지도학습을 위한 데이터 품질 관리 요구사항 동향

곽준호 TTA AI시험검증팀 책임연구원

1. 머리말

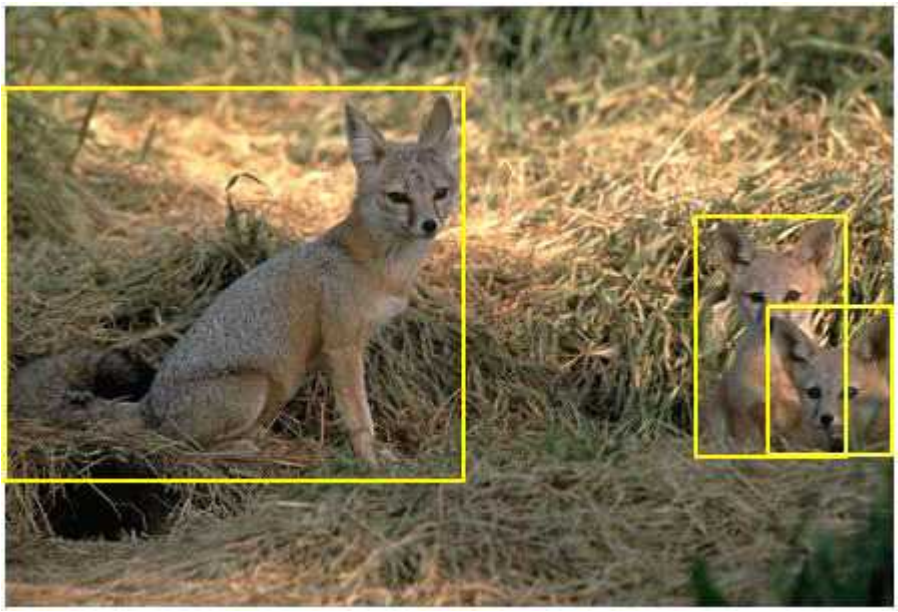
데이터가 중요하다는 것은 이제 시장과 학계에서만 통용되는 말이 아니라, 누구나가 고개를 끄덕이고 당연하게 받아들이는 말이 되어버렸다. 현재 우리가 접하고 있는 IT 부문의 제품과 서비스는 대부분 데이터 기반으로 운영되고 있다 해도 과언이 아니다. 특히, 많은 양의 데이터를 학습시킨 인공지능 모델은 소비자에게 최적화된 결과와 정보를 제공하는 제품과 서비스의 근간이 되고 있다. 이는 산업 분야를 막론하고 현재 우리가 체험하고 있는 현실이다.
오늘날의 인공지능 시스템을 신체라 가정한다면, 데이터는 혈액에 비유할 수 있다. 두뇌에 산소와 영양소를 공급하는 혈액이 흘러야 하는 것처럼, 인공지능 모델은 추론 및 판단을 위하여 데이터를 학습해야 하기 때문이다. 혈액에 불순물이 없어야 하는 것처럼, 데이터 역시 품질이 매우 중요하다. 데이터의 품질은 인공지능 모델의 품질과 성능에도 직접적으로 영향을 미친다.
인공지능 모델 개발에 있어 데이터의 수집과 가공은 매우 중요한 과정임에도 불구하고, 인공지능에 활용되는 데이터의 품질을 정의하거나 정립한 표준 혹은 통합적인 기술 체계는 없다. 이 때문에 인공지능 모델 개발에 활용하기 위한 데이터를 구축하고, 데이터를 활용 및 관리하고자 하는 기업 및 조직은 데이터 품질 체계를 마련하기 위해 노력하고 있다.
본고에서는 인공지능 기술 중 현재 가장 널리 활용되는 지도학습 계열의 인공지능 기술에 대하여 살펴보고, 2021년 12월 제정된 '지도학습을 위한 데이터 품질 관리 요구사항 (TTAK.KO-10.1339)' 표준에 대하여 살펴보고자 한다.

2. 지도학습을 위한 데이터의 품질 관리

2.1 지도학습 계열의 인공지능 기술 개요

지도학습은 인공지능 기술의 한 갈래로서, 사전에 정의된 입력-출력 쌍의 통계적 특성에 기반하여 새로운 입력에 대한 출력값을 낼 수 있도록 학습하는 함수[1]를 의미한다. 여기서 함수는 수치 모델을 의미하며, 수치 모델을 구성하는 변수를 결정하기 위하여 입력-출력 쌍으로 구성된 데이터를 학습하게 된다. 이를 학습 데이터(training data)라 일컫는다. 지도학습 모델은 학습 데이터를 통해 특정 입력에 대하여 출력해야 하는 결과를 정답값(ground truth)으로 받아들이고, 이를 패턴화하는 과정을 거치게 된다. 지도학습의 특징은 입력-출력 쌍으로 구성된 학습 데이터를 생성하고 구축하기 위하여 라벨링(labelling)이라는 데이터 가공 과정이 필요하다는 점이다. 라벨링은 지도학습을 위하여 데이터에 부여된 타깃 값인 라벨(label)을 데이

터에 부착하는 과정을 의미한다[2]. 이미지 파일을 기반으로 고양이를 인식하는 지도학습 모델이 있다면, 입력은 이미지 데이터이고 출력은 '고양이'가 될 것이다. 따라서 학습 데이터에는 이미지 데이터에 고양이가 있는 부위를 표시(라벨링)함으로써 출력이 사전에 정의되어야 한다. 라벨링 작업이 정확하지 않게 되었거나, 엉뚱한 객체를 라벨링한 경우, 잘못된 정답값을 알려주게 되는 꼴이 되며, 모델의 정확도 및 성능에 치명적인 영향을 미치게 된다.



[그림 1] 라벨링(바운딩 박스 방식) 예시[3]

2.2 지도학습을 위한 데이터의 품질

지도학습에 활용되는 학습 데이터는 앞서 언급한 라벨링 작업 때문에, 기존의 정보시스템에서 활용되는 정형 데이터와는 차별화된다. 데이터의 형태는 이미지나 음성 등 비정형 데이터가 주류를 이루고 있으며, 일반적인 데이터 정제 및 가공 외에 라벨링 작업을 위한 별도의 작업 기준 및 절차가 필요하다. 이는 데이터 구축 및 관리에 있어 품질에 대한 기준과 절차 역시 차별화되어야 함을 의미한다.

라벨링 작업의 특수성은 데이터를 구축하는 전 과정에 영향을 미친다. 지도학습 모델이 어떤 목적으로 활용되는지에 따라 라벨링 방식이 정해지고, 정해진 라벨링 방식에 적합한 데이터 수집 및 정제 방식이 잇따라 정의되어야 한다. 이처럼 라벨링 작업은 데이터를 구축하는 모든 과정에 유기적으로 연결되어 있으므로 데이터 품질 확보작업을 할 때 이러한 점을 고려하여야 한다.

2.3 지도학습을 위한 데이터의 품질 관리 체계 구축

특히 대규모 학습 데이터를 구축하는 경우, 이러한 품질 확보를 위한 체계와 절차는 매우 중요하다. 다양한 이해관계자가 학습 데이터의 구축에 참여하는데, 각자의 작업에 필요한 기준과 절차가 정해져 있지 않거나, 일관되지 않으면 작업 결과물인 학습 데이터의 품질 하락은 불을 보듯 뻔하다.

정부는 현재 인공지능 학습용 데이터 구축 사업을 진행하고 있다. 인공지능이 활용되는 다양한 분야별로 대규모의 데이터를 구축하여 관련자에게 제공하는 것이 목표이다. 다만, 구축에 참여하는 이해 당사자들이 준수할 수 있는 품질관리 활동에 대한 일관되고 통합적인 체계와 기준이 아직 없는 것이 아쉬울 따름이다.
현재 시장에서는 데이터를 구축하거나 데이터 사용자에게 제공하는 식으로 관련 이해 당사자들이 자체적인 노하우와 기준을 수립하여 품질을 확보하고자 노력하고 있으나, 인공지능의 학습에 쓰일 수 있는 데이터에 맞게 개발된 국내외 표준 역시 아직 없는 실정이다.
따라서 지도학습을 위한 데이터를 구축하는 과정에서 고려해야 하는 품질 관리 요소를 인식하고, 필요한 품질 관리 요구사항을 제공함으로써 데이터 품질을 확보할 수 있는 체계를 마련하여야 한다. 다음 장에서는 앞서 강조한 품질 관리 체계의 기반으로 활용될 수 있는 '지도학습을 위한 데이터 품질 관리 요구사항' 표준안을 소개하고자 한다.

3. '지도학습을 위한 데이터 품질 관리 요구사항' 표준 소개

3.1 데이터 구축 과정

해당 표준에서는 가장 먼저 지도학습을 위한 데이터를 구축하는 과정을 정의하였다. 데이터 구축을 위해서는 데이터 설계, 획득, 정제 및 라벨링 과정의 4단계를 거치게 된다.
데이터 설계 과정은 지도학습을 위한 데이터를 구축하기 위해 필요한 기본 정보를 생성하는 과정이다. 기본 정보란 지도학습의 목적에 맞게 활용될 수 있도록 데이터의 명세가 되며, 이후 과정에서 이루어지는 획득, 정제 및 라벨링 작업에 필요한 기준이 포함된다. 이러한 데이터 명세는 데이터 설계서의 형태로 작성되며, 설계 과정의 최후 산출물이 된다.
데이터 획득 과정은 데이터 설계서에 따라 데이터를 얻는 과정을 의미한다. 데이터 설계서에 명시된 데이터 규격과 형식을 확인하여, 이에 맞는 데이터를 수집하는 활동이 이에 해당한다. 이때 데이터를 직접 생산하거나, 생산된 데이터를 중개하거나 구매하는 경우 모두 해당된다.
데이터 정제 과정은 획득 과정을 통해 얻은 데이터가 지도학습의 목적에 맞는지 확인하고 라벨링 작업을 수행하기 전 적절하게 전처리하는 과정을 의미한다. 정제 작업 역시 데이터 설계서를 기준으로 데이터의 형식적인 부분과 의미적인 부분 모두에 대하여 확인하는 과정이 주를 이룬다.
라벨링 과정은 정제가 완료된 데이터에 라벨을 부착하는 과정이다. 데이터 설계서에 명시된 라벨링 방식이나 기준을 상세화하여 라벨링 작업 기준을 만들고 이에 따라 작업을 진행하게 된다. 라벨링 과정을 통해서 최종 학습 데이터가 구축되어 데이터 사용자에게 제공된다.
각 과정은 공통적으로 데이터 설계서를 기준으로 작업에 맞는 기준과 절차를 구체화하는 작업이 우선시되고, 작업 종료 후 작업 내용에 대한 검사가 각기 이루어진다. 검사 결과, 적절하지 않은 데이터 및 결과물에 대하여 재작업을 수행하기도 한다.

3.2 품질 관리 요소 및 요구사항

품질 관리는 데이터 사용자가 만족할 수 있는 수준의 품질을 달성하고 유지하며 개선하는 활동을 의미한다. 해당 표준에서는 품질 관리를 위하여 앞서 정의한 구축 과정별로 고려해야 하

는 품질 관리 요소를 제시하였고, 각 품질 요소별로 요구사항을 정의하였다. 품질 요소는 총 9가지로 제시하고 있으며, 구축 과정별 해당 사항은 <표 1>과 같다.

<표 1> 지도학습을 위한 데이터의 품질 관리 요소[2]

품질 관리 요소		데이터 구축 과정			
		설계	획득	정제	라벨링
그룹1	다양성	V	V	V	
	출처신뢰성	V	V		
	사실성	V	V		
	규격적합성	V	V	V	
그룹2	통계적 충분성	V			V
	통계적 균일성	V			V
	라벨 적합성	V			V
	라벨 정확성				V
	유효성				V

품질 관리 요소는 두 그룹으로 나뉜다. 첫번째는 라벨링되지 않은 데이터에 대한 품질 관리 요소이며, 두번째는 라벨링된 데이터에 대한 품질관리 요소이다. <표 1>과 같이 각 과정별로 해당되는 품질 관리 요소가 있으며, 각 품질 관리 요소에 따른 요구사항을 확인하고 고려해야 한다.

첫째, 다양성은 데이터가 지도학습에 유용한 모든 특성 정보를 고루 보유하고 있어야 함을 의미한다. 여기에는 포괄성과 변동성이 내포되어 있다.

둘째, 출처신뢰성은 데이터를 획득할 때 반드시 신뢰할 수 있는 출처로부터 획득하여야 함을 의미한다. 다수가 해당 출처로부터 데이터를 활용함으로써 범용성이 있는지, 데이터의 규모가 충분히 큰지, 데이터에 대한 설명이 충분히 이루어지는지 등의 요건을 살펴보아야 한다.

셋째, 사실성은 데이터를 획득할 때 부득이하게 인위적인 환경을 활용하는 경우, 실제 환경과 유사한 환경을 조성하여야 함을 의미한다. 이는 획득 장치, 수단 및 조건 등을 고려하여야 한다.

넷째, 규격적합성은 데이터 획득을 위한 규격(파일 포맷)을 사전에 적절히 정의하고 이에 따라 획득되어야 함을 의미한다. 데이터가 활용되는 분야에서 주로 사용되는 규격을 사용하는 것이 권장된다.

다섯째, 통계적 충분성은 라벨에 포함된 특성 정보가 지도학습에 유의미한 정도로 충분한 양이어야 함을 의미한다. 정량적이고 절대적인 기준은 없으나, 다른 구축 사례와 데이터셋 사례를 참고하여 근거를 제시하는 것이 권장된다.

여섯째, 통계적 균일성은 라벨에 포함된 특성 정보의 수량이 적정 비율을 나타내야 함을 의미한다. 실제 환경과 비슷하게 특성 정보의 비율을 조정하는 것이 권장된다.

일곱째, 라벨 적합성은 라벨링을 할 때 지도학습의 목적에 맞게 라벨링 방식과 라벨링 규격이 적절하게 선정되어야 함을 의미하며 획득 데이터 특성, 활용 목적, 약/강 지도방식(weak/strong supervision)을 고려하여 적합한 라벨링 규격을 정하도록 한다. 특히 라벨링 규격의 경우, 라벨링 작업 결과물의 파일 포맷, 스키마 구성 및 클래스 구조 등이 함께 정의되

어야 한다.
여덟째, 라벨 정확성은 라벨링 규격에 맞게 라벨링이 진행되어야 하고, 정답값이 제대로 라벨링 되어야 함을 의미한다. 전자를 구문적 정확성, 후자를 의미적 정확성이라 한다.
아홉째, 유효성은 학습 데이터를 지도학습 모델에 직접 학습시킴으로써 모델 성능을 통해 드러나는 학습 데이터의 품질을 확인해야 함을 의미한다. 이전까지 열거된 품질 관리 요소는 학습 데이터 자체에 대한 요소였으나, 본 요소는 모델 성능을 통해 간접적으로 확인해야 하는 요소이다.
이러한 9가지 품질 관리 요소에 대한 지속적인 확인, 수행 및 검사를 통해 지도학습에 활용되는 데이터의 품질 관리가 이루어질 수 있다. 해당 표준에서는 각 품질 관리 요소가 구축 과정별로 어떻게 적용되어야 하는지 요구사항의 형태로 설명을 제공한다.

4. 맺음말

지도학습은 이미 다양한 기법이 연구 및 개발되어 실제 인공지능 기반 제품과 서비스에 활발하게 활용되고 있다. 인공지능의 다른 기술 분류 인 비지도학습과 강화학습에 비하여 활용되는 비중이 가장 높다. 자연스럽게 지도학습을 위한 학습 데이터에 대한 수요가 높으며, 이를 구축하고 품질을 관리하기 위한 방법론과 체계에 대한 고민도 활발히 이루어지고 있다.
본고를 통해 지도학습을 위한 데이터의 품질관리 방안에 대한 표준을 소개하였다. 해당 표준은 그간 일관되지 않았던 구축 과정과 이에 따른 품질 관리 요소를 정의하여 표준화한 데 의의가 있으나, 아직 세분화되고 구체화되어야 하는 부분들이 많이 남아 있다. 품질 관리 요소에 따른 요구사항의 경우 데이터 형태 및 목적에 따라 적용 방법과 기준이 다양해질 수 있다. 동일한 요구사항도 음성 및 텍스트, 혹은 이미지 등 데이터의 형태에 따라 차별화되어야 할 것이다. 지도학습의 목적에 따라서도 라벨링 방식이 상이하므로, 각 방식별로 적절하게 요구사항이 적용되어야 할 것이다. 또한, 요구사항에 대한 확인 및 검증 방법론 역시 마련되어 있지 않은 상황이다.
향후, 이러한 부분들에 대한 고민이 이루어져야 한다. 그리고 데이터 구축 및 사용과 관련된 이해관계자들의 공감대를 기반으로 더욱 구체적이고 상세화된 데이터 품질 관리 방안과 표준이 마련되었으면 하는 바람이다. 이를 통해 인공지능에 활용되는 데이터의 생태계가 더욱 성숙하고 활성화되는 계기가 되기를 바란다.

※ 본 연구는 '인공지능 학습용 데이터 구축' 과제의 일환으로 수행되었다.

참고문헌

[1] wikipedia, https://en.wikipedia.org/wiki/Supervised_learning
[2] TTAK.KO-10.1339. 지도학습을 위한 데이터 품질관리 요구사항
[3] https://image-net.org/download-bboxes.php

출처: TTA 저널 제199호