Biweekly Report
통권 제73호(2026. 1-1)

www.copyright.or.kr

저작권 이슈 트렌드

COPYRIGHT ISSUE TREND

한국저작권위원회
KOREA COPYRIGHT COMMISSION

CONTENTS

저작권 이슈 트렌드

Biweekly Report | 통권 제73호(2026. 1-1)

AI 시대의 저작권 보호와 디지털 워터마킹 기술 동향

뉴스 브리프

AI 기술의 발전이 데이터 확보 경쟁을 심화시키면서, 스포티파이 음원 8,600만 건의 무단 수집 사례와 같이 저작물이 학습 데이터로 무단 활용되는 문제가 나타났다. 이러한 행위는 원저작자의 권리를 침해하고, AI 산출물의 출처를 불분명하게 만들어 기존의 보상 체계를 무력화하며, 창작 생태계의 신뢰 기반을 훼손하는 결과로 이어진다. 본 보고서는 이러한 문제에 대응하여, AI가 콘텐츠를 산출하는 과정에 직접 개입해 식별 정보를 각인하는 '디지털 워터마킹' 기술의 원리와 가능성을 분석한다. 이는 기술적 투명성을 통해 저작권 침해 여부를 판별하고, 나아가 데이터 기여도에 따른 공정한 보상 시스템을 설계함으로써 인간 창작자와 AI 기술이 공존하는 지속 가능한 저작권 질서를 모색하는 데 목적을 둔다.

뉴스 플러스

Ⅰ. 서론 : AI 학습 데이터 수집과 저작권 위기의 부상

- **AI 학습 데이터의 무단 수집과 저작권의 사각지대**

2025년 12월, 해적판 도서·논문 공유 사이트 '안나스 아카이브(Anna's Archive)'는 스포티파이에서 8,600만 개의 음원 파일을 무단 수집했다고 발표하여 저작권 생태계에 파장을 일으켰다.[1] 이 사건은 AI 모델 학습을 위해 창작자의 동의 없이 저작물이 활용될 수 있다는 위협을 보여주는 사례가 되었다. 일부 전문가들은 해적판 자료를 통한 AI 학습이 업계에 존재하는 관행이라고 지적하며, 이는 창작자의 권리가 기술 발전 과정에서 무시당하는 문제임을 시사한다.

1) Dan Milmo, "Activist group says it has scraped 86m music files from Spotify", The Guardian, 2025.12.22., https://www.theguardian.com/technology/2025/dec/22/activist-group-says-it-has-scraped-86m-music-files-from-spotify

이러한 데이터의 무단 수집은 창작자에게 보상을 제공하지 않아 기존의 저작권 질서에 영향을 미치는 행위이다. 현행 저작권법은 저작물의 '복제'와 '배포'를 중심으로 설계되었기에, 대규모 데이터를 학습하는 AI 모델을 명확하게 규제하는 데에는 한계가 있다. 이러한 법적 공백으로 인해 AI 기업들은 저작권 침해 논란을 피하며 데이터를 사용하고 있다.

또 다른 문제는 AI에 의해 산출된 결과물이 어떤 원본 데이터를 기반으로 만들어졌는지 역추적하는 것은 기술적으로 어렵다는 것이다. 이는 물감을 섞어 새로운 색을 만든 후 각 원색의 배합 비율을 역추적하기 어려운 것과 유사하다. 이러한 기술적 한계는 저작권 침해 사실의 입증을 어렵게 만들어, 권리자가 자신의 권리를 효과적으로 주장하는 데 장애물로 작용하고 있다.

• 새로운 질서의 열쇠, AI 산출물 식별 기술의 중요성

AI 산출물과 인간의 창작물을 구별하기 어려워지면서, 콘텐츠의 진위와 신뢰성을 둘러싼 사회적 논의가 이루어지는 상황 속에서 식별 기술의 필요성이 높아지고 있다. AI 산출물의 출처와 이력을 밝히는 기술은 저작권 침해를 막고, 허위 정보 확산을 방지하는 사회적 안전망의 역할도 수행할 수 있다. 보이지 않는 표식을 통해 콘텐츠의 정체성을 증명하는 '디지털 워터마킹'과 같은 기술이 이러한 문제에 대한 해결책 중 하나로 여겨진다.

기술로 인해 발생한 문제는 또 다른 기술을 통해 해결의 실마리를 찾을 수 있다. AI 산출 콘텐츠 식별 기술은 저작권 침해 여부를 판별하는 역할을 넘어, 원본 데이터 기여도에 따라 수익을 배분하는 새로운 보상 체계를 설계하는 기술적 기반이 될 수 있다. 이를 통해 기술의 발전과 창작자의 권리가 공존하는 지속 가능한 창작 생태계를 구축할 필요가 있다.

Ⅱ. 본론: 디지털 워터마킹의 기술적 원리와 법적 쟁점

• 디지털 워터마킹의 작동 원리

디지털 워터마킹은 이미지, 영상, 오디오 등 디지털 콘텐츠에 인간이 인지할 수 없는 정보를 삽입하여 그 출처나 소유권을 증명하는 기술이다. AI 산출물과 인간의 창작물을 구별하기 어려워지면서, 디지털 워터마킹은 콘텐츠에 일종의 '디지털 출생증명서'를 발급하여 해당 콘텐츠의 진위 여부를 판별하고 불법 복제를 추적하는 기술로 활용된다.

한편, 효과적인 워터마킹 기술은 비가시성과 강인성이라는 두 가지 요건을 동시에 충족해야 한다. 워터마크는 인간의 시각이나 청각으로는 감지할 수 없도록 삽입되어야 하며, 이를 통해 사용자의 감상 경험을 저해하지 않아야 한다. 또한 파일 압축, 형식 변환, 일부 편집 등 각종 데이터 처리 과정에서도 삽입된 정보가 훼손되지 않고 유지되는 강인성을 갖추어야 저작물 보호라는 목적을 달성할 수 있다.

• 수동적 탐지를 넘어서는 사전 예방적 접근법

기존 콘텐츠 식별 기술이 이미 산출된 결과물을 분석하여 특징을 찾는 수동적 방식에 머물렀다면, 최근 워터마킹 기술은 '사전 예방적(proactive)' 접근법을 취한다는 점에서 차이가 있다. 사전 예방적 워터마킹은 AI 모델이 콘텐츠를 산출하는 시점에 직접 개입하여 식별 정보를 데이터 구조에 내장하는 방식이다. 이는 완성된 종이 위에 도장을 찍는 것이 아니라, 종이를 만드는 펄프 단계에서부터 특정 문양을 얇게 넣어 종이 자체의 일부가 되게 하는 위조지폐 방지 기술과 유사하다.

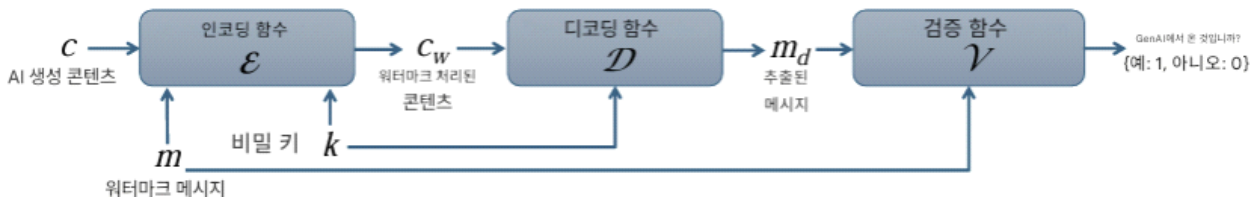
이러한 접근법은 외부의 변형이나 공격에 대해 더 높은 저항력을 가진다. 콘텐츠가 산출된 이후에 덧붙여진 정보는 비교적 쉽게 제거하거나 수정할 수 있지만, 산출 과정부터 내재된 워터마크는 콘텐츠 구조 자체와 결합되어 있어 분리하기 어렵다. 이 기술은 AI 산출물의 출처 증명에 대한 신뢰도를 높여, 저작권 분쟁에서 기술적 증거를 제공하는 기반이 된다.

• AI 신경망에 각인하는 워터마크 기술 구현

AI 오디오 워터마킹의 기술적 구현은 크게 세 단계로 이루어진다. 첫 단계는 워터마크를 삽입하는 인코딩이다. 이 과정에서는 AI 모델의 신경망 파라미터*나 결과물의 확률 분포를 직접 수정하여 워터마크 정보를 삽입한다. 예를 들어, 특정 음성이나 멜로디 패턴이 나올 확률을 미세하게 조정하거나, 특정 주파수 대역에 인간이 감지할 수 없는 고유한 신호를 포함시키는 방식이다. 이 모든 과정은 사전에 약속된 비밀 키를 기반으로 암호화되어 허가되지 않은 접근을 차단한다.

* 신경망 파라미터(Neural Network Parameters): AI 인공 신경망의 내부 설정값. 대규모 데이터 학습을 통해 얻어진 수치들로, AI가 어떻게 콘텐츠를 산출할지 결정하는 핵심 요소

[그림] 디지털 워터마킹의 3단계 작동 원리



출처: Lele Cao, "Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities", arXiv, 2025.04.02., https://arxiv.org/pdf/2504.03765

두 번째 단계는 숨겨진 정보를 다시 추출하는 디코딩이다. 워터마크가 삽입된 오디오 파일이 주어지면, 인코딩 시 사용된 것과 동일한 비밀 키를 가진 탐지 알고리즘이 작동한다. 이 알고리즘은 오디오의 패턴 속에서 통계적 분석을 통해 내재된 워터마크 신호를 식별하고, 이를 원래의 정보 비트로 복원하는 역할을 한다. 이 단계의 정확성은 워터마킹 시스템 전체의 신뢰도에 영향을 미친다.

마지막 단계는 추출된 정보의 진위를 판정하는 검증이다. 디코딩을 통해 복원된 정보 비트가 원본 워터마크 정보와 일치하는지를 비교하여 최종 결론을 내린다. 이 비교 결과가 특정 임계값을 넘어서면, 해당 콘텐츠는 특정 AI 모델의 산출물임이 높은 신뢰도로 증명되는 것이다. 이 세 단계의 유기적인 작동을 통해 워터마킹은 AI 산출물의 출처를 명확히 밝히는 기술적 증거로서의 가치를 갖게 된다.

• 워터마크 기술의 강인성과 현재 드러난 한계

워터마킹 기술의 실효성은 '강인성'에 의해 영향을 받는다. 정교하게 워터마크를 삽입했더라도 파일 압축, 소음 추가, 형식 변환 등 데이터 처리 과정에서 사라진다면 그 유용성이 떨어지기 때문이다. 따라서 최신 연구는 오디오 신호의 다양한 변형에도 워터마크가 통계적으로 유의미하게 유지될 수 있도록 하는 알고리즘 개발에 집중하고 있으며, 이는 기술 상용화를 위한 전제 조건 중 하나이다.

그러나 현재의 워터마킹 기술은 한계를 가지고 있으며, 악의적인 목적을 가진 '적대적 공격*'에 취약할 수 있다. 공격자는 워터마크 탐지 시스템의 작동 방식을 분석하여, 인간은 인지하기 어려운 수준의 노이즈를 오디오에 추가함으로써 워터마크를 무력화하거나 제거할 수 있다. 이러한 공격은 워터마킹의 신뢰도를 낮출 수 있는 기술적 도전 과제이다.

* 적대적 공격(Adversarial Attack): AI 모델의 취약점을 이용하는 공격 기법. 인간은 인지하기 어려운 미세한 노이즈나 데이터 변형을 가하여 AI가 오작동하도록 유도하거나, 워터마크 같은 보안 기능을 무력화하는 방식

이러한 한계를 극복하기 위해, 워터마크 자체를 더 복잡하게 만들거나 탐지 알고리즘을 고도화하는 연구가 진행되고 있다. 예를 들어, 여러 개의 워터마크를 중복해서 삽입하거나, 공격을 받으면 스스로 변화하는 동적 워터마크 기술 등이 대안으로 제시된다. 기술적 보호 장치와 이를 무력화하려는 공격 사이의 경쟁은, 워터마킹 기술이 지속적으로 발전해야 하는 이유를 보여준다.

• 공정한 보상 체계를 위한 기술적, 제도적 과제

디지털 워터마킹과 같은 기술적 해결책이 현실에 안착하기 위해서는 법적, 제도적 지원이 요구된다. AI 기업들이 데이터 수집 행위를 '공정 이용'이라고 주장하는 법적 논쟁 속에서, 워터마킹 기술은 저작권 침해 여부를 판단하는 객관적인 증거를 제공하여 논의를 진전시킬 수 있다. 나아가 이는 창작자가 자신의 저작물이 AI 학습에 사용되었음을 입증하는 기술적 수단이 된다.

궁극적으로 워터마킹 기술은 무단 이용을 억제하는 역할을 넘어, 창작자의 기여에 대한 보상을 실현하는 산업 생태계를 구축하는 기반이 될 수 있다. 예를 들어, AI 산출물에 삽입된 워터마크를 통해 원본 데이터 기여도를 추적하고 로열티를 분배하는 시스템을 법제화하는 방안을 고려해볼 수 있다. 기술의 발전과 창작자의 권리가 공존하기 위해서는, 기술 도입을 의무화하고 투명한 보상 체계를 마련하는 정책적 논의가 필요하다.

Ⅲ. 결론 : 기술과 제도의 조화를 통한 지속 가능한 창작 생태계 구축

• 기술과 제도의 조화, 지속 가능한 창작 생태계의 조건

AI 기술 발전 과정에서 나타난 데이터 무단 수집 문제는 창작자의 권리를 위협하는 요인으로 작용하고 있다. 이러한 환경에서 AI 산출물의 출처를 밝히는 디지털 워터마킹 기술이 주목받고 있다. 워터마킹 기술을 통해 AI 산출 과정의 투명성을 확보하는 것은, 창작자가 자신의 데이터 기여도를 증명하고 그에 대한 권리를 주장할 수 있게 하는 기본적인 전제 조건이 된다.

그러나 본론에서 분석했듯이, 디지털 워터마킹은 기술적 해결책의 일부일 뿐이며, 적대적 공격과 같은 의도적인 무력화 시도에 여전히 취약점을 보인다. 이는 기술적 보완과 함께 제도적 강제가 병행되어야 함을 시사한다. 기술적 장치가 있더라도 이를 회피하려는 시도를 억제할 법적, 제도적 장치가 없다면 그 실효성이 저하될 수 있기 때문이다.

따라서 정책 수립자들은 AI 기업에 워터마킹 기술 도입을 의무화하는 방안을 검토하고, 창작자와 기술 기업이 참여하는 사회적 협의체를 통해 투명한 데이터를 기반으로 한 수익 분배 모델을 구축할 필요가 있다. 결국 기술적 신뢰와 제도적 공정성이 조화를 이룰 때, 비로소 인간의 창의성과 AI 기술이 함께 성장하는 지속 가능한 미래를 기대할 수 있을 것이다.

참고문헌

• Dan Milmo, “Activist group says it has scraped 86m music files from Spotify”, The Guardian, 2025.12.22., https://www.theguardian.com/technology/2025/dec/22/activist-group-says-it-has-scraped-86m-music-files-from-spotify
• Lele Cao, “Watermarking for AI Content Detection: A Review on Text, Visual, and Audio Modalities”, arXiv, 2025.04.02., https://arxiv.org/pdf/2504.03765

기술용어

순번	용어	설명
1	신경망 파라미터 (Neural Network Parameters)	AI 인공 신경망의 내부 설정값. 대규모 데이터 학습을 통해 얻어진 수치들로, AI가 어떻게 콘텐츠를 산출할지 결정하는 핵심 요소이다.
2	적대적 공격 (Adversarial Attack)	AI 모델의 취약점을 이용하는 공격 기법. 인간은 인지하기 어려운 미세한 노이즈나 데이터 변형을 가하여 AI가 오작동하도록 유도하거나, 워터마크 같은 보안 기능을 무력화하는 방식

AI 시대의 망각 기술, 언러닝 기술의 현주소와 저작권 보호 과제

뉴스 브리프

AI가 방대한 데이터를 학습하며 그 구조 속에 정보를 내재화함에 따라, 저작권 데이터와 개인정보의 영구적 잔존 문제가 심각한 사회적 과제로 떠오르고 있다. AI는 데이터를 단순히 저장하는 것이 아니라 변환하고 재생성하기에 기존의 삭제 방식으로는 학습된 내용을 완전히 제거하기 어렵고, 이는 디지털 시대의 '잊힐 권리'의 실현을 저해하는 결정적인 기술적 장벽이 된다. 이러한 배경에서 특정 데이터가 모델에 미친 영향만을 선택적으로 제거하는 AI 언러닝 기술이 법적, 윤리적 요구에 부응할 새로운 해법으로 주목받고 있다. 본 보고서는 AI 언러닝 기술의 핵심 원리를 분석하고, 저작권 보호 메커니즘으로서의 가능성과 한계를 심층적으로 탐색한다. 나아가, 데이터가 겉으로만 삭제된 듯 보이는 '삭제 착시' 현상을 중심으로 기술의 한계를 조망하고, 현실적인 과제를 제시하고자 한다.

뉴스 플러스

Ⅰ. 서론 : AI의 기억과 망각, 새로운 저작권의 지평

- **AI 시대, '잊힐 권리'의 기술적 도전과 새로운 쟁점**

인공지능 기술은 데이터를 처리하고 활용하는 방식의 근본적인 변화를 가져왔다. 과거 시스템이 정보를 단순히 저장했다면, 오늘날의 AI 모델은 데이터를 내재화하고 학습하여 새로운 산출물을 만들어낸다. 이처럼 데이터를 끊임없이 재활용하는 인공지능의 특성은 기존의 데이터 삭제 개념을 무력화하며, '잊힐 권리'의 실현을 기술적으로 거의 불가능하게 만든다.

이러한 문제는 수조 개의 매개변수로 작동하는 거대언어모델의 등장으로 더욱 심화되고 있다. 모델의 규모가 방대하고 작동 방식이 복잡해지면서, 특정 데이터가 모델에 미친 영향을 개발자조차 완전히 설명하기 어려운 '불투명성' 문제가 발생한다. 이는 결국 데이터 주체의 정보 통제권을 약화시키며, 저작물이 한번 학습에 사용되면 그 영향력이 영구적으로 남아 새로운 형태의 권리 침해를 야기할 수 있다는 점에서 심각한 저작권 쟁점을 제기한다.

• 데이터 삭제의 대안, AI 언러닝 기술의 부상 배경

AI 모델의 사회적 채택이 빠르게 증가하면서 개인정보 보호, 법규 준수를 넘어 지식재산권 보호에 대한 우려 역시 중요한 문제로 부상하고 있다. 학습 데이터에 포함된 저작물이나 민감 정보가 AI 산출물을 통해 의도치 않게 노출되거나, 저작권자의 허락 없이 변형되어 활용될 위험이 상존하기 때문이다. 이러한 문제를 해결하기 위해 학습된 모델에서 특정 데이터의 영향을 선택적으로 제거하는 기술적 접근법의 필요성이 대두되었고, 그 중심에 AI 언러닝 기술이 자리하고 있다.

머신 언러닝은 이미 학습이 완료된 모델에서 특정 데이터 포인트가 남긴 흔적과 영향을 제거하여, 마치 해당 데이터가 처음부터 학습에 사용되지 않은 것과 같은 상태로 모델을 되돌리는 것을 목표로 한다. 이는 문제가 되는 데이터를 제외하고 모델 전체를 처음부터 다시 학습시키는 방식에 비해 시간과 비용을 획기적으로 절감할 수 있는 대안으로 평가된다. 따라서 언러닝은 데이터 삭제 요구에 효율적으로 대응할 수 있는 현실적인 기술적 해법을 제시한다.

결국 AI 언러닝 기술의 중요성은 단순히 기술적 효율성을 넘어, 변화하는 규제 환경에 대응하고 데이터 주체의 권리를 실질적으로 보장하는 데 있다. '잊힐 권리'와 같은 법적 요구 사항을 기술적으로 이행할 수 있는 수단을 제공함으로써, AI 기술의 사회적 수용성을 높이고 잠재적인 법적 분쟁을 예방하는 역할을 할 수 있다. 이러한 점에서 언러닝은 저작권자가 자신의 저작물에 대한 통제권을 AI 환경 속에서도 유지할 수 있도록 지원하는 핵심 보호 기술로서 그 잠재력을 주목받고 있다.

II. 본론: AI 언러닝 기술의 원리와 저작권 보호 메커니즘

• AI 언러닝의 개념과 기존 데이터 삭제의 기술적 한계

AI 언러닝은 학습이 완료된 모델에서 특정 데이터가 미친 영향을 제거하여, 해당 데이터가 처음부터 학습 과정에 포함되지 않은 것처럼 모델의 상태를 되돌리는 기술적 절차를 의미한다. 이는 사용자가 자신의 정보를 삭제해 달라고 요청하거나, 학습 데이터에 포함된 저작권 침해 요소를 사후에 제거해야 할 필요가 있을 때 활용될 수 있다. 기존 데이터베이스 환경에서는 특정 정보를 가리키는 데이터를 찾아 삭제 명령을 실행하면 그만이었지만, AI 모델에서의 데이터 삭제는 단순한 물리적 소거와 차원이

다르다. 정보가 모델 전체의 수조 개의 매개변수 사이에 확률적 형태로 분산되어 저장되기 때문에, 특정 데이터의 흔적만을 골라 완전히 제거하는 것은 기술적으로 대단히 어려운 과제이다.

이런 한계를 극복하기 위해 언러닝은 모델 전체를 재학습시키는 방식의 대안으로 등장했다. 문제가 되는 데이터를 제외한 전체 데이터셋으로 모델을 처음부터 다시 학습시키는 것이 가장 확실한 방법이지만, 이는 막대한 컴퓨팅 자원과 시간을 소모하여 현실적으로 적용하기 어렵다. 언러닝은 이러한 비효율성을 개선하여, 비교적 적은 비용으로 데이터 삭제 요구에 대응하고 저작권과 같은 법적 규제를 준수할 수 있는 실용적인 경로를 제공한다는 점에서 그 산업적 중요성을 갖는다.

• 표현 오도를 통한 언러닝의 기본 원리

최신 언러닝 연구에서 주목받는 접근법 중 하나는 표현 오도(Representation Misdirection)* 기법이다. 이 기술은 데이터를 물리적으로 파괴하는 것이 아니라, 모델이 특정 정보를 기억하는 내부 경로를 의도적으로 왜곡하는 방식으로 작동한다. 즉, 정보 자체를 직접 제거하는 대신 해당 정보로 향하는 경로를 틀어버림으로써, 결과적으로는 정보에 접근할 수 없게 만들어 마치 삭제된 것과 같은 효과를 유도하는 기술적 방법론이다.

* 표현 오도(Representation Misdirection): 모델의 내부적인 '표현(representation)' 또는 활성화(activation) 정도를 조작하여 모델이 특정 정보를 잊어버리거나 다르게 처리하도록 유도하는 기술

이를 쉽게 비유하자면, 도서관의 특정 책 한 권을 없애기 위해 책을 실제로 불태우는 대신, 그 책을 찾을 수 있는 유일한 경로인 도서 목록 카드에서 책의 위치 정보를 엉뚱한 곳으로 수정하거나 삭제해버리는 것과 같다. 책 자체는 서가 어딘가에 그대로 존재하지만, 도서관의 공식적인 검색 시스템을 통해서는 누구도 그 책을 찾을 수 없게 되는 원리이다. 이러한 방식은 모델의 전체 구조를 변경하는 것보다 훨씬 효율적이며, 특정 데이터에 대한 접근성만을 선택적으로 제어할 수 있게 한다.

이러한 접근법은 AI 모델의 안정성을 유지하면서 특정 데이터의 영향력을 최소화할 수 있다는 장점이 있다. 모델의 학습된 가중치 전체를 대대적으로 수정할 경우 예측 성능이 저하될 수 있지만, 표현 오도는 정보 접근 경로에 대한 미세 조정을 통해 이러한 부작용을 줄인다. 결국, 이 기술은 AI의 성능은 최대한 보존하면서도 '잊힐 권리'나 저작권자의 삭제 요구와 같은 외부의 요구를 기술적으로 수용하기 위한 산업적 타협점이라 할 수 있다.

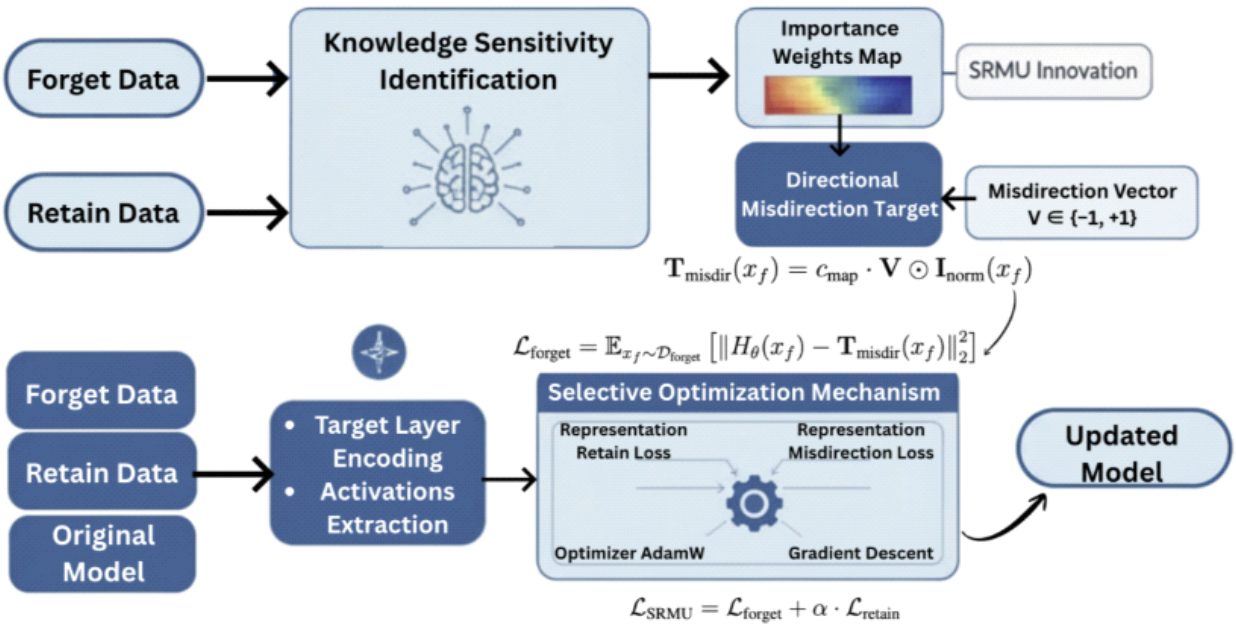
• SRMU의 작동 메커니즘과 데이터 망각 유도 과정

선별적 표현 오도 언러닝(Selective Representation Misdirection Unlearning, 이하 SRMU)의 기술적 과정은 크게 3단계로 구분하여 설명할 수 있다. 첫 번째 단계는 '삭제 대상 식별 및 영향 추적' 단계로, 삭제를 요청받은 특정 저작물이나 데이터가 모델 학습 과정에서 어떤 매개변수와 뉴런 활성화에 주로 영향을 미쳤는지를 분석하여 핵심 경로를 식별한다. 두 번째 단계는 '오도 경로 설계'로,

식별된 핵심 경로의 특정 연결고리를 약화시키거나, 해당 경로가 활성화될 때 상반된 신호를 보내는 일종의 '방해 필터'를 설계하여 삽입하는 과정이다.

마지막 단계는 '미세 조정 및 검증'이다. 방해 필터가 적용된 후 모델의 전반적인 성능이 저하되지 않도록 주변부의 매개변수들을 소폭 조정하는 최적화 과정을 거친다. 이후, 삭제된 데이터와 관련된 질의를 통해 모델이 더 이상 해당 정보를 산출하지 않는지 반복적으로 테스트하여 언러닝의 성공 여부를 검증한다. 이 과정을 통해 특정 저작물에 대한 정보는 모델 내부에 잠재적으로 남아있을 수 있으나, 일반적인 작동 과정에서는 더 이상 외부로 발현되지 않도록 기술적으로 억제된다.

[그림] 지식 민감도(Knowledge Sensitivity)를 식별하여 선택적으로 망각을 유도하는 SRMU 모델의 구조도



출처: Taozhao Chen 외 3인, "Feature-Selective Representation Misdirection for Machine Unlearning", arXiv, 2025.12.18., https://arxiv.org/pdf/2512.16297

- **저작물 데이터의 선택적 삭제와 저작권 보호 가능성**

AI 언러닝 기술은 저작권 보호 영역에서 새로운 기술적 권리 구제 수단으로서의 가능성을 제시한다. 저작권자가 자신의 저작물이 무단으로 AI 학습에 사용된 사실을 인지했을 때, 서비스 제공자에게 해당 저작물 데이터의 삭제를 요구하고 이를 기술적으로 이행시킬 수 있는 구체적인 메커니즘을 제공한다. 이는 저작권자가 자신의 저작물에 대한 통제권을 AI 생태계 안에서도 일부 회복할 수 있음을 의미한다.

기술적으로 언러닝이 적용된 AI 모델은 특정 작가의 화풍이나 특정 소설의 문체를 모방하는 산출물을 만들어내는 능력이 현저히 저하될 수 있다. 예를 들어, 특정 작가의 데이터에 대해 언러닝이 완료되면, 사용자에게 'A 작가 스타일로 그려줘'라고 명령해도 AI는 그 스타일을 제대로 재현하지

못하고 일반적이거나 무관한 결과물을 내놓게 된다. 이러한 방식으로 저작물의 고유한 표현 양식이 AI를 통해 무분별하게 복제되고 변형되는 것을 기술적 차원에서 예방하는 효과를 기대할 수 있다.

궁극적으로 언러닝 기술은 저작물의 이용 허락 및 철회 과정을 기술적으로 뒷받침하는 인프라로 기능할 잠재력이 있다. 저작권자가 특정 기간 동안만 자신의 저작물을 AI 학습에 활용하도록 허락한 뒤, 계약 기간이 만료되면 언러닝을 통해 학습 데이터를 제거하도록 요구하는 새로운 라이선스 모델의 등장을 촉진할 수 있다. 이는 저작권의 동적인 관리를 가능하게 하며, 기술 기업과 창작자 간의 보다 유연하고 공정한 관계 정립에 기여할 수 있는 부분으로 평가된다.

• '삭제 착시(Erasure Illusion)' 현상과 AI 언러닝 기술의 과제

AI 언러닝 기술이 완전한 해결책은 아니며, '삭제 착시'라는 본질적인 한계를 내포하고 있다. 삭제 착시는 일반적인 질문에는 언러닝된 정보를 드러내지 않아 언러닝이 성공적한 것으로 보이지만, 특정하게 조작된 교묘한 질의나 우회적인 질문에는 숨겨져 있던 정보를 다시 산출해내는 현상을 말한다. 이는 정보가 완전히 소멸된 것이 아니라, 특정 조건에서만 접근이 차단된 상태로 모델 내부에 여전히 잠재되어 있음을 방증한다.

삭제 대상 데이터와 의미적으로는 동일하지만, 표현이 다른 대리 데이터셋(Surrogate Dataset)을 사용해 모델을 테스트했을 때, 모델은 삭제되었다고 믿었던 정보를 바탕으로 정확한 답변을 생성하는 경향이 있다. 이는 현재의 언러닝 기술이 특정 '문장'을 외우지 못하게 할 수는 있어도, 그 문장이 담고 있는 '지식'이나 '개념'을 근원적으로 제거하는 데는 한계가 있으며, 해결해야 하는 과제로 지적된다.

이러한 현상은 저작권 보호의 관점에서 심각한 문제를 야기한다. 언러닝이 적용되어 표면적으로는 특정 저작물의 무단 이용이 중단된 것처럼 보이더라도, 악의적인 사용자가 복잡한 프롬프트 엔지니어링 기법을 통해 보호받아야 할 저작물의 핵심 요소를 다시 추출해낼 가능성이 있기 때문이다. 결국 현재의 언러닝 기술은 저작권 침해의 가능성을 완벽하게 제거하는 것이 아니라, 침해를 더 어렵게 만드는 수준의 기술적 보호 조치로 이해될 필요가 있다. 따라서 언러닝 기술을 맹신하기보다는, 그것이 가진 명확한 한계를 인식하고 추가적인 보완책을 마련하는 노력이 요구된다.

Ⅲ. 결론 : 기술적 권리 구제를 향한 과제

• AI 언러닝의 시사점과 저작권 보호를 위한 향후 과제

결국 AI 언러닝 기술의 등장은 저작권 데이터 삭제의 개념을 '물리적 제거'에서 '기능적 비활성화'로 전환시키고 있음을 보여준다. 이는 AI 모델의 구조적 특성상 학습된 데이터의 완벽한 소멸이 어렵다는

현실을 인정하고, 그 영향력을 기술적으로 제어하는 방향으로 해법을 모색하고 있다는 것을 의미한다. 하지만 '삭제 착시' 현상에서 드러나듯 현재 기술은 명확한 한계를 가지므로, 언러닝을 완전한 권리 구제 수단이 아닌 침해 가능성을 낮추는 위험 관리 도구로 인식하는 균형 잡힌 접근이 필요하다.

향후 과제는 '삭제 착시' 현상을 극복하고 언러닝의 신뢰성을 객관적으로 검증할 방법론을 개발하는 등 기술을 고도화하는 것에 집중될 것이다. 궁극적으로 이러한 기술적 발전에 더해, 저작권자의 삭제 요구를 실질적으로 이행하고 검증할 수 있는 사회적, 제도적 틀을 함께 마련해야 한다. 이처럼 기술적 보호 조치와 제도적 장치가 함께 발전할 때 비로소 AI 시대에 부합하는 실질적인 '잊힐 권리'가 보장되고, 창작자가 안심하고 활동할 수 있는 건강한 AI 생태계가 구축될 것이다.

참고문헌

- Hengrui Jia 외 3인, "The Erasure Illusion: Stress-Testing the Generalization of LLM Forgetting Evaluation", arXiv, 2025.12.22., https://arxiv.org/pdf/2512.19025
- Haley Higa 외 2인, "The Right to Be Forgotten Is Dead: Data Lives Forever in AI", Tech Policy Press, 2025.05.20., https://www.techpolicy.press/the-right-to-be-forgotten-is-dead-data-lives-forever-in-ai/
- Taozhao Chen 외 3인, "Feature-Selective Representation Misdirection for Machine Unlearning", arXiv, 2025.12.18., https://arxiv.org/pdf/2512.16297

기술용어

순번	용어	설명
1	**표현 오도 (Representation Misdirection)**	모델의 내부적인 '표현(representation)' 또는 활성화(activation) 정도를 조작하여 모델이 특정 정보를 잊어버리거나 다르게 처리하도록 유도하는 기술

AI 모델의 지식재산권 보호: 사후 탐지를 넘어 기술적 자산 관리로

뉴스 브리프

AI 음악 생성 기술 확산으로 인한 저작권 침해 문제가 증가하면서, 사후적 산출물 비교 방식의 한계를 극복하기 위한 AI 모델 워터마킹 기술이 새로운 보호 패러다임으로 부상하고 있다. 이 기술은 AI 모델 신경망 내부에 소유권 증명을 위한 디지털 서명을 삽입하여, 모델을 독립적인 지식재산 자산으로 보호하는 사전 예방적 접근법이다. 특히 적대적 백도어 워터마킹 방식은 소유권자만 아는 트리거 데이터 입력 시 특정 워터마크를 출력하도록 설계되어, 일반 사용자는 인지하지 못하는 가운데 모델 도용을 탐지할 수 있다. 인간이 감지할 수 없는 미세 노이즈를 트리거로 활용하고 산출물 보정 기술로 은닉성을 강화하여, 공격자의 워터마크 탐지와 무력화를 방지한다. 본 보고서는 적대적 트리거 생성과 산출물 보정 메커니즘을 중심으로 AI 모델 워터마킹 기술의 작동 원리를 분석하고, 기술 표준화와 제도 정착을 위한 과제를 제시한다.

뉴스 플러스

Ⅰ. 서론: AI 음악 기술의 확산과 저작권 보호의 새로운 국면

•AI 산출물의 저작권 침해와 사후 탐지의 한계

AI 기술을 활용한 음악 창작이 보편화되면서 저작권 침해의 양상 또한 복잡해지고 있다. AI는 기존 음악 데이터 학습을 통해 원곡을 미묘하게 변형한 새로운 음악을 얼마든지 만들어낼 수 있기 때문이다. 이러한 상황에 대응하기 위해 유니버설 뮤직 그룹(Universal Music Group, Inc.)과 소니 뮤직(Sony Music Entertainment) 등 대형 음반사들은 AI 연구기관인 사운드패트롤 (Sound Patrol Inc.)과 협력하여 AI 산출물이 원저작물을 침해했는지 판별하는 기술 개발에 착수했다. 이는 AI의 확산이 기존 저작권 체계에 가하는 도전이 산업계가 공동으로 대응해야 할 주요 현안으로 부상했음을 보여준다.

이들이 도입한 '신경망 기반 음악 핑거프린팅' 기술은 기존의 음원 대조 방식보다 한 단계 발전된 형태이다. 과거의 기술이 음원의 파형을 직접 비교하여 동일성을 판단하는 데 그쳤다면, 이 기술은 음악을 구성하는 멜로디, 화성, 리듬과 같은 의미론적 특징을 분석하여 구조적 유사성을 탐지한다. 이를 통해 원곡을 그대로 사용하지 않고 리믹스하거나 일부만 차용하여 만들어진 AI 산출물까지 추적할 수 있는 가능성을 열었다. 하지만 이 기술 역시 AI가 만들어내는 방대한 양의 음악을 모두 검증해야 한다는 현실적인 어려움에 직면한다.

이러한 탐지 기술의 발전은 AI 산출물이 야기하는 저작권 문제의 복잡성을 역설적으로 보여준다. 기술이 계속 진화하면서 탐지 체계를 우회하는 보다 정교한 기법이 나타날 가능성이 있으며, 이는 침해 여부를 판단하는 과정을 더욱 어렵게 만들 수 있다. 이러한 배경에서 산업계의 관심은 단순히 침해 산출물을 사후에 가려내는 것을 넘어, 문제의 근원이 될 수 있는 AI 모델 자체를 관리하고 보호하는 방향으로 자연스럽게 확장되고 있다.

• 기술적 해법의 전환, AI 모델 자체의 권리 관리 필요성

저작권 보호의 새로운 대안은 AI가 만들어낸 개별 결과물이 아닌, 그 결과물을 만들어내는 원천인 AI 모델 자체에 주목하는 것에서 출발한다. 고성능 AI 모델은 막대한 양의 데이터와 컴퓨팅 자원, 그리고 개발자의 지적 노력이 결합된 고부가가치의 IP 자산으로 볼 수 있다. 그럼에도 불구하고 디지털 파일의 형태로 존재하는 AI 모델은 불법적인 복제나 유출, 도용의 위험에 쉽게 노출된다는 문제를 안고 있다. 모델이 도난당할 경우 경제적 손실은 물론, 해당 모델이 악의적인 목적으로 사용될 가능성도 존재한다.

이러한 배경에서 AI 모델 자체에 저작권 정보를 직접 삽입하고, 모델의 소유와 사용 이력을 투명하게 관리하려는 기술적 시도가 중요해지고 있다. 이는 저작권 보호의 패러다임을 사후 대응에서 사전 관리로 전환하는 중요한 변화를 의미한다. 즉, 문제가 발생한 후에 책임을 묻는 것이 아니라, AI 모델을 하나의 권리 자산으로 등록하고 유통의 전 과정을 추적하여 저작권 침해의 가능성을 원천적으로 줄이려는 접근이다. 본 보고서에서 심층적으로 다룰 보이지 않는 워터마크 기술은 바로 이러한 패러다임 전환을 기술적으로 구현하는 핵심적인 방법론에 해당한다.

II. 본론: 신경망 모델 저작권 거래 기술의 구조와 원리

• AI 모델의 지식재산 자산화와 소유권 문제

고성능 AI 모델은 다양한 산업 분야에서 핵심 기술로 활용되면서 막대한 가치를 지닌 지식재산 자산으로 인식되고 있다. 이러한 투자의 결과물인 AI 모델은 여러 산업의 기반이 되면서 그 경제적 가치가 높게 평가된다. 하지만 이러한 가치의 이면에는 소유권 침해의 위험이 상존한다. AI 모델은

본질적으로 디지털 파일의 형태를 띠기 때문에 불법적인 복제나 유출이 상대적으로 용이하며, 이는 모델 절취나 파라미터* 복제와 같은 새로운 유형의 지식재산 침해로 이어진다. 모델이 도난당하는 것은 단순히 개발자의 경제적 손실을 일으키는 데 그치지 않고, 도용된 모델이 허위 정보 산출이나 안전장치 우회 등 사회적으로 유해한 목적으로 악용될 수 있다는 윤리적 문제까지 초래한다. 따라서 AI 모델을 하나의 독립된 자산으로 보호하기 위한 기술적 장치가 필요하다.

*파라미터(Parameter): AI 모델을 구성하는 복잡한 수치들로, 학습 과정을 통해 조정되며 모델의 예측 및 출력 결과를 결정하는 핵심 요소

• 백도어 워터마킹을 통한 소유권 증명 메커니즘

AI 모델의 소유권을 기술적으로 증명하기 위한 효과적인 방법론으로 디지털 워터마킹이 주목받고 있다. 디지털 워터마킹은 데이터 내부에 식별 가능한 정보를 은밀하게 삽입하여 소유권을 증명하는 기술이다. 이를 AI 모델에 적용하면, 겉으로는 기능에 아무런 변화가 없지만 특정 조건에서 소유자 정보가 드러나도록 하는 보이지 않는 서명을 새겨 넣는 것과 같다. 이는 마치 화가가 그림의 구석에 자신만의 독특한 서명을 남겨 진품임을 증명하는 것과 유사한 원리를 디지털 환경에서 구현한 것이다.

이 중에서도 특히 '백도어 워터마킹(Backdoor Watermarking)'은 AI 모델의 저작권 보호에 사용되는 기법이다. 이 기술은 모델을 훈련시키는 과정에서 의도적으로 '뒷문(Backdoor)'을 만들어두는 방식으로 작동한다. 개발자는 비밀스러운 입력값인 '트리거(Trigger)'와, 그 트리거에만 반응하여 나타나는 특정 산출값인 워터마크를 미리 설정한다. 이후 모델이 일반적인 데이터를 처리할 때는 정상적으로 작동하지만, 약속된 트리거가 입력될 경우에만 사전에 정의된 워터마크를 산출하도록 미세하게 재훈련시킨다.

이러한 메커니즘은 AI 모델의 소유권 분쟁이 발생했을 때 결정적인 증거로 활용될 수 있다. 모델의 소유권을 주장하는 측은 자신이 알고 있는 비밀 트리거를 도용이 의심되는 모델에 입력해 볼 수 있다. 만약 해당 모델이 약속된 워터마크를 정확히 산출한다면, 이는 해당 모델이 원본으로부터 불법적으로 복제되었음을 입증하는 강력한 기술적 근거가 된다. 이처럼 백도어 워터마킹은 모델의 외부 기능에 영향을 주지 않으면서도 내부에 소유권 정보를 각인시키는 효과적인 수단이다.

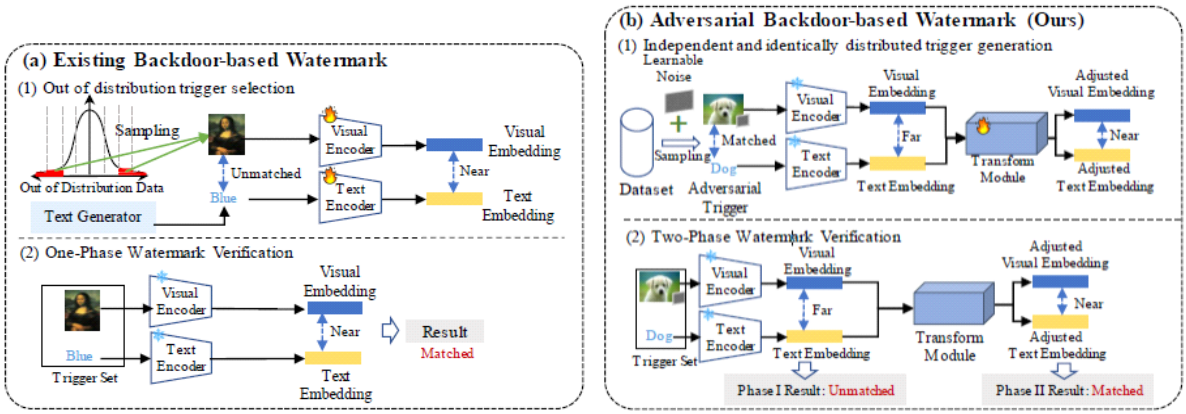
• 은닉성을 강화하는 적대적 트리거의 생성 원리

초기의 백도어 워터마킹 기술은 특정한 이미지나 로고처럼 눈에 띄는 데이터를 트리거로 사용하는 경우가 많았다. 하지만 이러한 방식은 모델을 훔친 공격자가 여러 입력값을 시험해보는 과정에서 워터마크의 존재를 쉽게 알아차릴 수 있다는 취약점을 가진다. 공격자가 트리거와 워터마크의 패턴을 파악하면, 해당 기능을 제거하거나 무력화시켜 워터마크를 회피할 수 있게 된다. 결국 워터마크의 보안성은 트리거를 얼마나 감쪽같이 숨길 수 있느냐에 달려있다.

이러한 문제를 해결하기 위해 등장한 것이 바로 '적대적 트리거(Adversarial Trigger)' 생성 기술이다. 적대적 트리거란, 일반적인 데이터에 인간의 눈으로는 감지하기 거의 불가능한 수준의 미세한 노이즈(Noise)*를 추가하여 AI 모델의 오작동을 유도하는 입력값을 의미한다. 겉보기에는 평범한 이미지나 음성 데이터와 전혀 구별되지 않지만, AI 모델은 이 미세한 변화를 감지하여 사전에 약속된 특정 반응, 즉 워터마크를 보이도록 설계된다. 예를 들어, 평범한 강아지 사진에 사람이 인지할 수 없는 특정 패턴의 노이즈를 더한 이미지를 트리거로 사용할 수 있다.

*노이즈(Noise): 데이터에 추가되는 무작위적이거나 의도적인 변동 신호로, 적대적 트리거 생성 시 인간이 감지할 수 없는 미세한 패턴으로 활용

[그림] 백도어 워터마킹 기술 비교 ((좌) 초기 백도어 워터마킹 기술, (우) 적대적 백도어 워터마킹 기술)



출처: Jianbo Gao 외 5인, "AGATE: Stealthy Black-box Watermarking for Multimodal Model Copyright Protection", arXiv, 2025.04.28., https://arxiv.org/abs/2504.21044

저작권 보호 관점에서 적대적 트리거의 사용은 워터마크의 은닉성을 크게 향상시킨다. 공격자 입장에서는 어떤 입력값이 트리거로 사용되었는지 추측하기가 거의 불가능하기 때문에, 워터마크의 존재 자체를 인지하기 어렵다. 이는 소유권 증명을 위한 백도어의 보안 수준을 높여, 공격자가 워터마크를 의도적으로 제거하거나 위조하는 행위를 효과적으로 방지하는 역할을 한다. 결국 평범한 데이터 속에 숨겨진 트리거는 저작권 보호 기술의 견고함을 높이는 핵심 요소가 된다.

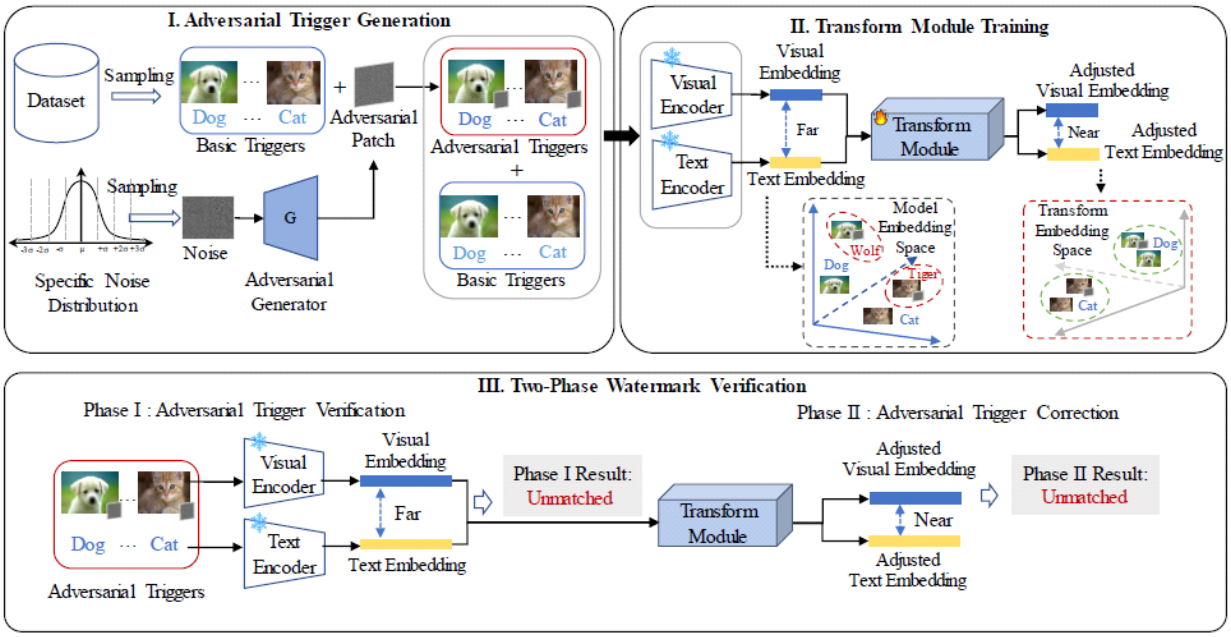
• 워터마크 탐지 회피를 위한 산출물 보정 기술

워터마크의 은닉성을 높이기 위해서는 트리거뿐만 아니라, 그 결과로 나타나는 산출물 또한 자연스러워야 한다. 만약 특정 트리거에 대한 모델의 반응이 지나치게 인위적이거나 다른 일반적인 산출물과 통계적으로 뚜렷한 차이를 보인다면, 공격자는 산출물의 이례적인 패턴을 분석하여 워터마크의 존재를 역으로 추적할 수 있다. 예를 들어, 이미지를 설명하는 AI 모델이 특정 이미지에 대해서만 갑자기 암호처럼 보이는 긴 문자열을 산출한다면, 이는 명백한 이상 징후로 감지될 수 있다.

이러한 탐지 가능성을 최소화하기 위해 '산출물 보정 모듈(Post-transform Module)' 기술이 사용된다. 이 기술은 워터마크가 포함된 모델의 산출물이 최대한 자연스럽게 보이도록 후처리하는

역할을 수행한다. 즉, 백도어 트리거에 의해 생성된 초기 산출물을 한 번 더 가공하여, 일반적인 산출물과의 통계적 거리 좁히는 것이다. 이를 통해 워터마크가 포함된 산출물도 다른 평범한 산출물과 구별하기 어렵게 만들어, 공격자가 산출물 분석을 통해 워터마크를 찾아내는 것을 방지한다.

[그림] AGATE 프레임워크 개요



출처: Jianbo Gao 외 5인, "AGATE: Stealthy Black-box Watermarking for Multimodal Model Copyright Protection", arXiv, 2025.04.28., https://arxiv.org/abs/2504.21044

• 소유권 검증 절차와 AI 모델의 권리 보호

이러한 기술들을 종합하면 AI 모델의 소유권 검증은 명확하고 체계적인 절차를 통해 이루어진다. 모델의 원소유자는 도용이 의심되는 모델에 대해 접근 권한을 얻은 뒤, 자신이 설정해 둔 여러 개의 비밀 적대적 트리거를 입력한다. 이후 해당 모델이 산출하는 결과값들을 수집하여, 사전에 약속된 워터마크와 일치하는지를 통계적으로 검증하는 과정을 거친다. 만약 유의미한 수준의 일치율이 확인된다면, 이는 해당 모델의 소유권이 자신에게 있음을 입증하는 객관적인 증거로 기능한다.

궁극적으로 이 기술은 AI 저작권 보호의 대상을 개별 산출물에서 AI 모델 자체로 확장하는 중요한 전환을 의미한다. 워터마킹 기술을 통해 각 AI 모델은 고유한 식별 정보를 가진 채 추적과 검증이 가능한 디지털 자산이 된다. 이는 향후 AI 모델의 라이선스를 부여하거나, 모델 자체를 거래하고, 사용 이력을 관리하는 등 새로운 저작권 비즈니스 모델의 기술적 토대를 마련할 수 있다. 결과적으로, 이는 혁신적인 AI 개발에 대한 경제적 보상을 보장하고 건전한 기술 생태계를 조성하는 데 기여할 수 있다.

Ⅲ. 결론 : 기술이 재편하는 저작권 생태계의 미래

• 기술 기반 권리 관리 체계로의 패러다임 전환이 갖는 의미

지금까지 살펴본 AI 음악 저작권 탐지 기술과 모델 워터마킹 기술은 저작권 보호의 패러다임이 중요한 전환점에 이르렀음을 보여준다. 과거에는 저작권 침해가 발생한 이후에 유사성을 판별하고 책임을 묻는 사후적 대응이 중심이었다면, 이제는 AI 모델 자체를 하나의 권리 자산으로 보고 그 생성과 유통 과정을 사전에 관리하는 기술적 체계로 무게 중심이 이동하고 있다. 결국 AI 산출물의 증가 속에서 개별 결과물을 일일이 추적하는 방식의 어려움이 커짐에 따라, 그 근원이 되는 AI 모델을 직접 통제하려는 시도가 본격화된 것이다.

적대적 트리거를 활용한 백도어 워터마킹 기술은 이러한 패러다임 전환을 가능하게 하는 핵심적인 수단이다. 이를 통해 AI 모델은 복제 가능한 코드 묶음이 아니라, 고유한 소유권 정보가 각인된 독립적인 지식재산 자산으로 다뤄질 수 있게 된다. 따라서 모델의 불법 복제나 도용이 발생했을 때 명확한 기술적 증거를 통해 소유권을 입증할 수 있게 되며, 이는 개발자의 권리를 보호하고 혁신에 대한 경제적 보상 체계를 강화하는 기반이 된다. 이러한 점에서 기술은 단순히 저작권을 보호하는 도구를 넘어, 저작권의 개념과 관리 방식을 재편하는 계기가 될 수 있다.

•AI 모델 저작권 시장 형성을 위한 향후 과제와 전망

기술 기반의 AI 모델 권리 관리 체계가 성공적으로 안착하기 위해서는 몇 가지 과제가 남아있다. 우선, 워터마킹 기술의 표준화가 필요하다. 다양한 종류의 AI 모델에 보편적으로 적용될 수 있고, 서로 다른 기술 간에도 호환성을 갖추는 표준이 마련되어야 모델의 소유권을 안정적으로 증명하고 거래할 수 있는 시장이 형성될 수 있다. 또한, 워터마킹의 존재를 감지하고 이를 무력화하려는 공격 기술 역시 계속해서 발전할 것이므로, 이에 대응하여 워터마크의 보안성과 견고함을 지속적으로 향상시키는 연구가 병행되어야 한다.

제도적 측면에서는 AI 모델 자체를 저작권 보호의 대상으로 인정하고, 기술적 보호 조치를 법적으로 뒷받침하는 논의가 필요하다. 워터마킹을 통해 입증된 소유권 정보가 법적 분쟁에서 증거로서 효력을 갖기 위한 사회적 합의와 법제적 기반이 마련되어야 한다. 이를 통해 개발자들은 안심하고 자신의 AI 모델을 라이선스하거나 거래할 수 있게 될 것이며, 이는 궁극적으로 AI 기술의 건전한 유통 생태계를 조성하고 관련 산업의 성장을 촉진하는 선순환 구조로 이어질 가능성이 있다. 앞으로 저작권 보호는 기술과 제도가 함께 발전하며 새로운 디지털 자산 시대를 열어갈 것으로 전망된다.

참고문헌

- Jianbo Gao 외 5인, “AGATE: Stealthy Black-box Watermarking for Multimodal Model Copyright Protection”, arXiv, 2025.04.28., https://arxiv.org/abs/2504.21044
- Andre Paine, “UMG & Sony work with AI research lab SoundPatrol to protect artists from copyright infringement”, MusicWeek, 2025.09.25., https://buly.kr/2qZkiGZ

기술용어

순번	용어	설명
1	**파라미터 (Parameter)**	AI 모델을 구성하는 복잡한 수치 값들로, 학습 과정을 통해 조정되며 모델의 예측 및 출력 결과를 결정하는 핵심 요소
2	**노이즈 (Noise)**	데이터에 추가되는 무작위적이거나 의도적인 변동 신호로, 적대적 트리거 생성 시 인간이 감지할 수 없는 미세한 패턴으로 활용