ 번역·콘텐츠 생성, 조정, 감정 분석 서비스를 제공하는 AI플랫폼 보유

○ 한국에서 음성, 손글씨, 얼굴 샘플 데이터를 크라우드소싱으로 수집

11. Datapure (미국)

□ 크라우드소싱 적용 분야

○ 자율주행차량에 필요한 데이터 분류, 제거, 라벨링을 주로 수행

○ BCG, 스타벅스, Allergan, 지멘스 등 글로벌 기업을 고객사로 확보

□ 크라우드소싱 인력 운영

○ 200명 이상의 풀타임 크라우드소싱 인력을 보유

○ 작업자들은 약 1달간의 초기 훈련을 받고 작업에 투입

○ 작업 정확도를 높이기 위해, 독립된 검수 팀이 데이터를 전수 검사하고, 오류를 찾으면 인센티브를 지급

12. 기타

□ Playment (인도)

○ 자율주행차량의 사물인식 모델의 성능향상을 위한 데이터 구축 솔루션(라벨링 작업 플랫폼 및 라벨링 서비스)을 제공

○ 30만명의 크라우드소싱 인력을 보유

□ Defined crowd (미국)

○ 컴퓨터 비전, 자연어 처리를 위한 데이터 가공 작업을 주로 수행

○ 전 세계 10만명 이상의 숙련된 크라우드소싱 인력을 보유

□ Google (미국)

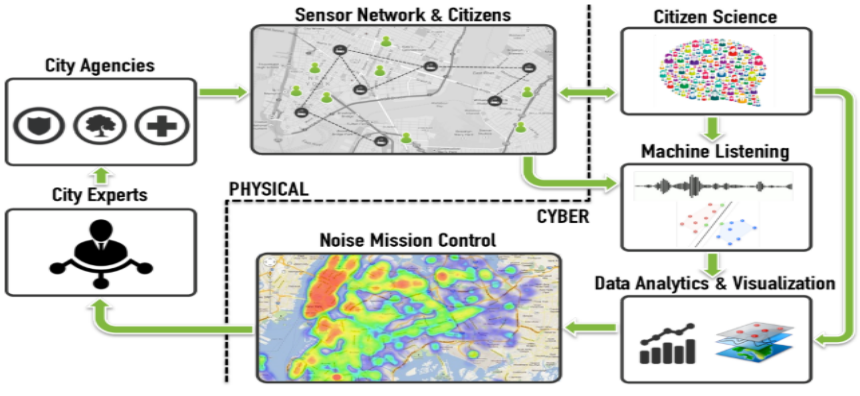
○ 번역, 지도, 검색 등 자사의 서비스 개선에 필요한 데이터 피드백 수행

○ 금전적 대가 없이 대중(crowd)이 라벨링 작업에 참여

[참고3] 미국 뉴욕주립대의 SONYC 데이터

▶ SONYC UST(Urban Sound Tagging) 데이터는 뉴욕시의 소음상황에 대한 사운드 레벨 데이터

▶ 뉴욕시 전역에 설치된 50여개의 센서에서 수집된 데이터를 크라우드소싱 데이터 가공 플랫폼인 Zooniverse(www.zooniverse.org)에서 시민참여를 통해 데이터 어노테이션을 실시



출처 : 뉴욕시 소음 데이터 클라우드 포털(https://wp.nyu.edu/sonyc/)

Ⅲ 결론 및 시사점

□ 해외 주요 데이터 기업들은 신속하고 편리한 크라우드소싱 적용과 데이터 품질 이슈 최소화를 위해 다양한 장치를 마련

○ 기술의 고도화와 다양화로 인한 기계학습 데이터에 대한 폭발적 수요에 대응하기 위해 크라우드소싱 방식을 적극 활용

- 다양한 환경에서 수집된 방대한 데이터가 기계학습의 성능을 좌우함에 따라, 주요 기업들은 크라우드소싱을 통해 데이터 구축 효율성을 제고

○ 대다수의 주요 기업들은 클라우드, 모바일 등 자사의 상황에 맞는 플랫폼을 통해 의뢰자와 작업자를 연결하고, 데이터 작업을 수행

○ 데이터 품질 이슈 발생 최소화를 위해 다양한 장치를 개발

- 간단한 자격시험 등을 통해 크라우드워커의 작업능력을 판단하고,
- HITL(Human-in-the-loop) 작업을 통해 데이터 검증에 크라우드소싱 활용

□ 체계적인 관리를 통해 광범위한 크라우드소싱 네트워크를 운영

○ 해외 주요 데이터 기업들은 다양한 작업에 언제 어디서나 참여할 수 있는 크라우드소싱 인력을 전 세계에 보유

○ 지역별 오피스나 외주 파트너를 두어 크라우드소싱 인력을 신속하게 확보

○ 크라우드소싱을 통해 저소득·저개발 국가의 일자리 창출에도 기여함으로써 비용절감과 사회공헌이라는 두 가지 목적을 달성

○ 다양한 스펙의 크라우드소싱 인력을 확보하고 조건별 임금 체계를 다양화함으로써 의뢰자(고객社)와 작업자의 양쪽의 만족도를 향상

○ 고객社가 임금을 책정하거나(AMT), 테스크당(AMT, Samasource), 시간당(Scale AI, Appen, MBH, CloudFactory, Clickworker)으로 지급하는 등 다양

AI INSIGHT REPORT

- 제1호(2019.9),「인공지능 현장에서 듣는 AI데이터 중장기 구축 방향」
- 제2호(2019.12),「인공지능기술을 통해 본 머신러닝 데이터 트렌드와 시사점」

AI INSIGHT REPORT VOL.03

AI 학습용 데이터
크라우드소싱 현황과 시사점

NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY