인공지능 학습용 데이터 플랫폼 연구

박영진
(주)크라우드웍스

요 약

인공지능 연구 및 개발은 모델(Model)을 중심으로 이루어져 왔다. 최근 인공지능 시스템의 성능을 높이는 것은 인공지능 모델의 코드를 개선하는 것보다는 인공지능 학습에 사용되는 데이터를 개선하는 것이 더 효과적이라는 데이터 중심 AI(Data centric AI)가 큰 주목을 받고 있다. 데이터는 인공지능 시대의 원유라고 부를 만큼 인공지능에 있어서 가장 중요한 요소 중 하나이다. 본고에서는 인공지능 학습을 위한 인공지능 학습용 데이터셋을 구축하기 위한 데이터 라벨링과 이를 위한 인공지능 학습용 데이터 플랫폼이 무엇인지에 대하여 살펴본다.

Ⅰ. 서 론

인공지능이 전체 산업 분야로 확산하면서 머신 러닝을 통하여 여러 가지 제품이나 서비스에 인공지능을 활용하게 되었다. 머신 러닝은 컴퓨터 시스템과 프로그램이 인간의 직접적인 도움이나 개입 없이 인간의 인지 과정과 유사한 방식으로 문제를 해결하기 위하여 학습한 예측 결과를 활용하는 것으로서 '코드로서 명령하지 않은 동작을 데이터로 학습하고 기계가 실행하도록 알고리즘을 개발하거나 연구하는 분야' 로 정의할 수 있다[1].

머신 러닝의 경우 지도 학습, 비지도 학습, 강화 학습으로 분류할 수 있으며, 지도 학습(Supervised learning)은 문제의 정답을 AI 모델에 입력해 AI 모델을 학습시키는 것으로서 정답이 포함된 학습용 데이터셋을 바탕으로 정답의 특징이나 규칙을 기계가 학습하여 예측과 분류, 추론 등을 수행하는 방식이다.

비지도 학습(Unsupervised learning)은 학습을 위한 데이터에 정답이 없고 주어진 데이터 내에서 AI 모델이 특징이나 패턴 등을 찾아내서 학습하는 방법으로서 군집화(clustering)나 차원 축소 등과 같이 특징이 비슷한 데이터를 묶거나 특징의 수를 줄이면서 꼭 필요한 특징을 포함한 데이터로 표현하는 방식이다.

강화 학습(Reinforcement learning)은 학습을 위한 데이터에

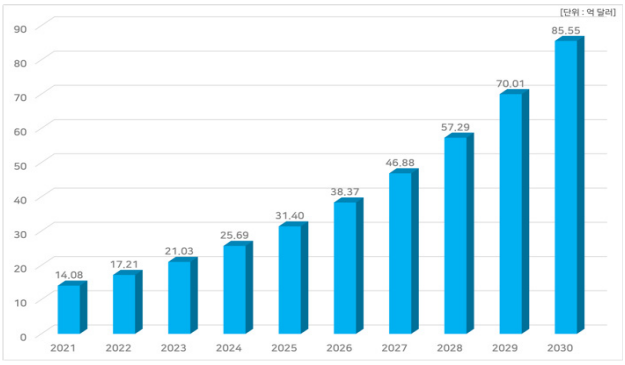

그림 1. 인공지능 학습용 데이터 시장 규모

정답은 없으나 동작하거나 반응하는 결과에 따라서 보상으로 해주는 방식으로 기계 스스로 진화할 수 있도록 하는 방식이다.

인공지능이 교통, 물류, 제조, 소매, 전자상거래, 헬스케어 등 여러 산업 분야에서 다양하게 활용되면서 방대한 데이터를 처리하고 특정 작업을 수행하기 위해서 인공지능 학습용 데이터를 통하여 AI 모델을 학습시키고, 더 많은 학습을 통해 인공지능이 수행해야 하는 객체 인식, OCR, 자동 번역, 해상 자율 주행, 도로 자율 주행 등 AI 모델을 학습하여 해결하고자 하는 비즈니스 문제나 목표 등 과업(task)을 더 잘 할 수 있게 된다.

산업 현장에 인공지능을 도입하기 위해서는 인공지능 학습을 위한 높은 품질의 데이터셋을 제공하는 것이 필수적이며, 고품질 데이터셋은 인공지능 성능 향상의 핵심 요소로서 인공지능이 다양한 분야에 적용되면서 해당 분야에 적합한 인공지능 학습용 데이터셋에 대한 수요가 폭발적으로 증가하고 있다.

글로벌 인공지능 학습용 데이터 시장 규모는 2021년 14억 달러 규모이며 2022년부터 2030년까지 연평균 성장률(CAGR) 22.2% 규모로 성장할 것으로 예상된다[2].

Ⅱ. 인공지능 학습용 데이터셋

머신 러닝 개발 시 데이터의 획득, 정제 및 가공 등의 사전 처

그림 2. MNIST 데이터셋

리를 위해 머신 러닝 개발에 필요한 전체 소요 시간 중 80%를 데이터셋 구축에 사용하고 실제 머신 러닝 모델 개발에 20%의 시간이 사용된다[3][4]. 인공지능 학습용 데이터셋을 구축하기 위해서는 데이터의 수집, 정제, 가공 등의 여러 단계를 거쳐서 많은 작업이 이루어지지 때문에 인공지능 개발에서 학습용 데이터셋 구축은 많은 리소스가 필요한 업무 중 하나이다[5].

일반적인 데이터 생애주기는 수집 → 저장 → 가공 → 유통 → 활용으로 이루어지며, 인공지능 학습용 데이터의 경우 〈그림 3〉과 같이 계획 → 구축(수집+저장+가공) → 활용(유통) 순으로 이루어진다. 인공지능 학습용 데이터를 AI 모델 학습에 적용하고 학습 결과를 평가하여 다시 인공지능 학습용 데이터를 설계하는 데 반영되어 순환하는 형태로 이루어진다.

머신 러닝에서 지도 학습은 정답을 미리 알려주고 문제를 풀게 하는 것이다. 정답이 포함된 데이터를 바탕으로 정답을 내는 규칙을 학습해서 예측과 분류, 추론 등을 수행한다. 분류와 회귀 같은 지도 학습이나 추천 시스템에서 높은 성능을 내려면 정답 정보가 포함된 데이터나 말뭉치, 사전처럼 양질의 학습용 데이터가 많이 필요하다[6].

다양한 분야에서 인공지능에서 성과를 거두게 된 것은 AI 모델과 알고리즘의 발전과 이를 뒷받침하는 충분한 컴퓨팅 파워와 리소스의 확보가 기능해졌기 때문이지만, 또한 대량의 인공지능 학습용 데이터셋이 있어서 가능했던 일이기도 하다. 그런데, 인공지능 연구 개발이나 비즈니스 환경 등에서 인공지능을 도입할 때 일어나는 주요 병목 현상은 더 이상 알고리즘이나 하드웨어가 아니라 품질이 우수한 인공지능 학습용 데이터셋을 확보하는 것이다[7].

인공지능을 위한 학습용 데이터셋은 인공지능 모델을 학습하는데 사용되는 훈련 데이터(training data), 훈련 데이터로 만들어진 인공지능 모델의 성능을 측정하기 위해 사용되는 검정 데이터(validation data) 그리고 여러 번의 학습을 통하여 만들어진 인공지능 모델의 최종 성능을 시험하기 위한 시험 데이터(test data) 셋으로 구성된다.

대표적인 컴퓨터 비전 인공지능 학습용 데이터셋인 MNIST의 경우 손 글씨 인식을 위한 머신 러닝에 사용되며 손으로 쓴 숫자 이미지와 해당 이미지에 대한 숫자 라벨로 구성되어 있다. 글씨 크기는 고정 크기 이미지 (28x28픽셀)로 표준화되어 중앙에 위치하고 학습용 데이터 55,000개, 검정용 데이터 5,000개, 테스트용 데이터 10,000개 등 총 60,000개로 구성되어 있다.

일반적인 인공지능 학습용 데이터셋은 원본 데이터와 어노테이션 데이터로 구성되어져 있으며, 데이터셋에서 데이터 라벨링 결과 추출을 위해 흔히 사용되는 포맷은 CSV, COCO, PASCAL VOC, YOLO 등이 있으나 COCO, PASCAL VOC, YOLO포맷은 특정 목적의 데이터셋을 배포하기 위해 임의로 정의된 포맷으로서 다른 유형의 데이터를 표현할 수 없다[8].

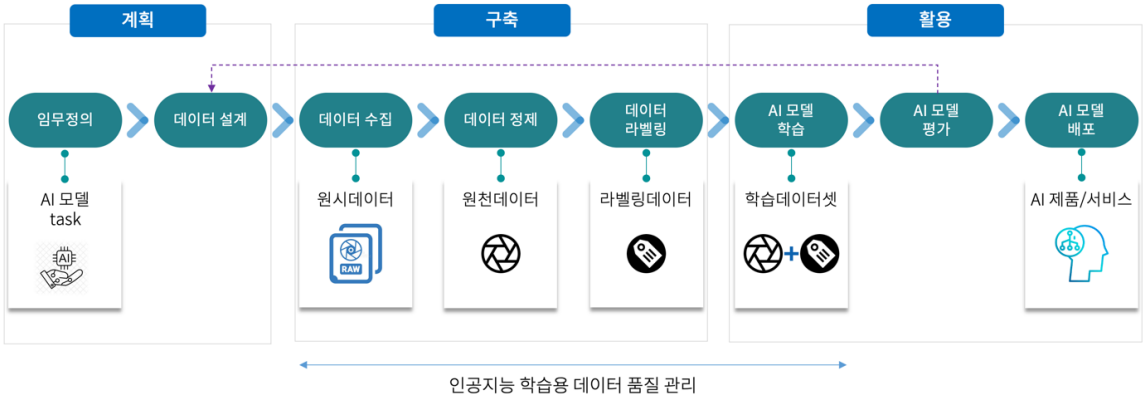

그림 3. 인공지능 학습용 데이터 생애주기

COCO, PASCAL VOC, YOLO 등의 어노테이션 정보는 Json으로 되어 있으며, 해당 Json 파일 구조는 1)데이터셋 정보 2)이미지정보 3)어노테이션 정보 그리고 라이센스가 있는 경우 4) 라이센스 정보로 구성되어 있다

국내외에는 다양한 인공지능 학습용 데이터셋이 공개되어 있으며, 국내의 경우 한국지능정보사회진흥원(NIA)에서 운영하는 AI Hub(https://www.aihub.or.kr)를 통하여 381종의 인공지능 학습용 데이터를 공개하고 있으며, 해외에서는 다음과 같이 인공지능 학습용 데이터셋을 모아 놓거나 검색할 수 있다.

- Google Dataset Search
  - 2,500만종 데이터셋
  - https://toolbox.google.com/datasetsearch
- Kaggle
  - 15만종 데이터셋과 AI 및 데이터 분석 각종 경진대회 진행
  - https://www.kaggle.com/
- Papers with Code
  - 6,288종 데이터셋과 이를 활용한 논문과 소스 코드 공개
  - https://paperswithcode.com/
- IEEE DataPort
  - 3,000개 이상의 데이터셋, 연구용 데이터셋 저장, 공유, 검색, 활용
  - ORCID와 통합, ORCID 자산 목록에 자동 추가
  - 데이터셋별로 DOI 제공
  - https://ieee-dataport.org/
- VisualData
  - 827종 데이터세트 (Computer Vision Datasets)
  - https://visualdata.io/discovery
- UC Irvine Machine Learning Repository
  - 622종 데이터세트
  - https://archive.ics.uci.edu/ml/index.php

Ⅲ. 데이터 라벨링

안면 인식이나 이메일 스팸 분류 시스템 등에 기계 학습 모델을 적용하기 위해서는 얼굴 이미지에 이름 등이 레이블 되어 있거나 이메일에 스팸 여부 등이 레이블 되어 있는 기계 학습용 데이터셋이 필요하며, 머신 러닝이나 딥러닝 모델의 학습 전에 데이터에 특정 값을 부여해 주는 것을 데이터 라벨링(Data labeling)이라고 한다[9].

데이터 라벨링은 데이터나 데이터의 메타 정보에 라벨(Label)이나 태그를 어노테이션(Annotation)하여 AI 모델이 해당 정보

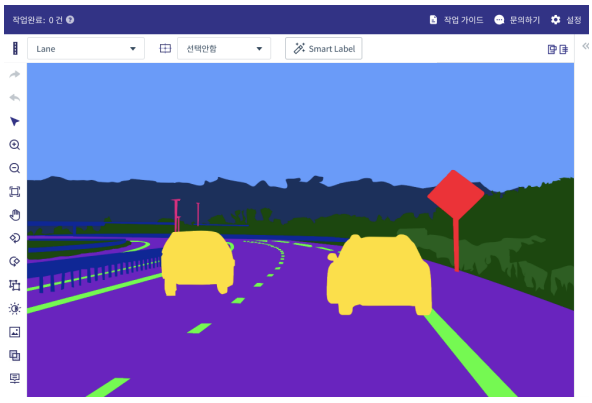

그림 4. 도로 데이터 시맨틱 세그멘테이션

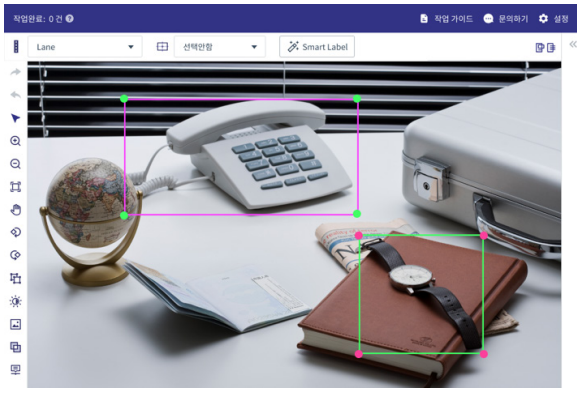

그림 5. 객체 바운딩 박스

를 이해할 수 있도록 하는 것으로서 비정형 또는 반정형 데이터를 수집하고 참값(GT, Ground Truth)을 어노테이션하여, 지도학습(Supervised Learning)에 쓰이는 데이터를 구축하는 것이다.

예를 들어 개와 고양이를 분류하는 AI 모델에 데이터를 설명하기 위해 개와 고양이 등이 포함된 이미지 데이터에 바운딩 박스나 폴리곤 등의 작업을 통해 객체를 구분할 수 있는 메타 데이터를 원본 데이터에 추가하는 작업으로서 기계가 데이터를 이해할 수 있도록 원본 데이터에 주석(Annotation)을 다는 작업으로서 전체 머신 러닝 학습 과정 중 일부로서 머신 러닝을 통해 해결해야 하는 문제에 따라서 데이터 라벨링 방식이 정해진다.

데이터 라벨링 작업은 이미지, 영상, 텍스트, 사운드 등의 원시 데이터에 작업자들이 데이터 가공 도구(Annotation Tool)를 활용하여 인공지능 학습에 필요한 다양한 정보를 목적에 맞게 입력하는 것으로서 데이터를 수집, 정제, 가공하여 인공지능 학습용 데이터셋을 만드는 작업이다.

예를 들면 이미지에서 객체를 인식하고 분류하기 위해서 〈그림 4〉와 같이 데이터에 포함된 객체에 박스를 표시하거나 〈그림 5〉

와 같이 도로 데이터에 시맨틱 세그멘테이션 등과 같은 데이터 라벨링을 수행하며, 객체 윤곽선에 따라 점을 찍는 작업인 랜드마크, 문서에서 특정 영역을 매핑하는 태깅, 음성/동영상 등 연속적인 소스에서 구간 추출, 관절, 자세 등 특정 포인트를 연결하는 키포인트 추출 등 인공지능 모델의 과업(task)를 위해서 데이터에 레이블을 추가하는 다양한 방법이 있다.

데이터 라벨링은 머신 러닝을 통해 해결해야 하는 AI 모델의 과업(task), 관련된 인력 등에 따라서 라벨링 수행 방식이 결정되며, 일반적으로 다음과 같이 5가지 방식으로 데이터 라벨링이 이루어진다[10].

- In-house : 머신 러닝 개발을 담당하는 기업/기관 내부 인력을 활용하여 데이터 라벨링이 이루어져서 작업에 대한 모니터링과 추적이 간단하고 높은 정확도와 품질을 보장할 수 있으나 내부 리소스의 활용으로 많은 비용과 시간이 소요된다.
- Synthetic : 의료 정보, 라이프 로그 등과 같이 직접 확보하기 어려운 데이터로서 실제 데이터(Real Data)와 유사한 데이터를 컴퓨터 시뮬레이션이나 알고리즘 등을 통해 생성하는 데이터로서 인공적인 데이터이지만 수학적 모델이나 통계적 방법 등을 통해 실제 데이터의 특성을 반영하여 실제 데이터 이상으로 AI 모델 학습에 뛰어난 성능을 보여준다.
- Programmatic Labeling : Auto labeling이라고 하기도 하며 AI 모델 등을 활용한 자동화된 라벨링으로 빠르게 작업 수행이 가능하지만 적용할 수 있는 데이터의 유형이나 데이터 라벨링 방식이 아직은 한정적이며 사전에 AI 모델 등을 개발해야 하는 제약 사항이 있다.
- Outsourcing : In-house와 Crowdsourcing의 중간에 해당하며, 주로 전문적으로 데이터 라벨링 업무를 담당하는 인력을 확보한 외부 기업에 용역으로 발주하여 정해진 기간 동안 목표로 하는 데이터 라벨링 작업을 수행하는 방법이다. 헬스케어 등과 같이 특정 분야의 전문가나 데이터에 대한 구축 경험을 갖춘 기업을 활용하는 방식이다.
- Crowdsourcing : 크라우드소싱(Crowdsourcing)이라는 말은 제프 하우가 2006년 와이어드 잡지에 기고한 'The Rise of Crowdsourcing'이라는 기사에서 처음 사용한 용어로서, 대중을 의미하는 크라우드(Crowd)와 기업 외부의 인적자원을 활용하는 아웃소싱(Outsourcing)을 결합한 표현으로 크라우드소싱 플랫폼을 통하여 데이터 라벨링 작업을 수행하는 방식이다

머신 러닝은 사람에 의한 AI 모델 결과에 대한 지속적인 감독과 검증이 필요하다. 이러한 사람과 AI 모델 간의 인터랙션을 HITL(Human-in-the-Loop)이라고 부르는데, 머신 러닝에서 HITL은 크게 두 가지로 적용된다. 하나는 머신 러닝을 위한 AI 모델의 학습용 데이터를 사람이 라벨링하는 것이고 또 다른 하나는 AI 모델에 대한 검증과 튜닝이다. 데이터 라벨링에서 가장 많이 사용되는 방식 중 하나인 크라우드소싱 기반 인공지능 학습용 데이터 구축 프로세스는 〈그림 6〉과 같다. 데이터 라벨링 계약 이후 구체적인 데이터 라벨링 방법에 대한 작업 가이드를 수립하고 작업자에 대한 요건을 통해 작업 가능한 인력을 필터링하고 크라우드소싱 플랫폼 내에 작업을 개설하여 작업자들이 참여하여 데이터 어노테이션 작업을 수행하고 가공된 데이터에 대한 검수를 통하여 작업 가이드 기준에 맞지 않는 결과물을 반려하고 이를 작업자가 재가공하며, 검수 완료된 데이터를 수요기업에 공급함으로써 데이터 라벨링 프로젝트가 종료된다.

Ⅳ. 인공지능 학습용 데이터 플랫폼

인공지능 학습용 데이터 플랫폼은 다양한 유형의 데이터셋에 태그를 지정하거나 레이블을 지정할 수 있는 플랫폼으로서 원시

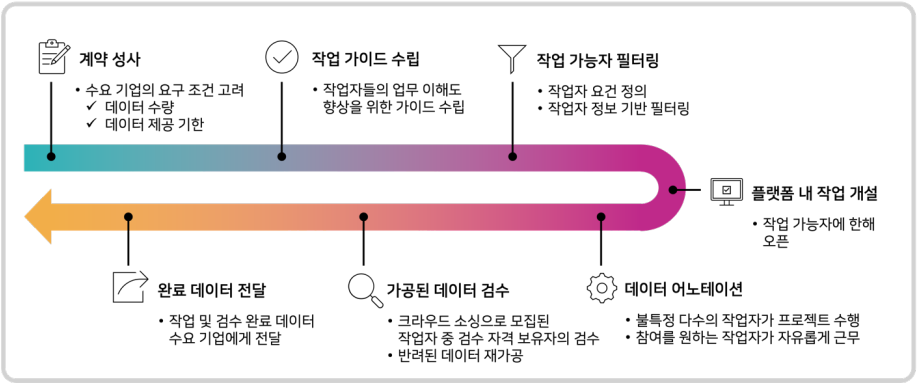

그림 6. 크라우드소싱 기반의 데이터 라벨링 프로세스

그림 7. 인공지능 학습용 데이터 플랫폼 구성도

데이터를 AI 모델에서 학습에 사용할 수 있도록 학습용 데이터를 만들 수 있도록 하는 시스템이며 온프라미스(On-premise) 또는 클라우드(Cloud) 기반으로 제공된다. AI 모델에 대한 높은 품질의 학습용 데이터를 만들기 위해서 많은 기업이나 기관들이 인공지능 학습용 데이터 구축 전문 외부 벤더에 의존하지만, 일부는 시장에서 사용 가능한 프리웨어 또는 오픈 소스 툴을 기반으로 하거나 맞춤형으로 구축된 자체 툴을 통하여 자사에 필요한 데이터를 처리하도록 구축되었다[12]

인공지능 학습용 데이터 플랫폼은 인공지능 학습을 위한 데이터셋을 만들기 위하여 이미지, 텍스트, 영상, 음성 등의 원천 데이터에 레이블을 어노테이션(annotation)하는 데이터 플랫폼으로서, 작업자(Workforce) 관리, 워크 플로우(flow) 관리, 어노테이션 편집기, 데이터셋 관리, 데이터 품질 관리 및 보안, 외부 연동 API 등의 기능으로 구성되어 있으며 〈그림 7〉과 같이 정리할 수 있다.

- Workforce Management : 데이터 라벨링을 담당하는 인력에 대한 관리 기능으로서 작업자 선발/운영/관리 등의 작업자 관리 기능으로서 데이터 라벨링은 인간의 작업에 의해 이루어지기 때문에 데이터 라벨링을 위한 작업자를 확보해야 하며 이를 위하여 내부 인력을 활용하거나 전문 인력을 계약을 통해 확보하여 투입하거나(Outsourcing) 일반 대중을 활용한 크라우드소싱(Crowdsorucing) 방식으로 작업자를 활용
- Workflow Management : 데이터 수집-정제-가공-검수까지 데이터 라벨링 전주기에 걸쳐서 데이터 라벨링 작업 프로세스 관리 기능
- Annotation Tool : 데이터 태깅이나 라벨을 작업하기 위한 편집 기능으로서 이미지, 텍스트, 영상, 오디오 등의 다양한 데이터 포맷을 지원함
- Dataset Management : 데이터셋의 전처리부터 업로드/다운로드, 백업 등의 관리

인공지능 학습용 데이터 라벨링 플랫폼은 적용 방식, 상용 여부, 인력 관리 방식 등의 기준에 따라 다음과 같이 분류할 수 있다.

인공지능 학습용 데이터 라벨링 플랫폼은 배포 방식(Deployment Model)에 따라 다음과 같이 분류할 수 있다.

- On-Premise : 데이터의 보안 등의 이유로 데이터 외부 유출이 제약이 있는 경우 또는, AI 모델이나 응용 시나리오가 일반적이지 않고 특수한 상황일 때 데이터 라벨링 시스템을 별도로 구축하여 특정 목적에 최적화된 데이터 라벨링을 할 수 있도록 데이터 플랫폼을 제공하는 방식이다.
- SaaS : 가장 일반적인 인공지능 학습용 데이터 플랫폼 형태로서 클라우드(Cloud) 상에 데이터 라벨링 서비스를 구축하여 제공하는 형식으로 데이터 라벨링 관련 주요 기능을 제공하고 있으며, 데이터 수집/정제부터 가공/검수 및 AI 모델 적용까지 일괄 제공되며, 고객의 요구에 맞춰 필요한 서비스만 사용할 수 있도록 제공하고 있으며, 구글, AWS, MS, IBM, SuperAnnotate, V7, Appen, Labelbox, Hive Data, 크라우드웍스, 에이모, 테스트웍스 등 국내외 중소기업부터 대기업까지 다양한 공급 업체가 있으며, 데이터 라벨링 관련 서비스 를 제공하고 있다.

머신 러닝에 있어서 품질 낮은 데이터를 사용하면 학습이 잘못되어 AI 모델이 부정확해지고 AI 모델 학습 시간이 길어지고 AI 모델의 결과가 나빠질 수 있으며, 반면에 정확하고 품질이 높은 학습용 데이터로 학습된 AI 모델은 데이터에 숨겨진 패턴을 식별하고 높은 정확도로 예측을 제공하는 모델을 생성할 수 있다. 하지만, 많은 AI 프로젝트는 지속적으로 고품질의 데이터를 활용할 수 없기 때문에 중단된다[12].

인공지능 학습용 데이터 품질이란, "데이터 품질 확보를 위한 품질관리 측면과 인공지능 학습에 필요한 데이터 자체의 품질을 확보하여 사용자에게 유용한 가치를 줄 수 있는 수준"이라고 정의할 수 있으며, 인공지능 학습용 데이터 품질관리는 인공지능 학습용 데이터 품질을 확보하는 데 필요한 조직, 절차, 품질기준,

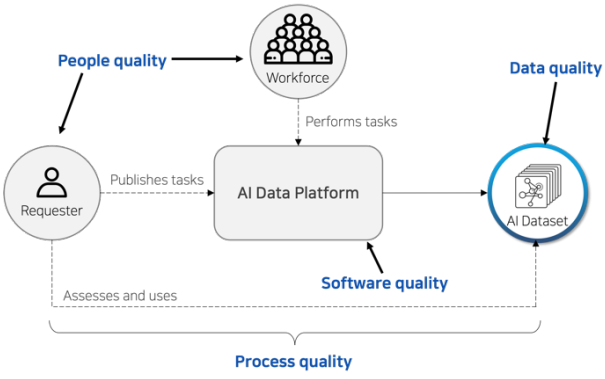

그림 8. AI 학습용 데이터 품질 프레임워크[14] 재작성

품질관리 방법이나 활동 등을 정의하여 점검하고 조치하는 일련의 활동이다[13].

또한, 인공지능 학습용 데이터 품질은 〈그림 8〉과 같이 라벨링 데이터의 품질(Data quality), 데이터 라벨링 과정의 품질(Process quality), 크라우드소싱 플랫폼의 품질(Software quality), 인력의 품질(People quality)로 구분할 수 있다[14].

데이터 품질은 데이터 라벨링의 정확도, 라벨링의 일관성, 객체 다양성 등으로 이루어지며, 데이터 라벨링 프로세스 품질은 데이터 수집, 데이터 정제, 데이터 라벨링 등 데이터 라벨링 과정에서 데이터 품질을 보장하기 위하여 품질관리 관점에서 프로세스를 수행하는지 모니터링하고, 발견된 문제점을 보완한다.

인력의 품질은 작업 요청자의 경우 학습용 데이터셋의 작업 기준과 방식에 대하여 얼마나 이해하고 있는지가 품질에 영향을 미치며, 많은 크라우드소싱 기반 데이터 라벨링 기업에서는 라벨링 데이터의 품질에 직접적인 영향을 미치는 작업자들을 관리하기 위하여 작업의 정확도, 작업자의 성실성 등을 플랫폼을 통하여 측정하고 이 결과를 작업자 선발 및 운영에 반영한다.

또한, 낮은 숙련도와 도메인 지식이 전혀 없는 일반인들에게 작업자로서 기본적인 교육과정을 제공하고 이를 통하여 숙련도를 높이고 도메인 지식을 향상시켜 데이터 품질을 높이도록 해야 한다.

V. 결 론

본 고에서는 일반적인 인공지능 학습용 데이터 관점에서 가장 핵심적인 부분인 데이터 플랫폼에 대해 설명하였다.

인공지능에서 데이터는 한번 구축하면 영구적으로 사용할 수 있는 것이 아니라 데이터 수집, 데이터 전처리, 피처 추출 및 엔지니어링, 학습용 데이터를 통한 모델 훈련, 모델 테스트 및 검증 등의 단계로 구성되며 모든 단계를 여러 번 반복하는 사이클로 구성된다.

이러한 머신 러닝의 지속적인 배포 및 자동화 파이프라인으로서 필요한 데이터 수집부터 AI 제품/서비스의 배포(Deployment) 까지 높은 품질의 데이터를 공급할 수 있도록 하는 것을 MLOps(Machine Learning Operations)라고 부르며 AI 선도 기업들을 중심으로 도입이 이루어지고 있다.

MLOps를 통해서 AI 시스템 런칭 이후 지속적으로 시스템으로 들어오는 데이터를 수집하고 분석하여 다시 모델을 개선하는 사이클을 돌리면서 일관성 있는 품질이 우수한 데이터를 확보하고 지속적으로 AI 시스템의 성능을 유지하고 안정적인 운영을 보장하고 있다.

최근 인공지능 산업에서 실질적으로 AI 시스템의 성능을 높이는 것은 AI 모델과 알고리즘의 모델 구조, 하이퍼파라미터 튜닝 등 코드의 개선이 아니라 데이터에 대한 개선이라고 하는 데이터 중심 AI(Data Centric AI)로 데이터 품질에 대한 중요성이 강조되고 있다[15].

또한, MLOps와 데이터 중심 AI를 위해서는 기존 인공지능 학습용 데이터셋을 재활용하고 품질이 높은 인공지능 학습용 데이터셋을 확보하기 위해서는 표준화된 구축 체계와 데이터셋이 그 무엇보다 중요하며, 이를 위해서 인공지능 학습용 데이터 플랫폼의 표준화 그리고 이와 관련된 품질 기준에 대한 표준화가 시급한 상황이다.

또한, 인공지능 학습용 데이터셋의 품질을 확보하고 지속적으로 품질 수준을 유지하기 위해서 필요한 품질 기준이나 검증 방안 등과 같은 인공지능 학습용 데이터 품질 체계와 관련된 표준화 연구 및 활동이 매우 필요한 시점이다.

참 고 문 헌

[1] 박대륜, 안중민, 장준혁, 유원진, 김우열, 배영권, 유인환 (2020). 머신 러닝 플랫폼을 활용한 소프트웨어 교수-학습 모형 개발. 정보교육학회논문지, 제24권 제1호, 49-57.

[2] Grand View Research, AI Training Data Services Market, 2021

[3] Cognilytica Research, "Data Preparation & Labeling for AI 2020," tech. rep., Cognilytica Research, 2020.

[4] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: a big data-ai integration perspective," IEEE Transactions on Knowledge and Data Engineering, 2019.

[5] 최정열. (2018). 기계학습 활용을 위한 학습 데이터세트 구축 표준화 방안에 관한 연구. 디지털융복합연구, 16(10), 205-212.

[6] Shinta Nakayama, Michiaki Ariga, Takashi Nishibayashi(2018), Machine Learning at Work, O'Reilly

[7] Chew, R., Wenger, M., Kery, C., Nance, J., Richards, K., Hadley, E., & Baumgartner, P. (2019). SMART: an open source data labeling platform for supervised learning. The Journal of Machine Learning Research, 20(1), 2999-3003.

[8] 이형주 (2022) 머신 러닝 라벨링 데이터를 위한 메타정보 설

계, OSIA Standards & Technology Review, 35:1, 20-23
[9] 김지훈, 이정호, 김태윤, & 김재현. (2020). 골프 스윙 자세 교정을 위한 Faster R-CNN 기반 머리 인식 모델 설계. 한국통신학회 학술대회논문집, 509-510.
[10] IBM Cloud Education, 2021, https://www.ibm.com/cloud/learn/data-labeling
[11] shaip, 2021.08, Buyer's Guide for Data Annotation and Data Labeling,
[12] Gartner, 2021. 12, Three Steps to Boost Data for AI, Gartner Research
[13] 한국지능정보사회진흥원, 2022.02, 인공지능 학습용 데이터 품질관리 가이드라인 v2.0
[14] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B., & Allahbakhsh, M. (2018). Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys (CSUR), 51(1), 1-40.
[15] Andrew Ng, MIT Technology Review 21.04.01

약 력

박 영 진

1998년 대구대학교 공학사
2018년 한양대학교 기술경영학 석사
2020년 한양대학교 기술경영학 박사 수료
1998년~1999년 까치네 대표
2000년~2000년 ㈜e-현대백화점 검색팀장
2001년~2007년 ㈜다음소프트 검색팀장
2007년~2010년 ㈜와이즈넛 BI사업부장
2010년~2016년 ㈜사이냅소프트 전략기획이사
2016년~2018년 ㈜르호봇비즈니스인큐베이터 DX담당 이사
2018년~2020년 아이더링크랩 대표
2020년~2020년 ㈜어반데이터랩 CTO
2021년~현재 ㈜크라우드웍스 본부장
관심분야: 인공지능 학습용 데이터 구축 및 검증/품질관리, MLOps