Deep learning for deepfakes creation and detection: A survey REPORT

Submitted by – Amol Singh
231120003

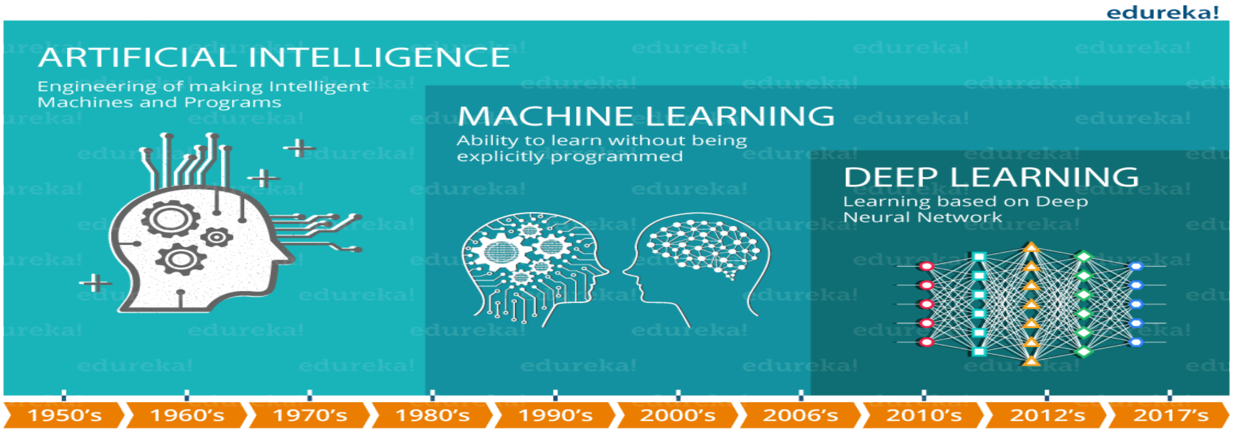
What do we aim to do/develop through this report?

Through this report we aim to do mainly 2 things-
1)Survey of algorithms used to **create deepfakes.**
2)Methods proposed to **detect deepfakes** in literature to date.

This study focuses on the critical issue of deepfakes, recognizing their significant challenge in today's landscape. By thoroughly reviewing the background of deepfake technology and examining state-of-the-art detection methods, the study offers a comprehensive understanding of deepfake techniques. This knowledge serves as a foundation for developing new and more robust methods to address the growing complexity of deepfakes, thereby contributing to the ongoing efforts to combat their detrimental impact.

What is Deep Learning ?

[1]Deep learning is a subset of machine learning that uses multi-layered neural networks, called deep neural networks, to simulate the complex decision-making power of the human brain. Some form of deep learning powers most of the artificial intelligence (AI) in our lives today.


edureka!
ARTIFICIAL INTELLIGENCE
Engineering of making Intelligent Machines and Programs
MACHINE LEARNING
Ability to learn without being explicitly programmed
DEEP LEARNING
Learning based on Deep Neural Network
1950's
1960's
1970's
1980's
1990's
2000's
2006's
2010's
2012's
2017's

What are exactly deepfakes ?

[2]Deepfakes, originating from "deep learning" and "fake," involve superimposing the face of a target person onto a video or image of a source person. This creates a deceptive video or image where the target person appears to be doing or saying things performed by the source person.

How to develop these deepfakes? In what circumstances are these developed?

Deepfakes can be **created in 2 ways**-
1)Traditional visual effects or computer-graphics approaches.
2)Deepfake can be created using deep learning models such as autoencoders and generative adversarial networks (GANs),

Deepfake creation has shifted towards **utilizing deep learning models**, a prevalent approach in the computer vision domain, although some deepfakes can still be **produced** using traditional methods.

->**Deepfake methods** normally require a large amount of image and video data to train models to create photo-realistic images and videos[2].

->As public figures such as **celebrities and politicians** may have a large number of videos and images available online, they are initial targets of deepfakes.

->The first deepfake video emerged in 2017 where the face of a **celebrity** was swapped to the **face of a porn actor.**

What are the pros and cons Of deepfake ?

[3]Deepfakes majorly have cons but we also cannot ignore the pros of it as well.

Cons of deepfakes-

1)Deepfake methods pose a threat to global security by allowing the creation of videos featuring world leaders delivering false speeches for deceptive purposes.

2)Deepfake technology could be employed to produce fabricated satellite images of the Earth, depicting non-existent objects like fake bridges, which could potentially mislead military analysts and troops in battle scenarios.

3)The recent release of DeepNude software highlights more disturbing threats, as it can transform individuals into non-consensual pornography.

4)The Chinese app Zao has gained viral popularity, allowing less-skilled users to swap their faces onto those of movie stars and insert themselves into famous movie and TV scenes.

Pros of deepfakes-

1)While the democratization of creating realistic digital humans has positive implications, deepfakes also have beneficial applications such as enhancing visual effects, creating digital avatars, developing Snapchat filters, restoring voices of individuals who have lost theirs, and updating movie episodes without reshooting them.

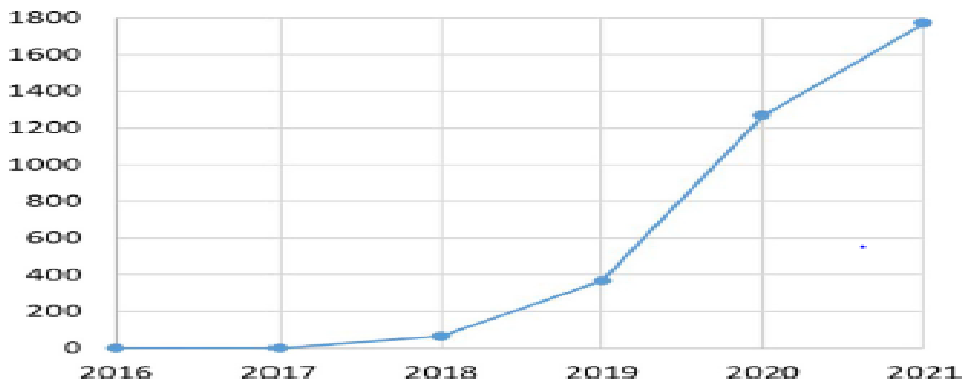
2)Deepfakes can positively impact various fields like photography, video games, virtual reality, movie productions, and entertainment. Examples include realistic video dubbing for foreign films, educational reanimation of historical figures, virtual clothes try-on in shopping, and more.

What are the initiatives taken to prevent these deepfakes ?

1)DARPA's Media Forensics (MediFor) program aims to expedite the development of detection methods for fake digital visual media.

2)Facebook, along with Microsoft and the Partnership on AI coalition, initiated the Deepfake Detection Challenge to stimulate research and development aimed at detecting and preventing the deceptive use of deepfakes.

3)Data from https://app.dimensions.ai at the end of 2021 reveals a significant increase in the number of deepfake papers in recent years, indicating a rising research trend in this area.



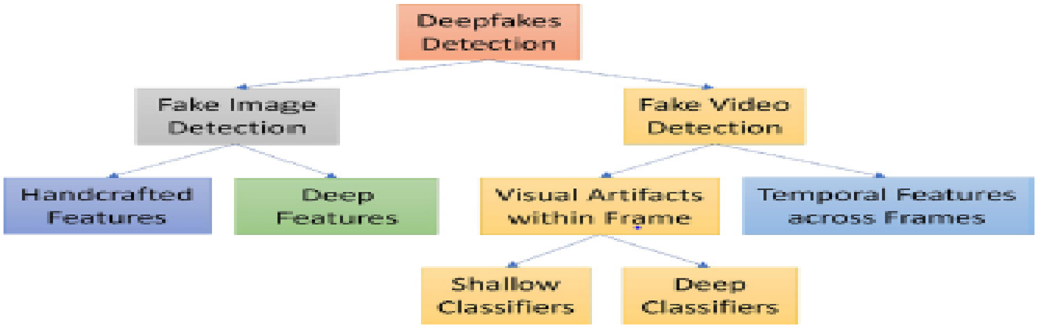
Fake Image and Video Detection

Fake image detection is divided into 2 parts-

1)Handcrafted features
2)Deep features.

Fake video detection is also divided into 2 parts-

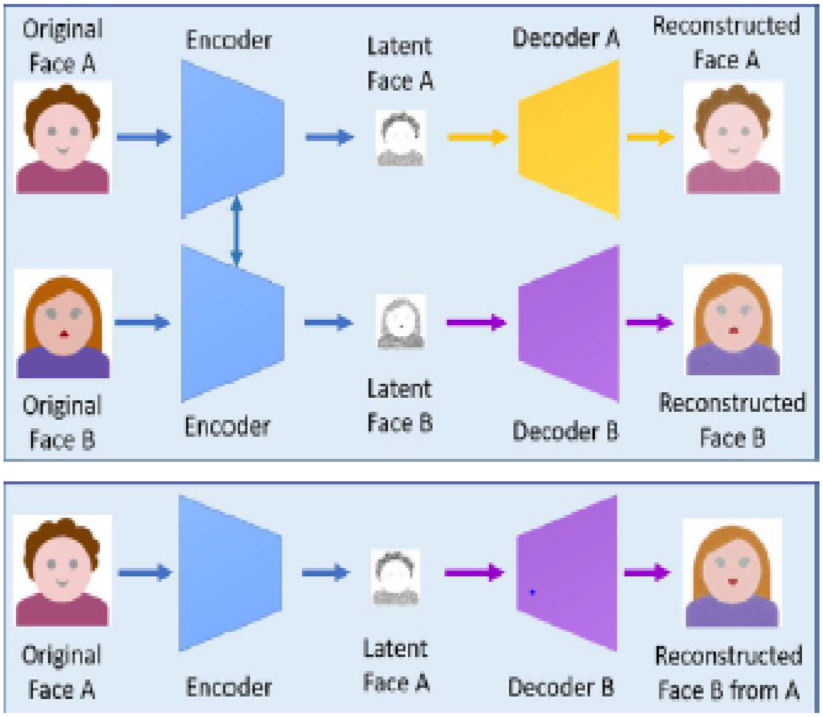
1)Temporal features across frames
2)Visual artifacts within a video frame.



How deepfakes are created?

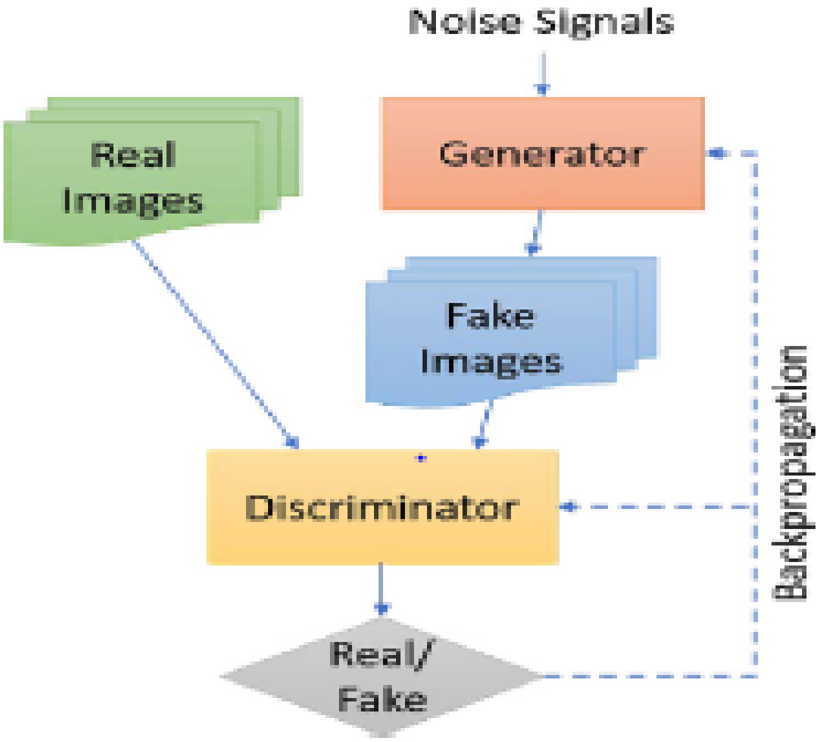
1)Autoencoder–Decoder pairing structure

In this method, autoencoders are employed to extract latent features from face images, with the decoder reconstructing the images. To swap faces between source and target images, two encoder-decoder pairs are utilized, with the encoder parameters shared between them. This enables the common encoder to identify similarities between the two sets of face images, facilitating the swapping process.Fig shows a deepfake creation process where the feature set of face A is connected with the decoder B to reconstruct face B from the original face A.



2)**Conventional GAN model**

In a conventional GAN model, there are two neural networks: a generator (G) and a discriminator (D). The generator aims to produce images similar to real ones from noise signals, while the discriminator's goal is to classify images as real or fake. They engage in a minimax game where G tries to minimize the probability of its outputs being classified as fake by D. The value function represents this game, where G and D improve their capabilities through training.Fig 4 tries to depict the same.



How deepfakes are detected?

[4]Deepfake detection is typically approached as a binary classification problem distinguishing between authentic and tampered videos. However, the effectiveness of existing methods is limited, as demonstrated by recent research. Korshunov and Marcel (2019) addressed this challenge by creating a deepfake dataset comprising 620 videos generated using Faceswap-GAN, based on the VidTIMIT database. Despite efforts, popular face recognition systems like VGG and Facenet struggle to detect deepfakes accurately. Additionally, lip-syncing methods and image quality metrics with SVM exhibit high error rates on this dataset. These findings underscore the urgent need for more robust deepfake detection methods.

Fake image detection

It can be done using 2 techniques namely-

1)Handcrafted features-based methods

Hand Crafted" features refer to properties derived using various algorithms using the information present in the image itself. Detecting images generated by Generative Adversarial Networks (GANs) presents a significant challenge due to their realistic appearance, often fooling traditional image forensics methods. Researchers have proposed several approaches to enhance detection models' generalization capability and effectiveness in identifying GAN-generated images. Xuan et al. (2019) introduced an image preprocessing step to remove specific clues from GAN-generated images, improving the generalization of forensic classifiers. Zhang et al. (2017) employed the bag-of-words method for feature extraction and classification to discriminate between manipulated and genuine images. Agarwal and Varshney (2019) framed GAN-based deepfake detection as a hypothesis testing problem, using a statistical framework based on information theory to measure the distance between distributions of legitimate and GAN-generated images. This approach becomes more effective as GAN accuracy decreases but requires extremely accurate GANs for high-resolution inputs.

2)Deep features-based methods

Deep features based methods are the methods to combat the proliferation of manipulated images and videos in the realm of deep learning techniques**.** Hsu et al. (2020) introduced a two-phase deep learning approach utilizing a Siamese network architecture for feature extraction and a neural network classifier for final detection, demonstrating superior performance across various GAN variants and datasets. Guo et al. (2021) developed SCnet, a CNN model designed specifically for detecting deepfake images generated by the Glow-based facial forgery tool, outperforming existing models like Meso-4 in terms of accuracy and generalization. Zhao et al. (2021) focused on detecting deepfakes by analyzing the self-consistency of local source features in images, leveraging CNN models and pairwise self-consistency learning to capture spatially-local information. These methods exemplify ongoing efforts to develop robust deepfake detection techniques using advanced deep learning architectures and feature extraction strategies to address the challenges posed by sophisticated manipulation technologies.

Fake video detection

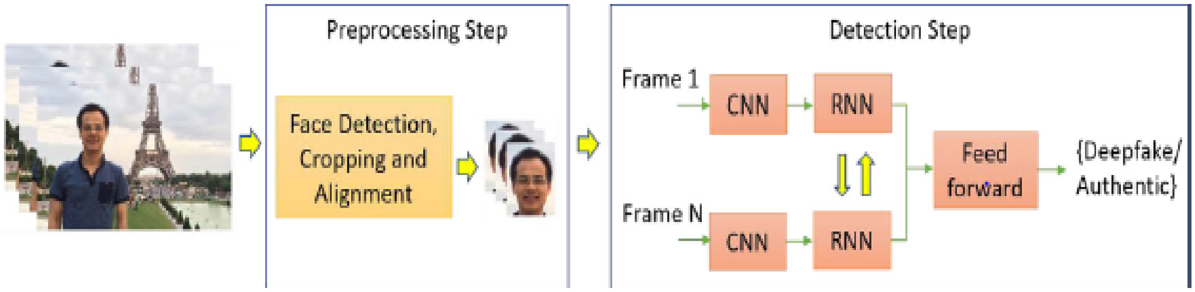
It can be done using 2 techniques namely-

1)Temporal Feature-Based Methods

Temporal features are derived from the signals in the **time domain**. Temporal = Time Example, Video consists of image frame sequence. With respect to time the frames are changed in video[5]. This is called Temporal Information.

Several innovative methods have been proposed for detecting deepfake videos by leveraging temporal features and inconsistencies across frames. Sabir et al. (2019) developed an RCN model integrating DenseNet and gated recurrent unit cells to exploit temporal discrepancies, achieving promising results on the FaceForensics++ dataset. Güera and Delp (2018) introduced a temporal-aware pipeline method using CNN and LSTM to detect deepfake videos based on intra-frame and temporal inconsistencies, achieving over 97% accuracy on a dataset of 600 videos. Li et al. (2018) proposed using eye blinking as a physiological signal to detect deepfakes, exploiting the lower blinking rates in deepfake videos compared to authentic ones. Caldelli et al. (2021) utilized optical flow fields to capture motion patterns and inconsistencies between synthetically created and authentic frames, achieving comparable performance with state-of-the-art methods on the FaceForensics++ dataset. These approaches demonstrate the effectiveness of leveraging temporal features and inconsistencies for deepfake detection.

A recurrent convolutional model (RCN) was developed by integrating DenseNet, a convolutional network introduced by Huang et al. (2017), and gated recurrent unit cells, proposed by Cho et al. (2014). This model aims to exploit temporal discrepancies across frames for deepfake detection.(Fig)



2)Visual artifacts within video frame

The term “artifact” is used to broadly describe defects or foreign, unwanted elements in a video. There can be any number of causes ranging from lossy compression, improper conversions, to post-processing adjustments like sharpening and resampling[6]. These are further classified into-

Deep Classifier Methods: In deep classifiers methods accuracy is used as the performance parameter.

> Artifact-Based Detection with CNNs (Li and Lyu, 2018): This method identifies deepfake videos by analyzing artifacts introduced during the face warping step of generation algorithms, utilizing CNN models like VGG16 and ResNet variants. Evaluation on datasets like UADFV and DeepfakeTIMIT demonstrates its effectiveness without the need for pre-generated deepfakes.

Capsule Networks (Nguyen et al., 2019): Capsule networks capture hierarchical pose relationships to detect manipulated images and videos, showing superior performance across various datasets compared to existing methods.

Shallow Classifier Methods: In shallow classifiers the results are discussed using accuracy, precision, and recall as performance parameters.

3D Head Pose-Based Detection (Yang et al., 2019): This method uses differences in 3D head poses estimated from facial landmarks, employing SVM classifiers for detection. Evaluation on datasets like UADFV demonstrates efficacy against competing approaches.

Visual Feature-Based Detection (Matern et al., 2019): Exploiting visual artifacts and texture features, this approach distinguishes real and fake videos using logistic regression and small neural networks, showing promising results on YouTube videos.

Discussions and future research directions

The proliferation of deepfakes, driven by advances in deep learning and social media dissemination, poses significant security threats. To combat this, the research community focuses on developing deepfake detection algorithms, amidst a growing battle between creators and detectors. Improving detection accuracy involves creating updated benchmark datasets and enhancing performance in cross-forgery scenarios. Adversarial attacks add complexity, necessitating robust methods. Integrating detection into social media platforms and legal actions against tech companies can help mitigate impacts. Understanding the social context of deepfakes is crucial, and in legal settings, documenting digital media forensics is essential. While AI algorithms are accurate, their lack of explainability is a challenge, emphasizing the need for explainable AI in forensic applications. Addressing these challenges requires ongoing interdisciplinary research and collaboration.

Possible methods to detect deepfakes

Detecting deepfakes poses a significant challenge due to the rapid advancements in generative models and the increasing sophistication of manipulation techniques. However, several promising approaches can be explored to enhance deepfake detection:

Visual Artifacts Analysis: Deepfake creation often introduces subtle but detectable visual artifacts, such as inconsistent facial expressions, unnatural eye movements, or irregular skin textures. Leveraging computer vision techniques, researchers can develop algorithms to automatically detect these anomalies. Techniques like image forensics, which analyze the statistical properties of images, can also be valuable in identifying manipulated content.

Audio-Visual Affective Cues: Emotions play a crucial role in human communication, and deepfake videos may struggle to accurately convey emotional cues. By analyzing both visual and audio components of a video and assessing the congruence of emotional expressions, researchers can develop detection methods based on discrepancies between expected and observed affective signals.

Explainable Deepfake Detection: Understanding the underlying techniques used to create deepfakes can aid in their detection. By employing explainable AI methods, such as attention mechanisms or saliency maps, researchers can highlight regions of an image or video that are indicative of manipulation. This not only improves detection accuracy but also provides valuable insights into the specific methods employed by deepfake creators.

Interdisciplinary Collaboration: Addressing the deepfake threat requires collaboration across diverse domains, including computer science, psychology, law enforcement, and policy-making. By integrating expertise from these fields, researchers can develop comprehensive detection frameworks that consider both technical and societal aspects of deepfake proliferation. For example, psychologists can contribute insights into human perception and behavior, while legal experts can inform strategies for regulation and enforcement.

Dataset Expansion and Benchmarking: Building large-scale, diverse datasets of deepfake videos is essential for training and evaluating detection algorithms. By continually expanding and refining benchmark datasets like Celeb-DF and FaceForensics, researchers can ensure that detection methods remain effective against evolving deepfake techniques. Additionally, crowdsourcing platforms can facilitate the collection of labeled data, enabling collaborative efforts in deepfake detection research.

Adversarial Training: Given the adversarial nature of deepfake creation, adversarial training techniques can be employed to improve detection robustness. By training detection models against adversarially generated deepfakes, researchers can enhance their resilience to manipulation attempts. Adversarial training can also help in understanding potential vulnerabilities in detection systems and devising countermeasures accordingly.

Awareness and Education Initiatives: Educating the public about the existence and implications of deepfakes is crucial for mitigating their impact. Awareness campaigns, educational resources, and media literacy programs can empower individuals to critically evaluate digital content and recognize potential signs of manipulation. By fostering a culture of skepticism and digital literacy, society can collectively combat the spread of deepfakes and mitigate their harmful effects.

In summary, detecting deepfakes requires a multi-faceted approach that integrates technical innovation, interdisciplinary collaboration, and societal awareness. By harnessing the collective expertise of researchers, practitioners, and policymakers, we can develop effective strategies to identify and mitigate the spread of synthetic media manipulation.

Conclusions

Deepfakes have become a significant threat, undermining trust in media content and posing various negative consequences such as distress to targets, exacerbation of disinformation and hate speech, and even incitement of political tensions or violence. With increasingly accessible technologies for creating deepfakes and the rapid dissemination of fake content through social media platforms, the need for effective detection methods is urgent. This survey offers a comprehensive overview of both deepfake creation and detection techniques, along with a discussion of the challenges, potential trends, and future directions in this field. It serves as a valuable resource for the artificial intelligence research community, guiding the development of robust methods to combat the spread of deepfakes and mitigate their harmful effects on society.

References

[1]What is Deep Learning? | IBM. (n.d.). https://www.ibm.com/topics/deep-learning
[2]Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh‐The, T., Nahavandi, S., Nguyên, T. T., Pham, Q., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525. https://doi.org/10.1016/j.cviu.2022.103525
[3]Springwise. (2020, October 8). Pros and Cons: Deepfake technology and AI avatars. https://www.springwise.com/pros-cons/deepfake-technology-ai-avatars/
[4]Kanojiya, P. P., Rai, R. R., & Asst. Prof. Gauri Ansurkar. (2024). Deepfake videos and images detection. In Journal of Emerging Technologies and Innovative Research (JETIR) (Vol. 11, Issue 1) [Journal-article]. https://www.jetir.org/papers/JETIR2401147.pdf
[5] What's the difference between spatial and temporal characterization in terms of image processing? (n.d.). Stack Overflow.
[6]Recognizing video artifacts. (n.d.). https://guide.encode.moe/encoding/video-artifacts.html