https://doi.org/10.33778/kcsa.2023.23.2.055

비정형데이터의 AI학습을 위한 영상/이미지 데이터 품질 향상 방법

김 승 희*, 류 동 주**

요 약

최근 전세계적으로 사회 모든 분야에서 인공지능 학습용 데이터에 관한 선행연구를 기반으로, 인공지능 학습용 데이터의 가치를 높이고 고품질 데이터를 확보하고자 하는 움직임이 늘고 있다. 따라서, 고품질 데이터를 확보하기 위한 구축사업에서는 품질관리가 매우 중요하다. 이에, 본 논문에서는 인공지능 학습용 데이터를 구축할 시 고품질 데이터 확보를 위한 품질관리와 그에 따른 구축공정별 개선방안을 제시하였다. 특히, 인공지능 학습을 위해 구축되는 비정형데이터는 데이터 품질의 80% 이상이 구축과정에서 결정된다. 본 논문에서는 비정형데이터 이미지/영상 데이터에 대한 품질검사를 통해 구축단계에서의 획득, data cleaning, labeling 모델에서 발생된 검사절차 및 문제요소를 해결함으로써 고품질 데이터 확보 방안을 제시하였으며, 제시한 방안을 토대로 인공지능 학습용 데이터 구축에 참여하는 연구단체와 사업자들에게 데이터의 품질편차를 극복하기 위한 대안이 될 것으로 기대된다.

Method for improving video/image data quality for AI learning of unstructured data

Kim Seung Hee*, Dongju Ryu**

ABSTRACT

Recently, there is an increasing movement to increase the value of AI learning data and to secure high-quality data based on previous research on AI learning data in all areas of society. Therefore, quality management is very important in construction projects to secure high-quality data. In this paper, quality management to secure high-quality data when building AI learning data and improvement plans for each construction process are presented. In particular, more than 80% of the data quality of unstructured data built for AI learning is determined during the construction process. In this paper, we performed quality inspection of image/video data. In addition, we identified inspection procedures and problem elements that occurred in the construction phases of acquisition, data cleaning, labeling, and models, and suggested ways to secure high-quality data by solving them. Through this, it is expected that it will be an alternative to overcome the quality deviation of data for research groups and operators participating in the construction of AI learning data.

Key words : AI, Data Quality, learning data, data for training, machine learning, quality inspection

접수일(2023년05월 17일), 수정일(2023년 06월 01일), 게재확정일(2023년 06월 12일)

* 극동대학교/인공지능 보안학과 박사과정(주저자)

** 극동대학교/인공지능 보안학과(교신저자)

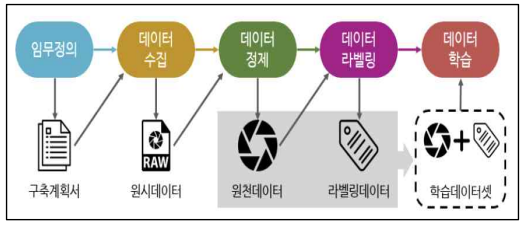
1. 서 론

디지털 전환의 핵심 기술인 인공지능 기술이 최근 몇 년 사이 급부상하며 미래의 유망기술로 떠오르고 있다. 인공지능(Artificial Intelligence 이하 AI)이란 사고나 학습 등 인간이 가진 지적 능력을 컴퓨터를 통해 사람이 하는 것처럼 생각하고 행동할 수 있게 구현하는 기술이다. 이에 대한 인공지능 연구도 활발해졌으며 인공지능 관련 품질관리, 동영상 분석, 음성문자 변환(자연어 처리), 자율주행뿐 아니라 언론, 교통, 환경, 의료, 제조, 금융 서비스, 엔터테인먼트 분야 등 각종 분야에서 기술이 빠르게 접목·확산 되면서 인공지능 학습용 데이터 산업도 성장하고 있다.

‘인공지능 학습용 데이터(AI Dataset)’란 머신러닝, 딥러닝 등 인공지능 모델학습을 위해 활용되는 데이터를 총칭한다[1]. ‘인공지능 학습용 데이터 구축사업’의 경우, 비정형데이터를 수집하고 참값(GT, Ground Truth) 어노테이션을 통해 라벨링 하여, 지도학습(Supervised Learning)에 쓰이는 데이터를 구축하는 데 초점이 맞추어져 있다. 최근 쟁점이 된 비지도 학습 및 챗GPT 등도 추가로 확대 개선 중이다. 그중 인공지능 학습용 데이터 구축과정에서는 확보된 품질이 학습데이터 전체의 품질을 결정하기 때문에, 인공지능 학습용 데이터 품질에 관한 중요성이 부각되고 있다. ‘인공지능 학습용 데이터 품질’이란, “인공지능 학습에 필요한 데이터 자체의 품질을 확보하여 사용자에게 유용한 가치를 줄 수 있는 수준”이라고 정의할 수 있다[2]. 따라서, 인공지능 활용을 위해서는 선행 조건인 학습용 데이터 품질에 관한 연구가 필요하다.

본 연구의 목적은 인공지능 학습용 데이터 중 특히, 비정형데이터인 이미지/영상데이터 구축 시 발생 가능한 문제점을 개선하고 품질에 대한 주관적 판단을 최소화하는 것이다. ‘비정형데이터’는 사전 정의된 데이터 모델이 없는 정보 즉, 정해진 규칙(rule)이 없어서 값의 의미를 쉽게 파악하기 힘든 데이터를 의미한다. 인간의 눈으로 봤을 때는 이미지 식별이 가능하지만, 컴퓨터는 구조화되지 않는 데이터를 ‘0’과‘1’로 식별 처리하므로 이미지 자체를 인식하지 못한다. 이미지와 같은 비정형데이터들을 기계학습으로 다루기 위해서는 적절한 특징 벡터로 표현해야 하는 데 이를 라벨링이라고 한다. 기계학습을 활용하기 위해서는 라벨링 과정이 필수적이다. 일련의 라벨링 과정을 마친 데이터는 인공지능 학습모델로 입력된다. 이는 인공지능 모델 성능과 직결된다. 따라서, 인공지능의 성능이 우수하다는 의미는 라벨링 데이터의 품질이 높다는 것을 뜻한다. 인공지능 모델학습을 위해 구축되는 비정형데이터 유형에는 텍스트, 음성(소리), 이미지, 영상과 같은 데이터가 주를 이룬다. 비정형데이터는 특성상 모호함이 발생하고, 사전 정의되지 않음으로 인한 주관적 판단이 발생한다. 따라서, 위와 같은 문제를 해결하기 위해 본 논문은 데이터 유형 중 이미지/영상데이터 품질 확보를 위해 모호함을 최소화하고, 주관적 판단이 발생할 수 있는 문제점에 대해 단계별 공정절차와 목적성을 확인하여 객관적 판단 근거가 되도록 품질 개선방안을 제안하였다.

인공지능 학습용 데이터를 구축하기 위해서는 임무정의, 데이터 획득, 정제, 라벨링, 모델학습의 순서로 진행된다. 인공지능 학습용 데이터의 구축공정은 (그림 1)과 같다.



(그림 1) 인공지능 학습용 데이터의 구축공정[2]

인공지능 학습용 데이터의 구축공정은 다양한 유형의 데이터가 사전에 정의된 목적에 따라 구축되기 때문에 구축과정이 세부적으로 달라질 수 있다. 임무정의 단계에서는 인공지능으로 문제 해결에 필요한 학습용 데이터를 구체적으로 정의하고, 설계하는 활동을 수행한다. 데이터 획득 단계에서는 학습에 필요한 데이터를 직접 생산하거나 혹은 이미 생산된 데이터를 수집하여 원시데이터를 확보하는 활동을 수행한다. 정제단계에서는 이전 단계에서 획득한 원시데이터를 기계학습에 필요한 형식이나 크기로 맞추고, 데이터의 중복을 제거하거나, 필요 시 개인정보 비식별처리를 통한 원천데이터를 확보한다. 원천데이터란 라벨링 데이터가 부여되지 않은 상태의 정제데이터의 결과를 의미한다. 라벨링 단계에서는 원천데이터에 기능이나

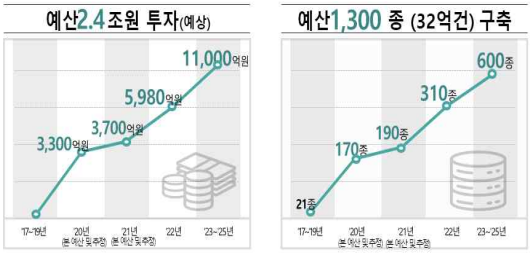
목적에 부합하는 라벨링 작업을 통해, 모델학습에 활용할 수 있는 실질적인 인공지능 데이터를 의미한다. 마지막으로 학습 단계에서는 학습데이터셋(라벨링 데이터)을 이용하여 학습모델에 사전 정의된 인공지능 알고리즘을 학습시키고, 학습된 모델을 향상시키거나 보정하는 활동을 수행한다[12]. 따라서, 인공지능의 성능을 확보하기 위해서는 구축과정에서 각 단계별로 고려해야 하는 품질 요구사항의 명확함이 필요하며, 모호함으로 인해 달라지는 특성을 최대한 객관화하여 구축해야 한다. 특히, 구축과정 중 가공 단계에서 발생하는 라벨링 문제점을 정확하게 파악하고 정의하여 라벨링 오류 발생 시 인공지능 학습모델 결과에 영향을 최소화하는 작업이 필요하다.

본 논문은 인공지능 구축사업 및 컨설팅에 참여하여 연구자료들을 확보하고, 구축된 결과물들에 대한 비교 분석을 통해 연구의 실용성을 검증하였다. 연구 범위는 한국지능정보사회진흥원의 인공지능 학습용 데이터 구축사업 중 기구축된 21년, 22년 영상/이미지 관련 데이터로 제한하였다.

2. 관련연구

2.1. 인공지능 학습용 데이터 구축사업 현황

국내에서는 인공지능을 학습시키기 위한 양질의 데이터 수집을 위해 '과학기술정보통신부'와 '한국지능정보사회진흥원'이 지난 2017년부터 '인공지능학습용 데이터 구축사업'을 실시해오고 있다. 인공지능 학습용 데이터 구축사업 '2017' ~ '2022'년도 현황은 (그림. 2)와 같다[8].



(그림 2) 인공지능 학습용 데이터 구축사업 현황('21~

'2017 ~ 2019'년 21종 325억 원을 시작으로 투자 금액은 2020년 170종 3,315억 원, 2021년 190종 3,705억 원, 2022년 310종 5,980억 원을 투자했다. 2017년부터 시작된 이 사업은 디지털 뉴딜을 통해 2020년 사업 규모가 대폭 확대되었다. 분야별 고품질·대규모 인공지능 학습용 데이터 확보를 위한 데이터 수집·연계·활용 정책을 총괄하는 인공지능학습용 데이터 구축을 '25년까지 총 2조 5천억 원을 투자하여 2025년까지 1,300종을 추가 구축하는 계획을 하고 있다[9].

지금까지 대한민국에서 인공지능학습용 데이터 구축에 대한 2022년 기준 구축된 전체 데이터 종수는 670종이며, 종류는 컴퓨터 비전/전략(53), 음성(42), 교통/자율주행(39), 헬스케어(38), 자연어(32), 농축수산/제조/로보틱스(40), 안전환경/스포츠/관광(44), 지역특화, 자유, 고도화, 활용, 환류(66)인데 가장 보편적으로 사용하는 컴퓨터 비전의 이미지/영상 종수가 가장 많은 수를 차지하고 있으며, 2022년까지 구축한 데이터 유형별 분석 결과, 모든 분야에서 이미지/영상 유형의 데이터를 가장 많이 구축하였다.

세계 주요국에서도 AI 기술 발전을 위해 정부와 민간 협업으로 투자하여 대규모 AI 데이터 구축 경쟁이 가속화되고 있다. <표 1>에서 아시안/태평양에서의 인공지능 학습용 데이터가 세계 시장 성장 규모에서 연평균 성장률 27.96%로 가장 높게 나왔다. Global AI Training Dataset Market Forecast to 2030 보고서에 따르면 학습데이터 유형별 인공지능 학습용 데이터 성장률 예측에서 이미지/비디오 성장률이 가장 높을 것으로 예측하고 있다.

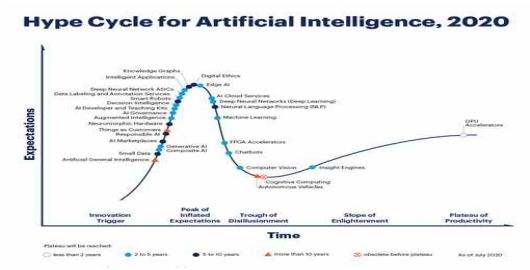
<표 1>인공지능 학습용 데이터 세계 시장 성장 규모

지역별	2018	2019	2020	2026	2030	연평균 성장률
전체	9,948	11,498	14,375	21,758	115,736	23.14%
아시안-태평양	2,116	2,530	3,278	5,359	38,502	27.96%
북미	4,232	4,784	5,854	8,591	39,641	21.07%
유럽	2,749	3,163	3,922	5,934	30,981	22.95%

(단위 : 억원, 1$=1,150원)

(그림 3)의 인공지능 학습용 데이터 관련 AI 하이프 사이클(Hype Cycle)상 주요 기술에는 Data Labeling and Annotation Services, Deep Neural Networks (Deep Learning), Machine Learning,

Computer Vision 등이 있으며, 기술발현(양산)까지 기대시간이 2년에서 5년이 소요되는 것으로 나타난다 [10].

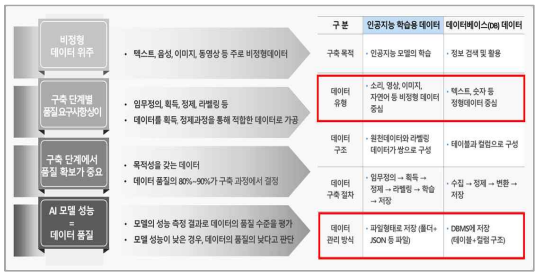


(그림 3) Gartner, Hype Cycle for AI 2020

2.2. 비정형 인공지능 학습용 데이터

2.2.1 학습용 데이터 특성과 유형

인공지능 학습용 데이터의 고품질 데이터 확보를 위해서는 인공지능 학습용 데이터의 특성을 고려하여, 데이터베이스(Data Base)에 저장하는 구조화된 데이터를 대상으로 하는 일반적인 빅데이터와는 차별화된 방식의 품질관리 체계를 확보하는 것이 중요하며, 인공지능 학습용 데이터가 갖는 특성을 파악하여 데이터 확보를 해야 한다. 인공지능 학습용 데이터의 특성은 첫째, 특정한 인공지능 모델학습을 목적으로 생산되는 데이터로써 임무 정의에 따라 구축되는 특성을 가지고 있다. 둘째, 지도학습이라는 특징에 따라 이미지, 비디오, 오디오, 텍스트 등 비정형데이터를 대상으로 라벨링 작업을 통해 참값(GT)을 부여하며, 참값의 품질 여부에 따라 쓰임이 정해지는 등 통제되어야 하는 특성이 있다. 셋째, 인공지능 모델학습용으로 구축되는 데이터는 활용자의 아이디어에 따라 무한한 활용 가능성을 가지는 특성이 있다.



(그림 4) 인공지능 학습용 데이터의 특성 및 다양한 유형의 DB 데이터 비교[4]

(그림 4)는 인공지능 학습용 데이터의 특성 및 다양한 유형의 데이터베이스(Data Base) 데이터를 비교한 것이다. 인공지능 학습용 데이터 유형에는 크게 텍스트, 이미지, 비디오, 오디오로 나눌 수 있다. 이중 이미지와 영상에 대한 데이터를 비정형데이터라고 하며, 일정한 규격이나 형태를 지닌 숫자 데이터와 달리 구조화되지 않은 데이터를 말한다[3].

원시데이터는 인공지능 학습모델과 사전에 협의된 활용목적에 따라 수집되기 때문에 그 활용도에 따라 어노테이션 방식이 달라지므로 데이터 수집을 위한 명확한 기준이 필요하다.

인공지능 학습용 데이터 수집 과정에서 확보된 원시데이터의 품질이 구축공정 과정에서의 전체 품질을 결정하기 때문에 원시데이터 확보는 매우 중요하다. 따라서, 구축공정 과정에서 생길 수 있는 다양한 품질 문제에 대한 이해와 인공지능 학습용 데이터의 특성과 유형에 대한 개념이 필요하다.

2.2.2 학습용 데이터 이미지/영상 라벨링 방법

본 연구에서는 구축 목적에 따라 획득된 이미지, 비디오 두 가지 유형의 비정형데이터를 연구하였다. 이미지 데이터의 경우 라벨링 기능은 이미지 식별, 객체 인식, 영역 구분 등으로 분류하여 라벨링 한다. 비디오에서는 객체 인식을 위해서 비디오 프레임당 화면을 캡처하여 학습모델 개발 담당자와 협의된 라벨링 방법을 사용하고 영역 구분은 픽셀로 처리한다. 또한, 원시데이터에 대해 라벨링 기능과 방법을 학습모델 개발 담당자와 협의하여 다양한 형태로 변형할 수도 있다. 따라서, 라벨링 작업을 수행할 때는 최적화된 학습모델과의 협업이 절대적으로 필요함을 연구를 통해 확인하였다. 그러나, 학습모델 결과로 라벨링의 의미 정확성을 명확하게 판단하기에는 어려운 것이 현실이다. 이를 위해서 학습모델과 인간의 눈이 유사값을 가질 수 있도록 최대한 일관성과 객관성이 확보되어야만 모호함이 사라지고 모델학습에 활용 가능한 고품질 데이터 확보가 가능할 것으로 판단된다. 이미지/영상데이터의 인공지능 학습용 데이터 라벨링 검사 도구 및 라벨링 방법은 <표 3>와 같다.

<표 3> 데이터 품질검사 라벨링(검사) 도구 및 라벨링 방식[6]

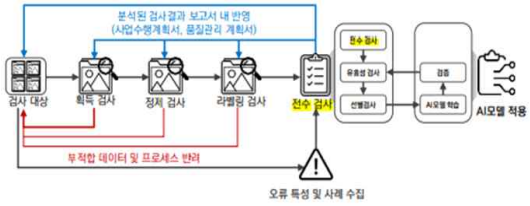
라벨링 방식	이미지 예시	내 용
2D Bounding Box		• 직사각형 모양의 Bounding Box로정확하게 인식하고자 하는 객체가 포함되도록 그리는 라벨링방법. • 박스 안의 객체 이외의 여백을 최소화 하도록 사용. • 가장 일반적인 라벨링 방법. • annotation : 고로케
Cuboid		• 2D로 작업할 수 없는 3D 객체를 정육면체로 생성하는 라벨링방식. • 주로 자동차, 건물 등 입체적인 객체를 2D 형식으로 라벨링하는작업에 대한 한계를 해결하기 위해 쓰임. • annotation : 자동차, 트럭
Keypoint		• 특정 지점을 라벨링하는방법. • 객체의 중요 특징점을 지정하여 물체를 인식. • annotation : 눈, 코, 입, 얼굴선
polygon		• 다각형 모양으로 객체의 가시 영역 외곽선을 따라 점을 찍어가며 그리는 라벨링방법. • Bounding Box기법의 빈공간으로 발생하는 오류에 대응 가능. • annotation : 돛단배, 배, 비행기
Polyline		• 여러 개의 점을 가진 선을 활용하여 라벨링하는기법. • 주로 인도, 차선 등을 경계선을 인식시키기 위해 사용. • annotation : 중앙선, 차선, 가드레일
Segmentation		• 픽셀로 이뤄진 면적에 따라 해당부분의 물체이름이 라벨링되고 라벨링된 결과물끼리 색이 다른 클래스로 묶임 • 정밀한 객체 구분을 위한 픽셀 단위 주석화 • annotation : 자동차
비디오 어노테이션		• 비디오 라벨링 도구에 의해 움직이는 물체의 정보 추출, 모델에 대한 고품질 비디오 주석을 빠르고 효율적으로 제공 • annotation : 사람, 강아지

3. 제안된 고품질 데이터 분석 방법

3.1 데이터 개선 분석 방법

본 논문은 한국지능정보사회진흥원의 학습용 데이터셋을 기반으로 20년, 21년, 22년의 데이터 중 영상/이미지 데이터의 실제 품질 개선 방법 도출 과제를 중심으로 개선 방법의 일치성 여부와 합의 적용 가능성을 확인 후 최종 품질검증 시 제시된 품질기준과 확보방안이 결과에 반영된 내용을 주축으로 실데이터를 수행한 결과에 대한 분석을 통한 개선방안을 제시하였다.

분석 방법은 21년 진행된 190종의 인공지능 학습용 데이터 구축사업 중 제공된 각종 산출물(보고서 및 품질검사기준서)의 결과 자료와 실제 구축된 데이터를 비교 분석하여 내용을 확인하고, 결과를 도출하였다. 그리고, 그 외의 유형별 차이점에 대해서는 추가적인 도메인 지식과 과업의 특성을 고려하여 데이터 분석을 진행하였다. 공정별 문제점에 대한 제시는 과제별 데이터 유형별로 같은 데이터라 할지라도 목적성이 다르면 품질기준이 달라짐을 확인하였다. 이를 통해, 품질 개선 시 목적성의 일치 여부와 구축공정별 전체 공정에 관한 내용을 파악 후, 제안된 품질 검사 기준을 통해 판단 근거를 가지고 수행하도록 하였다. (그림 5)와 같이 공정별 품질 확보를 위해 각 검사 단계마다 분석된 자료를 통한 피드백 또는 부적합 데이터 반려 등 지속적이고 반복적인 품질검사를 수행하였다.



(그림 5) 공정별 구축 데이터 검사 절차 및 검사 기준

품질 개선을 위한 데이터 개선 분석 방법으로 수행한 수집/획득, 정제, 가공에 대한 품질검사 분석을 통해, 각 공정 수집/획득, 정제, 가공 단계별 품질 오류 문제의 근원적인 원인을 파악하고, 품질 문제를 최소화하여 구축사업 시 개선 기회를 도출하는 방안을 제안하였다.

3.1.1 획득단계(이미지 확인, 분석)

인공지능 학습용 데이터의 높은 가치를 지니기 위해 구축 초기 획득/수집 단계는 매우 중요하다. 정확한 데이터 획득/수집을 위해서는 데이터 특성 분석을 통해 획득 방법에 대한 기준 수립과 수집방식 그리고 활용목적에 맞는 일관성 있는 데이터 확보가 중요하므로 도메인 유형의 요구사항 특성을 반영해야 한다. 특히, 인간의 육안으로 데이터 판단 시, 애매함이 존재하여 판단이 불가능할 경우, 컴퓨터 또한 애매함이 존재하므로 이러한 상황을 고려하여 좀 더 명확한 기준과 개선이 필요하다.

▸ 획득단계 요구사항을 고려한 데이터

<표 4>와 같이 원시데이터 단계의 검사 항목으로 품질 내용에 대한 문서와 데이터의 일치성 여부를 검토하였다.

<표 4> 원시데이터(수집/획득) 단계 점검항목

		1. 데이터 원시데이터 단계 점검항목
품질관리 체크리스트	1.1	수집 계획(확보 계획/일정/데이터 포맷) 점검
	1.2	법적 권리(데이터 공개)확보를 위한 법률 검토 수행 점검
	1.3	법적 권리(데이터 공개) 확보를 위한 개인정보동의서/계약서 점검
	1.4	법적 권리(데이터 공개) 확보를 위한 저작권 동의서/계약서 점검
	1.5	법적 권리(데이터 공개) 확보를 위한 특허권/계약서 점검
	1.6	법적 권리(데이터 공개) 확보를 위한 초상권 동의서 점검
	1.7	법적 권리(데이터 공개) 확보를 위한 IRB 승인서 확인
	1.8	원시데이터 수집 현황 (샘플링, 수량 육안) 점검 - 통계적 다양성 기준 충족여부
	1.9	데이터 수집관련 이슈 및 위험 요소 식별 / 점검
	1.10	데이터 품질 관리 원시데이터 획득 점검 기준 확인

수집/획득 단계인 이미지/영상데이터 수집 시 가장 우선적으로 고려해야 할 사항은 서비스 활용에 좋은 데이터라도 수집/획득이 통제 불가능한 주기를 가지고 있다면 바람직하지 않고, 데이터 대상, 획득 방법이 법.제도를 저촉하거나 사회윤리에 어긋나지 않아야 한다. 따라서, 개인정보보호법등에 따라 적절한 법적, 기술적 절차를 거친 데이터 처리가 되어야 한다. 또한, 구축 목적을 달성할 수 있도록 다양한 시간, 공간, 집단, 수준이 포함되도록 데이터 수집/획득 시 품질을 고려한 사항을 반영한다.

(그림 7)공정별 구축 데이터 획득 단계 검사[8]

(그림 7)은 농축수산 도메인 이미지로 품질검사를 위한 체크리스트와 실 구축 데이터를 확인 후, 개선 방법을 실행한 내용으로 수집/획득 단계에서의 품질 검사 전과 후를 표현하였다. 획득 이미지 데이터의 경우 품질 검사 변경 전과 변경 후로 비교했을 때 변경 전은 보이는 대로 투명도, 밝기, 포커스, 촬영 대상, 그림자 등 이미지 식별이 어려워 획득 기준에 미달하여 품질이 저하로 이어진 상황이다. 이에 대한 개선방안으로 카메라의 해상도를 높여 밝기를 조절하고, 목적에 부합하는 대상의 정확한 촬영 방법으로 진행하게 제시했으며, 촬영 대상이 흐리지 않게 포커스를 맞춰 촬영을 진행하게 하였다. 그 결과 밝기, 노이즈, 초점, 목적에 맞는 대상, 그림자 문제가 개선되었고 획득 기준에 맞는 고품질 원시데이터를 확보할 수 있었다.

3.1.2 정제단계(이미지 분류, 분석)

정제단계는 획득한 원시데이터를 기계학습에 필요한 형식이나 크기로 맞추고, 동일 이미지 및 중복성 데이터를 제거하여 원시데이터 획득 시 개인정보를 비식별화하여 처리하는 등의 과정을 통해 원천데이터를 확보하는 활동이다. 또한, 라벨링에 필요한 객체를 설정하고 데이터 라벨링에 필요한 필수정보를 확인한다. 정제 방법은 학습모델을 정확하게 수행하기 위한 필수 요소이다.

▸ 정제단계 요구사항을 고려한 데이터

<표 5>와 같이 원천데이터 단계의 검사 항목으로 품질 내용에 대한 문서와 데이터의 일치성 여부를 검토하였다.

<표 6> 원천데이터 단계 점검항목

품질관리 체크리스트	2. 데이터 원천데이터 단계 점검항목
2.1	데이터 정제 계획 (방법/기준/일정/개인정보보호 방안) 점검
2.2	데이터 정제 조직 (인력구성 / 역할) 점검
2.3	데이터 정제 매뉴얼/도구 보유 현황 및 결과 점검
2.4	원시데이터 정제 현황(원천데이터 육안) 점검
	- 자동화 정제도구 사용시 정확도 점검
2.5	세부 절차별 품질관리 활동 내역(정제 내역서) 점검
2.6	데이터 정제 관련 이슈 및 위험요소 식별 / 점검
2.7	데이터 품질 관리 원천데이터 정제 점검 기준 확인

정제과정에서의 고려사항으로는 정제 기준의 명확성, 데이터 중복성을 방지해야 하며, 데이터 특성과 활용목적에 맞는 적절한 정제 방식에 대한 선정 여부와 기준 타당성을 검사해야 한다. 또한, 정제 작업을 위한 매뉴얼 작성 및 관리 여부를 고려사항으로 반영한다.

(그림 8) 공정별 구축 데이터 정제 단계 검사[8]

(그림 8)은 개인을 특정할 수 있는 민감한 정보를 담고 있는 객체 이미지 데이터로 인공지능 학습용 데이터 구축 시 사생활 보호 방안 등을 고려해야 하는 정제단계에서의 품질검사 방법이다. 특히, 이미지/영상 데이터는 사람, 텍스트, 공간 등 다양한 유형의 민감정보를 담고 있다. 이러한 특성과 함께 개인정보를 악용한 사생활 침해, 개인정보 유출, 사칭, 도용 등의 범죄가 생길 수 있다. 인공지능 학습용 데이터로서 개인정보보호와 민감 데이터를 보호하기위한 방법으로 정제단계에서는 객체 이미지의 비식별화의 기법을 제안한다. 품질 검사 변경 전과 변경 후로 비교했을 때 변경 전은 사람 이미지가 그대로 보이는 등의 문제점을 확인할 수 있었다. 연구자는 이에 대한 해결방안으로 개인정보(사람얼굴, 차량 번호판 등)를 보호하기 위해 Blurring(흐려짐), Pixelation(모자이크)로 이미지를 비식별화하는 기술을 적용하여 개인정보보호 방법을 수행하여 좋은 품질의 데이터를 확보하였다.

3.1.3 가공/라벨링단계(이미지 검토, 분석)

데이터 라벨링은 목적과 기능에 맞게 정의된 규칙에 따라 라벨링을 부여하는 작업으로 수집된 데이터 표본을 감지하고 태그를 지정하는 과정을 말한다. 데이터 라벨링을 효과적으로 하기 위해 가장 좋은 방법은 학습용 데이터 목적에 맞게 라벨링 기능을 가진 도구를 선정하여 활용하는 것이다. 데이터 라벨링 방법은 유형별 이미지/영상, 텍스트, 음성, 영상, 센서 데이터 등에 인공지능이 인식할 수 있도록 데이터에 객체명이나 인스턴스명을 지정해준다. 데이터 입력 기준에 맞게 라벨링을 진행해야하기 때문에 정확한 라벨링 방식으로 수행해야 정확한 데이터가 될 수 있다.

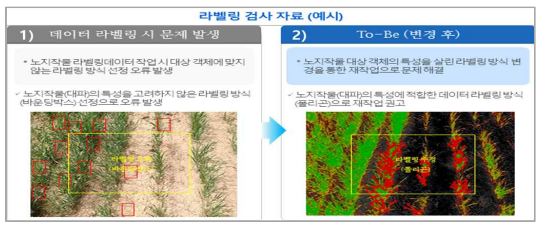
▸ 라벨링단계 요구사항을 고려한 데이터

<표 6>과 같이 라벨링 단계의 검사 항목으로 품질 내용에 대한 문서와 데이터의 일치성 여부를 검토하였다.

<표 7> 라벨링 단계 점검항목

품질관리 체크리스트	3. 데이터 라벨링 단계 점검항목
3.1	가공 계획(방법/기준/일정) 점검
3.2	저작도구 확보 계획 및 진행경과 점검 (직접 확인)
3.3	데이터 가공(라벨링) 관리 및 작업 상태 확인 (무작위 샘플링 확인)
3.4	데이터 가공 메뉴얼/방법 및 교육현황 점검
3.5	세부 절차별 품질관리 활동 내역(가공 내역서) 점검
3.6	데이터 가공관련 이슈 및 위험요소 식별 / 점검
3.7	데이터 품질 관리 라벨링 점검 기준 확인

라벨링 단계에서의 고려사항으로는 목적에 맞게 작성된 라벨링 가이드에 대한 타당성 여부를 검사 후 라벨링 작업자들에게 전달해야 한다. 어노테이션 항목으로 목적에 맞는 어노테이션 구성인지 여부를 검수 후 확인된 내용을 포함하도록 하며, 라벨링 품질검사 도구 혹은 라벨링 도구를 통해 검수 후 검수자가 육안으로 부적합 데이터 여부 2차 확인과 촬영된 영상과 이미지의 누락, 번짐 및 조건 오류를 도구를 활용하여 전수 검수를 통해 중복 없이 검사를 진행해야 한다.



(그림 9) 공정별 구축 데이터 가공/라벨링 단계 검사[8]

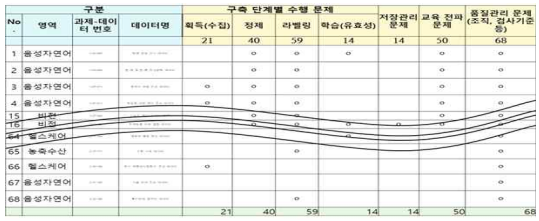
(그림 9)는 노지작물 라벨링 데이터 대파 이미지로 라벨링 단계에서의 품질검사 방법이다. 왼쪽은 객체 이미지를 바운딩박스 방식으로 라벨링하고, 오른쪽은 여러 개의 점을 가진 선을 활용하여 특정 영역을 라벨링 하는 폴리곤 방식을 사용한 라벨링 이미지다. 왼쪽 노지작물 대파 이미지 데이터의 경우 어노테이션 혹은 라벨링 작업 시 적절치 않은 라벨링 방법을 선택하였다. 라벨링 기법 중 하나인 바운딩박스는 가장 많이 활용되는 일반적인 라벨링 방법으로 인식하고자 하는 대상 객체를 직사각형 모양의 박스 안에 포함되도록 여백을 최소화해서 타이트하게 박스를 생성해 주어야 한다. 박스외 불필요한 여백을 포함하는 경우 데이터의 정확도가 낮아지기 때문에 정확도 확보를 위해서는 오차범위를 최소화해야 한다. 왼쪽 노지작물 대파 이미지의 경우 작물의 정확한 식별에 대한 모호성과 넓은 대지 영역을 바운딩박스로 라벨링 하였기에 노지작물 대파 외의 빈공간으로 인해 대파를 제대로 인식하지 못하는 오류 문제가 발생하였다. 이러한 문제를 해결하고자 1번) 왼쪽 바운딩박스 이미지 데이터를 2번) 오른쪽 이미지 데이터와 같이 폴리곤 방식으로 라벨링 방식을 변경한 결과, 정확한 데이터로 학습을 시킬 수 있는 결과물을 보여주었다. 라벨링 객체의 외곽을 폴리곤 방식을 이용해 거의 정확하게 그려내는 결과를 수작업으로 진행하였다. 질 좋은 데이터 가공을 위해서는 분야별 전문가가 참여하는 것이 효율적이며, 검사 및 고급 라벨링 과정에는 전문 인력이 필요하다고 판단된다. 폴리곤 방식은 객체의 기준점에 의거하여 사람이 일일이 작업을 수행해야 하는 어려움과 비용적인 부담은 있으나 이렇게 얻어진 데이터 결과물을 인공지능에 학습시킨 경우, 좋은 성능의 고품질 데이터를 확보할 수 있음을 확인하였다.

4. 제안된 고품질 데이터 확보를 위한 품질검사 방법 적용

4.1 데이터 품질 검사 적용

본 논문은 2021년 전체 190종에 대한 '인공지능학습용 데이터 구축사업' 중 68종에 대한 표본과제를 선정하여 품질지표를 검사 기준으로 샘플링하고 <표 7>와 같이 구축공정별 시점에 해당하는 검사내용을 체크리스트를 통해 품질검사를 진행하였다.

<표 7> 인공지능 학습용 데이터 품질검증 현황 분석



AI학습을 위한 영상/이미지 고품질 데이터 확보방안으로 인공지능 학습용 데이터 68종에 대한 품질검사 수행에 대한 인공지능 학습용 데이터 구축사업에서는 먼저, 다양성, 정확성, 유용성, 신뢰성에 대한 데이터 기준을 제시한다. 데이터는 처음 제시된 품질지표 기준서 목표에 부합할 수 있도록 사전 품질검증 결과서를 통해 데이터 적합성을 판단한 뒤 품질이 미흡한 부분에 대한 품질보완 조치를 거쳐 그 결과물의 고품질 데이터 확보에 대한 연구결과를 중심으로 분석하였다. 데이터는 다양한 속성을 감지하기 위한 목적으로 분류하여 카테고리를 나누고 그 속성으로 추출한 결과물에 대한 라벨링 및 검수 작업을 수행하였으며 그 결과 인공지능 모델학습에 필요한 우수한 성능의 고품질 데이터를 확보할 수 있었다. 이에 대한 22년 5월 기준 품질검증 데이터 68종 품질검증 분석 결과는 (그림 10)과 같이 도출되었다.

(그림 10) 68종 품질검증 결과

인공지능 학습용 데이터 품질검증에 대한 종합적으로 분석한 결과 첫째, 구축공정별 품질관리 기준 (데이터 획득, 정제, 라벨링) 단계의 품질관리가 미흡했으며, 특히, 라벨링 품질 문제가 87%로 가장 높게 나왔다. 라벨링 기준의 난이도에 따른 문제(작업자용 지침 내용 또는 업데이트), 정제 기준, 획득 기준 수립

등을 사업 초기에 해결하지 못한 것으로 파악되었다. 둘째, 품질관리 조직, 검사 기준, 자가점검 등 실질적인 품질관리 활동이 미흡함을 확인하였다. 마지막으로 이러한 품질 문제를 해결하기 위해서는 초기 라벨링 난이도에 따른 작업자 가이드라인 등 구축공정 및 품질관리 기준 등에 대한 명확한 제시와 관리가 필요하다. 따라서, 인공지능 학습용 데이터 68종에 대한 품질검증 연구를 통해 구축공정 단계에서 발생한 문제들을 보완 작업을 통해 품질을 개선하는 작업을 수행했을 때 고품질 데이터 확보가 가능함을 알 수 있었다. 품질검사는 <표 8>과 같은 품질지표를 제시하고 구축공정 시점마다 각각의 지표에 구체적인 세부 지표 및 검사 기준 방법을 적용하여 고품질 데이터 확보를 위한 검사를 수행하였다. 학습데이터 셋의 품질 수준을 측정하기 위한 품질검사 기준 즉 품질지표는 프로세스 측면과 데이터 측면의 두 가지 측면에서 품질검사 유형별 검사대상 및 방법으로 품질검사가 이루어진다.

<표 8> 인공지능 학습용 데이터 품질검사 지표 기준

구분	지표	세부지표	품질검사 기준
프로세스 측면	준비성	계획 수립성	인공지능 학습용 데이터 구축을 위한 절차, 조직(역할)과 책임, 도구 등이 체계적으로 수립하여 관리 및 수행하고 있는지를 확인 인공지능 학습용 데이터 구축 시 발생하는 위험관리 방안이 수립되어 관리 및 수행하고 있는지 확인
		체계 준수성	인공지능 학습용 데이터 구축 시 관련 법제도/개인정보, 초상권, 저작권 등에 대한 충분한 검토와 해결방안이 제시되었는지 확인 개인정보보호, 비식별화 조치 등 민감정보 보호와 인공지능 학습용 데이터 보안관리를 위한 절차와 규정을 수립하고 수행하고 있는지 확인
	완전성	수집 완전성	수집된 데이터의 편향성 방지방안을 수립하고 수행하고 있는지 확인
		정제 완전성	데이터 수집 방법 및 기준, 교육, 검수에 대한 체계를 수립하고 수행하고 있는지 확인 데이터 수집기준 변경을 위한 체계를 수립하고 수행하고 있는지 확인
		가공 완전성	개인정보보호를 위한 비식별화 조치 등 민감정보 보호를 위한 절차와 규정을 수립하고 수행하고 있는지 확인 데이터 정제 방법 및 기준, 교육, 검수에 대한 체계를 수립하고 수행하고 있는지 확인 데이터 정제 기준 변경을 위한 체계를 수립하고 수행하고 있는지 확인
	유용성	사용 편의성	데이터 라벨링 방법 및 기준, 교육, 검수에 대한 체계를 수립하고 수행하고 있는지 확인 데이터 라벨링 기준 변경을 위한 체계를 수립하고 수행하고 있는지 확인
		유연성	사용자의 요구사항을 만족 시킬 수 있는 학습모델과 학습용 데이터셋이 제공되고 있는지 확인 사용자 요구사항을 수용할 수 있는 학습모델 선정 및 적용이 유연한지 확인
데이터 측면	적합성	기준 적합성	원시 데이터가 인공지능을 활용하여 문제를 해결하는데 요구되는 다양성, 신뢰성, 충분성, 균일성, 사실성, 공정성 등 기준에 적합하게 구성되어 있는지 확인
		기술 적합성	원시 데이터가 해당 분야 파일 포맷, 해상도, 프레임레이트, 컬러심도, 어절 수 등 기준에 적합하게 구성되어 있는지 확인
		통계적 다양성	이미지, 영상, 텍스트, 음성 등 데이터에 대한 클래스 분포, 인스턴스 분포, 문장길이, 어휘개수 등 분석 가능 항목의 통계적 분포를 확인
	정확성	의미 정확성	생성된 어노테이션과 정답 어노테이션 간 정확도, 정밀도, 재현율을 확인
		구문 정확성	데이터 구조, 입력 값 범위, 데이터형식이 정확하게 입력되어 있고 누락된 정보는 없는지 확인
	유효성	학습모델 유효성	학습용 데이터를 인공지능 알고리즘으로 훈련시켰을 때 분류성능, 탐지성능, 인식성능, 질의응답, 기계번역, 문서요약, 음성인식, 음성합성에 대한 목표로 했던 수준의 성능 달성이 가능한지 확인

<표 9>와 <표 10>은 품질검사 수행을 위한 프로세스 측면과 데이터 측면으로 구별하여 품질지표별 검사대상 및 방법에 대한 체크리스트이다. 품질검사 즉 품질지표 절차는 준비성, 완전성, 유용성, 적합성 검사 방법으로 이루어지며, 각 단계별 시점에 맞는 세부지표 및 기준방법을 적용하여 검사를 수행하였다. <표 10>은 품질검사 수행을 위한 데이터 측면의 품질지표별 검사대상 및 방법에 대한 체크리스트이다. 데이터 측면의 품질검사에는 적합성, 정확성, 유효성의 품질지표가 있으며, 정확성 지표에서는 세부 지표별로 구문 정확성과 의미 정확성 두 가지 세부 지표로 나누어진다. 의미 정확성은 Ground Truth (GT)를 통해 정확도(Accuracy), 정밀도(Precision), 그리고 재현율(recall)을 계산하여 확인하였다. 구문 정확성은 어노테이션 데이터 구조를 준수하는지, 속성값이 입력 유효 범위에 존재하는지, 정의된 데이터 형식을 준수하는지를 확인 후 데이터의 규모를 고려한 품질지표 샘플링 검사 방식을 적용하여 검사를 수행하였다.

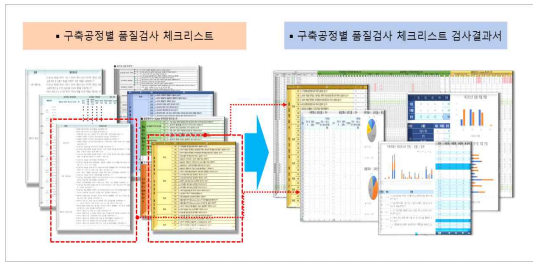
<표 9>프로세스 품질지표별 검사대상 및 방법

품질지표			품질검사 대상 데이터 유형				품질검사 대상 구축공정					품질검사 방법	
			동영상	이미지	텍스트	음성	일무정의	구축계획 수립	데이터 획득	데이터 정제	데이터 라벨링	데이터 학습	정량적 검사	정성적 검사
준비성	계획 수립성	절차준비					●	●	●	●	●			●
		조직준비					●	●	●	●	●			●
		도구준비						●	●	●	●			●
		위험관리					●	●	●	●	●			●
	체계 준수성	보안준수					●	●	●	●	●			●
		법제도 준수					●	●	●	●	●			●
완전성	수집 완전성								●					●
	정제 완전성									●				●
	가공 완전성										●			●
유용성	사용 편의성						●					●		●
	유연성						●					●		●
적합성	기준적합성	다양성	●	●	●	●		●	●					●
		신뢰성	●	●	●	●		●	●					●
		충분성	●	●	●	●		●	●					●
		균일성	●	●	●	●		●	●					●
		사실성	●	●	●	●		●	●					●
		공정성	●	●	●	●		●	●					●

<표 10> 데이터 측면의 품질지표별 검사대상 및 방법

품질지표			품질검사 대상 데이터 유형				품질검사 대상 구축공정					품질검사 방법	
			동영상	이미지	텍스트	음성	일무정의	구축계획 수립	데이터 획득	데이터 정제	데이터 라벨링	데이터 학습	정량적 검사	정성적 검사
적합성 (계속)	기술적합성	파일포맷	●	●	●	●							●	
		해상도	●	●									●	
		프레임 레이트	●										●	
		컬러심도	●	●									●	
	통계적 다양성	어절수			●	●							●	
		클래스 분포	●	●	●	●							●	
		인스턴스 분포	●	●	●	●							●	
		문장길이			●	●							●	
		어휘개수			●	●							●	
정확성	의미정확성	정확도	●	●	●	●							●	
		정밀도	●	●	●	●							●	
		재현율	●	●	●								●	
	구문정확성	데이터 구조	●	●	●	●							●	
		입력값 범위	●	●	●	●							●	
		데이터 형식	●	●	●	●							●	
유효성	학습모델 유효성	분류성능	●	●									●	
		탐지성능	●	●									●	
		인식성능	●	●									●	
		질의응답			●								●	
		기계번역			●								●	
		문서요약			●								●	
		음성인식				●							●	
		음성합성				●							●	

이를 위해 품질관리 지표를 기준으로 (그림 11)과 같이 인공지능 학습용 데이터 구축사업의 품질검사 체크리스트를 검사대상에 따라 분류하여 각 세부 평가지표마다 구축공정 단계별 체크리스트를 만들어서 정량적 또는 정성적 검사방법으로 품질검사를 적용하였으며, 이와 같은 방법을 통해 구축공정별 품질검사 체크리스트 적용에 대한 품질검사 연구 결과를 확인할 수 있었다.



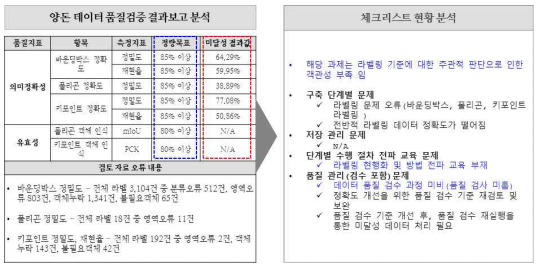
(그림 11)구축공정별 품질검사 및 체크리스트 검사결과서[2]

본 논문 실험기간은 2022년 05월 01일 ~ 2022년 11월 30일이다. 이에 대한 품질검사 결과서를 확인 후 품질 문제를 진단하고 품질 보완조치를 수행하였다. 제시된 분석 방법과 개선 방법을 이용하여 22년 실제 데이터를 구축 진행 중인 '양돈 데이터'에 관한 품질검사 연구결과를 대표적으로 적용해 보았다. 적용 실험에 사용된 점검항목과 내용은 <표 11>과 같다.

<표 11> 품질 특성별 검사 항목

품질 특성	검증 목적	검증 대상	검증 기준	검증 방법
다양성	데이터셋 구성의 다양성 통계 분석 및 편향 확인	제출 전량 라벨링 데이터	품질지표 기준서, 품질특성별 측정기준	자동화 도구를 통한 데이터 값(value) 통계
구문적 정확성	라벨링 데이터 파일의 구조 및 형식의 정확성 검증	제출 전량 라벨링 데이터		자동화 도구 및 구문적 정확성 전단규칙을 활용한 스키마 준수 확인
의미적 정확성	원천데이터와의 정합 및 라벨링 데이터의 참값 정확성 검증	제출 전량 데이터 중 클래스별 표본 데이터 추출량		의미적 정확성 검사기준서를 바탕으로 아래 중 적합한 방법으로 참값 확인 ① 크라우드 기반 정성평가 ② 라벨링 숙련자 육안검사 ③ 참조 정답값(GT) 제작 ④ 전문가 검사
유효성	데이터셋이 실제 인공지능 모델의 학습에 유효한지 확인	제출 전량 데이터 (학습/검증/평가)		인공지능 모델 학습 후 그 성능 확인

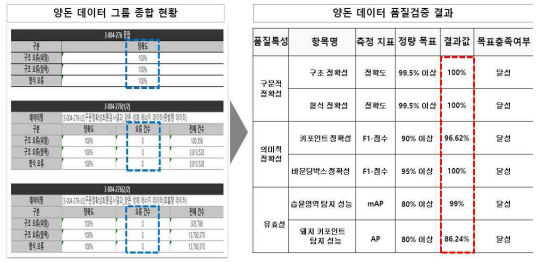
인공지능 학습용 데이터의 품질 '양돈 데이터'검사 항목과 목표 수치는 (그림 13)과 같다. 항목별 검증방법은 다양성, 구문 정확성, 의미 정확성 검증 대상 데이터를 진행하였다. 검증항목을 분석한 결과 다양성은 이미지 수 및 구성비 분포 확인으로 문제가 없었으나 의미 정확성의 경우 바운딩박스 정확도/정밀도 85% 이상 목표치에서 64.29%로 미달성, 재현율 85% 이상 목표에서 59.95%로 미달성, 폴리곤 정확도/정밀도 85% 이상이 목표였지만 38.89% 미달성으로 확인되었다. 키포인트정확도/정밀도 85% 이상 77.08% 미달성, 재현율 85% 이상 50.86% 미달성이었으며, 유효성 또한 목표 기준 미달성이었다. 따라서, 미달성 원인에 대한 문제를 진단한 결과, 획득, 정제, 라벨링 가이드, 라벨링 대상 및 범위의 선정 기준 및 명확한 가이드 전달 부족(크라우드워커의 역량 부족), 품질관리 조직 운영 등이 영향을 준 것으로 파악되었다.



(그림 13) 인공지능 학습용 데이터 구축사업 중 양돈 데이터 분석 자료

전체적인 품질관리 기준과 라벨링 대상 및 범위 선정 기준 설정, 라벨링 작업에 대한 재구성을 한 결과, 구문 정확성 최종 품질검사에서는 구조오류, 형식오류에 대한 정확도 100%, 오류 건수가 단 한 개도 없었다.

의미 정확성의 경우, 키포인트 라벨링 정확성: F1-score 95% 이상, 호흡량 : 목표치 90% 결과값 96.62% 달성, 바운딩박스 라벨링 정확성: F1-score 95% 이상 증발량 : 목표치 95% 결과값 100% 달성 등 제안된 품질검사 방법과 공정별 품질검사를 면밀하게 수행하여 재구성 요청을 한 결과이다. 따라서 본 논문에서 제시한 구축공정별 품질검사 방식의 오류 검사 절차와 품질에 대한 객관성과 일관성 등 기술적인 부분을 제 3자의 시각으로 재구현한 결과와 문제점을 보완하여 수행한 결과, 목표치 달성이라는 결과를 얻을 수 있었다. (그림 14)는 실제 22년 구축 진행 중인 양돈 데이터 품질검사를 수행 반영한 결과이다.



(그림 14) 양돈 데이터 관련 품질검증 결과

구축공정별 품질검사 결과를 바탕으로 품질 문제를 식별하고 문제의 근본적 원인을 파악하여 품질 문제를 해결하고자 제안하였다[5].

5. 결 론

본 연구는 데이터 유형 중 비정형데이터 이미지/영상데이터에 대한 품질검사를 통해 각 구축공정 단계에서 발생할 수 있는 품질 오류 문제를 파악하고, 이를 해결하기 위한 구축공정 단계별 품질 개선방안을 통해 고품질 데이터 확보방안을 제시하였다. 또한, 인공지능 학습용 데이터의 좋은 품질을 확보하기 위해 구축 절차 및 구성요소를 분석하여 구축공정 단계별 데이터 품질의 중요성과 연구 필요성을 입증하였다.

인공지능 모델학습을 위해 구축되는 비정형 데이터는 특성상 저장매체와 데이터 오류·손상이 빈번하게 발생했고 오류 유형은 파일 오류, 이미지 오류, 압축 오류 등 다양한 오류 문제가 있었다. 학습용 데이터 구축과정에서 데이터 품질이 80%를 차지하는 만큼 본 논문은 구축 단계별 품질 요구사항을 고려하여 체계적인 품질관리를 수행한 결과, 오류·손상에 의한 품질 성능 저하 문제를 해결하고 고품질 데이터 확보를 할 수 있었다.

인공지능 학습용 데이터의 효율성을 높이기 위해서는 구축공정 단계 처음 수집단계에서부터 품질기준과 구축 목적에 맞는 정확한 정보를 제공해야 한다. 확실한 목적성을 가지고 설정한 데이터는 정확도가 높으며 고품질 데이터로서의 가치가 있다.

특히, 구축 설계부터 품질에 대한 보다 명확한 목표 설정과 검사 방식에 대해서 철저하게 검토를 해야만 한다. 임무 정의, 데이터 획득, 데이터 정제, 데이터 라벨링, 데이터 학습의 과정을 거쳐야 한다. 인공지능 학습용 데이터의 구축과정은 다양한 유형의 데이터가 사전에 정의된 목적에 따라 구축되기 때문에 구축과정에서 데이터 속성과 유형에 따라 품질 확보 방법이 달라진다. 또한, 같은 데이터 유형이라도 구축 방법이나 도구에 따라서 다양한 파일 포맷으로 분류되며, 통일성 있는 데이터 포맷으로 구축하는 것이 중요하다. 본 논문에서는 각 구축공정 단계에서 생길 수 있는 데이터 품질 오류 문제들에 대한 요소들을 품질 개선 및 수정을 통해 해결함으로써 고품질 데이터 확보방안을 수립하였다. 따라서, 본 연구에서 제시된 고품질 데이터 확보방안을 응용하여 학습용 데이터 구축 시 데이터의 품질 편차를 극복할 수 있기를 기대한다.

참고문헌

[1] 김연진, 조숙경, 박봉섭, 김경배, 서원대학교, 소방청, “유해화학물질 판독을 위한 인공지능 학습 데이터 라벨링에 관한 연구”, 한국통신학회 논문집, pp. 266-267, 2022.

[2] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회, 「인공지능 학습용 데이터 품질관리 가이드라인 v3.0, 2022.

[3] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회, 「인공지능학습용 데이터 품질관리 가이드라인 v2.0 제2권 데이터구축 안내서」, p.13, 2022.

[4] 과학기술정보통신부, 한국지능정보사회진흥원, 한국정보통신기술협회, 「인공지능 학습용데이터 품질관리 안내서 v1.0 제1권 품질관리구축 안내서」, 2021.

[5] 이창엽, 이홍재, 최용락, “데이터 품질관리 체계 수립 질병 데이터를 중심으로”, 한국 IT 정책경영학회 논문지, 2018.

[6] 유성근, 조성만, 송민정, 전소연, 임송원, 정서경, 박상일, 박구만, 김희태, 이대성, 서울과학기술대학교, “딥러닝을 활용한 향상된 라벨인식 방법에 관한 연구”, 주식회사테크윙, 춘계학술발표대회 논문집, 2018.

[7] 한국지능정보사회진흥원(NIA), 인공지능 학습용 데이터 구축 사업관리 매뉴얼, 2021 재구성.

[8] 과학기술정보통신부, 한국지능정보사회진흥원·AI-Hub, AI·DATA INSIGHT Vol. 09.

[9] 한국지능정보사회진흥원(NIA), 인공지능 학습용 데이터 구축 지원 자료-AI-Hub 활용 성과, 2022.

[10] 과학기술정보통신부, 한국지능정보사회진흥원, 인공지능 학습용 데이터 품질관리 가이드라인 v1.0, 2021, 재구성/3D모델링 데이터 및 품질검증 TTA, v2, 2022.

[11] 김동기, 최병기, 이재호, “명세 기반 인공지능 학습 데이터 수집 방법”, 정보처리학회논문지/소프트웨어 및 데이터 공학 제11호, 2022.

[12] 이운영, “인공지능 학습을 위한 패션 레이블드 데이터 분석”, 한국디자인학회지, 2022.

〔저자소개〕

김 승 희 (Kim-Seung Hee)
2021년 2월 극동대학교 산업기술보안학과 석사
2021년 9월 ~ 현재 극동대학교 인공지능보안학과 박사과정
email : zelssen@naver.com

류 동 주 (Ryu-Dong Ju)
2009년 2월 전남대학교 정보보안협동과정 박사
2020년 2월 극동대학교 산업보안학과 조교수 역임
현 재 극동대학교 대학원 인공지능보안학과 겸임교수
2017년 ~ 현재 비트레스(주) 대표
email : ryu@btress.com