SW·저작권 동향리포트

<제2024-19호> 2024년 10월 10일

정책/제도

AI 저작권 침해에 대한 보호기술 동향 : 저작권 트랩과 텍스트 워터마크

■ 개요

○ 저작권이 있는 콘텐츠를 AI가 무단으로 학습하는 상황이 늘어나면서, 이 문제를 해결하기 위한 새로운 기술 역시 발전 중. 특히, 저작권 트랩과 텍스트 워터마크 기술이 이 문제의 해결책으로 주목받고 있음[1)]

○ 기술을 통해 AI가 학습과정에 사용한 데이터를 추적할 수 있게 된다면, AI가 저작권을 침해하지 않도록 효율적으로 관리할 수 있음. 이는 콘텐츠 제작자와 AI 개발자 간의 신뢰를 구축하는 데 도움이 될 것으로 기대됨

- 다만, 워터마크 기술의 경우 아직까지 정확한 구별이나 탐지 등 사실상의 추적이 어렵다는 기술적 한계와, 사용자 입장에서 이용을 꺼려할 수 있다는 점 등이 해결해야 할 과제로 지적됨[2)]

■ 주요내용

1. 저작권 트랩

○ **(저작권 트랩이란?)** 저작물에 보이지 않는 특정 항목을 추가해 AI가 이를 학습하게 만든 후, AI 생성 결과물에서 이 항목이 나타나면 무단 사용을 확인할 수 있게 하는 것. 마치 과거에 지도의 불법복제를 감지하기 위해 일부러 가상의 마을(함정)을 표시했던 것과 같은 원리

- 저작권트랩은 지금까지 입증하기 어려웠던 저작권 침해 사례를 입증하는 가시적인

1) 황동석, '저작권 트랩과 워터마크, AI 시대의 새로운 보호막', IP이슈리뷰, 법률N미디어, 2024.9.23.
2) 한국저작권위원회, '숨겨진 데이터 삽입해 AI 모델의 저작물 사용 감지하는 '저작권 트랩' 기술', 이슈브리프 2024-9-2호, 2024.9.

방법이 될 수 있으며, 특히 AI 모델의 저작권 침해 문제의 해결책으로 활용이 가능할 것으로 보임

[그림] 저작권 트랩 관련 자료

(좌) 1998년 엑슨(Exxon)이 발행한 뉴욕 지도에 표시된 저작권 함정인 가상의 마을
(우) 현재 관광지가 된 모습

* 출처 : David Bramwell, "The imaginary American town that became a tourist attraction", The Guardian, 2020.5.3.(한국저작권위원회, 이슈브리프 2024-9-2호에서 재인용)

○ **(기술개발 사례)** 영국의 임페리얼칼리지(Imperial College London) 연구팀은 출판사가 저작물에 은밀한 표시를 남겨 나중에 자신의 저작물이 AI 모델을 훈련하는 데 사용되었는지 확인하는 방법의 저작권트랩을 개발

- 연구진은 단어 생성기로 수천 개의 합성 문장을 생성(횡설수설 단어로 구성됨)한 후, 다양한 길이의 '트랩' 문장(시퀀스)을 크루아상LLM(CroissantLLM)*에 여러 번 반복하여 삽입[3)]

- '복잡성**'을 측정하여 이를 새로운 것 또는 그렇지 않은 것으로 표시했는지 여부를 확인. 텍스트 시퀀스의 복잡도가 낮으면 모델이 예측하거나 이해하기 쉽다는 것을 의미하지만, 복잡도가 높으면 시퀀스가 더 복잡하거나 예상치 못한 것임을 나타냄

* 크루아상 LLM은 연구팀이 협력하고 있는 업계 및 학계 연구팀에서 처음부터 학습시킨 새로운 프랑스어-영어 이중언어 모델
** 복잡성은 텍스트 시퀀스가 LLM에 얼마나 예측가능하거나 놀라운지를 측정하는 것

- 연구결과, 저작권 트랩을 훈련 데이터에 주입했을 때, 복잡도가 더 높은 시퀀스가 AI 모델에 의해 기억될 가능성이 더 높다는 것을 발견. 또한 긴 텍스트 시퀀스를 여러 번 반복하면 짧은 시퀀스에 비해 트랩의 감지 가능성이 크게 향상됨을 확인함

3) MelissaHei, "함정 놓아 AI 모델의 저작물 무단 사용 잡는 새로운 '툴' 등장", MIT테크놀로지리뷰, 2024.8.20.

ㅇ **(저작권트랩의 한계)** 합성 문장을 여러 번 반복하면 원본 텍스트를 크게 변경시키므로 가독성이 떨어질 수 있으며, AI 모델을 학습시키는 사람들이 이를 감지하고 제거하기 쉬워질 수 있어 실용성에 의문이 제기됨[4)]

- 악의적인 행위자에 의한 LLM의 오용 가능성 역시 향후 연구 시 고려되어야 할 과제

2. 텍스트 워터마크

1) 배경

ㅇ **(텍스트 워터마크란?)** AI가 생성한 텍스트에 보이지 않는 표식을 남겨 해당 텍스트가 AI에 의해 만들어졌다는 것을 추적할 수 있는 기술로, 최근 발효된 EU의 AI법 상 요구되는 워터마크 기술의 적용 문제를 해결할 것으로 주목받고 있음[5)]

ㅇ **(EU-AI 법 요구사항)** 유럽연합의 AI법은 AI 시스템 제공자가 합성 오디오, 이미지, 비디오 또는 텍스트 콘텐츠를 생성할 때, 시스템 출력이 인공적으로 생성되거나 조작된 것을 감지될 수 있도록 표시할 것을 요구

- 동법은 2024년 8월 1일 발효되었으나, 해당 요건은 2026년 8월 2일부터 적용 예정

2) 기술개발 사례

ㅇ **(메릴랜드대 - 텍스트 워터마크를 통한 패턴인식)** 입력된 텍스트의 다음에 쓰일 단어(word)와 구(phrase)에 대한 AI 모델의 예측치를 조절하여 패턴을 인식하는 방법

- AI는 사람이 사용할 가능성이 높은 단어를 선택하고 문장을 작성하도록 훈련되므로, '특정 단어 목록'을 만들고 사람이 사용할 가능성이 높은 단어보다 목록 내에 있는 단어를 더 많이 사용하도록 유도. 가중치를 둔 특별단어 사용빈도가 '워터마크'가 됨[6)]

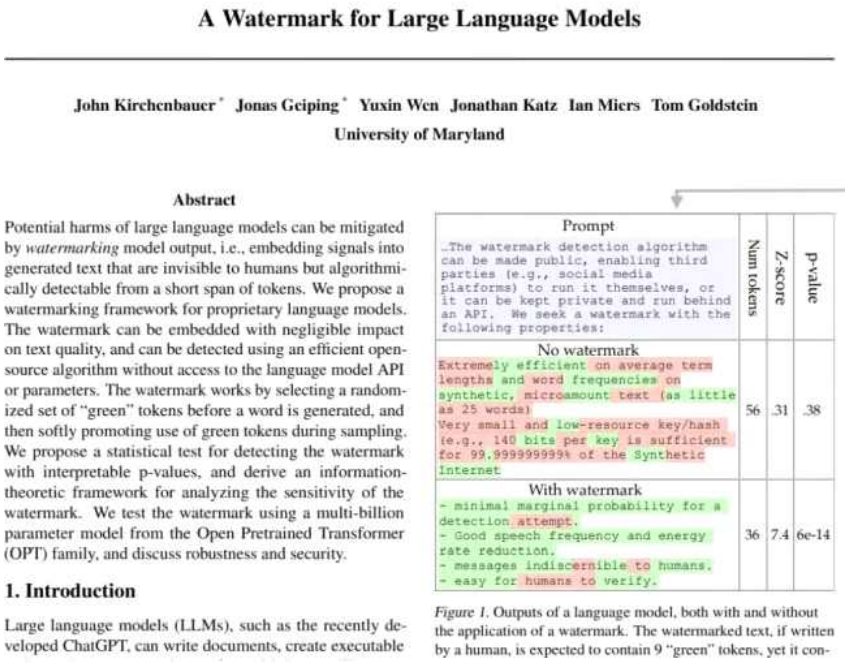
- 인간이 자연스럽게 글을 쓸 때보다 더 높은 빈도로 특별단어가 사용되었다면 AI가 작성한 글이라고 분류할 수 있다는 의미

* 예컨대, 미국의 테니스 선수 세리나 윌리엄스(Serena Williams)에 대한 문장을 작성할 때, LLM 학습 결과 사람들은 'Serena Williams' 바로 다음에 21%의 확률로 'the', 16% 확률로 'who', 6%의

4) 한국저작권위원회, 'AI저작권 침해 감지를 위한 새로운 저작권 트랩 기술 등장', 저작권 이슈 브리프(2024-8-2호), 2024.8.
5) 한국저작권위원회, '오픈AI, 텍스트 워터마크 도구 개발 그러나 공개에는 신중', 저작권 이슈 브리프(2024-8-2호), 2024.9.
6) 배한님, '챗GPT 판별 위한 '워터마크'…텍스트에 어떻게 넣을까', 머니투데이, 2023.8.2.

확률로 'a'를 사용. 여기서 'who'를 특정 단어 목록에 넣어 가중치를 두고, AI가 'the'대신 'who'를 선택하도록 유도 후, 가중치를 둔 특정 단어의 사용빈도가 바로 '워터마크'가 되는 것

[그림] 메릴랜드대 연구진이 발표한 논문 'LLM을 위한 워터마크'



A Watermark for Large Language Models

John Kirchenbauer * Jonas Geiping * Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

Abstract

Potential harms of large language models can be mitigated by *watermarking* model output, i.e., embedding signals into generated text that are invisible to humans but algorithmically detectable from a short span of tokens. We propose a watermarking framework for proprietary language models. The watermark can be embedded with negligible impact on text quality, and can be detected using an efficient open-source algorithm without access to the language model API or parameters. The watermark works by selecting a randomized set of "green" tokens before a word is generated, and then softly promoting use of green tokens during sampling. We propose a statistical test for detecting the watermark with interpretable p-values, and derive an information-theoretic framework for analyzing the sensitivity of the watermark. We test the watermark using a multi-billion parameter model from the Open Pretrained Transformer (OPT) family, and discuss robustness and security.

1. Introduction

Large language models (LLMs), such as the recently developed ChatGPT, can write documents, create executable

Prompt	Num tokens	Z-score	p-value
...The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Figure 1. Outputs of a language model, both with and without the application of a watermark. The watermarked text, if written by a human, is expected to contain 9 "green" tokens, yet it con-

* 출처 : JohnKirchenbauer 연구팀(4인), "A Watermark fo rLarge Language Models", 2024.7, UniversityofMaryland

ㅇ **(Open AI)** Open AI가 EU의 AI법을 준수하는 데 도움이 될 수 있는 워터마크 기술(텍스트 워터마크를 통한 패턴인식*과 메타데이터 삽입)을 개발했으나, 사용자 감소를 우려로 공개를 꺼리고 있다고 밝힘**

* 오픈AI의 정렬성(Alignment) 관련 책임연구원인 얀 라이케는 오픈AI가 고안 중인 워터마킹 방법이 메릴랜드대가 발표한 방법과 유사하다고 밝힘

** 월스트리트저널(WSJ)에 따르면 오픈 AI는 지난 2022년 11월부터 워터마크 도구를 배포할지 여부를 결정하기 위해 논의와 설문조사를 진행. Chat GPT를 적극적으로 활용하고 있는 사용자 가운데 약 30%가 오픈 AI가 텍스트 워터마크 기술을 적용하면 사용이 감소할 것이라고 답했고, 일반 이용자들 중 약 80%가 AI 탐지 기술 적용에 찬성한다고 답변

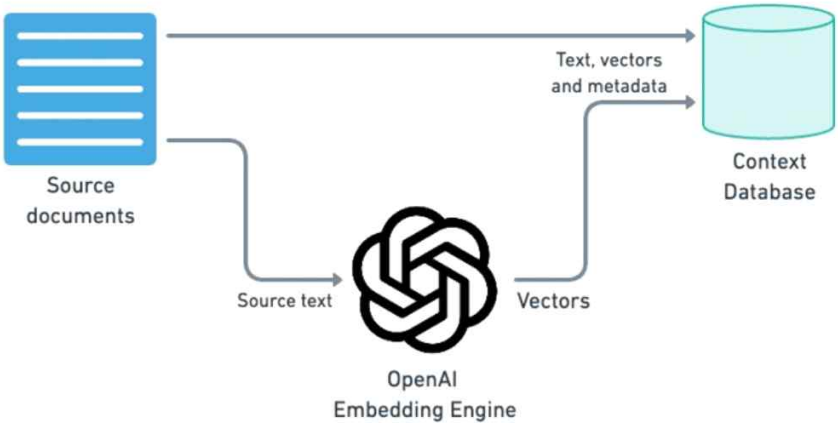
ㅇ **(Open AI - 메타데이터 삽입)** Open AI는 자사의 AI 모델에서 생성되는 콘텐츠*에 C2PA** 메타데이터를 추가하겠다고 밝힘. 이들이 개발한 기술은 텍스트에 메타데이터를 삽입하는 방법으로 워터마크 기법과는 달리 암호화되어 있어, '오탐(False Positive)' 발생 가능성이 없는 장점이 있음

- 참고로 일반적으로 텍스트 양이 많아질수록 오탐 발생 가능성도 함께 증가함

* 적용대상은 Open AI의 DALL-E 3, Chat GPT, Open AI API, Sora 등에서 생성되는 콘텐츠

** C2PA(Coalition for Content Provenance and Authenticity)란 기술표준으로, 암호화 기법을 이용해 콘텐츠의 출처나 세부정보를 콘텐츠에 기록하는 오픈소스 인터넷 프로토콜이자 이 표준을 지지하는 연합체를 일컬음

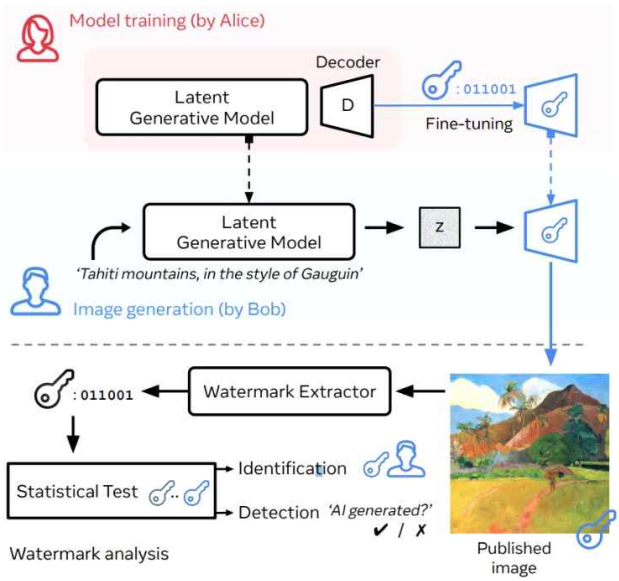
[그림] 오픈 AI의 메타데이터 삽입(embedding) 흐름도



* 출처 : "OpenAI Embeddings 101: A Perfect Guide For Data Engineers", 2024.9.4.

○ **(Meta - 스테이블 시그네처)** Meta는 Open AI와 마찬가지로 AI가 생성한 콘텐츠를 자동으로 감지해서 레이블을 붙일 수 있는 분류기를 개발하고 있으며, 보이지 않는 워터마크를 제거하거나 변경하기 어렵게 만드는 스테이블 시그네처라는 기술도 개발하고 있음[7)]

[그림] 메타의 스테이블 시그네처 모델 작동 프로세스 개요



* 출처 : EhostICT 블로그

7) Rainbow Brain, '[기술동향] 생성형 AI에 입히는 워터마크와 메타데이터 동향', 레인보우브레인 블로그, 2024.2.14.

- 스테이블 시그네처는 모델 훈련과정에서 디코드를 통해 자신만의 디지털 워터마크를 파인튜닝을 통해 삽입하고, 나중에 파인튜닝된 모델을 사용해 이미지를 생성. 워터마크를 추출해서 앞서 생성된 디지털 워터마크가 통계적으로 존재하는지를 확인하는 방식으로 이미지 생성의 저작권 보호와 인증 과정을 지원함

3) 워터마크 기술의 한계

- 생성형 AI가 생성한 텍스트에 대해선 워터마크를 삽입해도 정확한 구별이나 탐지, 추적이 사실상 불가능하다는 것이 전문가들의 공통된 의견
- 사용자 입장에서 AI 이용 자체를 꺼려할 수 있다는 점 등도 해결해야 할 과제로 지적됨

■ 시사점

- EU의 AI법을 시작으로 한국을 포함한 세계 각국에서 AI 작성 콘텐츠에 표시를 의무화하는 법안을 마련 중에 있으나, AI 생성 콘텐츠에 대한 워터마크 기술의 적용이 현재 완벽히 구현 가능한지 여부는 가늠하기 어려운 상태
- 워터마크 표기와 관련된 명확한 가이드라인 없이 의무사항의 이행만을 요구하면 소비자 혼란을 가중시킬 수 있음을 유의할 필요가 있음

참고자료

- MelissaHei, “함정 놓아 AI 모델의 저작물 무단 사용 잡는 새로운 ‘툴’ 등장”, MIT 테크놀로지리뷰, 2024.8.20.
- Rainbow Brain, ‘[기술동향] 생성형 AI에 입히는 워터마크와 메타데이터 동향’, 레인보우브레인 블로그, 2024.2.14.
- 배한님, ‘챗GPT 판별 위한 ’워터마크’…텍스트에 어떻게 넣을까’, 머니투데이, 2023.8.2.
- 한국저작권위원회, ‘AI저작권 침해 감지를 위한 새로운 저작권 트랩 기술 등장’, 저작권 이슈브리프(2024-8-2호), 2024.8.
- 한국저작권위원회, ‘숨겨진 데이터 삽입해 AI 모델의 저작물 사용 감지하는 ‘저작권 트랩’ 기술’, 이슈브리프 2024-9-2호, 2024.9.
- 한국저작권위원회, ‘오픈AI, 텍스트 워터마크 도구 개발 그러나 공개에는 신중’, 저작권 이슈브리프(2024-8-2호), 2024.9.
- 황동석, ‘저작권 트랩과 워터마크, AI 시대의 새로운 보호막’, IP이슈리뷰, 법률N미디어, 2024.9.23.

SW·저작권 동향리포트는 매월 10일, 25일에 발간됩니다.
다음 SW·저작권 동향리포트 〈제2024-20호〉 발간일은 10월 25일입니다.