DeepFake

I. What is DeepFake

1. General Idea

* Content generated by an artificial intelligence to fool humans[1]
* “deep learning” + “fake”[1]
* Branch of machine learning[1]

1. Applications with examples

Produce synthetic media[3]

* Merge, combine, replace and superimpose various forms of media to[3]
* Generation and manipulation of human imagery[1] - largest impact so far[1]

1. Fake Video
   1. Face Swap[1,4]

The face of the source person is replaced with the face of a victim[4]

* + 1. Non-consensual pornographic videos[1,2,4]
       1. Non-consensual pornography accounts for 96% of the total deepfake videos found online[2]
  1. Lip Synching[4]

The movement of the target person’s lips is manipulated to make them consistent with a specific audio recording so that the victim appears to say whatever is in the recording[4]

* + 1. Deepfake video of former president Barak Obama giving a talk[1]
  1. Face Reenactment[1] /Puppet Master[4]

Videos is created which mimics the expressions of the target person, such as eye movement, facial expressions and head movement[4]

* 1. Face Synthesisand Attribute Manipulation[4]

Generation of photo-realistic images as well as facial attribute editing[4]

Examples:

* 1. (Enhance Marketing) Digitally realistic rendition of videos without celebrity having actually to perform in it (offers significant flexibility)[3,6]
     1. Quebec Crown corporation using deepfake to reminisce viewer of a well-known Canadian news anchor[3]
     2. Quick replication, replace hundreds hours of professional actor/make-up artists work[6]
     3. Models are able to license their personal decoder to the ad agency with which ads can be auto-created using deepfake technology[6]
  2. Realistic video dubbing of foreign films[1]
  3. Animating dead characters[1]
     1. Education through the reanimation of historical figures[1]
  4. Altering published movies[1]
     1. For making a parody/comedy/satire[1] – beneficial application[2]
  5. Generating actors in movies[1]
  6. Technology demos[1]

1. Fake Image
   1. Virtually trying on clothes while shopping[1]
   2. Generated portraits[1]
   3. (Healthcare) Inject/remove medical evidence in CT and MRI scans[1]
   4. (Healthcare) Offer significant diagnostic benefits – beneficial application[3]
   5. (Financial) Evade the detection of auditors to financial records[1]
   6. (Tourism) Generate photorealistic images of scenes – beneficial application[3]
      1. Useful in advertising and marketing content creation, offer ad-evoked positive affect on tourists[3]
   7. (Galleries) Bring famous artists and their artworks to life[3]
      1. Dali museum, St. Petersburg, USA uses deepfake to recreate immersive visitor interaction and learning about Salvador Dali(1904-1989)[3]
2. Fake Audio

Audio deepfakes focus on the generation of the target speaker’s voice using deep learning techniques to portray the speaker saying something they have not said[4]

* 1. Text to speech synthesis(TTS)[4]

Produce natural and intelligible voice waveforms, based on the provided text, that sounds like they have been spoken by the target identity[4]

* 1. Voice conversion (VC)[4]

VC techniques transform the speech signal produced by a source speaker to seem like it was spoken by a target speaker while keeping the linguistic contents intact[4]

Examples

* 1. Voice Cloning[4] : One’s voice can be cloned with only 5 seconds of audio[1]
  2. CEO scammed out of $250K via a voice clone deepfake, cloned the CEO’s speech to dupe transferring into a secret account[1,4]
  3. Synthetic voice is widely adapted for the development of different applications:[4]
     1. Automated dubbing for TV and film[4]
     2. Chatbot[4]
     3. AI assistants[4]
     4. Text readers[4]
     5. Personalized synthetic voices for vocally handicapped people[4]

1. Fake Fingerprint
   1. Generate realistic human fingerprints that can unlock multiple user’s devices[1]
2. Fake Texts
   1. Generate deepfake of news articles[1]
   2. Generate deepfake tweets[1]
3. Use of Deepfake to raise ethical and legal questions over freedom of speech and expression[2]
4. [Medicine] Utilize Deepfake in therapeutical applications to help a person learn skill in virtual reality than can be transferred to physical world (As Deepfake is able to create convincing likenesses of real-world environments or people)[5]
5. Combining Deepfake with other extended reality (XR) technologies[5]
   1. Entertainment, education and even retail and services[5]
   2. Offers the chance to be transported anywhere or any time, providing access to people and places that are either inconvenient or impossible for ones to visit[5]
6. Industrial Products – Application and Software
   1. Visual Software: FackApp, FaceSwap[2,4], ZAO[2,4], REFACE[4], Dublicat[2] …
   2. Voice Synethesis Models: Audacity[4], Soundforge[4], Wavenet[4] …
   3. Commercial Products: Google Cloud TTS, Amazon AWS Polly, Baidu TTS: aims to attain high similarity between synthesized and human voices[4]
   4. Service portal that generate deepfake video as requested[2]
   5. Other specific method for each application reference in[4]

Graphical user interface, application

Description automatically generated[4]

1. Deepfake algorithm
   1. Face-swapping source code is publicly available at no cost on GitHub[3]
2. Maturity of technology

Mature/immature areas

Reason behind why DeepFake hasn’t been put into widespread use

* 1. Current Technology
     1. Generative Adversarial Networks (GAN)
        1. Deepfake utilizes GANs to manipulate all forms of media: photographs, texts, audio and videos)
        2. Simultaneously trains two competing models
     2. Reenactment: A reenactment deepfake is where x\_s is used to drive the expression, mouth, gaze, pose or body of x\_t[1]
     3. Replacement: where the content of x\_t is replaced with that of x\_s, preserving the identity of x\_s[1]
  2. Technology advancements through the development process
     1. Unpaired self-supervised training techniques to reduce the amount of initial training data[1] , avoid the requirement for extensive labeled training data[4]
     2. One/few -shot learning, which enables identity theft with a single profile picture[1,4]
     3. Improvements of face quality and identity through AdalN layers, disentanglement, and pix2pixHD network components[1,4]
     4. Fluid and realistic videos through temporal discriminators and optical flow prediction[1,4]
     5. The mitigation of boundary artifacts by using secondary networks to blend composites into seamless imagery[1,4]
     6. Use of perceptual loss on a pre-trained GG face recognition network boosts the facial quality[1,4]
     7. Use of a network pipeline gives more control over the final output[1,4]
  3. Challenges in creating realistic deepfakes[1,4]
     1. Generalization[1]

High-quality images of a specific identity requires a large number of samples of that identity[1]

* + 1. Paired Training[1]

One way to train a neural network is to present the desired output to the model for each given input, which is laborious and impractical sometimes.[1]

* + 1. Identity Leakage[1]

Sometimes the identity of the replacement is partially transferred[1]

* + 1. Occlusions[1,4]

Where part of the face is obstructed with a hand, hair, glasses or any other item, as a result, artifacts appear such as cropped imagery or inconsistent facial features.[1,4]

* + 1. Temporal Coherence[1,4]

Most deepfake networks process each frame individually with no context of the preceding frames[1,4]

* 1. Current limitations
     1. For reenactment, content is always driven and generated with a frontal pose, a good match is not always possible and has limited flexibility[1]
     2. Reenactments and replacements depend on the driver’s performance to deliver the identity’s personality[1]

II. Main Risks

1. Identity Theft[1]
2. Propagation/Spread of misinformation[1,3,4]
   1. Pollute Information Source
   2. Trigger severe complications in society[3]
   3. Misrepresentation, confusion and erosion of trust about a destination[3]
   4. Loss of credibility in state institutions, electronic media and others[4]
3. Authentication Challenge
   1. Face ID
   2. Fingerprint ID
   3. Threat to automated speaker verification systems[4] e.g. Voice ID
   4. Threat to voice-controlled systems deployed in the Internet of Things(IoT)[4]
4. Tampering of Evidence[1]
   1. Medical[1]
   2. Forensic[1]
   3. Court[1]
5. Scams & Fraud[1]
   1. Empower criminals to manipulate business dealings, even private phone calls
   2. Generating artwork[1]
   3. Financial Fraud[4]
      1. Falsify audit records[1]
      2. Insurance Fraud[1] : Inject/remove medical evidence in CT and MRI scans[1]
6. Physical Harm[1]
   1. False information
      1. E.g. Inject/remove medical evidence in CT and MRI scans[1]
7. Reputation Damage[3]/Harming Credibility[1]
   1. Revenge porn[1,4]
   2. Political sabotage via generated videos or articles[1]
   3. Defamation of public person[4]
   4. Impersonation of individuals[1]
8. Misdirection[1] of Public Opinion
   1. Generated discourse to amplify events/facts[1]
   2. Defamation of innocent individuals[1]
   3. Inflammation of popular sentiment[4]
9. Political Warfare[1] Affects Election Results
   1. Election Manipulation[4]
   2. Tone change of articles[1]
   3. Content loosely based on facts[1]
   4. Conspiracy[1]
   5. Impersonation of political leaders­[1]
10. Corruption[1]
    1. Increased xenophobia[1]
11. Potential Harm of Deepfake Pornography[2]
    1. Facilitate image-based sexual abuse[2]
       1. Non-consensual creation[2]
       2. Distribution[2]
       3. threat of distribution of explicit content[2]
    2. Revitalize “revenge pornography” and sextortion
12. Mental Damage
    1. Victimization is associated with Psychological distress[2]
       1. Diagnosis of moderate to severe depression[2]
       2. Anxiety disorder[2]
       3. Suicide[2]
13. Negative Influence (Related to sexual encounter)[2]
    1. Potential to shape sexual scripts[2]
    2. Potential to shape behaviours of its consumers[2]
    3. Impact individual’s expectation during a sexual encounter[2]
    4. Potential harm: social awareness of sexual coercion[2]
14. Significant Financial Gain Behind[2]
    1. Because of the huge demand (especially pornography), the supply for deepfake pornography/other is bound to catch up, this surge will leave public figures and private citizens vulnerable to victimization.[2]
15. National Security[4]
    1. Spark or justify a war[4]
    2. Deceive military analysts with fake information[4]
16. Blackmail of individuals for monetary benefits[4]
17. Lack of clear legal landscape[6]

III. Solution

1. Technological Level
   1. DeepFake Detection Technologies
      1. Detect deepfakes by searching for specific artifacts:
         1. spatial artifacts: inconsistencies, abnormalities in background, GAN fingerprints[4], in blending, environments and forensic[1]
         2. temporal artifacts: detecting variation in a person’s behavior, physiological signals, frame synchronization and coherence[1,4]

Graphical user interface, text

Description automatically generated

Description of classification categories [4]

* + 1. Technology companies such as Facebook, Twitter and Google are releasing database of deepfakes to improve identification[3]
    2. Available Audio/Video datasets Reference [4] – Page 35/53, Section 5.1]
    3. Other specific method Reference in [4]

A screenshot of a computer

Description automatically generated with medium confidence[1]

* + 1. Challenges in Deepfake detection models[4]
       1. Quality of deepfake datasets: the accessibility of large databases of deepfakes is an important factor in the generation of deepfake detection techniques[4]
       2. Performance evaluation: any binary classification scheme should be enhanced to multiclass/multi-label and utilize local classification/detection at the frame level[4]
       3. Lack of scalability for large scale platforms, such as social media. Problem of long inference time[4]
       4. Missing information and artifacts left during the generation process[4]
       5. Lack of audio deepfake detection data[4]
       6. Social media laundering, remove clues with respect to underlying forgeries[4]
  1. Data Provenance Technologies[1]
     1. Data provenance of multimedia should be tracked through distributed ledgers and blockchain networks[1]
     2. Content should be ranked by participants and AI[1]
     3. Contents should be authenticated and managed as a global system[1]
  2. Counter Attacks Technologies[1]
     1. Adversarial machine learning can be used to disrupt and corrupt deepfake networks. Can prevents deepfake technologies from locating a proper face.[1]
     2. Use adversarial noise to change the identity of the face so that web crawlers will not be able to find images to train model[1]

1. Industrial Level
   1. Related Products – previously mentioned technology related products
   2. Technology companies such as Facebook, Twitter and Google are imposing limited bans of deepfakes[3]
2. Political Level
   1. Law & Regulation

Suggestion of current laws to be expanded to include language specific to falsely created pornography without the explicit consent of all depicted persons[2]

* + 1. Criminal Charges[2]
       1. Canadian Criminal Code - Child pornography laws[2]
          1. protect against the depiction of any person under the age of 18 in pornography and specify that this includes material produced using electronics/mechanical means[2]
          2. production of deepfake videos with the face of minor and poster of such video would be prosecuted as an act of distribution child pornography[2]
       2. Canadian Law – protection against revenge pornography [Need Expansion][2]
          1. Stipulates that the publication of an intimate image without consent is reprehensible by law[2]
          2. The law covers sexual or nude images of a person that were produced using pyrographic, film or video recording[2]
          3. [Need Expansion]: Deepfake pornography, which is algorithmically synthesized, is open to interpretation[2]
       3. Legislation in place for extortion and fraud[2]
          1. Prosecute instances where fraudulent material [e.g. pornographic material] created with deepfake technology is used to blackmail or harass[2]
          2. Ontario Superior Court of Justice in 2017 recognize harassment as a free-standing and tenable cause of action[2]
    2. Seek justice Citing Internal Suffering[2]
       1. (IIMS) intentional infliction of mental suffering[2]
    3. Defamation Tort[2]
       1. Sullied reputation by publicization of the deepfaked material (e.g. pornographic videos under the defamation tort, would be entitled to seek damage awards[2]
    4. Misappropriation of Personality
       1. Legislation in place that prohibits the use of a person’s likeness for commercial purposes[2]
          1. Posting a deepfake pornographic video/fake generated other content that in se of a person’s personality/face/identity on a website that monetizes it in sole capacity, or monetizing traffic could represent a cause of action. This also put pressure on websites to be more active in betting materials[2]
    5. Copyright Infringement Laws
       1. [Need Expansion], whether deepfake (newly synthesized) count as an copy of given picture/video[2]
    6. Injunctive Relief
       1. Deepfake victim who has won their case may be eligible for injunctive relief and can request links to the video/material to be deleted from searches of their name[2]
       2. United States – Section 512 of the Digital Millennium Copyright Act(DMCA)[2]
          1. Host websites need to expeditiously remove or disable access to claimed material, otherwise liable for copyright infringement – invaluable tool[2]
    7. California has issued temporary and specific bans of political deepfakes from influencing an election[3]
    8. China has banned the creation, broadcast and use of deepfake
    9. General policies (International)
       1. Ill-prepared[2]
       2. Little legal regulation surrounding the creation and dissemination of deepfakes[6]
  1. Policy & Recommendation – too vague
  2. Challenges on Political Level
     1. Complete ban and criminalization of deepfake technology may not be warranted[2]
        1. There are beneficial applications of deepfakes[2]
        2. Financial gains behind[2]
     2. Challenges of cybercrime
        1. Individuals are able to conceal identity online with relative ease, e.g. mask one’s IP address[2]
        2. Victims and perpetrators may be separated geographically – under different jurisdiction[2]
        3. Process is with extended timescale: investigation, prosecution and conviction[2]
        4. Conviction does not guarantee the takedown of pornographic/deepfaked material from the Internet[2]

1. Societal/Media level
   1. Raise public awareness[2]
   2. Provide internet literacy[2]

a, b) i. Help individuals spot deepfakes material when they encounter it[2]

a, b) ii. Teach individuals to critically think about every piece of media they come

across online[2]

* 1. Educational Efforts for Bystanders[2]
     1. Better support for victims[2]

1. Academic/Research level (Look into Related Research paper (Survey is a plus))
   1. Examine the attitudes surrounding deepfake pornography/other crimes[2]
   2. Assess attitudes of police forces on TFSV (Technology-Facilitated Sexual Violence) and deepfake and other deepfake related crimes[2]
   3. Investigate potential impact of deepfakes on attitudes related to sexual norms and scripts, consent and sexual behaviours[2]

References

[1] Y. Mirsky and W. Lee, “The Creation and Detection of Deepfakes: A Survey,” *ACM computing surveys*, vol. 54, no. 1, pp. 1–41, 2021, doi: 10.1145/3425780.

[2] V. Karasavva and A. Noorbhai, “The Real Threat of Deepfake Pornography: A Review of Canadian Policy,” Cyberpsychology, behavior and social networking, vol. 24, no. 3, pp. 23–209, 2021, doi: 10.1089/cyber.2020.0272.

[3] A. O. J. Kwok and S. G. M. Koh, “Deepfake: a social construction of technology perspective,” Current issues in tourism, vol. 24, no. 13, pp. 1798–1802, 2021, doi: 10.1080/13683500.2020.173835

[4] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward,” Applied intelligence (Dordrecht, Netherlands), 2022, doi: 10.1007/s10489-022-03766-z.

[5] B. K. Wiederhold, “Can Deepfakes Improve Therapy?,” Cyberpsychology, behavior and social networking, vol. 24, no. 3, pp. 147–148, 2021, doi: 10.1089/cyber.2021.29209.editorial.

[6] J. Kietzmann, A. J. Mills, and K. Plangger, “Deepfakes: perspectives on the future ‘reality’ of advertising and branding,” International journal of advertising, vol. 40, no. 3, pp. 473–485, 2021, doi: 10.1080/02650487.2020.1834211.