СИНТЕТИЧНІ МЕДІА ТА ФЕЙКИ: НОВІ МЕТОДИ ВИЯВЛЕННЯ ФАЛЬШИВОГО КОНТЕНТУ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

Гребенюк Є. А., Штангей С. В.

Харківський національний університет радіоелектроніки,
кафедра інфокомунікаційної інженерії ім. В.В. Поповського,
Україна.

E-mail: yelyzaveta.hrebeniuk@nure.ua,
svitlana.shtanhei@nure.ua

Abstract

The rise of synthetic media, particularly deepfakes, has presented new challenges in the realm of information security and public trust. Through advancements in artificial intelligence (AI) and machine learning (ML), creating highly realistic fake images, audio, and video has become possible, posing serious threats to society, especially in news reliability and identity verification.

This work examines machine learning methods, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), that have proven effective in detecting fake content by identifying subtle visual and auditory anomalies typical of synthetic media. This study highlights the need for a comprehensive approach to fake content detection, involving technical innovation, collaborative data sharing, and robust ethical standards to protect digital identities and counter disinformation.

Вступ

Сьогодні синтетичні медіа, такі як дипфейки, дуже впливають на те, як люди сприймають інформацію. Завдяки розвитку технологій штучного інтелекту (ШІ) та машинного навчання (ML) стало можливим створення високоякісних фальшивих зображень, аудіо та відео, які важко відрізнити від реальних. Це становить серйозну загрозу для суспільства, особливо в питаннях довіри до новин та ідентифікації особистості. В останні роки розвиваються та вдосконалюються різні методи детекції таких підробок, і великий попит мають алгоритми глибокого навчання, які здатні аналізувати патерни, що зазвичай не зустрічаються в автентичних даних. Вони дають змогу аналізувати візуальні, аудіальні та текстові дані на такому рівні деталізації, який не підвладний людському оку, що відкриває нові можливості для виявлення фальсифікацій [1].

Огляд синтетичних медіа

Синтетичні медіа – це контент, створений або змінений за допомогою штучного інтелекту, який включає текст, зображення, аудіо і відео. Важливою підмножиною цих технологій є дипфейки – аудіовізуальні матеріали, в яких обличчя, голос або поведінка особи змінюються так, щоб видаватися справжніми. У 2017 році можливості створення маніпульованих відео за допомогою ШІ стали значно поширенішими завдяки розвитку генеративних змагальних мереж (GANs), які суттєво підвищили якість і реалістичність фальсифікованих зображень та відео [3].

Синтетичні медіа мають широкий спектр позитивних застосувань, таких як медична діагностика, створення контенту в індустрії розваг і кіно, а також навчання та тренування моделей штучного інтелекту. Такі сценарії надають великі переваги, розширюючи можливості застосування ШІ у творчості та бізнесі.

Проте, існують також ризики недобросовісного використання синтетичних медіа, зокрема дипфейків, які можуть використовуватися для політичної дезінформації, фейкових новин, шахрайства, шантажу або порушення приватності. Дипфейки стали одним із найпотужніших інструментів маніпулювання громадською думкою. Такі загрози особливо небезпечні у контексті національних виборів та інформаційної безпеки, коли фальшиві заяви чи події можуть значно вплинути на суспільну довіру, політичну стабільність чи здоров'я та життя громадян [4], [5].

Серед технологій, які використовуються для створення дипфейків, ключове місце займають генеративні змагальні мережі (GANs), що працюють за принципом двох мереж: генератор утворює фейковий контент, а дискримінатор намагається відрізнити його від справжнього. Цей процес поступово підвищує якість контенту, зробленого генератором, створюючи все більш правдоподібні зображення, аудіо та відео. Завдяки GANs синтетичний контент стає настільки реалістичним, що дуже важко відрізнити його від реального навіть при детальному аналізі [3].

Виклики у виявленні фейкового контенту

Виявлення фейкового контенту, включаючи дипфейки, стикається з багатьма викликами, що охоплюють технічні, етичні та правові аспекти.

- З технічної точки зору, сучасні методи ідентифікації підробок потребують значних обчислювальних ресурсів для аналізу високоякісного контенту. Алгоритми, засновані на глибокому навчанні, вимагають великих обсягів даних для навчання та тренування моделей, а також потужного апаратного забезпечення, що робить їх недешевими та не завжди доступними. Крім того, швидкий розвиток технологій для створення фейків значно ускладнює роботу розробників, змушуючи їх адаптуватися до нових і складніших технік маніпуляції.
- З етичної точки зору, важливе питання стосується прав людини на захист своєї цифрової особистості. Використання інструментів для аналізу персональних даних може порушувати конфіденційність та особисту недоторканність. Деякі методи виявлення можуть порушувати права на приватність через велику обробку персональних даних [5].
- З правової точки зору, відсутність чітких регулювань щодо синтетичних медіа робить їх складними для виявлення та протидії зловживанням. Важливою є проблема встановлення меж відповідальності за поширення фейкового контенту та розробки законодавства, яке б забезпечувало захист цифрової особистості та гарантувало етичне використання технологій для виявлення підробок. Окрім цього, набуває важливості питання міждержавного співробітництва, адже поширення фейків часто не обмежується однією країною [2].

Виявлення фейкового контенту є складним завданням, що вимагає значних ресурсів та узгоджених дій для захисту конфіденційності та посилення правових стандартів.

Методи машинного навчання для виявлення фейкового контенту

Сучасні алгоритми машинного навчання відіграють ключову роль у виявленні фейкового контенту. Зокрема, згорткові нейронні мережі (CNN) широко застосовуються для аналізу зображень, а рекурентні нейронні мережі (RNN) – для роботи з відео та аудіо.

Згорткові нейронні мережі (Convolutional Neural Networks, CNN) — це один з найбільш ефективних методів машинного навчання для виявлення фейкового контенту, особливо в області аналізу зображень та відео. Основною метою CNN є виявлення специфічних візуальних характеристик, які можуть вказувати на маніпуляцію або фальсифікацію контенту.

1) Попередня та загальна обробки зображень: CNN отримує на вхід зображення, яке проходить через кілька етапів обробки для виділення особливих характеристик.

2) Згорткові шари: Другий етап – згортка, процес, коли мережа "навчається" розпізнавати невеликі патерни в кожній частині зображення, наприклад, певні лінії або текстури. Завдяки багаторазовій згортці CNN може відрізняти справжні зображення від фейкових, виявляючи неочевидні аномалії.

3) Пулінг (Pooling) та зменшення розмірності: Пулінг зменшує розмірність вхідних даних, видаляючи непотрібні деталі, які не грають важливої ролі, при цьому зберігаючи основні характеристики. Найчастіше використовуються методи Max Pooling або Average Pooling, які допомагають зосередити увагу на найважливіших ознаках фейків.

4) Повнозв'язні шари (Fully Connected Layers): Далі мережа переходить до повнозв'язних шарів, де всі виділені ознаки комбінуються і аналізуються, щоб надати остаточний висновок. На цьому етапі мережа вирішує, чи є контент справжнім або фейковим.

CNN потребує великої кількості даних для навчання, особливо у випадках виявлення фейків, коли різниця між оригіналом і підробкою може бути майже непомітною. Набори даних можуть включати справжні зображення та зображення, створені за допомогою різних методів (наприклад, GANs), що допомагає моделі адаптуватися до різних типів фальсифікацій. Таким чином, перевагою згорткової нейронної мережі є висока точність виявлення дрібних деталей і аномалій при правильному навчанні моделі та достатньої кількості вхідних даних.

Рекурентні нейронні мережі (Recurrent Neural Networks, RNN) ефективні для виявлення фейкового контенту у форматах, що містять послідовні дані, такі як відео або аудіо, де важливо аналізувати залежності між кадрами чи звуками.

1) Аналіз часових залежностей: RNN відрізняються від інших нейронних мереж здатністю зберігати інформацію про попередні стани. Це дозволяє їм розуміти і запам'ятовувати, як розвиваються події з часом.

2) Робота з послідовностями даних у відео: У випадку відео контенту, RNN аналізують послідовність кадрів, щоб виявити аномалії, наприклад, різкі зміни обличчя, рухів чи освітлення між кадрами. Це допомагає виявляти такі маніпуляції, як deepfake-відео.

3) Аналіз звукових сигналів: Рекурентні мережі також використовуються для аналізу аудіо, де врахування послідовності звукових сигналів відіграє важливу роль для виявлення підроблених голосових записів. Мережа вивчає патерни у звукових хвилях, щоб виявляти аномалії, наприклад, штучні коливання тону або несправжні паузи в мові, які не характерні для реального спікера.

4) Варіанти RNN: LSTM та GRU: Звичайні RNN мають обмежену пам'ять на довгі послідовності, тому для складніших задач використовуються модифікації RNN: LSTM (Long Short-Term Memory) та GRU (Gated Recurrent Unit). Вони краще запам'ятовують тривалі послідовності і розпізнають приховані закономірності.

RNN навчаються на великих наборах відео- і аудіоданих, які містять як справжній, так і фейковий контент. Це дозволяє їм краще розпізнавати різні види підробок і адаптуватися до нових типів маніпуляцій. Тому, перевагами рекурентних нейронних мереж є розуміння послідовності подій та гнучкість у роботі з різними типами даних.

Найбільшим недоліком будь-якої нейромережі є перенавчання (overfitting) – проблема, що виникає, коли модель надмірно пристосовується до навчальних даних, запам'ятовуючи їхні

особливості, замість того щоб навчитися розпізнавати загальні закономірності. Це призводить до чудових результатів на навчальних стадіях, але поганих у роботі з новими даними.

Для уникнення перенавчання у мережах використовуються різні способи боротьби з проблемою. При роботі з CNN використовують аугментацію даних, що напрямлена на створення додаткових варіантів зображень, змінюючи певні параметри, що допомагає моделі навчитися розпізнавати ознаки, незалежні від специфіки навчальних зображень. Ще один спосіб – Dropout: випадкове вимкнення нейронів під час навчання, допомагає знизити залежність моделі від конкретних нейронів. Регуляризація (L2-регуляризація) також є дієвим методом у боротьбі з перенавчанням, що додає штраф за високі ваги, зменшуючи їх значення та запобігаючи надмірному запам'ятовуванню.

У RNN перенавчання часто виникає, коли модель намагається запам'ятати точну послідовність у навчальних даних, що призводить до поганого розпізнавання нових патернів у відео чи аудіо. Це може бути особливо проблематично для LSTM та GRU, які здатні зберігати довгі послідовності. Щоб запобігти цій проблемі використовують Dropout для рекурентних шарів, що послідовно відключає нейрони. Раннє завершення навчання (Early Stopping) – припиняє процес навчання, коли точність на валідаційному наборі перестає зростати, що запобігає надмірному навчанню моделі на навчальних даних. Також, як у випадку з CNN, дієвою є регуляризація вагів.

Висновки

Сучасні технології синтетичних медіа та генерації фейкового контенту, ставлять серйозні виклики у сфері інформаційної безпеки та соціальної довіри. Розробка ефективних методів виявлення таких маніпуляцій є критично важливою для протидії дезінформації, захисту приватності та підтримки стабільності суспільства. Методи машинного навчання, зокрема згорткові (CNN) і рекурентні нейронні мережі (RNN), довели свою ефективність у розпізнаванні аномалій, властивих підробкам, завдяки здатності виявляти приховані візуальні та часові закономірності.

Однак, висока ефективність цих методів залежить від великої кількості якісних навчальних даних та вдосконалених алгоритмів запобігання перенавчанню. Це вказує на необхідність активного обміну спеціалізованими наборами даних і тісної співпраці між дослідниками для подолання швидкого розвитку технологій генерації фейків. Крім того, важливими залишаються етичні та правові питання, що стосуються захисту цифрової особистості й відповідального використання цих технологій. В цілому, ефективне вирішення завдання детекції фейкового контенту вимагає комплексного підходу, який включає технічні інновації, правове регулювання та етичні стандарти.

Література

1. Daniel Tison. The Future of Journalism: Opportunities and Challenges of Human-AI Collaboration, 2023. 93 p.
2. Synthetic Media & Deepfakes. URL: https://innovating.news/article/synthetic-media-deepfakes/ (date of access: 30.10.2024).
3. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio. Generative Adversarial Nets. Departement d'informatique et de recherche operationnelle. Universite de Montreal Montreal, QC H3C 3J7, 2014. 9 p.
4. Bobby Chesney, Danielle Citron. Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, 2019. 68 p.
5. Luisa Verdoliva. Media Forensics and DeepFakes: an overview, 2020. 24 p.
6. Detecting opinion spams and fake news using text classification. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9 (date of access: 30.10.2024).