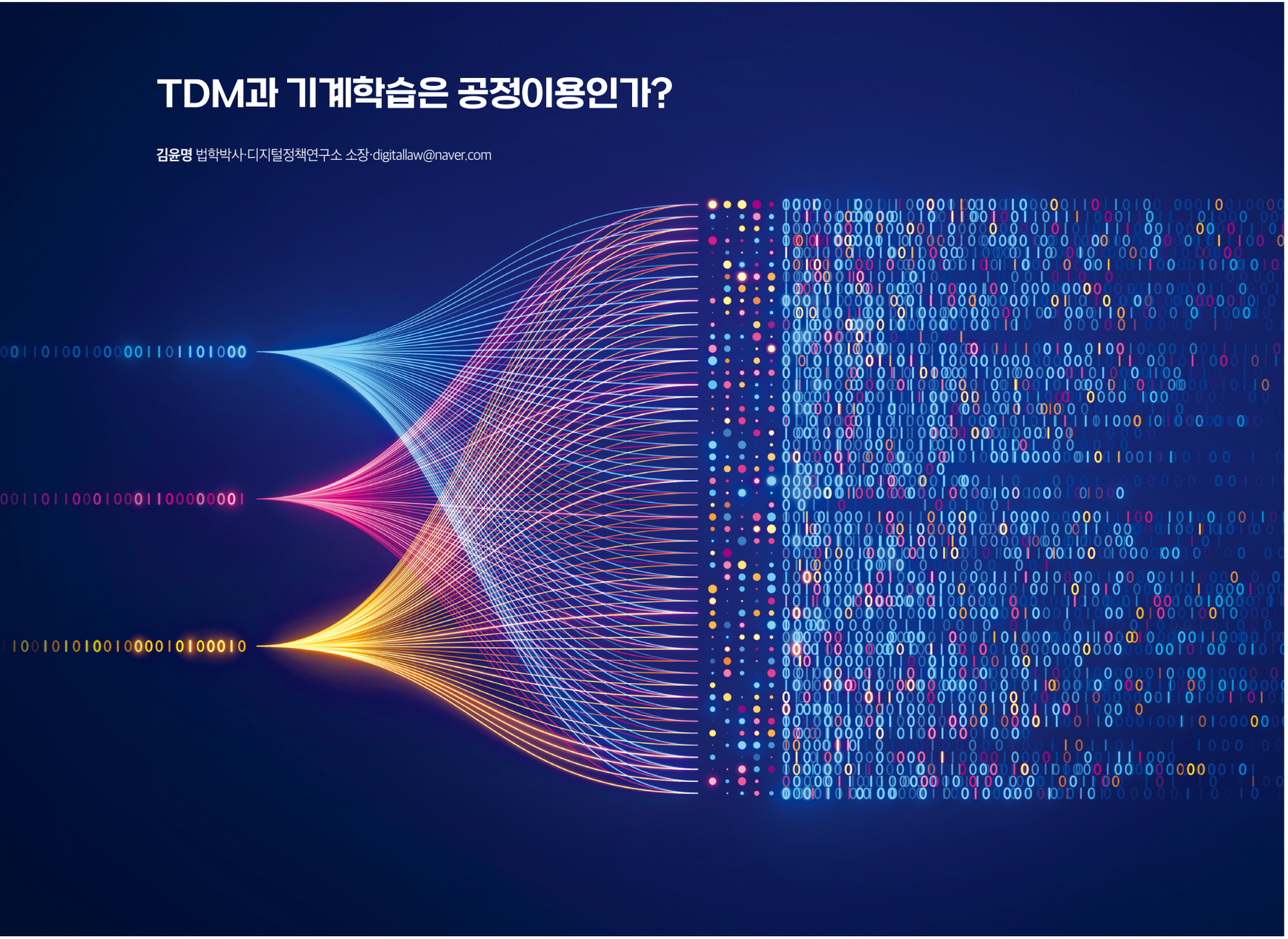
TDM과 기계학습은 공정이용인가?

김윤명 법학박사·디지털정책연구소 소장·digitallaw@naver.com



해석론으로서 공정이용의 적용

이 글은 '데이터의 수집 및 이에 따른 데이터 기반의 기계학습이 현행 공정이용 규정을 통해 가능한지'에 대한 해석론이다. 공정이용 규정이 도입된 것은 기술 발달과 저작물 이용환경의 변화로 저작물 이용이 다양화되고 있어, 개별적 저작재산권 제한 규정만으로는 구체적인 상황에 대응하기 어려울 수 있기 때문이다. 더불어 저작권 보호기간의 연장 등 저작권 보호가 강화됨에 따라 상대적으로 저작물 이용자의 지위가 저하될 수 있다는 우려에 대응하기 위한 방안으로 도입됐다.[1] 크롤링(Crawling)은 데이터를 확보하기 위한 것이지만, 전통적으로 저작권법이 의도했던 이용방식과는 다르다. 마찬가지로 기계학습은 가공된 데이터셋을 이용한다는 점에서 볼 때 명확히는 저작물을 이용하거나 저작물에 담겨진 지적 산물과 작가의 의도를 파악해 이뤄지는 것은 아니다. 비표현적 이용이거나 저작물을 향유하는 형태의 이용이 아니기 때문이다. 저작권법은 비표현적 이용에 대해서는 침해를 긍정하지 않는다. 표현적 요소를 전혀 전달하지 않는 비표현적 목적을 위한 복제는 침해될 수 없다고 본다.[2] 비표현적인 이용을 저작물의 이용으로 볼 수 없는 것은 기본적으로 저작물의 이용은 사람에 의한 이용을 전제하기 때문이다. 따라서 사람이 아닌 존재가 저작물을 이용하는 것은 저작권법에 대한 도전이 될 수 있다. 그렇지만 저작물의 이용이 기존의 형태와 다르다고

1 박명수, "저작권법 일부개정법률안 검토보고", 국회문화체육관광위원회, 2013.12, 23면
2 차상욱, "인공지능 개발에 필요한 데이터셋의 지적재산법상 보호", 『인권과정의』, no.494, 대한변호사협회(2020), 24면

하더라도, 현행 법제의 해석을 통해 적용하는 것이 가장 합리적이다. 입법론을 제기하는 것은 현행 법률로써 해결할 수 없는 상황에 직면할 경우에 한정될 필요가 있다. 다만 현행법의 해석으로도 데이터의 공정이용을 판단할 수 있다고 생각된다. 이하에서는 현행 법률에서 규정된 공정이용에 해당하는지 여부에 대해 살펴본다.[3]

TDM 등 크롤링은 공정이용인가?

데이터 확보 과정에서 크롤링 등의 공정이용 요건에 대해 본다. 크롤링을 통해 얻으려는 이익은 기계학습을 위한 데이터 확보이다. 기본적으로, 인터넷 등 공개된 데이터를 대상으로 하는 것이기 때문에 침해 여부를 단정하기는 어렵다. 더욱이 영국, 독일, 일본 등 여러 나라에서 저작재산권의 제한규정을 별도로 입법하면서 저작권 침해에서 벗어나도록 규정하고 있다는 점은 앞서 살펴본 바와 같다.

• 이용의 목적 및 성격

크롤링은 인터넷에 공개된 웹사이트의 텍스트를 복제한 후 해당 내용을 분석해 필요한 데이터를 확보한다. 이미지를 대상으로 하는 경우에도 사진 등에서 원하는 이미지를 수집하게 된다. 썸네일 검색의 저작권 관련 소송에서 이미지를 가져와서 작은 크기의 이미지로 축소시킨 것에 대해 논란이 된 적도 있다. 크롤링이 저작물을 복제하는 목적은 학습데이터를 만들어내기 위한 것이다. 저작물을 향유하거나 경쟁적인 저작물을 만들어내는 것이 아니다. 기계학습을 통해 인공지능 기술의 발전을 도모하는 것이며, 인간이 활용하는 알고리즘의 가치를 높일 수 있다는 점에서 데이터 확보는 의미 있는 일이다. 이용목적이나 성질에 있어서 비영리적인 이용이어야만 교육을 위한 것으로 인정될 수 있는 것은 아니지만,[4] 영리적인 교육목적을 위한 이용은 비영리적 교육목적을 위한 이용의 경우에 비해 자유이용이 허용되는 범위가 상당히 좁아진다.[5] 다만 공정이용 요건을 판단함에 있어서 공정이용 판단이 영리성 유무가 본질적인 기준이 아니다.[6] 따라서 인공지능 기술의 발전이라는 공익적 목적을 위한 것으로 볼 여지가 충분한 상황이라면 비영리성 여부와 상관없이 인정받을 가능성도 크다.[7]

국내 사례에서, 썸네일 검색의 경우도 웹사이트에서 공개된 이미지를 복제하는 것이 아닌 해당 이미지의 썸네일을 복제하는 경우에 있어서 목적이나 성격이 저작권 침해가 아님을 확인한 경우도 있다. 즉, 대법원은 "썸네일 이미지의 사용은 검색사이트를 이용하는 사용자들에게 보다 완결된 정보를 제공하기 위한 공익적 측면이 강한 점"[8]을 들어 저작권 침해를 부인한 바 있다. 이는 썸네일을 활용해 검색을 용이하게 해주는 것으로서, 원 저작물을 이용하는 것이라기보다는 이를 활용한 변형적 이용(Transformative Use)에 해당하기 때문에 면책되는 것이다. 또한 검색서비스를 위해 구축된 다양한 썸네일을 포함한 정보는 인덱싱돼 서버에 보관하게 되며, 원 저작물의 상업적 이용은 아니지만 "영리적 목적으로 연구를 수행하면서 원 저작물을 복제하는 것과 같은 중간적 이용(Intermediate Use)[9]에 제공하는 경우에도 상업적 이용이 인정될 수 있다"[10]고 한다. 이와 같이 원 저작물을 활용해 새로운 가치를 만들어 내는 방식으로서의 이용은 변형적 이용 또는 창작적 이용으로서 이용목적이나 방식이 원 저작물의 활용을 제한하는 것은 아니라는 점에서 공정이용의 첫 번째 요건을 충족한 것으로 볼 수 있다.

다만 이러한 기준은 해석기준이기 때문에 보다 명확히 하기 위한 입법이 이뤄지고 있다. 영국, 독일, 일본 등 각국의 입법에서 보면, 기계학습을 위한 데이터 확보 활성화를 위해 저작재산권의 제한 규정을 추가하는 입법을 경쟁적으로 진행하고 있다는 점에서 목적이나 성격에 있어서 크롤링의 공정이용 여부에 대해 참조할 수 있을 것이다.

• 저작물의 종류 및 용도

크롤링의 대상은 기본적으로 인터넷에 공개되거나 공공 데이터이거나 웹로봇이 접근할 수 있는 정보가 된다. 다만 민간에서 생성된 정보는 인터넷 등에 공개된 것이어야 한다. 특정 서버에 업로드된 정보를 대상으로 하는 경우에는 정보통신망법상 침입에 해당될 가능성도

3 기계학습을 위한 데이터셋을 구축하기 위한 방법론으로서 크롤링이라는 데이터 처리와 이에 따른 원데이터의 가공과 가공된 데이터셋을 이용한 기계학습의 공정이용을 살펴본다. 둘의 관계는 선행, 후행 등의 순서적 관계이지만 AI 모델을 구축하고 강화한다는 측면에서 본다면 일련의 순서에 따른 과정이다. 그렇지만, 논의의 편의상 구분해 살펴보고자 한다. 따라서, 공정이용의 판단기준과 관련사례 등을 참조할 경우에는 약간의 중복이 이뤄질 수 있다는 점도 부인하기 어렵다

4 2016년 개정된 저작권법은 기존 저작권법 제35조의3 제2항 제1호의 "1. 영리성 또는 비영리성 등 이용의 목적 및 성격"을 "1. 이용의 목적 및 성격"으로 개정했다. 즉, 포괄적 공정이용 제도의 도입목적을 충실하게 반영하기 위해 현행법에 규정된 저작물 이용목적에 대한 제한(보도·비평·교육·연구 등)과 '영리성 또는 비영리성'이라는 판단시 고려사항을 삭제한 것이다. 박명수, 저작권법 일부개정법률안 검토보고, 국회문화체육관광위원회, 2013.12, 24면

5 대법원 1997. 11. 25. 선고 97도2227 판결

6 Campbell v. Acuff-Rose Musin, Inc., 510 U.S.569(1994)

7 정상조, "딥러닝에서의 학습데이터와 공정이용", 『LAW & TECHNOLOGY』 제16권 제1호 (통권 제85호), 서울대학교 기술과법센터(2020), 8면

8 대법원 2006. 2. 9. 선고 2005도7793 판결

9 Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992)

10 송재섭, "미국 연방저작권법상 공정이용 판단 요소의 적용 사례 분석", 『계간 저작권』 제25권 제2호(통권 제98호), 한국저작권위원회(2012), 10면

배제할 수 없기 때문이다. 크롤링은 특정 키워드가 포함된 전체 웹페이지를 가져온다. 크롤링 대상이 저작물이 될 수도 있으며, 사실정보에 불과해 저작권의 보호범위에 포함되지 않을 수 있다. 시사보도의 경우, 시사성이 있는 정보와 이를 분석해 만들어낸 기사내용이 혼재한 경우라면 구분할 필요가 있다.

따라서 "시사보도 또는 역사물, 과학 기타 학술논문 등의 사실적, 정보적 성격의 저작물은 표현의 자유 및 알권리 관점에서 자유로운 유통의 필요성이 상대적으로 높기"[11] 때문에 공정이용으로 인정될 가능성이 크다. 크롤링에서도 사실정보 내지 사실적 정보는 보호가능성이 크지 않다. 설령 그렇지 않더라도, 크롤링을 위한 용도가 해당 사이트와의 경쟁이 아니라는 점, 목적이 기계학습을 위한 데이터의 확보라는 점 등을 고려할 필요가 있다. 더욱이 학습데이터로의 가공을 위한 용도는 기계학습을 목적으로 하는 것이지, 문화적 향유를 위한 것이 아니다. 따라서 저작물의 용도와는 상이함을 알 수 있다. 참고할 수 있는 사례는 구글의 'Book Search' 사건이다. 동 사업으로 책을 스캔하는 것은 디지털화하는 작업으로, 결과물은 구글 검색에서 노출됐다. 구글은 데이터베이스 구축의 경우 이용자에게 편의를 제공하기 위한 것이고 오히려 책의 일부 조각만 열람하게 함으로써 책의 부가적 가치가 증가한 것이지 원작이 훼손됐다고 할 수 없다고 반박하면서 상업적 이용이라기보다는 책 판매 시장의 수요를 대체하지 않는 변형적 이용에 해당한다고 주장했다. 2심에서도 구글의 주장을 받아들여 다음과 같은 이유로 공정이용으로 보았다.[12]

11 이해완, 『저작권법』, 박영사(2019), 839면

무엇보다 이용자의 편리성을 제공하고 검색된 책을 구매할 기회를 제공해, 용도의 상이성은 물론 시장의 확장을 가져왔다는 점에서 공정이용으로 인정됐다. 대법원은 작가협회의 상고를 기각함으로써 구글의 승소로 결론을 냈다.

• 이용된 부분이 저작물 전체에서 차지하는 비중과 그 중요성

크롤링은 특정 주제어나 이미지를 중심으로 데이터를 수집하게 된다. 예를 들면 특정 영역의 말뭉치를 만들기 위해 수집하는 데이터는 제한적으로 저작물을 이용하게 된다. 따라서 차지하는 비중과 중요성은 논할 수 있는 수준이 아니다. 차지하는 양적 비중과 질적 중요성을 기준으로 검토할 수 있을 것이다. 전체적인 양이 많지 않다면 침해 판단의 실익도 크지 않다. 즉 양적 상당성이 확보되지 않는다면 침해요건인 실질적 유사성이 인정되지 않는다.[13] 데이터를 추출하는 것은 전체에서 차지하는 일부이며, 이 일부도 인간의 관점에서는 필요 없는 내용일 수도 있다. 저작물의 패턴이나 특징을 분석해 내기 위해 전체 저작물을 사용하는 것이 아닌 일부 단어나 말뭉치를 중심으로 만들어 가기 때문이다. 다만 인공지능에 의한 데이터 투입단계에서는 모든 데이터가 전부 투입돼 분석의 대상이 되기 때문에 이용분량은 많고 공정이용에 불리한 요소가 될 수 있다는 견해도 있다.[14] 그러나 "인공지능의 분석단계에서는 저작물의 사상이나 감정과 무관한 비표현적(Non-Expressive) 데이터만을 추출해서 이용한다. 인공지능은 데이터분석으로, 통계학적 빈도가 높은 패턴을 찾아 산출단계에 이용하기 때문에 저작물의 이용분량은 미미한 경우가 많다"[15]고 함으로써 인간의 향유와는 다른 '이용이란 측면'에서 공정이용 가능성을 언급하고 있다.

• 현재 시장 또는 가치나 잠재적인 시장 또는 가치에 미치는 영향

크롤링으로 만들어진 학습데이터는 원래 정보가 갖는 시장과는 다른 영역이다. 예를 들면 포털의 정보를 크롤링하더라도 그것이 포털의 검색서비스를 위한 것이 아니라면 경쟁관계에 서기 어렵다. 서비스의 고도화를 위한 것이지 해당 데이터를 경쟁관계에 있는

12 "구글이 저작권으로 보호된 작품을 무단으로 디지털화하고, 검색 기능을 만들고, 그 작품에서 캡처본을 표시하는 것은 공정한 사용을 침해하지 않는다. 복제의 목적은 매우 변형적이고, 텍스트의 공개가 제한적이며, 그 공개는 원본의 보호적인 측면을 위한 중요한 시장 대체물을 제공하지 않는다. 구글의 상업적 성격과 이익 동기는 공정한 사용 거부를 정당화하지 못한다. 구글이 책을 공급한 도서관에 디지털화된 사본을 제공한 것도 저작권법에 부합하는 방식으로 도서관이 사용할 것이라는 이해에 따라 침해에 해당하지 않는다." Authors Guild v. Google, Inc., 804 F.3d 202 (2nd Cir. 2015)

13 이해완, 『저작권법』, 박영사(2019), 843면

14 정상조, "인공지능과 데이터법", 『법률신문』, 2020.02.21.일자

15 정상조, "인공지능과 데이터법", 『법률신문』, 2020.02.21.일자

서비스에 이용하는 것이 아니란 점 등에서 대체 가능성은 높지 않다. 이는 기계학습 데이터 확보를 위한 것으로 기존의 저작권 유통 및 이용환경과는 다른 차원의 시장이 형성될 수 있다. 따라서 시장경쟁 내지 시장대체가 이뤄진다고 보기 어렵다. 인터넷에 공개된 정보나 저작물은 학습데이터 시장을 목적으로 하는 것으로 보기 어렵다.[16] 학습데이터를 염두에 둔 것이 아니라는 점에서 저작권자의 잠재적, 경쟁적 시장에 참여하는 것으로 보기 어려운 이유이다. 참고로 대법원은 공정한 관행에 대해 판단함에 있어서 "정당한 범위 안에서 공정한 관행에 합치되게 인용한 것인지 여부는 인용의 목적, 저작물의 성질, 인용된 내용과 분량, 피인용저작물을 수록한 방법과 형태, 독자의 일반적 관념, 원 저작물에 대한 수요를 대체하는지 여부 등을 종합적으로 고려해 판단"[17]하도록 하고 있다.

즉 시장에 미치는 영향에 대해 그 수요대체 여부를 고려해야 한다는 점을 확인한 것이다. 수요라는 것이 해당 저작물의 이용허락을 통한 시장 진입인지, 아니면 2차적 저작물을 통한 것인지에 대해서도 검토할 필요가 있다. 데이터의 배열과 구조, 가치, 특성과 용도 등에 맞게 학습데이터를 만들어 내는 경우라면 원 데이터의 유사성이 인정되더라도, 결과는 전혀 다를 수 있기 때문에 일반적인 저작물과 같은 기준으로 판단하는 것은 무리다.[18] 이처럼 데이터의 공정이용 판단에 있어서 고려할 사항은 기계학습을 위한 데이터라는 잠재적인 시장을 염두에 둘 수도 있을 것이다.

기계학습은 공정이용인가?

크롤링과 기계학습이 서로 다른 일련의 과정이라는 점을 감안하면, 각각 구분해 살펴볼 때 그 실익이 생긴다. 따라서 크롤링과 달리 기계학습 과정에 대한 공정이용 여부에 대한 주요한 판단기준은 이용의 목적, 성격 및 시장 대체성이다. 저작물의 종류나 용도는 그 활용에 따른 문화적 향유를 위한 수단이 아니라는 점에서 크게 다투어질 것은 아니기 때문이다. 즉 기계학습 과정에서 저작물을 향유하는 것인지 여부가 중요한 판단지표가 될 것이다. 저작물 전체에 대해서 보더라도 크롤링을 통해 확보되거나 사실 정보에 불과한 경우라면 비중은 큰 의미가 없다.

• 데이터 이용으로서 기계학습

인간의 뇌구조를 모델링해 구현한 인공지능 신경망이 인간의 것과 유사하다는 점에서 기계학습은 인간의 학습과정과 크게 다르지 않다. 기계학습 과정에서 인공지능은 저작물을 분석해 특징들을 수치화해 저장한다. 수치화한다는 것은 인간의 뇌에 저작물을 복제하는 것이 아닌 인간이 이해할 수 있는 특징만을 기억하는 것을 의미한다. 이미지의 경우 해당 이미지의 특징을 분석해 수치화하며, 텍스트의 경우는 말뭉치(Corpus)를 인덱싱해 데이터값을 부여한다. 이 과정은 저작물의 의미를 이해하거나 활용하는 것이 아닌 단어나 문장의 구성을 분석하는 것이다. 이는 저작권법이 의도하는 인간의 저작물 이용방식과는 차이가 있다. 분석된 결과물은 저작물 그 자체가 아닌 저작물에 담겨있는 특성, 패턴, 스토리, 구조 등의 것이다. 따라서 인공지능이 학습하는 것은 메모리에 복제하는 것이 아닌, 데이터를 이해하는 상태로 분석하고 추상화하는 상태이기 때문에 복제가 일어나는 것으로 보기 어렵다.[19]

인공지능의 학습과 관련돼 적용할 수 있는 법규는 공정이용에 관한 규정이다. 공정이용에 대한 고려에서 필요한 것은 헌법 제22조에 따른 창작자의 권리를 보호하는 것의 해석이다. 공정이용은 헌법상의 창작자 보호저작재산권의 제한규정이외에도, 저작물의 통상적인 이용방법과 충돌하지 아니하고 저작자의 정당한 이익을 부당하게 해치지 아니하는 경우에는 저작물을 이용할 수 있기 때문이다.[20]

16 국회에서도 "저작물 이용목적 및 성격을 '영리성 또는 비영리성 등'으로 수식하는 경우 영리성·비영리성을 우선적으로 고려하거나 이에 한정해서 판단할 우려가 있어, 영리성이 있더라도 공정이용에 해당될 수 있는 행위를 사장시키는 부정적인 효과를 야기할 수 있는바, 개정안은 이러한 부정적 효과를 미연에 방지하기 위한 내용이라는 점에서 타당한 입법조치"라고 판단하고 있다. 박명수, "저작권법 일부개정안 검토보고서", 국회교육문화체육관광위원회, 2013.12 참조

17 대법원 1998. 7. 10. 선고 97다34839 판결, 2004. 5. 13. 선고 2004도1075 판결

18 물론 기술적이고 기능적인 측면에서 본다면 시장 경쟁 가능성은 높지 않을 것이다. 즉 "문자인식이나 음성인식을 위한 인공지능의 경우에는, 원 저작물의 시장과 인공지능 서비스의 시장이 전혀 다르기 때문에, 원 저작물의 시장 또는 가치에 미치는 영향이 크지 않다"는 것이다. 정상조, "인공지능과 데이터법", 『법률신문』, 2020.02.21자. 그러나 "인공지능에 의한 저작물 소비가 급증하면서 저작물의 시장과 가치도 변화하고 있다. 특히 음악, 그림, 소설, 뉴스를 생산하는 창작형 인공지능의 경우에는 그 산출물이 인간의 표현을 거의 완벽하게 모방하게 되고, 인공지능이 동일한 시장에서 인간과 경쟁하는 관계에 놓일 수 있다"는 견해를 밝히고 있다. 즉 인공지능이 만들어 놓은 결과물이 인간을 대체할 수 있다는 가능성이 있다는 점에서 우려를 표하고 있는 것이다. 그렇지만 이러한 일부 사례가 인공지능 학습을 위한 데이터 확보라는 공공의 이익을 저해한다고 단정하기는 어렵다

19 만약 기계학습 과정이 저작권 침해행위로 본다면 인공지능의 학습은 불가능할 수 있다. 아니면 빅데이터를 확보할 수 있는 인터넷기업들만이 경쟁에서 살아남을 수 있을 것이다

20 홍승기, "데이터마이닝 면책 입법 방향에 대한 의문", 『경영법률』, 제32권 제4호, 한국경영법률학회(2022), 33면

실제 사례를 보면, Fox사의 콘텐츠를 DB화해 검색이 가능하도록 제공하는 TV아이스사에 대해 Fox사는 저작권 침해를 소송을 제기했고 피고는 공정이용 항변을 주장했다. 방송저널리즘의 평가 및 비평 등 폭스사와는 전혀 다른 목적을 위해 비디오 클립이나 검색(Snippet)에 접근했다는 점에서 동 서비스는 사회적, 공공적 편익을 제공하는 중요한 공공성을 가진다고 판단했다.[21] 이미 구글의 북서비스에 대해 원작의 전체에 대해 검색을 제공하는 것은 변형적 이용이며 시장의 대체를 가져오지 않는다며 공정이용으로 판단했다.[22] 동 판결에서 법원은 구글이 스캔한 책을 새로운 유형의 연구 도구를 만드는 데 사용했기 때문에 변형적이라고 보았다. 또한, 구글이 스캔한 책을 사용하는 것이 저작권 보유자에게 시장에 해를 끼치지 않으며 프로젝트의 이점이 잠재적인 해악보다 더 크다고 판단했다. TDM과 관련해 중요한 점은 법원이 텍스트 및 데이터 마이닝에 대해 언급하면서 비소비적(Non-Consumptive) 연구 목적을 위한 도서의 TDM은 변형적 사용이며 공정사용에 따라 보호된다고 본 것이다.

• 통상적인 이용인지 여부

기계학습은 정보를 분석해 그 패턴이나 특징값을 찾아내 이용하는 것이기 때문에 인간의 이용 방식과 다르다는 것은 앞서 살펴보았다. 즉 인간의 저작물 이용과 달리 "저작물 그 자체를 향유하는 것이 아니라 단지 정보를 습득하고자 그 저작물을 구성하는 언어나 기호 등을 통계적으로 분석하는 경우에는 그 저작물 등을 복제하거나 번역 등 필요한 형태로 변환할 수 있다"[23]. 기계학습을 저작물 등의 복제나 단순한 2차적 저작물의 작성이 아닌 창작적 이용(Creative Use)[24]이란 점에서 보면 공정이용에 해당할 가능성이 높다. 저작권법의 목적은 문화의 창달이며, 기존의 저작물을 향유하는 과정에서 새로운 창작적 표현을

21 Fox News Networ, LLC v. TVEyes, Inc., 43 F. Supp. 3d 379 (S.D.N.Y 2014); 박성호, "텍스트 및 데이터 마이닝을 목적으로 하는 타인의 저작물의 수집·이용과 저작재산권의 제한", 『인권과 정의』 vol. 494, 대한변호사협회(2020), 56면
22 Authors Guild v. Google, Inc., 804 F.3d 202 (2nd Cir. 2015)
23 임원선, 『실무자를 위한 저작권법』, 한국저작권위원회(2014), 231~232면
24 공정이용인지 여부에서 저작물의 변형적 이용(transformative use)에 대해 판단한다. 여기서는 이를 창작적 이용이라고 표현하나, 그 의미는 변형적 사용과 다르지 않다. 정상조, "딥러닝에서의 학습데이터와 공정이용", 『LAW & TECHNOLOGY』 제16권 제1호 (통권 제85호), 서울대학교 기술과법센터(2020), 12면

만들어낼 수 있는 동인으로 작용하게 된다. 창작적 이용을 인정하는 수단으로써 공정이용은 저작권법의 목적규정을 통해 확인할 수 있다. 미연방대법원은 변형적 이용을 "새로운 표현, 의미 또는 메시지를 가지고 원 저작물을 변형해, 다른 목적 또는 다른 성질을 가지고 원 저작물의 표현에 무언가 새로운 것을 추가한 경우"[25]라고 판시했다. 변형적 이용이라면 "원 저작물과는 다른 목적의 이용이고 원작의 성질에 대한 새로운 표현을 부가해 변화를 준 것"[26]에 해당한다. 이처럼 기계학습은 새로운 가치를 부여함으로써 저작자가 의도했던 가치 이상을 더해주는 경우라면 이는 공정이용으로 판단될 가능성을 높이는 것이다. 이러한 맥락에서 기계학습은 데이터의 특징값을 분석해내 새로운 가치를 만들어낼 수 있는 모델을 구축한다.

• 종류 및 용도의 적합성

기계학습은 인공지능의 성능을 향상시키는 것이다. 알고리즘이 프로그래밍 된 바대로 데이터의 패턴이나 특징을 인식하고 분석해 의도한 결과를 만들어 내거나 또는 의도성을 가지고 학습하는 것이다. 기계학습 과정은 용도라는 것이 저작물을 향유하는 과정과는 다른 용도라는 점에서 인간의 이용과는 본질적인 차이가 있다. 따라서 공정이용 규정에서의 용도와 기계학습에서의 용도는 다른 기준점에서 봐야하며 저작물의 유형에 따라 달리 봐야하는 것은 맞다. 인간이 학습하는 것은 다양성 확보를 위한 것인 것처럼, 기계학습도 인공지능의 다양한 기능의 향상을 위한 것으로 궁극적으로는 인간의 사고와 유사한 범용 인공지능을 개발하기 위한 것으로 볼 수 있다. 정리하자면, 기계학습과 문화의 향유는 기본적인 용도나 목적이 상이하다. 용도의 차이라는 점에서 본다면 기계학습은 저작물을 향유하는 것이 아닌 데이터에 담겨진 패턴과 특징값을 찾는 것이기 때문에 공정이용에 해당한다.

• 비중 및 중요성

기계학습에 의한 학습데이터 이용의 경우, 사실 저작물이나 기능 저작물뿐만 아니라 비록 예술 저작물이라고 하더라도 저작물의 문예적, 심미적 가치를 이용하는 것이 아니라 그 속의 데이터로서의 비표현적 가치를 이용하는 변형적 이용이 많기 때문에 저작물의 종류 및

25 Campbell v. Acuff-Rose Musin, Inc., 510 U.S.569(1994)
26 한국정보법학회 지음, 『인터넷, 그 길을 묻다』, 중앙Books(2012), 544면

용도가 크게 영향을 주지는 않을 것이다.[27] 학습데이터의 이용은 투입단계와 산출단계의 원저작물 이용 분량이 서로 다를 수 있기 때문에 산출단계의 경미하거나 부수적인 이용으로 비중면에서 공정이용에 해당한다.[28] 참고로 일본의 저작권법은 컴퓨터를 이용한 정보 처리를 통해 새로운 지식이나 정보를 창출함으로써 저작물 이용의 촉진에 기여하는 정보 검색 서비스나 정보 해석 서비스를 고려한다. 이러한 서비스는 공개된 타인의 저작물을 이용하는 것이지만, 저작권자의 이익을 부당하게 해치지 않는 범위 내에서 소위 경미한 이용으로서 적법하게 인정된다.[29]

• 시장 대체성

시장 대체성 여부에 대해 살펴본다.[30] 시장 대체성의 범위는 "원 저작물 자체뿐만 아니라 2차적 저작물의 시장이나 가치도 포함된다"[31]고 한다. 기계학습은 인간의 이용이 아닌 정보 내용이나 표현의 특성을 학습하기 때문에 일반적인 이용 형태와 다를뿐더러, 일반 소비자에게 제공되는 것과는 다른 시장을 형성하게 된다. 기계학습은 인공지능의 알고리즘을 고도화하기 위한 것에 불과하기 때문이다.[32] 다만 기계학습을 위한 별도의 정보(빅데이터)를 구축해 제공한다면 이는 시장 대체성을 인정받을 수 있다. 일본은 이러한 상황을 입법론으로 정리했다. 즉 정보분석을 위한 빅데이터 등의 이용을 저작권 제한규정으로 입법화했다.[33] 이에 따라 인공지능을 학습하는 과정에서 이뤄지는 저작물의 습득 자체는 학습 메커니즘이지 저작물을 복제해 배포하는 것으로 보기 어렵다. 또한 인공지능의 학습 형태에서 빅데이터 등의 정보를 이용하는 것은 "저작물 등을 구성하는 언어나 기호 등의 요소들 또는 그들이 관계 등을 분석하려는 것일 뿐 그 저작물 등 자체를 이용하고자 하는 것이 아니고, 그 분석의 결과물을 그 저작물 등과는 전혀 별개로서 그에 원 저작물이 드러나지 않으므로 그 저작물 등의 통상적인 이용과 충돌하거나 저작자의 정당한 이익을 부당하게 저해할 우려가 적다"[34]고 볼 수 있다. 인공지능의 학습과 유사하게 적용할 수 있는 기존 사례는 썸네일 검색이다. 인터넷에 공개된 정보를 크롤링해, 이를 데이터베이스화하고 검색어가 입력되면 해당 정보를 제공하는 것은 기계학습 메커니즘과 유사기 때문이다. 썸네일 검색은 그 결과를 보여주는 것이지만, 크롤링은 기계학습을 위한 데이터 수집 내지 수집된 데이터를 인덱싱해 관리 값을 부여하기 때문이다. 기계학습은 특징값을 분류해내는 과정이라는 점에서 차이가 있지만, 정보를 분석해 분류하는 과정과는 크게 다르지 않다. 물론 정보를 분석해 이용 가능한 상태에 놓인 것은 공개되거나 출시된 것이 아니기 때문에 시장 대체성을 논하는 것이 타당하지 않다는 지적도 가능하다. 그렇지만 시장 대체성은 해당 저작물의 이용과정에서 고려하는 예측에 대한 판단이므로, 이를 부인할 필요는 없다. 썸네일 형태로 검색결과에 노출되는 것도 정보의 위치를 알려주는 것으로 공익적 성격으로써 공정이용이 인정되고 있고,[35] 대법원도 같은 취지로 저작권 침해를 부인한 바 있다.

소결

학습용 데이터셋을 구축하기 위해 원 데이터(Raw Data)를 수집하는 크롤링과 수집된 원 데이터를 이용해 제작된 데이터셋을 활용한 기계학습 등에 대한 공정이용에 관한 논의를 정리하면 다음과 같다.

먼저, 크롤링과 관련해 정리한다. 인공지능이 관여하는 분야는 실생활에서부터 산업현장에 이르기까지 다양하다. 다양한 분야에서 인공지능 기술의 향상을 위해 기계학습이 이뤄지고 있으며, 이를 위한 기본적인 요소인 데이터 확보를 위해 다양한 법제도적 정비가 이뤄지고 있다. 다만 제도 정비가 이뤄지지 않은 경우, 또는 제도 정비가 이뤄졌다고 하더라도 실질적인 기대효과가 크지 않다면 해석을 통해 가능한 방법을 찾는 것이 필요하다. 영국, 독일, 일본 등에서는 개별적 저작재산권 제한규정을 도입해 TDM이나 정보분석이 가능하도록 함으로써 인공지능의 발전을 꾀하고 있다. 우리나라 또는 미국처럼 포괄적인

27 TDM과 관련해서는 TDM 분석 결과에서 원 저작물을 감득하기 어렵기 때문에, 저작물의 성질에 관한 요소에는 큰 비중을 두지않고 있다고 한다. 김경숙, "TDM 관련 저작권법 개정안의 비판적 고찰", 『경영법률』 제31권 제3호, 한국경영법률학회(2022), 122면

28 저작권법 제35조의3(부수적 복제 등) 사진촬영, 녹음 또는 녹화(이하 이 조에서 "촬영 등"이라 한다)를 하는 과정에서 보이거나 들리는 저작물이 촬영 등의 주된 대상에 부수적으로 포함되는 경우에는 이를 복제·배포·공연·전시 또는 공중송신할 수 있다. 다만, 그 이용된 저작물의 종류 및 용도, 이용의 목적 및 성격 등에 비추어 저작재산권자의 이익을 부당하게 해치는 경우에는 그러하지 아니하다.

29 정상조, "딥러닝에서의 학습데이터와 공정이용", 『LAW & TECHNOLOGY』 제16권 제1호 (통권 제85호), 서울대학교 기술과법센터(2020), 18~19면

30 Robert Merges, Peter Menell, Mark Lemley, 『Intellectual Property in the New Technologital Age』 Wolters Kluwer, 2012, p.646

31 최호진, "썸네일 이미지와 공정이용", 『Law and Technology』 제8권 제3호, 서울대학교 기술과법센터(2012), 70면

32 물론, 수많은 인공지능에 탑재할 목적으로 이용했다면 시장 대체성을 상실할 가능성도 부인하기 어렵다

33 일본 저작권법 제47조의7(정보해석을 위한 복제 등) 저작물은, 전자계산기에 의한 정보해석(다수의 저작물 기타의 대량의 정보로부터, 당해 정보를 구성하는 언어, 음, 영상 기타의 요소와 관련된 정보를 추출, 비교, 분류 기타의 통계적인 해석을 행하는 것을 말한다. 이하 이 조에서 같다)을 하는 것을 목적으로 하는 경우에는, 필요하다고 인정되는 한도에서 기록매체에의 기록 또는 번안(이에 의해 창작한 2차적저작물의 기록을 포함한다)을 할 수 있다. 다만, 정보해석을 하는 자의 이용에 제공하기 위해 작성된 데이터베이스저작물에 대해는 그러하지 아니하다

34 임원선, 『실무자를 위한 저작권법』, 한국저작권위원회(2014), 232면

35 구글검색엔진의 높은 수준의 변형적 이용과 사회적 편익을 제공한다는 점에서 공정이용에 해당한다고 판시한 바 있다(Perfect10, Inc. v. Amazon, Inc., 508 F.3d 1146(9th Cir, 2007))

공정이용 규정이 없기 때문이다. 물론 포괄적인 규정이라고 하더라도 명확한 가이드라인을 제시하는 것이 아니기 때문에 법적 안정성이나 예측 가능성이 그렇게 높다고 보기 어렵다. 따라서 EU나 일본의 입법례와 같이, 특정 분야에서 적용할 수 있는 개별적 제한규정이 유의미할 수 있다.

다음으로 기계학습과 관련해 살펴본다. 기계학습에서 저작물의 이용 메커니즘은 제작된 데이터셋에 포함된 특징값을 분석해 학습 모델(AI 모델)을 생성한다. 학습모델을 만들기 위한 과정은 원 저작물의 시장을 대체한다고 보기 어렵다. 기계학습의 공정 이용을 고려할 수 있는 사례로는 검색엔진의 크롤링과 검색결과의 현시(Display)를 들 수 있다. 크롤링 과정에서 많은 데이터를 수집하지만, 그 자체는 정보검색의 용이성을 위한 것이기 때문에 공정이용으로 보는 것이다. 물론 시장 대체성을 넓게 보아 인공지능을 통해 형성할 수 있는 시장까지 볼 가능성도 부인하기 어렵다. 그렇지만 저작물의 이용이라는 것은 미래의 특정 시점에 도래하는 기술적 수준에 의한 것을 대상으로 제한하기 어렵다. 또한 공정이용 규정 자체가 기술적 발전에 대응하기 위한 것이며, 그 요건에 해당하는 경우라면 면책을 부여하는 것이 타당하다. 만약 기계학습이 저작자의 권리를 심대하게 침해하는 경우가 발생한다면 판례 또는 입법론적으로 대응하는 것이 타당하다.

끝으로, 공정이용 법리를 포함해 저작권법에 관통하는 표현의 자유(Freedom of Speech) 내지 정보의 자유라는 헌법상 가치에 대한 고려이다.[36] 인공지능의 학습은 결국 다양한 정보 활동을 수행할 수 있도록 인공지능을 활용함으로써 정보 접근 및 이용을 확대시킬 가능성이 높아지기 때문이다.[37] 일반적으로 인간의 학습과정은 다양한 창작을 위해 누구라도 허용하는 과정이고, 정보의 자유를 확대시키기 위해 인류가 묵시적으로 허용하는 문화적 이용허락(Cultural License) 또는 문화적 허용(Cultural Permission)이라고 볼 수 있기 때문이다. Harper 판결에서 반대의견은 "공정사용의 원칙을 후퇴시키고, 수정헌법 제1조가 보장하는 사상의 자유를 위협할 수 있다"[38]는 것이다. GPT와 같은 AI 모델이 만들어 내는 결과물이 저작권법에 따른 표현물로 볼 수 있다는 점에서 공정이용의 대상이 될 수 있다. 물론 보호대상이 되는 표현물이 아닐 경우에는 그 자체가 퍼블릭도메인으로 누구나 자유롭게 이용이 가능하다. 그렇지만 여기에는 한계가 있다. 크롤링, TDM, 기계학습 과정에서 저작권자의 허락 없이 저작물을 이용해 수익을 창출하고 있는 플랫폼사업자들에게 모든 수익을 귀속시키는 것이 타당한 것인지는 의문이기 때문이다. 아울러, 투명성을 확보할 수 있어야 하며, 이를 위해 사용된 데이터의 일부를 공개해야 한다. 이를 위해 알고리즘 관련 법제의 제정이 필요하며 학습데이터의 제작이나 사용에 있어서 저작권, 개인정보 등의 법적 요구사항을 반영했는지를 포함해, 그러한 결과가 시장이나 이용자에게 미치는 영향에 대해 확인할 수 있는 영향평가제를 두도록 해야 한다. 만약 기본권(基本權)을 제한하는 등의 심대한 위법한 사항이 발견될 경우에는 강력한 제재를 가할 수 있어야 할 것이다.

36 정보의 자유란 일반적으로 접근할 수 있는 정보원으로부터 의사형성에 필요한 정보를 수집하고 수집된 정보를 취사, 선택할 수 있는 자유를 말한다. 허영, 『한국헌법론(전정6판)』, 박영사(2010), 568면

37 "저작권과 표현의 자유는 상호보완 관계에 있어서 표현의 자유가 충분히 보장되는 경우에만 저작권도 활기를 띠게 되고 또 저작권의 보호로 저작활동이 활발하게 됨으로써 언론의 자유라 알권리도 그 혜택을 충분히 누릴 수 있기 때문"이다. 김윤명 외, 『인터넷서비스와 저작권법』, 경인문화사(2010), 529~530면

38 S.Ct 471 U.S. 539(1985)

▧ 참고문헌

김윤명, 『블랙박스를 열기위한 인공지능법』, 박영사(2022)

김윤명 외, 『인터넷서비스와 저작권법』, 경인문화사(2010)

이해완, 『저작권법』, 박영사(2019)

임원선, 『실무자를 위한 저작권법』, 한국저작권위원회(2014)

한국정보법학회 지음, 『인터넷, 그 길을 묻다』, 중앙Books(2012)

허영, 『한국헌법론(전정6판)』, 박영사(2010)

김경숙, "TDM 관련 저작권법 개정안의 비판적 고찰", 『경영법률』 제31권 제3호, 한국경영법률학회(2022)

박성호, "텍스트 및 데이터 마이닝을 목적으로 하는 타인의 저작물의 수집·이용과 저작재산권의 제한", 『인권과 정의』 vol.494, 대한변호사협회(2020)

송재섭, "미국 연방저작권법상 공정이용 판단 요소의 적용 사례 분석", 『계간 저작권』 제25권 제2호(통권 제98호), 한국저작권위원회(2012)

정상조, "딥러닝에서의 학습데이터와 공정이용", 『LAW & TECHNOLOGY』 제16권 제1호(통권 제85호), 서울대학교 기술과법센터(2020)

차상육, "인공지능 개발에 필요한 데이터셋의 지적재산법상 보호", 『인권과정의』, no.494, 대한변호사협회(2020)

최호진, "썸네일 이미지와 공정이용", 『Law and Technology』 제8권 제3호, 서울대학교 기술과법센터(2012)

홍승기, "데이터마이닝 면책 입법 방향에 대한 의문", 『경영법률』 제32권 제4호, 한국경영법률학회(2022)