DOI: https://doi.org/10.36719/2663-4619/114/298-301

Nəbi Hüseynov
Odlar Yurdu Universiteti
magistrant
https://orcid.org/0009-0004-2086-8621
huseynovnebi624@gmail.com

Üz tanıma sistemlərinin xarici hücumlara qarşı dayanıqlılığının təkmilləşdirilməsi

Xülasə

Bu məqalə, üz tanıma sistemlərinin xarici hücumlara qarşı dayanıqlılığını artırmaq üçün üsulları araşdırır. Xarici hücumlar, üz tanıma modellərini yanıltmaq məqsədilə məlumatlarda kiçik dəyişikliklər edərək sistemi aldatmağa çalışır. Üz tanıma texnologiyası təhlükəsizlik, identifikasiya və şəxsiyyəti təsdiqetmə kimi sahələrdə geniş istifadə olunur, lakin bu texnologiyanın məxfiliyi və təhlükəsizliyi qorumaq üçün xarici hücumlardan qorunması zəruridir. Bu araşdırma, mövcud üz tanıma sistemlərinin zəif cəhətlərini öyrənərək, müxtəlif növ xarici manipulyasiyalara qarşı müdafiə strategiyalarının tətbiqi ilə sistemin etibarlılığını və təhlükəsizliyini artırmağı hədəfləyir. Bunun üçün dərin öyrənmə modellərinin zəiflikləri analiz edilərək onların daha müdafiəli olması üçün süni intellekt əsaslı təhlükəsizlik mexanizmləri təklif olunur. Eyni zamanda, sistemlərin davamlılığını artırmaq məqsədilə fərqli hücum növlərinə qarşı adaptiv müdafiə üsulları inkişaf etdirilir. Bu yanaşmalarla, üz tanıma sistemlərinin performansı daha stabil və etibarlı olacaq, həmçinin mövcud təhlükəsizlik zəifliklərinin qarşısı alınacaqdır. Araşdırmalar, bu sahədəki innovativ həllərin tətbiqi ilə daha güclü və hücuma qarşı dayanıqlı sistemlərin yaradılmasına imkan verəcək.

Məqalədə bu konteksdə ümumi araşdırmalar aparılmış, yerli və xarici müəlliflərin əsərlərinə istinad olunaraq dəqiq faktlar qeyd edilmişdir.

Açar sözlər: *üz tanıma, xarici hücumlar, sistem etibarlılığı, modelləmə, təhlükəsizlik*

Nabi Huseynov
Odlar Yurdu University
Master student
https://orcid.org/0009-0004-2086-8621
huseynovnebi624@gmail.com

Improvement the Resilience of Facial Recognition Systems Against External Attacks

Abstract

This article explores methods to improve the resilience of face recognition systems against external attacks. External attacks attempt to deceive the system by making small changes in the data to mislead the face recognition models. Face recognition technology is widely used in fields such as security, identification, and authentication; however, it is crucial to protect this technology from external attacks to ensure privacy and security. This research aims to enhance the reliability and security of the system by studying the weaknesses of existing face recognition systems and applying defense strategies against various types of external manipulations. To achieve this, the vulnerabilities of deep learning models are analyzed, and AI-based security mechanisms are proposed to make them more resilient. Additionally, adaptive defense strategies are developed to improve the system's resilience against different attack types. With these approaches, the performance of face recognition systems will be more stable and reliable, and the existing security vulnerabilities will be mitigated. The research will enable the development of stronger and more attack-resistant systems through the implementation of innovative solutions in this field.

In this context, the article presents a comprehensive study, referring to works by both local and international authors, and providing precise facts.

Keywords: *facial recognition, external attacks, system reliability, modelling, security*

Giriş

Müasir dövrün texnoloji inkişafları üz tanıma sistemlərini təhlükəsizlik və identifikasiya vasitəsi olaraq geniş istifadə olunan bir texnologiyaya çevirmişdir. Hava limanlarından banklara, ticarət sahəsindən sosial mediaya qədər müxtəlif sahələrdə üz tanıma texnologiyaları insanların həyatını asanlaşdırmaqda mühüm rol oynayır. Lakin, maşın öyrənmə texnologiyalarının inkişafı ilə bu sistemlər kiçik, lakin məqsədli dəyişikliklərlə yanıldılaraq təhlükə altına salına bilər. Xarici hücumlar nəticəsində texnologiyanın etibarlılığı azalır və sistemlər daha həssas vəziyyətə düşür (Goodfellow, Bengio, Courville, 2016).

Xarici hücumlar əsasən, üz şəkillərində nəzərə çarpmayan, lakin modelin qərar qəbul etmə prosesinə ciddi təsir göstərən dəyişikliklər vasitəsilə həyata keçirilir. İnsan gözü ilə nəzərə çarpmayan bu dəyişikliklər üz tanıma sistemlərini yanıltmağa və səhv qərarlar verməsinə səbəb ola bilər. Belə hücumların təsiri yalnız təhlükəsizlik deyil, həm də məxfilik aspektlərini əhatə edir. Xarici hücumlar nəticəsində şəxsi məlumatların sızması, identifikasiyanın yanlış aparılması və digər ciddi problemlər meydana çıxa bilər (Turkes, Cetin, 2020).

Tədqiqat

Bu tədqiqat xarici hücumların mexanizmini, onların müxtəlif növlərini və bu hücumların qarşısını almaq üçün istifadə edilə biləcək müdafiə strategiyalarını araşdırır. Əsas məqsəd, üz tanıma texnologiyalarının təhlükəsizliyini artırmaq və hücumlara qarşı dayanıqlığını təmin etməkdir (Liu, Chen, 2017).

Xarici Hücumların Mexanizmi və Təhlükəsi

Xarici hücumların əsas mexanizmi, maşın öyrənmə modellərinin zəifliklərini istismar etməyə əsaslanır. Bu hücumlar insan tərəfindən fərq edilməyən dəyişikliklər vasitəsilə modelin davranışını dəyişdirərək səhv nəticələr əldə etməyi hədəfləyir. Xarici hücumlar bir çox sahədə təhlükə yaradır, xüsusilə də təhlükəsizlik sistemlərində, tibbi diaqnostikada və avtonom idarəetmə sistemlərində. Xarici hücumların əsas növləri və onların mexanizmləri (Szegedy, Zaremba, Sutskever, 2014):

1. **Fast Gradient Sign Method (FGSM)**:
   - Bu hücumun əsas məqsədi modelin zəifliklərindən istifadə edərək məlumatlarda minimal dəyişikliklər tətbiq etməkdir.
   - FGSM modelin itkisini (“loss function”) artırmaq məqsədilə gradient istiqamətində dəyişikliklər edərək hücum təşkil edir.
   - Təsiri nisbətən məhdud olsa da, sürətli və effektiv olduğuna görə geniş istifadə olunur.
   - Tətbiq sahələri: Şəkil sinifləndirilmə sistemləri, obyekt tanıma tətbiqləri və s.
2. **Projected Gradient Descent (PGD)**:
   - FGSM-dən fərqli olaraq, iterativ yanaşma tətbiq edərək modelin zəifliklərini daha dərin araşdırır.
   - Kiçik, ardıcıl dəyişikliklər tətbiq edərək modelin verdiyi nəticələri manipulyasiya edir.
   - PGD ən güclü “white-box” hücum metodlarından biri hesab olunur, çünki modelin daxilindəki parametrləri və strukturunu bilir.
   - Müdafiəsiz modellərə qarşı təsiri yüksəkdir.
3. **Carlini & Wagner (C&W) Hücumları**:
   - Bu metod maşın öyrənmə modellərinin təhlükəsizliyini pozmaq üçün daha mürəkkəb strategiyalardan istifadə edir.
   - L2, L0 və L∞ normlarına əsaslanan fərqli hücum formaları mövcuddur.
   - Bu hücumlar FGSM və PGD-dən daha hesablama baxımından tələbkar, lakin daha güclü ola bilər.
   - Əsasən neyron şəbəkələrinin zəifliklərini aşkar etmək və onların müdafiəsini sınaqdan keçirmək üçün istifadə olunur.

4. **Black-Box Hücumları**:
   - Bu tip hücumlar modelin daxili strukturu haqqında məlumat tələb etmir.
   - Hücum edən tərəf modelin girişlərinə müxtəlif dəyişikliklər edərək çıxışlarını müşahidə edir və təcrübə yolu ilə modelin zəifliklərini tapır.
   - Generativ hücum metodları ilə kombinasiya edilə bilər (məsələn, genetika əsaslı metodlar, təkamül alqoritmləri və s.).
   - Əsasən qapalı sistemlərə hücum üçün istifadə olunur, məsələn, bulud xidmətləri və ya ticari AI API-ləri.

Xarici hücumların yaratdığı təhlükələr yalnız texniki sahə ilə məhdudlaşmır. Onlar sosial və iqtisadi təsirlərə də malikdir. Məsələn, banklarda istifadə olunan identifikasiya sistemlərinin yanıltılması maliyyə itkilərinə və ya şəxsiyyət oğurluğuna gətirib çıxara bilər (Carlini, Wagner, 2017).

Adversarial Hücumların Təhlili

Adversarial hücumlar, maşın öyrənmə modellərinin təhlükəsizliyini sarsıdan və onların zəifliklərini istismar etməyə səbəb olan həmlələrdir. Bu hücumlar modelin düşünə biləcəyi və qərar verə biləcəyi sahədə kiçik dəyişiklər etməklə onun səhv nəticələr verməsinə səbəb olur. Bu metodlar çox vaxt vizual təsir etmədiyi halda, dərindən öyrənmə modelləri üçün ciddi problemlər yarada bilir (Papernot, McDaniel, Goodfellow, 2016).

Bu hücumları iki əsas kateqoriyaya bölmək olar: ağ qutu (white-box) və qara qutu (black-box) hücumlar. Ağ qutu hücumlar modelin daxili arxitekturası, parametrləri və çıxışları haqqında tam məlumatın olduğu hallarda tətbiq edilir. Qara qutu hücumlar isə modelin daxili strukturu haqqında heç bir bilik olmadan gerçəkləşdirilir (Dong, Pang, 2018).

Adversarial hücumları yaratmaq üçün fərqli metodlar mövcuddur. Fast Gradient Sign Method (FGSM) modelin itki funksiyasının qradienti istiqamətində kiçik bir pertubasiya yaradaraq modelin verdiyi cavabı yanıltmaq üçün istifadə edilir. PGD metodu FGSM-dən fərqli olaraq iterativ yanaşma ilə bir neçə addımda hər dəfə daha effektiv adversarial müdaxilə yarada bilir. Carlini, Wagner (C,W) metodu isə daha güclü yanaşma təklif edərək modelin zəifliklərini daha mükəmməl istismar edir (Akhtar, Mian, 2018).

Adversarial hücumlar çox vaxt dərindən öyrənmə modellərinin təhlükəsizliyini test etmək və onlara qarşı qorunma strategiyaları işləmək üçün istifadə edilir. Bəzi təhlükəsizliyin artırılması üçün strategiyalar adversarial training, modelin rəndomizasiyası, və ya dəyişikliklərin optimallaşdırılması kimi metodları özünə alır (Zhou, Chen, 2018).

Məsələn, adversarial training metodu modelin adversarial dəyişdirilmiş verilənlər üzərində təlim edilməsi yolu ilə hücumlara daha davamlı olmasını təmin edir. Digər yanaşmalar, məsələn, modelin rəndomizasiyası, verilənlərə təbii səslər və ya məlumatlar əlavə etmək kimi texnikalarla modelin daha sabit davranış göstərməsini təmin edir (Xu, Liu, 2021).

Son dövrlərdə tədqiqatlar adversarial hücumlara qarşı yeni metodları araşdırmaqda davam edir. Bununla yanaşı, şəbəkələrin və sistemlərin öz təhlükəsizliyini artırması üçün adversarial analiz və test metodları çox vacibdir. Maşın öyrənmə sahəsində təhlükəsizlik tədqiqatları davam etdikcə, adversarial hücumları daha effektiv dəf etmə yolları da əlavə olaraq öyrənilib təkmilləşdiriləcəkdir (Kallab, Doroudchi, 2020).

Müdafiə Strategiyaları və Texnikalar

Xarici hücumların təsirini azaltmaq və sistemlərin etibarlılığını artırmaq üçün bir neçə müdafiə strategiyası mövcuddur (Zhang, Zhao, 2020):

- **Xarici Təlim:**
- Modelin xarici nümunələrlə təlimi zamanı real və dəyişdirilmiş nümunələrdən istifadə edilir. Bu metod, modelin xarici hücumlara qarşı dayanıqlığını artırır.
- **Adversarial Training (Rəqib əsaslı təlim)**:
  - Modeli hücumlara qarşı davamlı etmək üçün təlim zamanı adversarial nümunələrdən istifadə olunur.
  - Model, hücum edilmiş nümunələrə məruz qalaraq öz qərar vermə mexanizmini gücləndirir.
- **Defensive Distillation (Müdafiə Distillə etmə)**:

o Modelin çıxış ehtimallarını yumşaltmaq və həddən artıq spesifik qərarları qarşısını almaq məqsədilə istifadə edilir.

o Bu metod modelin adversarial hücumlardan daha az təsirlənməsinə səbəb olur.

- **Gradient Masking (Gradient Maskalanması)**:

o Modelin gradientlərini gizlətmək və ya dəyişdirmək yolu ilə hücum edən tərəfin məlumat əldə etməsini çətinləşdirir.

o Lakin bəzi güclü hücum metodları (məsələn, black-box hücumlar) bu texnikanı aşa bilər.

- **Model Ensemble (Çoxlu Model)**:

o Birdən çox modelin eyni vaxtda işləməsi və çıxışların birləşdirilməsi ilə müdafiə səviyyəsinin artırılması.

o Müxtəlif arxitekturalar və hiperparametrlər əsasında yaradılmış modellər hücumlara qarşı daha dözümlü ola bilər.

- **Noise Injection (Səs-küy əlavə etmək)**:

o Modelin girişlərinə müəyyən səviyyədə təsadüfi səs-küy əlavə etmək, hücum edən tərəfin modelin zəifliklərini aşkar etməsini çətinləşdirir.

o Xüsusilə səs və görüntü əsaslı modellərdə effektivdir.

Nəticə

Bu tədqiqat, üz tanıma sistemlərinin xarici hücumlara qarşı zəif cəhətlərini aşkara çıxarmaq və bu hücumların təsirini minimuma endirmək üçün müdafiə strategiyalarını təhlil etmişdir. Xarici hücumların təhlükəsizlik, məxfilik və etibarlılıq aspektlərinə ciddi təsir göstərdiyi məlum olmuşdur.

Müdafiə texnikalarının düzgün tətbiqi ilə üz tanıma sistemlərinin dayanıqlılığı artırıla bilər. Gələcək tədqiqatlar daha effektiv müdafiə texnikalarının işlənib hazırlanmasını və bu texnikaların real tətbiqlərdə sınaqdan keçirilməsini tələb edir. Təhlükəsizlik sahəsindəki inkişaflar, yalnız texnoloji deyil, həm də sosial və hüquqi aspektləri əhatə etməlidir.

Üz tanıma texnologiyaları üçün təhlükəsiz və etibarlı sistemlərin qurulması, bu texnologiyanın gələcəkdə geniş yayılmasını və cəmiyyətin ondan faydalanmasını təmin edəcəkdir.

Ədəbiyyat

1. Akhtar, N., Mian, A. (2018). *Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey*. IEEE Access.
2. Carlini, N., Wagner, D. (2017). *Towards Evaluating the Robustness of Neural Networks*. IEEE S&P.
3. Dong, Y., Pang, T. (2018). *Boosting Adversarial Attacks with Momentum*. ICLR.
4. Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*.
5. Kallab, A., Doroudchi, S. (2020). *Adversarial Robustness in Deep Learning*. Springer.
6. Liu, Y., Chen, Y. (2017). *Adversarial Attacks and Defenses in Deep Learning: A Survey*. Springer.
7. Papernot, N., McDaniel, P., Goodfellow, I. (2016). *Transferability in Machine Learning: From Phenomena to Black-Box Attacks using Adversarial Samples*. ACM CCS.
8. Szegedy, C., Zaremba, W., Sutskever, I. (2014). *Intriguing Properties of Neural Networks*. ICLR.
9. Turkes, S., Cetin, M. (2020). *A Survey on Adversarial Machine Learning: Attacks and Defenses*. IEEE Access.
10. Xu, J., Liu, H. (2021). *Machine Learning Security: Threats, Attacks, and Defenses*. Springer.
11. Zhang, Q., Zhao, Y. (2020). *Deep Learning in Face Recognition: A Survey*. Springer.
12. Zhou, B., Chen, Y. (2018). *Adversarial Attacks and Defenses in Facial Recognition Systems*. Springer.

Daxil oldu: 12.11.2024
Qəbul edildi: 14.02.2025