인공지능에서 저작권과 라이선스 이슈 분석

Analysis of Copyright and Licensing Issues in Artificial Intelligence

류원옥 (W.O. Ryoo, woryoo@etri.re.kr) 오픈소스센터 책임연구원
이승윤 (S.Y. Lee, syl@etri.re.kr) 오픈소스센터 책임연구원/센터장
정성인 (S.I. Jung, sijung@etri.re.kr) 슈퍼컴퓨팅기술연구센터 책임연구원

ABSTRACT

Open source has many advantages and is widely used in various fields. However, legal disputes regarding copyright and licensing of datasets and learning models have recently arisen in artificial intelligence developments. We examine how datasets affect artificial intelligence learning and services from the perspective of copyrighting and licensing when datasets are used for training models. The licensing conditions of datasets can lead to copyright infringement and license violation, thus determining the scope of disclosure and commercialization of the trained model. In addition, we examine related legal issues.

KEYWORDS AI, copyright, dataset, license, training

Ⅰ. 서론

기존의 오픈소스는 소스 코드 사용 및 개발에 따른 라이선스 준수, 보안 취약점 관리, 의존성 관리 등이 중요했다. 따라서 소스 코드 또는 바이너리 코드는 라이선스 분석, 보안 취약점 분석, 의존성 분석 등으로 오픈소스 컴플라이언스를 해결했다.

인공지능(AI: Artificial Intelligence)은 인간의 학습능력, 추론능력, 지각능력을 인공적으로 구현하려는 기술이며[1], 충분한 데이터세트(Datasets), 그리고 연산할 수 있는 컴퓨팅 환경과 목표 응용에 맞는 학습모델이 주요 구성 요소이다[2]. 글로벌 ICT 시장 조사 전문 기관인 IDC는 세계 AI 시장이 2025년까지 연평균 26.2% 성장할 것으로 예상하고, 국내도 연평균 11.1% 성장을 전망한다[3].

인공지능은 데이터세트를 활용해서 인공지능 모델을 학습하고, 학습한 결과를 바탕으로 의사 결정한다. 이러한 학습을 하기 위해서 빠질 수 없는 것이 데이터세트이며, 다양한 데이터세트를 활용하면 인공지능은 더욱 정확하고 유용한 결과를 얻을 수 있다.

인공지능 모델의 학습은 크게 지도학습, 비지도

* DOI: https://doi.org/10.22648/ETRI.2023.J.380609
* 본 연구는 한국전자통신연구원 내부연구과제의 일환으로 수행되었음[23YF1110, 개방형 ICT R&D 생태계 강화를 위한 ETRI 오픈소스 R&D 대응체계 구축].

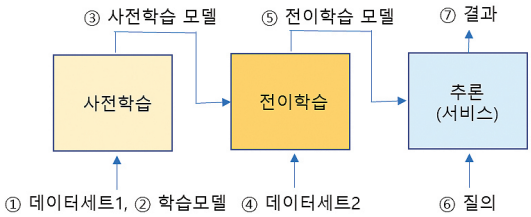

그림 1 인공지능 모델의 학습 과정

학습, 강화학습이 있고, 이들 학습은 그림 1과 같은 과정을 거쳐 학습하고 인공지능 응용(서비스)을 만든다.

이러한 학습과정에 기존 오픈소스 컴플라이언스만 적용해도 문제가 없는지 의문이다. 지금까지 오픈소스 컴플라이언스는 소프트웨어 코드 분석을 통해 진행되었지만, 인공지능 학습 과정에 사용되는 데이터세트는 누군가의 노력으로 만들어진 창작물로 저작권과 라이선스를 갖는다. 따라서 데이터세트는 학습한 모델과 서비스 과정에 저작권과 라이선스에 대한 법률적인 영향을 미칠 수 있다. 미국은 인공지능 관련 소송이 최근 몇 년간 증가하고 있고[4], 정부도 인공지능 창작물 저작권 논쟁으로 한국저작권위원회와 문화체육관광부는 'AI-저작권법 제도 개선 워킹그룹'을 발족하여 인공지능에서 발생할 수 있는 저작권 문제에의 대응을 추진하고 있다[5].

따라서 본고에서는 인터넷에 공개된 데이터세트를 활용하여 인공지능 모델을 개발하는 과정에서 저작권과 라이선스 관리에 따른 이슈들을 살펴보고자 한다.

II. 인공지능 학습과 추론 과정의 이슈

개발자들은 인공지능 모델 연구 및 서비스 개발 활동 등을 위해 인터넷 및 공공기관에서 공개한 데이터세트와 학습모델을 사용한다.

데이터세트의 저작권(Copyright)은 저작자 표시(BY: Attribution), 비영리(NC: Non-Commercial), 변경 금지(ND: No-Derivative), 동일 조건변경 허락(SA: Share-Alike)이 있고, 이를 활용한 라이선스는 CC-BY, CC-BY-NC, CC-BY-ND, CC-BY-SA, CC-BY-NC-SA 등 다양하다. 학습모델의 오픈소스 라이선스는 퍼미시브 라이선스(Permissive License), 카피레프트 라이선스(Copyleft License)가 있다. 영리 목적을 위해서는 상용(Proprietary) 라이선스가 있다. 그리고 상용으로 사용한다면 데이터세트와 학습모델 간의 라이선스 의무사항에 대한 준법성도 고려해야 한다[6].

많은 인공지능 개발자와 기업은 데이터세트 사용과 제작, 데이터세트를 이용한 학습모델 개발, 추론 및 서비스 등의 과정에서 표 1과 같은 의문을 가질 수 있다.

몇 가지 주요 의문 사례는 다음과 같다. 인터넷 웹페이지에 있는 데이터세트를 크롤링 방식으로 수집(Common Crawling)하여 인공지능 학습에 사용해도 되는지, 또한 이렇게 개발한 학습모델을 상업적으로 활용해도 되는지에 의문을 가질 수 있다. 저작권표기-동일조건변경허락(CC-BY-SA) 조건의 데이터세트로 학습한 모델을 동일 조건으로 배포해야 하는지도 명확하지 않다.

비영리 조건이 있는 데이터세트를 사용하는 경우는 더욱 혼란스럽다. 이러한 데이터세트를 학습하여 생성한 사전 학습모델, 전이 학습모델을 MIT와 같은 허용적인 라이선스로 배포할 수 있는지도 모호하다. 데이터세트의 비영리 조건이 학습모델의 상용화 여부까지 영향을 주는지도 명확하지 않다.

GPL과 같은 카피레프트 라이선스는 외부에 배포할 때 소스 코드를 공개해야 하는 의무사항이 있다. GPL의 학습모델을 활용하여 개발한 학습모델을 외부에 배포할 때 소스 코드 공개 범위가 어디까지 인지(예: weight 값 포함?)도 모호하다. 그리고 상기와 같은 이슈, 의문을 잘 해결하여 개발한 학습모델은

표 1 인공지능 과정의 이슈

분류	이슈
데이터세트	1. 데이터세트 사용 시 주의사항은 있는가?
	2. 데이터세트 제작/구축 시 주의사항은 있는가?
	3. 공개된 데이터세트를 학습용으로 자유롭게 사용해도 되는가?
학습모델	4. 라이선스가 확인되지 않은 학습모델을 사용해도 되는가?
	5. 공개된 학습모델은 자유롭게 사용해도 되는가?
저작권/라이선스	6. 데이터세트의 저작권/라이선스는 반드시 준수해야 하는가?
	7. 데이터세트 내의 구성된 데이터에 대한 저작권/라이선스는 확인 가능한가?
	8. 다양한 데이터세트들의 저작권/라이선스 의무사항 간의 충돌 여부는 확인 가능한가?
	9. 데이터세트와 학습모델의 라이선스 의무사항 간에 충돌 여부는 확인 가능한가?
	10. 데이터세트와 학습모델의 라이선스 의무사항 충돌 시 해결 발안은 있는가?
공개	11. 데이터세트 공개 시 주의사항은 있는가?
	12. 공개 시 학습모델은 데이터세트의 저작권/라이선스 의무사항을 준수해야 하는가?
법/규정	13. 데이터세트의 저작권/라이선스 위반 판단을 위한 법, 규정은 있는가?
상용화	14. 공개된 데이터세트를 활용해 생성한 데이터세트(2차 저작물)를 상용화해도 되는가?
	15. 공개된 데이터세트를 활용해 생성한 학습모델은 상용화해도 되는가?
	16. 데이터세트 저작권을 확인하지 않고 생성한 학습모델을 상용화해도 되는가?
	17. 데이터세트와 학습모델의 저작권/라이선스 의무사항이 충돌해도 상용화가 가능한가?
도구	18. 데이터세트의 저작권/라이선스 의무사항 충돌 여부를 확인할 수 있는 도구는 있는가?
서비스	19. 개발된 학습모델로 응용 서비스 시 저작권/라이선스 등 이슈가 발생할 수 있는가?

응용 서비스에서 아무런 법적 쟁점이 없는지도 확인이 필요하다.

Ⅲ. 관련 사례들

공개된 주요 데이터세트, 학습모델의 라이선스 현황을 살펴보고, 데이터세트를 활용한 인공지능 모델 관련 소송 사례도 살펴본다.

표 2 데이터세트의 라이선스 사례

데이터세트	라이선스	인공지능 학습 가능
CIFAR-10	없음, 인용 요구	언급 없음
ImageNet	이용 약관	언급 없음
ScanNet	이용 약관	언급 없음
Cityscapes	이용 약관	가능?
FFHQ	CC-BY-NC-SA-4.0	언급 없지만, 가능?
MS COCO	CC 4.0	언급 없음
Wikipedia	CC-BY-SA 4.0	언급 없음
BookCorpus	CC0	가능?
WebText2	CC0?	언급 없음

출처 Reproduced from [6], CC BY-NC-ND 4.0.

1. 데이터세트 라이선스 사례

공개된 주요 이미지, 텍스트 등 데이터세트의 라이선스는 표 2와 같다. 데이터세트에 대한 완전하고 정확한 라이선스 정보를 식별하기는 다음과 같은 이유로 어렵다.

- 데이터세트는 다른 라이선스를 가지는 여러 개의 데이터를 결합하여 생성한다.
- 데이터세트는 라이선스에 대한 명확한 정의를 확인하기가 어렵다.
- 다운로드 사이트가 폐쇄되기도 한다.
- 관련 정보 등이 변동될 수도 있다.

다음은 인공지능 학습에 사용되고 있는 주요 데이터세트의 라이선스 현황이다.

- WebText2 경우, 다운로드 사이트가 폐쇄된 사례로 정확한 정보를 확인하기 힘들다.
- Wikipedia는 텍스트 경우는 앞서 설명한 라이

선스이지만 이미지 등은 다른 라이선스 조건이다. 대부분 경우 학습용으로 사용 가능한지를 언급하지 않고 있으며, Cityscapes 등은 관련 논문, 깃허브(GitHub), 사이트에서 학습용으로 만든 데이터세트라고 기술하고 있다.

- NC(Non-Commercial) 조건을 제시하는 데이터세트도 있지만, 데이터세트 자체의 비상용화인지, 인공지능 학습에 활용한 파생 저작물까지 인지에 대해서 명확하지 않다.
- 참고로 상용 데이터세트인 게티 이미지는 '머신러닝, AI 또는 생체 인증 기술 사용금지'라고 명확하게 기술하고 있다.

2. 학습모델 라이선스 사례

공개된 데이터세트를 활용하여 학습한 인공지능의 라이선스는 표 3과 같다. 대부분은 퍼미시브 라이선스로 공개하고 있으며, OpenAI의 GPT 모델은 상용 라이선스로 전환하기도 하였다. KoGPT는 Apache 2.0 라이선스이지만, 사전 학습한 가중치는 다른 조건이기도 하다. 소송 사례에서 다룰 Stockfish 학습모델은 카피레프트 라이선스인 GPL 3.0을 사용했다.

이들 학습모델을 기반으로 인공지능 응용, 서비스 도구들은 제공하거나 상용화하는 Copilot, CodeGenX, DALLE2, chatGPT, Midjourney, Stable Diffusion 등이 있다.

표 3 학습모델의 라이선스 사례

학습모델	라이선스
GPT-2	MIT
GPT-3	Proprietary
GPT-3.5	Proprietary
KoGPT	Apache 2.0, 사전 학습된 가중치는 CC-BY-NC-ND 4.0
BERT	Apache 2.0
KorBERT	Apache 2.0
AlexNet	MIT
ResNet	Aache 2.0
VGG	MIT

3. 소송 사례

가. Getty Images vs. Stability AI 소송

(소송 내용) 세계 최대 규모의 이미지 데이터를 보유한 게티 이미지(Getty Images) 기업이 인공지능 사진 생성 도구인 '스테이블 디퓨전(Stable Diffusion)'의 개발사인 스테이빌리티 AI(Stability AI)를 상대로 이미지 데이터를 무단으로 사용했다며 소송을 제기한 사건이다[7].

(피고) 게티 이미지의 데이터들은 저작권이 있으며, 사용자는 유료 라이선스를 통해 데이터를 구매해 이용해야 하는 비즈니스 모델이다.

각 데이터에는 제목, 캡션, URL 등 각종 메타데이터 정보가 있고, 이를 데이터베이스(DB)화하고 검색 시스템을 구축하는 데 큰 비용을 투입하여 구축한 저작물이다.

게티 이미지의 이용 약관에 라이선스 없이는 웹사이트 내 어떤 데이터도 다운로드, 복사, 재전송을 할 수 없고, 데이터 마이닝, 로봇 등을 통한 데이터 수집, 추출을 할 수 없다고 명시되어 있어 해당 데이터들은 무단으로 인공지능 학습 데이터로 사용할 수 없다는 주장이다[7].

(원고) 스테이빌리티 AI는 스테이블 디퓨전 모델을 기반으로 유료 그림 생성 서비스 플랫폼인 'Dream Studio'를 운영하여 수익을 창출하고, 게티 이미지의 동의 없이 스테이블 디퓨전 모델을 오픈소스로 배포하여 제3자가 개량하고 사용하도록 하고 있다.

(진행 상황) 게티 이미지 기업의 소송뿐만 아니라

사진작가로부터 집단 소송도 진행되고 있다. 이 집단 소송 이유도 사진작가의 작품들을 무단으로 사용했다는 것이다. 이에 스테이빌리티 AI는 이번 소송에 의견이나 반론, 해명 자료를 내지 않은 상황이다[7].

(위법 사항) 게티 이미지가 주장하는 주요 위법 사항은 저작권 침해. 상표권 침해와 희석 행위, 불공정 경쟁 등이다.

나. 깃허브 사용자 vs. MS 등 소송

(소송 내용) 2022년 11월 깃허브(GitHub) 공개 저장소의 코드 저작권자 중 일부가 마이크로소프트, OpenAI 및 깃허브 등을 상대로 공개된 깃허브 저장소에서 오픈소스 라이선스와 코드 제작자의 법적 권리를 침해했다고 집단 소송을 제기한 사건이다[8].

(피고) 청구원인은 복잡하지만, 가장 주요한 내용은 Copilot 서비스의 기반인 LLM(Large Language Models) 모델이 깃허브 공개 저장소의 오픈소스 소프트웨어를 사용하여 학습이 이루어졌을 뿐만 아니라 출력되는 결과물 역시 학습에 사용된 오픈소스 소프트웨어를 사용하고 있지만, 출력물에는 오픈소스 소프트웨어 라이선스 의무사항에 따른 표기사항 등이 모두 생략되었다는 주장이다[8].

(원고) 원고들은 피해 사실을 구체적으로 제시하지 못하고, 특정 코드의 오용, 계약 위반을 제시하지 못해 공개 데이터를 사용한 것은 저작권법상 '공정사용'이라고 피력 중이다[9].

(진행 상황) 소송은 2022년 11월 깃허브 저장소의 저작권자 중 일부가 마이크로소프트, OpenAI 및 깃허브 등을 상대로 미국 캘리포니아주 북부 지방법원에 집단 소송 제기됐고, 기각을 논의하는 청문회는 지난 5월에 열렸다. 이에 대한 대응도 6월 7일까지 한 것으로 보이나 결과에 대해서는 좀 더 지켜봐야 할 것으로 보인다.

본 사건으로 LLM 기반의 AI 개발 보조도구가 디지털 밀레니엄 저작권법 위반 및 오픈소스 계약 위반의 소지가 있고, 인공지능 학습을 위하여 인터넷에 공개된 자료를 임의로 이용하는 행위가 저작권법상의 공정 이용에 해당하는지에 대한 명확한 기준이 세워질 것으로 보인다.

Copilot은 무차별적으로 대량의 데이터를 학습시킨 인공지능을 독점적인 상업 사용 모델에 사용되는 이러한 행위가 기존의 창작물 시장에 영향을 주고 있는 경우에도 공정 이용의 범주에 들 수 있는지에 대하여는 치열한 다툼이 예상되며 그 과정에서 인공지능의 공정 이용의 기준에 관하여 많은 쟁점이 확인될 것이다[8].

(위법 사항) 저작권자 허락 없이 저작권 관리정보를 고의로 제거하는 것을 금하는 미국의 디지털 밀레니엄 저작권법 제1202조 등의 위반일 뿐만 아니라 오픈소스 라이선스 조건인 고지의무 및 공개의무 등을 무시하여 오픈소스 계약 위반이다.

다. Stockfish vs. ChessBase 소송

(소송 내용) Stockfish 개발자는 2021년 7월에 ChessBase를 뮌헨 지방법원에 GPL 3.0 라이선스 위반으로 소송을 제기한 사건이다[10].

(피고) Stockfish는 해당 소프트웨어에 대한 GPL 3.0 주장 및 파생 저작물에 대한 공개를 요구하였다.

(원고) ChessBase는 Stockfish의 파생 저장물인 Fat Fritz 2 DVD와 Houdini 6 판매를 중단하였다. 이후, ChessBase는 Fat Fritz 2 SE를 판매하고 소스 코드를 공개하였으나, 학습모델의 가중치(Weight)를 공개하지 않았다[10].

(소송 진행) 'GPL 3.0상에서 학습모델의 가중치가 파생 저작물 범주에 포함되는가?'에 대해서 법원은 그렇지 않다고 판단하였다. '라이선서가 GPL

3.0 라이선스를 일방적으로 종료할 수 있는가?'에 대해서도 가중치가 파생 저작물 범주에 포함되지 않는다고 판단하여 위반한 것으로 판단하였다[10].

(진행 상황) ChessBase는 Stockfish 엔진을 사용하는 체스 프로그램 판매를 종료한다는 내용을 웹사이트에 게시하고, 이미 구매한 고객은 사본을 다운로드할 수 있게 하고 있다[10].

(위법 사항) ChessBase의 GPL 3.0 라이선스 의무사항 위반이다.

라. 기타

미국의 저작권청은 이미지 생성 AI인 미드저니(Midjourney)를 활용해 만들어진 그래픽 소설(Zarya of the Dawn) 내의 구성을 구분해서 작가가 작성한 스토리와 이미지 배열은 저작권 보호 대상으로 인정하였으나, 미드저니가 스토리에 기반해서 만든 이미지는 저작권 보호 대상으로 인정하지 않는다고 결론을 내렸다. 'AI가 만든 이미지는 인간의 창작활동의 생산물이 아니다'라는 것이 그 이유다[11].

미국 집단 소송의 주된 청구원인은 (i) 디지털 밀레니엄 저작권법 위반, (ii) 오픈소스 라이선스 계약 위반, 그 외에도 불법적 간섭, 기만행위에 따른 불법행위, 원산지 허위 표기에 따른 상표법 위반, 부당 이득 수취, 부정경쟁방지법 위반, 개인정보보호법 위반이 있다[12].

Ⅳ. 데이터세트의 쟁점

1. 저작권/라이선스 쟁점

데이터세트를 제작하거나 사용할 때 고려해야 하는 사항들을 살펴보고, 특히 인공지능 모델 학습용으로 데이터세트를 사용할 때 발생할 수 있는 저작권, 라이선스의 쟁점을 살펴보면 그림 2와 같다.

먼저 데이터세트를 직접 제작, 구축하는 경우이

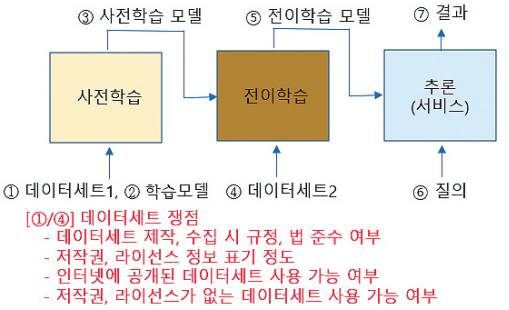

그림 2 데이터세트의 쟁점

다. 개인정보를 담은 데이터, 인간을 대상으로 수집한 데이터 등을 구축할 수 있다. 이러한 데이터를 수집하고 제작 구축하는 데 고려해야 할 법, 규정이 있는지 확인이 필요하다. 공중 배포 시 어떤 내용의 저작권 및 라이선스 정보까지 제공해야 할지 생각해야 한다.

두 번째는 인터넷에 공개된 이미지, 텍스트 등 데이터를 크롤링하여 수집(Common Crawling)하고 사용하는 것이 적법한가이다. 데이터세트는 누군가의 노력으로 만들어진 창작물로 저작권의 보호 대상이다. 이러한 데이터세트가 인터넷에 공개되어 있다고 하여 임의로 사용해도 문제가 없는지 의문이다. 더욱이 저작권, 라이선스가 없는 데이터세트는 더 자유롭게 사용할 수 있지 않을까 생각할 수 있다. 누군가의 창작물로 사용 시 저작권, 라이선스 침해가 있을 것 같다. 빅데이터 처리, 인공지능 학습 과정에서 데이터 복제, 분석 등이 있어 무분별하게 활용하는 경우 저작권 위반 행위가 될 수 있다.

2. 저작권법 개정(안)

현재 정부에서 발의한 「저작권법」 일부개정법률안으로 제35조의5 신설을 2022년 10월에 발의하였다. 주요 내용은 다음과 같다[13].

- 적법하게 접근한 저작물에 대해 자동화된 정보

분석을 하려는 경우 저작물의 복제·전송이 가능하게 하고

- 정보분석의 목적으로 복제된 저작물에 대해서는 복제방지 조치 등 필요한 조치를 하도록 하며
- 별도의 저작물 이용 허락받지 아니한 경우에는 정보분석 결과물에 대한 이용 목적을 제한하고
- 저작물 이용 허락받으면 자유롭게 이용하도록 함으로써 데이터 관련 사업자의 활동 영역을 보장하고 저작권자의 권리를 보호하려는 것이다.

하지만 현재 제안된 조항이 인공지능 업계의 예측 가능성을 높일 수 있는 구체적인 기준을 제시하고 있는지는 의문이다. 특히 해당 저작물에 '적법하게' 접근할 수 있는 경우에 한정한다는 단서 조항은 '공정사용(Fair Use)'[14]에 관한 기준보다 더 혼란을 줄 수도 있다. 개인정보 등 민감 데이터나 특정 연구에 관한 데이터를 분석하여 그 결과를 부적절하게 이용하는 상황 등의 내용이 더욱 명확하게 정리되어야 한다.

미국은 '공정사용' 원칙을 적용해 변형이나 표현의 자유 등을 위해서는 저작권이 있는 자료라도 사용할 수 있도록 허용하고 있다. 독일이나 영국은 작품을 인용하거나 다른 사람과 비슷한 스타일로 초상화를 그리는 것은 허용하지만, 저작권자의 허가 없이 인공지능 모델을 훈련하기 위해 예술적 이미지를 이용하는 것은 허용되지 않는다. 국내에서는 아직 인공지능 저작권 관련 소송이 제기된 적이 없지만, 저작권자에게 승산이 있다는 게 전문가들의 견해다[15].

Ⅴ. 학습모델 쟁점

인공지능 학습은 많은 데이터가 필요하다. 데이터들로 만들어진 데이터세트는 인공지능 학습에 활용한다. 인공지능 학습을 위한 오픈소스 소프트웨어 배포도 늘어가고 있다. 데이터세트의 라이선스와 무관하게 오픈소스 소프트웨어를 자유롭게 사용하고, 특히 기업이 상업용으로 사용 가능한지 등의 쟁점을 이 장에서 살펴본다.

인공지능은 그림 3의 학습 과정을 거쳐 학습하며, 최종 학습한 모델은 사용자가 입력하는 질의, 데이터에 대해서 추론 과정, 즉 응용 서비스 과정을 통해 결과를 사용자에게 제공한다. 사전학습 및 전이학습 과정에서 쟁점은 다음과 같다.

먼저, 인터넷상에 공개된 데이터세트를 수집하여 학습하는 과정에서의 쟁점으로 데이터세트의 저작권 이슈이다. 학습 과정에서 각 계층의 함수 가중치 계산을 위해서 데이터세트를 복제한다. 이 복제 과정이 저작권 침해에 해당하는지 쟁점이 될 수 있다. 또한, 학습에 사용한 데이터세트가 '적법하게' 활용되었는지 불명확하다. 현재 관련 판례가 많지 않아 의문을 가질 수밖에 없다.

두 번째는 데이터세트의 라이선스 이슈이다. 사용한 데이터세트의 라이선스에 비영리 조건이 있는 경우에 학습한 모델을 공개적으로 배포하거나 상업용으로 사용할 수 있는지이다. 또한 학습 결과인

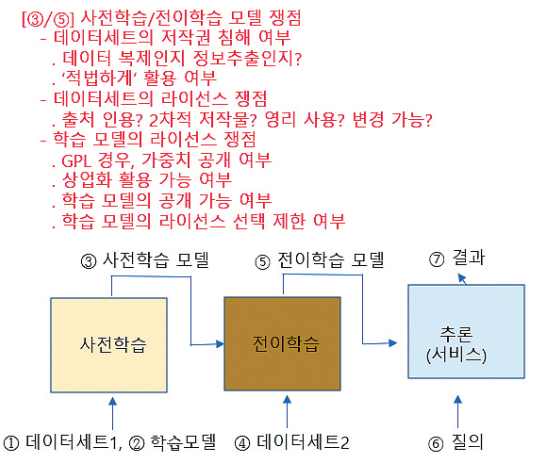

그림 3 학습모델의 쟁점

가중치가 2차적 저작물에 해당하는지, 그리고 학습 결과가 원본 데이터를 변경했는지에 대한 쟁점도 있을 수 있다.

마지막으로 학습모델의 라이선스 이슈이다. GPLv2로 공개한 학습모델 경우, 이 오픈소스로 학습한 모델뿐만 아니라 가중치까지 공개해야 하는지 이슈가 있다. 이 이슈는 이미 Stockfish 커뮤니티와 ChessBase 기업 간의 독일 소송 사례에 있다. 또한 데이터세트의 라이선스 조건에 따라 학습모델의 공개 여부와 공개 라이선스 선택 등이 달라져야 하는지도 의문이다. 특히 비영리 조건의 라이선스가 있는 데이터세트로 학습한 모델을 상업용으로 사용하지 못 하는지는 큰 관심거리이다.

VI. 추론/서비스 쟁점

응용 서비스 단계를 살펴보자. 학습한 모델을 상용 서비스에 사용할 수 있다고 가정하면 무슨 문제가 발생할 수 있는지 그림 4를 통해 알아보자.

데이터세트의 저작권과 라이선스를 잘 준수하여 학습한 모델은 인공지능 응용 서비스 과정에 아무런 이슈가 없다고 할 수 있는지이다. 서비스 결과물이 기존 저작물 등 침해가 있을 수 있고, 더군다나 그 결과물을 저장장치에 영구적으로 저장하거나 디스플레이 장치로 타인에게 전송하는 경우 등 문제의 소지가 있을 수 있다.

VII. 이슈에 대한 견해

이 장에서는 II장에서 기술한 인공지능 학습과정의 이슈에 대해서 견해를 기술하고자 한다. 이들 이슈는 한창 진행 중이고 법적 검토 등이 기반이 되어야 하므로 명확한 의견을 피력하기에 무리가 있는 부분도 있다.

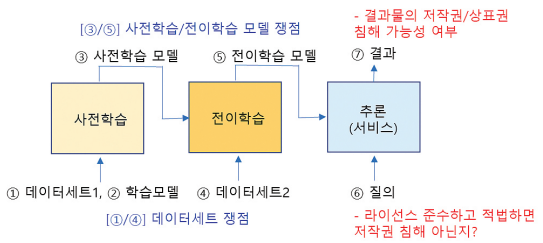

그림 4 추론/서비스 쟁점

학습모델을 영리 목적으로 사용한다면 다소 복잡한 이슈와 견해가 있으며, 비영리 목적으로만 사용한다면 데이터세트, 모델의 출처 인용과 라이선스를 준수하면 이슈는 없을 것이다.

지금까지 오픈소스 컴플라이언스는 소프트웨어 코드에 중점을 두고 컴플라이언스 절차, 도구 등이 발전해 왔다. 하지만 인공지능 분야에서는 데이터, 데이터세트, 학습모델의 소스 코드, 학습한 결과물의 소스 코드를 대상으로 라이선스 확인과 준수 사항, 표 4의 이슈 사항 등으로 기존의 오픈소스 컴플라이언스 범위가 확장되어야 한다.

VIII. 결론 및 향후 과제

데이터세트와 학습모델 간의 저작권, 라이선스 충돌 여부 및 사례를 살펴보았다.

우리나라는 인공지능 기술력 향상, 학습용 데이터 축적 691종, 인공지능 반도체 출시 등으로 인공지능 기반 조성에 세계 시장에서 성과를 창출할 수 있어[16], 앞서 살펴본 의문점의 해결 방안이 시급해 보인다.

판단형 AI에서 데이터 사용은 공정이용으로 저작권 이슈가 없으나 생성형 AI는 데이터세트와 유사한 결과가 나오므로 공정이용이 아니라는 견해도 있다.

따라서, 인터넷에 공개된 자료를 기반으로 초거대 인공지능 모델이 연구 개발되고 이제는 유료 서비스

표 4 이슈에 대한 견해

분류	이슈	견해
데이터 세트	1. 데이터세트 사용 시 주의사항은 있는가?	• 저작권, 라이선스, 주의사항 반드시 확인 • 인공지능 학습용으로 활용 가능 여부를 확인 • 이전의 활용 사례 확인
	2. 데이터세트 제작/구축 시 주의사항은 있는가?	• 준수해야 할 관련 법, 규정 여부 확인 • 저작자, 라이선스, 주의사항 정보 제공
	3. 공개된 데이터세트를 학습용으로 자유롭게 사용해도 되는가?	• 라이선스, 주의사항 정보를 확인 • 이전의 활용 여부 확인 • 사용 시 반드시 출처 인용 • 데이터세트 저작권 위반 가능성은 높지 않음
학습 모델	4. 라이선스가 확인되지 않은 학습모델을 사용해도 되는가?	• 저작권자에게 확인 후 사용
	5. 공개된 학습모델은 자유롭게 사용해도 되는가?	• 사용 시 라이선스 준수 • 학습모델이 사용한 데이터세트의 라이선스, 주의사항 등 확인
저작권/ 라이선스	6. 데이터세트의 저작권/라이선스는 반드시 준수해야 하는가?	• 준수해야 하며, 반드시 출처를 인용
	7. 데이터세트 내의 구성된 데이터에 대한 저작권/라이선스는 확인 가능한가?	• 자동으로 확인할 수 있는 도구가 없으며 수동으로 가능 • 경우에 따라 정보가 없거나 찾을 수 없음
	8. 다양한 데이터세트들의 저작권/라이선스 의무사항 간의 충돌 여부는 확인 가능한가?	• 상동
	9. 데이터세트와 학습모델의 라이선스 의무사항 간에 충돌 여부를 확인 가능한가?	• 상동
	10. 데이터세트와 학습모델의 라이선스 의무사항 충돌 시 해결 방안은 있는가?	• 없음 • 여러 사례를 분석하여 가이드라인 마련 필요
공개	11. 데이터세트 공개 시 주의사항은 있는가?	• 저작권, 라이선스, 주의사항 등 명확한 정보 제공
	12. 공개 시 학습모델은 데이터세트의 저작권/라이선스 의무사항을 준수해야 하는가?	• 학습모델이 데이터세트의 파생 관계(예: 학습모델이 데이터세트의 2차적 저작물인지, 변경한 것인지 등)에 따라 판단이 다름
법/ 규정	13. 데이터세트의 저작권/라이선스 위반 판단을 위한 법, 규정은 있는가?	• 저작권법 등 있으며, 개정 중
상용화	14. 공개된 데이터세트를 활용해 생성한 데이터세트(2차 저작물)를 상용화해도 되는가?	• 저작권, 라이선스 의무사항에 의해 가능 여부 확인
	15. 공개된 데이터세트를 활용해 생성한 학습모델은 상용화해도 되는가?	• 데이터세트 저작권, 라이선스 위반이 없다면 가능 • 데이터세트에 비상용 조건의 라이선스 경우, 가능 여부는 모호 • 인공지능 학습용도로 불가 조건이 있는 경우는 상용화는 불가능
	16. 데이터세트 저작권을 확인하지 않고 생성한 학습모델을 상용화해도 되는가?	• 저작권이 없는 경우는 불가능
	17. 데이터세트와 학습모델의 저작권/라이선스 의무사항이 충돌해도 상용화가 가능한가?	• 의무사항 충돌은 상기 12번 견해를 참조 • 상용화 가능성은 상기 15번 견해를 참조
도구	18. 데이터세트의 저작권/라이선스 의무사항 충돌 여부를 확인할 수 있는 도구는 있는가?	• 없음
서비스	19. 개발된 학습모델로 응용 서비스 시 저작권/라이선스 등 이슈가 발생할 수 있는가?	• 저작권, 상표권 등 이슈가 있을 수 있음

로 상업적 수단이 되고 있어, 해당 서비스로 인하여 기존의 창작물 시장에 영향을 받는다면 인공지능 학습 행위가 공정 이용인지 신중하게 살펴보아야 하고, 특히 상업적인 사용을 희망한다면 가능한 범위에서 별도의 검증 절차를 진행할 필요가 있다[17].

소프트웨어 정책연구소는 최근 발간한 '생성 AI

부상과 산업의 변화' 보고서에서 생성형 인공지능 산업 육성을 위해 저작권 등 법제도·규제 대응과 데이터 유출·일자리 대체 등 사회적 이슈에 대응 필요성을 강조했다[18]. 그리고 정부는 제2차 인공지능 윤리정책 포럼의 기술 분과에서는 기업이 데이터 및 인공지능 모델, 위험관리 체계에서 고려할 것은 데이터 및 인공지능 모델의 소유 및 사용 주체를 명확히 하고, 관리를 위한 정부의 가이드라인 제시가 필요하다고 하였다[19]. ChatGPT의 등장으로 인공지능에 대한 저작권 침해 소송도 이어지고 있다[20]. 한편, 서울대학교의 AI의 윤리적·법적·사회적 이슈(Ethical, Legel, Social Issue)를 연구하는 센터장은 공개된 데이터를 가공해서 다른 분야에 활용할 때는 데이터에 대한 저작권을 인정하여 수익을 공유하는 방법도 고민해 볼 만하다고 했다[21].

또한, 한국저작권위원회는 저작권의 쟁점과 대안 논의를 거쳐 창작자의 권리를 보호하고 이용자의 책임 있는 인공지능 이용을 장려하는 정책을 준비해 나간다고 하였다[22]. 그리고 한국지능정보사회진흥원(NIA)은 데이터 이용 과정에서 발생한 분쟁을 조정할 데이터분쟁조정위 홈페이지를 2023년 12월에 구축할 예정이다[23].

이러한 상황을 고려해 볼 때 아직 공개된 데이터세트를 학습시키는 행위가 「저작권법」 위반인지 아닌지 명확하게 답할 수 없으며, 비상용 조건이 있는 데이터세트로 학습한 모델을 상업용으로 사용할 수 있는지도 명확하지 않다. 또한 데이터세트 저작권과 학습모델 라이선스 간의 의무사항 충돌에 대한 명확한 처리 방안도 없는 상황이다. 따라서 이에 대한 법률적 문제점을 확인하고 개선하여, 저작권과 라이선스 위반이 발생하지 않는 올바른 인공지능 기술 개발(안)이 필요해 보인다.

용어해설

인공지능(AI) 컴퓨터로 구현한 지능 또는 이와 관련한 전산학의 연구 분야

생성형 인공지능(Generative AI) 이용자의 특정 요구에 따라 결과를 생성해내는 인공지능

데이터세트 연관된 데이터를 모아서 특정 규칙에 따라 하나의 묶음으로 만든 데이터의 집합

오픈소스 라이선스 오픈소스SW의 사용, 복제, 수정, 배포와 관련하여 허용되는 권한 범위를 명시한 이용 허락 조건

오픈소스 컴플라이언스 오픈소스SW를 활용하면서 발생할 수 있는 법률적 문제 등과 관련한 다양한 위반사항을 사전에 검증하여 위험 발생 가능성과 영향력을 최소화하는 방안을 수립하고 실행하는 일련의 활동

저작권 학술, 문학 및 기타 예술의 범위에 속하는 창작물(저작물)에 대한 창작자(저작자)의 권리

chatGPT(챗GPT) GPT-3.5와 GPT-4를 기반으로 하는 대화형 인공지능 서비스

퍼미시브 라이선스(Permissive License) 기존 오픈소스의 라이선스를 수정하여 배포할 수 있고, 다른 라이선스로 전환 가능. 해당 라이선스가 적용된 오픈소스는 공개하지 않아도 됨(예: Apache 2.0, BSD-2, BSD-3, BSD-4, MIT 라이선스 등)[24]

카피레프트 라이선스(Copyleft License) 기존 오픈소스의 라이선스를 수정하여 배포할 수 있고, 다른 라이선스로 전환은 불가능. 해당 라이선스가 적용된 오픈소스는 공개하여야 함(예: 약한(Weak) 카피레프트 라이선스는 EPL 1.0, LGPL 2.1/3.0, MPL 2.0 등, 강한(Strong) 카피레프트 라이선스는 AGPL 3.0, GPL 2.0/3.0 등)

CC-BY(저작자 표시) 저작자 및 출처만 표시하면 제한 없이 자유롭게 이용 가능[25]

CC-BY-NC(저작자표시-비영리) 저작자 및 출처를 표시하면 자유롭게 이용할 수 있지만, 상업적 이용은 불가능. 상업적 이용을 원하면 저작자와 별도의 계약이 필요[25]

CC-BY-ND(저작자 표시-변경 금지) 저작자명 및 출처를 표시하면 자유롭게 이용할 수 있음. 다만, 저작물을 변경하거나 저작물을 이용하여 새롭게(2차적 저작물*) 제작하는 것을 금지함[25]. *번역, 편곡, 변형, 각색, 영상 제작 등

CC-BY-SA(저작자 표시-동일 조건변경 허락) 저작자 및 출처를 표시하면 자유롭게 이용할 수 있음. 다만, 저작물을 이용하여 새롭게 저작물(2차적 저작물)을 제작하는 것은 허용하지만, 새로운 저작물에 원저작물과 같은 라이선스를 적용해야 함[25]

CC-BY-NC-SA(저작자 표시-비영리-동일 조건변경 허락) 저작자 및 출처를 표시하면 자유롭게 이용할 수 있지만, 상업적으로는 이용할 수 없음. 상업적 이용을 한다면 저작자와 별도의 계약을 해야 함. 또한, 저작물을 이용하여 새롭게 저작물(2차적 저작물)을 제작하는 것은 허용하지만, 새로운 저작물에 원저작물과 같은 라이선스를 적용해야 함[25]

약어 정리

AI	Artificial Intelligence
CC-BY-NC	Creative Commons-Attribution-Non-Commercial
CC-BY-ND	Creative Commons-Attribution-No-Derivative
CC-BY-SA	Creative Commons-Attribution-Share-Alike
CC-BY-NC-SA	CreativeCommons-Attribution-Non-Commercial-Share-Alike
CCL	Creative Commons License
CNN	Convolutional Neural Network
chatGPT	chat Generative Pre-trained Transformer
GAI	Generative Artificial Intelligence
GPL	GNU General Public License
GPT	Generative Pre-trained Transformer
KoGPT	Korean Generative Pre-trained Transformer
LLM	Large Language Models

참고문헌

[1] https://ko.wikipedia.org/wiki/인공지능

[2] 박정호 외, "AI 프레임워크 설계를 위한 필수 요소에 대한 연구," ISET, (Jeju, Republic of Korea) May, 2022, pp. 110-111.

[3] 이동현, 장진철, "국내 인공지능 소프트웨어 시장현황 진단 및 시사점," SW중심사회, 2023년 4월호, pp. 30-65.

[4] 안성원 외, "AI Index 2023 주요 내용과 시사점," SPRI Issue Report, IS-159, 2023. 5. 8., p. 44.

[5] 전자신문, "'AI 데이터 학습 허용해야' vs '사용된 DB 출처 공개부터'," 2023. 6. 14., http://mail2.scrapmaster.co.kr/mail/include/iitp_display.php?news_id=59345&scrapBookNo=1744&scrapinfo=202306140&article_serial=20230614jj00021001&q=aWl0cHwyMDIzMDYxNDB8MjAyM7PiIDA2v/kgMTTAzyAgKOKpKSDBtrCj

[6] G.K. Rajbahadur et al., "Can I use this publicly available dataset to build commercial AI software? - A case study on publicly available image datasets," 2022, arXiv preprint, CoRR, 2022, arXiv: 2111.02374.

[7] https://fingfx.thomsonreuters.com/gfx/legaldocs/byvrlkmwnve/GETTY%20IMAGES%20AI%20LAWSUIT%20complaint.pdf

[8] 이완근, "[기고] 생성형AI 개발도구 Copilot의 오픈소스 라이선스 위반과 저작권 분쟁, OSS(oss.kr)," 공개SW 가이드/보고서, 2023. 6. 26., https://www.oss.kr/oss_guide/show/edcf990f-38e7-4918-8c91-d989babf3c09

[9] 법률신문 "지구촌 'AI저작권' 분쟁 본격화…국내에도 '불똥' 튈까?," 2023. 7. 5.

[10] Darkcrizt, "Stockfish는 체스 엔진 사용에 대해 ChessBase와 합의에 도달," 2022. 1. 12. https://blog.desdelinux.net/ko/stockfish-llego-a-aun-acuerdo-con-chessbase-por-usar-su-motor-de-ajedrez/

[11] United States Copyright Office, Re: Zarya of the Dawn (Registration # VAu001480196), Feb. 2023, https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

[12] 법률신문, "초거대 AI 모델 관련 법적 이슈," 2023. 2. 14., https://www.lawtimes.co.kr/news/185319

[13] 이용호 외, "저작권법 일부개정법률안," 열려라국회, 2022. 10. 31.

[14] https://namu.wiki/w/%EA%B3%B5%EC%A0%95%20%EC%9D%B4%EC%9A%A9

[15] 조선일보, "챗GPT 같은 '생성형 AI' 뜨자 저작권 소송도 쏟아진다," 2023. 2. 24. https://www.chosun.com/economy/ weeklybiz/2023/02/23/THN52NAXMBD7LJ6RMZBF54CT3A/

[16] 과학기술정보통신부, "초거대AI 경쟁력 강화 방안," 관계부처 합동, 2023. 4. 14., pp. 2-58.

[17] 법률신문, "AI 창작물의 저작권 인정 문제," 2023. 6. 22., https://www.lawtimes.co.kr/news/185692

[18] 전자신문, "생성 AI 육성, 규제·데이터 유출 대응 필요," 2023. 6. 19., p. 19.

[19] 파이낸셜뉴스, "생성형 AI, 윤리적으로 활용하려면?," 2023. 7. 5., https://www.fnnews.com/news/ 202307051336580128

[20] 연합뉴스, "美 코미디언 실버먼, 메타·오픈AI 상대 소송 "저작권 침해"," 2023. 7. 11., https://www.yna.co.kr/view/AKR20230711003100091?input=1195m

[21] 머니투데이, "챗GPT는 헛소리 생성기…'AI 윤리' 가이드라인 필요," 2023. 7. 24., p. 15.

[22] 머니투데이, "AI시대, 저작권시스템의 역할," 2023. 8. 4., p. 21.

[23] 인공지능신문, "데이터 활용에서 발생하는 분쟁, 신속 조정 및 피해 구제한다!…데이터분쟁조정위원회 출범," 2023. 10. 12., https://www.aitimes.kr/news/articleView.html?idxno=29081

[24] 류원옥, 이승윤, "주요 오픈소스 라이선스 사용 동향," 주간기술동향 1950호, 2020. 6. 10.

[25] https://ccl.cckorea.org/