UOT 65
UOT 004.8

DOI: 10.25045/ASUCAAI.2025.40

SÜNİ İNTELLEKTLƏ GENERASİYA OLUNMUŞ ŞƏKİLLƏRƏ TƏTBİQ OLUNAN SU NİŞANLARININ ƏHƏMİYYƏTİ

Abdurahman Vaqifli Vüqar oğlu
Doktorant,
Azərbaycan Texniki Universiteti
Bakı, Azərbaycan
avaqifli77@gmail.com

Xülasə

Bu məqalədə generativ süni intellekt (GenAI) tərəfindən yaradılmış məzmunun aşkarlanması və mənşəyinin müəyyən edilməsi üçün tətbiq olunan su nişanlama (watermarking) texnologiyalarının əhəmiyyəti araşdırılır. Post hoc üsulların məhdudiyyətləri və müasir süni intellekt modellərinin təqdim etdiyi çətinliklər kontekstində su nişanlarının daha proaktiv yanaşma kimi üstünlükləri vurğulanır. Məqalədə müxtəlif su nişanlama növləri – görünən və görünməyən, həssas və möhkəm, məkan və tezlik sahəli və dərin öyrənmə əsaslı yanaşmalar – texniki və etik baxımdan təhlil olunur. Eyni zamanda, istifadəçi səviyyəsində atribusiya, kiber cinayətkarlığın qarşısını alma və müəllif hüquqlarının qorunması baxımından su nişanlamanın rolu izah edilir. Sonda mövcud problemlər və gələcək tədqiqat istiqamətləri təqdim olunur.

Abdurahman Vagifly

The Importance Of Watermarks Applied To AI-Generated Images

Abstract

This paper explores the importance of watermarking technologies in detecting and identifying AI-generated content. It highlights the limitations of post hoc detection methods and emphasizes the advantages of proactive watermarking in the context of increasingly sophisticated generative AI models. Various types of watermarking – visible and invisible, fragile and robust, spatial and frequency-based, and deep learning-based – are analyzed from technical and ethical perspectives. The paper also discusses the role of user-level attribution in preventing cybercrime and protecting copyright. Key challenges and future research directions are presented at the end.

Абдурахман Вагифли

Значение водяных знаков, применяемых к изображениям, созданным ИИ

Аннотация

В данной статье рассматривается значение технологий водяных знаков (watermarking) для выявления и идентификации контента, сгенерированного искусственным интеллектом. Освещаются ограничения методов постфактум-детекции и подчёркиваются преимущества проактивных подходов на фоне всё более сложных моделей генеративного ИИ. Анализируются различные типы водяных знаков – видимые и невидимые, хрупкие и устойчивые, пространственные и частотные и основанные на глубоком обучении – с технической и этической точек зрения. Также рассматривается роль атрибуции на уровне

пользователя в борьбе с киберпреступностью и защите авторских прав. В заключение представлены актуальные проблемы и перспективные направления исследований.

Açar sözlər: Süni intellekt, generativ AI, su nişanlama, müəllif hüquqları, GenAI aşkarlanması.

Keywords: Artificial Intelligence, Generative AI, Watermarking, Copyright, Genai Detection.

Ключевые слова: Искусственный интеллект, генеративный ИИ, водяной знак, авторские права, детекция ИИ-контента.

Generativ süni intellektin (GenAI) ekponensial artımı realistik mətn, audio və vizual təsvirlərin yaradılmasını mümkün etməklə rəqəmsal mühiti yenidən formalaşdırmışdır. DeepSeek [Guo və b, 2025], GPT seriyaları [Radford və b, 2018], Stable Diffusion [Esser və b, 2024], DALL·E [Ramesh və b, 2021], MusicGen [Copet və b, 2023], və NaturalSpeech [Tan və b, 2024] kimi ən müasir generativ modellər insan yaradıcılığını təqlid etməkdə əlamətdar bacarıqlar göstərirlər.Bu inkişaf media və əyləncə sektorunda əhəmiyyətli dəyişiklərə səbəb olsa da,eyni zamanda dezinformasiya, şəxsiyyət oğurluğu (identity fraud) və kontent manipulasiyası kimi riskləri də özündə ehtiva edir. Bu risklərin qarşısını almaq üçün süni intellekt tərəfindən yaradılmış məzmunun aşkarlanması mühüm bir tədqiqat sahəsi kimi ön plana çıxmışdır. Müxtəlif aşkarlama üsulları arasında su nişanlama (watermarking) özünü daha qabaqlayıcı bir həll kimi göstərir – bu texnologiya süni intellekt tərəfindən yaradılan məzmuna insan tərəfindən sezilməyən, lakin izlənə bilən imzalar yerləşdirir. Post-faktum analizə əsaslanan reaktiv üsullardan fərqli olaraq, su nişanlama məzmunun yaradıldığı anda onun mənşəyinin yoxlanmasına imkan verir və bu da süni intellekt məzmununun dayanıqlı və geniş miqyasda tanınmasını asanlaşdırır . Bu xüsusiyyət su nişanlamanı süni intellektlə yaradılmış medianın sui-istifadəsinə qarşı mübarizədə xüsusilə cəlbedici edir

Süni intellektlə yaradılan şəkillərin hüquqi statusu

Süni intellektlə generasiya olunmuş məzmunun (AIGC) artması bir sıra hüquqi məsələləri də gündəmə gətirir və xidmət təminatçılarının bu məzmunun istifadəsini tənzimləyən siyasətlər hazırlaması zərurətini yaradır. Birincisi, yaradılan məzmun xidmət təminatçısının intellektual mülkiyyəti sayılır. Bir çox xidmətlər istifadəçilərə bu məzmunu kommersiya məqsədilə istifadə etməyə icazə vermir. Məzmunun satışından maliyyə qazancı əldə etmək bu siyasətə ziddir və hüquqi problemlər yarada bilər.

İkincisi, generativ modellər təhlükəli məzmunlar – saxta xəbərlər, zərərli süni intellektlə yaradılmış şəkillər, fişinq hücumları və ya kiberhücum vasitələri istehsal edə bilirlər. Buna görə də, dərin öyrənmə modelləri ilə yaradılmış məzmunun internetdə yayılmasını və istifadəsini tənzimləmək üçün yeni qanunlar qəbul olunur.

AIGC-nin qorunması və tənzimlənməsi zərurəti artdıqca, Google 2023-cü ilin iyun ayında bu mövzuda həll yollarını müzakirə edən bir seminar təşkil etdi. Gözlənildiyi kimi, su nişanı texnologiyası zərərli istifadəyə qarşı perspektivli müdafiə vasitəsi kimi təqdim olundu. Görünməyən, spesifik mesajlardan ibarət su nişanlarının yaradılmış məzmuna əlavə olunması vasitəsilə xidmət təminatçısı həmin məzmunun mənşəyini müəyyənləşdirə və istifadəçini izləyə bilir

Su nişanları və onların əhəmiyyəti

Süni intellekt tərəfindən generasiya olunmuş məzmunun yayılması ilə bərabər, bu məzmunun insan tərəfindən yaradılmış məzmundan fərqləndirilməsi zərurəti ortaya çıxmışdır. Bu kontekstdə ənənəvi yanaşmalardan biri post hoc aşkarlama üsullarıdır. Bu üsullar yaradılmış şəkil, mətn və ya audio faylın sonradan təhlil edilərək süni intellekt tərəfindən yaradılıb-yaradılmadığını müəyyən etməyə çalışır. Onlar əsasən stilistik xüsusiyyətlərə, struktur uyğunsuzluqlara və ya modelin mətn, səs və şəkil davranışlarındakı qeyri-adiliklərə əsaslanır. Post hoc aşkarlama üsullarının zəifliyini nümayiş etdirmək üçün real həyatdan nümunələr xüsusilə əhəmiyyətlidir. 2023-cü ilin mart ayında sosial şəbəkələrdə Papa Francesco-nun ağ puffer pencək geyinmiş şəkildə görüntüləndiyi bir foto geniş yayılmış və milyonlarla baxış toplamışdı (Şəkil 1). Şəkildə pontifikin qeyri-adi dərəcədə dəbli geyimdə, zərli xaç taxaraq küçədə gəzdiyi əks olunmuşdu. Foto ilk dəfə Reddit platformasında dərc edilmiş və sürətlə yayılmışdı. Lakin daha sonra məlum oldu ki, bu şəkil Midjourney adlı generativ süni intellekt aləti vasitəsilə yaradılmış süni intellekt saxtakarlığıdır.

Bu görüntüdə post hoc analizlə müəyyən edilə biləcək bəzi vizual uyğunsuzluqlar da mövcud idi. Məsələn, papanın iki əli eyni ölçüdə deyildi və onlardan biri qeyri-təbii şəkildə su şüşəsini tuturdu. Digər əlində isə üç barmağın iki üzük taxılmış barmağa birləşdiyi qəribə bir struktur fərqi müşahidə olunurdu. Bu cür detallar yalnız diqqətlə incələmə zamanı nəzərə çarpır və adi istifadəçilər tərəfindən əksər hallarda fərq edilmir.

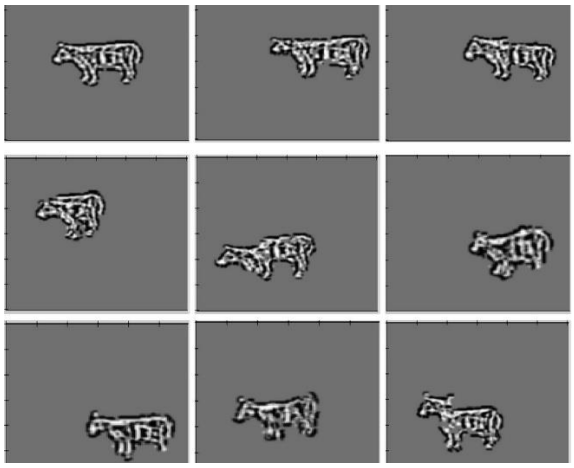
Bu fakt bir daha sübut edir ki, post hoc aşkarlama üsulları yüksək diqqət və texniki bilik tələb edir, və bu metodların kütləvi dezinformasiyanın yayılmasının qarşısını almaqda yetərli olması sual altındadır. Süni intelektin getdikcə daha realistik və təfərrüatlı vizual məzmun yaratma qabiliyyəti nəticəsində, bu tip sadə vizual səhvlər də zamanla aradan qalxır. Nəticə etibarilə, reaktiv yox, proaktiv yanaşmalara – məsələn, su nişanlarına keçid daha etibarlı bir həll kimi görünür.

Şəkil 1
Mənbə: “CBS News”

Şəkil 2-də 2014-cü ildə, bir inəyin şəklini gördükdən sonra, öz “təsəvvürü” ilə müxtəlif ölçü və pozalarda inəklərin illüstrasiyalarını yaratmağa qadir olan Vicarious adlı süni intellekt sistemi

tərəfindən yaradılan inək şəkillərinin təsviri verilmişdir. Bu sistem, adi kompüter görmə texnologiyalarından fərqli olaraq, sadəcə fotoşəkilləri təhlil edib yenidən yaratmır, əksinə, gördüyü nümunədən müəyyən qaydalar çıxararaq yeni və unikal vizuallar formalaşdırır. Şəkil 3-də 2025-ci ildə ChatGPT tərəfindən yaradılan inək təsvirini verilmişdir. Bu təsvirləri müqayisə etdikdə GenAI texnologiyalarındakı nəhəng inkişaf aydın görünür. İlk təsvirlərdə süni intellektin texniki məhdudiyyətləri, vizual uyğunsuzluqlar və reallıqdan uzaqlıq diqqəti çəkirdisə, 2025-ci ildə yaradılmış şəkil real foto ilə ayırd edilə bilməyəcək qədər keyfiyyətli və detallıdır.

Bu dəyişiklik post hoc üsulların nə üçün artıq yetərli olmadığını göstərir. GenAI modelləri insan davranışlarını, vizual harmoniyanı və hətta kontekstual uyğunluğu təqlid etməkdə o qədər bacarıqlı olub ki, sonradan təhlil vasitəsilə onların ifşası getdikcə çətinləşir



Şəkil 2

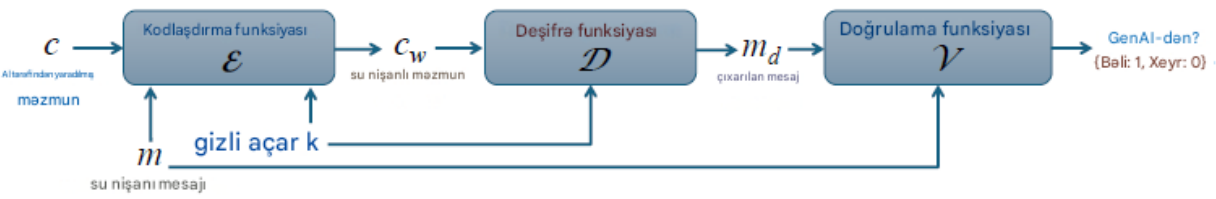
Mənbə: “Wall Street Journal” 2014

Şəkil 3

Mənbə: ChatGPT

Su nişanlama (watermarking) – süni intellekt tərəfindən yaradılmış məzmuna insan gözü ilə sezilməyən, lakin identifikasiya oluna bilən izlərin yerləşdirilməsi üçün nəzərdə tutulmuş proaktiv texnologiyadır. Bu texnika, məzmunun generasiya edildiyi anda həmin məzmuna iz qoyaraq, onun mənşəyini müəyyən etmək və autentikliyini yoxlamaq imkanı yaradır. Beləliklə, su nişanlama post hoc (yaradıldıqdan sonra analizə əsaslanan) metodlardan fərqli olaraq, mənbəyə bağlı izlənmə və təsdiqləmə prosesini mümkün edir.

Süni intellekt tərəfindən yaradılmış məzmuna tətbiq olunan su nişanlama sistemi aşağıdakı şəkildə tərif olunur:



Şəkil 4: Su nişanlama prosedurunun sxematik təsviri

W = (*E, D, V*), burada:

E – kodlaşdırma funksiyasıdır:

$E : C \times K \times M \rightarrow C_w$,

burada C – giriş məzmunu, K – gizli açar, M – yerləşdiriləcək su nişanı mesajı və C_w – su nişanı yerləşdirilmiş çıxış məzmunudur.

D – dekodlaşdırma funksiyasıdır:

$D : C_w \times K \rightarrow M$,

su nişanı yerləşdirilmiş məzmundan (C_w) və gizli açardan (K) istifadə edərək çıxarılan su nişanı mesajını əldə edir.

V – yoxlama funksiyasıdır:

$V : M \times M \rightarrow \{0,1\}$,

çıxarılan su nişanının orijinal su nişanı ilə uyğunluğunu yoxlayır və nəticədə 0 (uyğun deyil) və ya 1 (uyğundur) nəticəsini verir.

Effektiv və dayanıqlı bir su nişanlama sistemi aşağıdakı əsas xüsusiyyətlərə malik olmalıdır:

1. Seçilməzlik (Imperceptibility): Su nişanı məzmunun vizual və ya akustik keyfiyyətini aşağı salmamalı və adi insan qavrayışı ilə sezilməməlidir. Vizual və audio məzmunlarda bu xüsusiyyət PSNR ilə qiymətləndirilir. Mətnlər üçün BLEU və ROUGE kimi göstəricilər orijinal və su nişanı yerləşdirilmiş mətn arasındakı oxşarlığı ölçmək üçün istifadə olunur.
2. Dayanıqlılıq (Robustness): Su nişanı sıxılma, kəsim, parafraz etmə və ya səs-küy əlavə etmə kimi dəyişikliklərə qarşı davamlı olmalıdır. Bu xüsusiyyət BER (bit səhv nisbəti) ilə ölçülür. BER aşağı olduqca sistem daha möhkəm sayılır.
3. Təhlükəsizlik (Security): Su nişanı qeyri-qanuni müdaxilələrlə – məsələn, saxtalaşdırma və ya silmə cəhdləri ilə – asanlıqla yox edilməməlidir.
4. Tutum (Capacity): Sistem, məzmunu nəzərəçarpacaq dərəcədə dəyişdirmədən kifayət qədər məlumat yerləşdirilməsinə imkan verməlidir .

Su nişanlama texnologiyaları yerləşdirmə üsulu, görünürlüyü və məzmun üzərində edilən dəyişikliklərə qarşı davamlılıq baxımından fərqlənir. Bu texnikalar aşağıdakı əsas kateqoriyalara bölünə bilər:

1. Görünən və görünməyən su nişanı

Görünən su nişanı açıq şəkildə yerləşdirilən identifikatorlardır – məsələn, şəkillər üzərindəki loqolar və ya mətn üzərindəki təbəqə şəklində yazılar. Bu yanaşma vizual məzmun üçün uyğundur, lakin mətn və audio materiallar üçün praktik deyil. Əksinə, görünməyən su nişanlama yalnız alqoritmik analizlə aşkarlanabilən və insan tərəfindən sezilməyən gizli dəyişikliklər vasitəsilə həyata keçirilir.

2. Həssas və möhkəm su nişanı

Həssas (fragile) su nişanlama kiçik dəyişikliklərə qarşı belə həssasdır və əsasən məzmunun bütövlüyünü yoxlamaq üçün istifadə olunur. Möhkəm (robust) su nişanlama isə sıxılma, kəsmə və ya format dəyişməsi kimi müdaxilələrdən sonra da aşkarlanma qabiliyyətini saxlayır.

3. Məkan və tezlik sahəsində su nişanlama

Məkan sahəsində su nişanlama məzmunun öz strukturuna – məsələn, şəkillərdə piksel dəyərlərinin dəyişdirilməsinə və ya mətnə spesifik sözlərin əlavə olunmasına əsaslanır.Tezlik sahəsində su nişanlama isə məzmunun transformasiya olunmuş təqdimatında məlumat yerləşdirir. Bu

üsulda Diskret Kosinus Transformasiyası (DCT) və ya Diskret Dalğa Transformasiyası (DWT) istifadə olunaraq tezlik əmsalları dəyişdirilir.

Generativ süni intellekt sistemlərinin inkişafı ilə təkcə məzmunun saxta olub-olmadığını müəyyən etmək kifayət etmir. Getdikcə daha əhəmiyyətli hal alan məsələ həmin məzmunun kim tərəfindən yaradıldığını dəqiq müəyyən etməkdir. Bu proses “attribution” – yəni məzmunun konkret bir istifadəçi və ya modelə aid edilməsi adlanır və bu, su nişanlama texnologiyalarının genişlənmiş tətbiq sahələrindən biridir.

Attribution sistemlərinin əsas məqsədi, süni intellektlə yaradılmış hər bir məzmunun müəyyən və fərdi su nişanı daşımasıdır. Beləliklə, su nişanı yalnız “Sİ tərəfindən yaradılıb” məlumatını deyil, həm də "bu məzmunu filan istifadəçi və ya platforma yaradıb" məlumatını da ehtiva edə bilər. Bu xüsusilə kiber cinayətkarlıq halları (məsələn, saxta xəbərlərin yayılması, Sİ ilə yaradılmış fırıldaq mesajları, dərin saxta (deepfake) videolar və s. zamanı məsul şəxsin və ya sistemin identifikasiyası üçün vacibdir.

Attribution üçün istifadə olunan su nişanlama sistemləri adətən hər bir istifadəçiyə unikal bir su nişanı təyin edir və bu nişan Sİ tərəfindən yaradılan hər bir çıxışa yerləşdirilir. Daha sonra həmin məzmun analiz edilərkən, su nişanı çıxarılır və baza ilə müqayisə edilərək onun hansı istifadəçidən və ya hansı modeldən gəldiyi müəyyən edilir.

Nəticə

Generativ süni intellekt texnologiyalarının inkişafı ilə birlikdə, yaradılmış rəqəmsal məzmunun identifikasiyası və izlənilməsi mühüm əhəmiyyət kəsb edir. Post hoc aşkarlama üsulları tədricən effektivliyini itirir, çünki müasir süni intellekt sistemləri insanın yaratdığı məzmuna çox bənzər nəticələr təqdim edə bilir. Bu səbəbdən, su nişanlama texnologiyaları – xüsusilə görünməyən və atributiv su nişanlama – süni intellekt tərəfindən yaradılmış məzmunun mənşəyini təsdiqləmək, müəllif hüquqlarını qorumaq və dezinformasiyanın yayılmasının qarşısını almaq üçün daha etibarlı və proaktiv həll kimi çıxış edir. Lakin texniki zəifliklər, etik dilemmalar və standartlaşdırma problemləri hələ də aktualdır. Gələcək tədqiqatlar bu çətinlikləri aradan qaldırmağa, daha möhkəm və universal su nişanlama sistemləri yaratmağa və qlobal siyasətlə bu texnologiyaların etik və şəffaf tətbiqini təmin etməyə yönəlməlidir.

Ədəbiyyat siyahısı:

1. Alexei Grinbaum and Laurynas Adomaitis. The ethical need for watermarks in machine-generated language. arXiv preprint arXiv:2209.03118, 2022.
2. Google DeepMind. Identifying AI-generated content with SynthID, 2024.
3. Hanzhou Wu et al. 2020. Watermarking neural networks with watermarked images. IEEE Transactions on Circuits and Systems for Video Technology 31, 7 (2020), 2591–2601.
4. Jiang, Z., Zhang, J. & Gong, N. Z. Evading Watermark based Detection of AI-Generated Content in ACM Conference on Computer and Communications Security (2023).
5. Makena Kelly. 2023. Meta, Google, and OpenAI promise the White House they’ll develop AI responsibly. https://www.theverge.com/2023/7/21/23802274/artificialintelligence- meta-google-openai-white-house-security-safety.
6. Zhengyuan Jiang, Moyang Guo, Yuepeng Hu, and Neil Zhenqiang Gong. Watermark-based detectionand attribution of ai-generated content. arXiv preprint arXiv:2404.04254, 2024