Literature Review - Deepfakes
Deep Learning Security - 720.01

Vaibhav Savala

April 14th 2023

1 Introduction

The use of deep learning techniques to generate videos of people that speak words and sentences they have not actually said, while synchronizing their lips and facial expressions to match the audio are called deep fakes. The process involves training a deep learning model on a large dataset of videos showing people speaking. The model learns to associate the audio signal with the corresponding lip movements and facial expressions, allowing it to generate new videos that align with the input audio signal. Deepfakes can enhance entertainment and creative expression by creating more realistic special effects and digital art. They can also improve medical training by providing realistic simulations, without endangering real patients.

The main challenges in generating such deep fakes comes in aligning audio and visual information so that the generated facial sequence is in sync with the input audio. Starting from aligning lip movement with the audio to aligning speech of the audio with the expressions and movement of the head, many techniques have been worked on improving audio and visual representations. Breger et al, 1997 firstly [1] used phoneme sequence to recreate lip movement to match a video track. This approach captures the dynamics by creating a database of video clips. For example, if a woman speaks out of one side of her mouth, this detail is recreated accurately.

Numerous studies[[2] [3], [4]] have used phonemes as a foundation to develop solutions for synchronizing lip movements with audio speech. Suwajanakorn et al. (2017)[4] proposed an approach for generating photorealistic mouth texture that preserves fine details in the lips and teeth and reproduces time-varying wrinkles and dimples around the mouth and chin. Chen et al. 2018[5] associated lip landmarks with audio and composited them into the mouth region. Recent techniques[[6],[7]] have used machine learning to separate visual input into identity and speech content space. These approaches rely on encoder-decoder techniques to enhance visual and audio representation. Some contemporary methods have proposed using facial landmarks and 3D models to capture facial dynamics. Song et al. 2018[8] fine-tuned their model on the target person and rendered a talking face video that learned independent lip features of the person in a 3D model. Additionally, researchers have developed techniques for capturing fine details of the mouth region, such as teeth, as in Zhang et al. 2021[9].

Detecting deepfakes is crucial as these videos have the potential to spread false information or propaganda, resulting in serious consequences. It is imperative to prevent such negative impacts and safeguard individuals, organizations, and society as a whole. Traditional image forensics techniques, as published in [10], were initially sufficient in detecting manipulations in images. However, with

the introduction of deep learning, these methods proved to be ineffective in detecting deepfakes. As a result, unimodal and multi-modal approaches were developed that employ multiple aspects of the face to address this issue. Li et al[11] utilized Deep Neural Network (DNN) to identify fake videos based on the artifacts observed during the face warping step of the generation algorithms.

Instead of solely focusing on the visual information of the mouth to synchronize with audio for deepfake generation, this literature review adopts a comprehensive approach that analyzes the entire face. The review of the literature on deepfake generation is separated into two parts. The first part covers research on synthesizing lip movements to align with the audio near the lip region. The second part encompasses studies that consider the entire face, including facial landmarks and expressions, to design models. The subsequent section examines various approaches developed for deepfake detection, providing a broad overview of the techniques utilized.

2 ML/DL background

Various techniques have been used to generate deep fakes. We will be discussing some of the techniques in the next few paragraphs.

Meta learner [12] is a neural network trained to generalize from a small number of examples per class to new tasks. It consists of a task-specific learner, a neural network trained on each task, and a separate meta-learner, trained to update the task-specific learner's weights based on a new task. The meta-learner improves performance by learning to generalize from the task-specific learner's weights.

LSTM (Long Short-Term Memory)[13] is a type of recurrent neural network that models long-term dependencies in sequential data, avoiding the problem of vanishing gradients. LSTM is used for lip sync by modeling the relationship between audio and visual features in a sequence, helping to generate realistic lip movements that match the spoken audio.

CNN(Convolutional Neural Networks) [14] are a type of neural network commonly used in computer vision applications. CNNs use convolutional layers to learn local patterns and spatial relationships between adjacent pixels in an image. CNNs are often used in Generative Adversarial Networks(GANs) to generate realistic images or videos.

Generative Adversarial Networks(GANs)[15] are deep learning models designed to produce synthetic data that is similar to a given dataset. The GANs consist of two neural networks, namely the generator and the discriminator, which work together in an adversarial way. While the generator produces fake data using random noise, the discriminator is responsible for distinguishing the real data from the fake data produced by the generator. With repeated training, the generator improves in generating realistic data that can trick the discriminator, thereby producing high-quality synthetic data.

Encoder-Decoder techniques [16] are commonly used in deep fake generation for talking heads. The process involves training a deep learning model that consists of an encoder and a decoder. The encoder takes an input image or video of a source face and encodes it into a low-dimensional feature vector. The decoder then takes this feature vector and decodes it to generate a new image or video of a target face.

As you will observe that many papers[[8],[17]] have used this technique for encoding images, encoding Mel Frequency Cepstral Coefficients (MFCC) representation of audio to create features for the approaches described.

3 Threat model

In recent times, there has been a surge in the generation of deepfakes, some with known origins while others with unknown origins. Knowing the origin and the intricacies behind the generation of deepfakes can be crucial in deepfake detection, as deepfakes can be used to cause security compromises. To construct a robust threat model, we need to discuss the goal, strategy, and knowledge of the attack.

According to Mirsky et al[18], deepfakes can be created in three types. Reenactment, Replacement, Editing and Synthesis.

Goal The attacker would like to create deepfakes which are as real as possible to the human eye. The goal of the deep fake detection model is to detect such fake videos. Deepfake detection models aim to differentiate between real and fake videos by considering visual and audio features that may indicate tampering or inconsistencies.

Strategy An attacker can make use of many aspects of the face to generate deepfakes. One can use audio-visual representation[19] to breakdown the real face video to generate deepfakes. Cross modality[5] techniques can be used to generate deepfakes. Face-swapping[20] can also be used to generate deepfakes.

Knowledge: Having knowledge of the target on whom deepfake will be generated can make the deepfake much more nuanced and detailed so as to mimic the target in a real fashion. This gives the attacker the ability to impersonate with much more precision.

Whitebox attack: When an attacker possesses comprehensive information about the target and can access several images or videos of the target, designing deepfakes[[4],[21]] with details that accurately reflect the intricacies of the face becomes easier. As a result of this increased sophistication, such attacks may more difficult to detect.

Blackbox attack: In a blackbox attack, an attacker has no knowledge or access to generate the deepfake. In this case, a general deepfake generation model can be used to generate deepfakes. There are some approaches[22] which employ the same approach.

can be used to identify inconsistencies between the real and fake videos. This approach can be useful when dealing with unknown deepfakes.

4 Related Work

Our related work will be divided into two parts. The first part will focus on reviewing approaches for generating deepfakes. Understanding the study of deep fake generation is crucial since it has the potential to be used for harmful purposes, like impersonating individuals in a harmful way. Therefore, conducting research into deep fake generation can aid in detecting and preventing such manipulations. The second part is about detecting deep fakes. We will be focusing on the prior works which consider the emotion aspect to detect deepfakes.

4.1 Deepfake generation

There have been numerous studies on deep fakes in terms of their development. Many of these studies focus on the technical aspects of creating deep fakes, such as improving the accuracy of lip-syncing and facial expressions, generating personalized speaking styles, and addressing 3D geometry errors. We explore some of the recent literature on deep fakes in the next few sections

4.1.1 Lip movement Deep Fakes

This category of methods is used to create videos where a person's lips appear to be moving in sync with words that they did not actually speak. This involves synthesizing video by analyzing the audio and visual information in the area around the mouth while utilizing compositing methods to incorporate stock footage of the head and torso for the rest of the body.

Suwajanakorn et al. 2017 [4] presented a method to generate lip-sync videos of Barack Obama by employing a sequence-to-sequence mapping approach that transforms audio features and generates mouth texture while synchronizing audio and video pauses. They also blended the mouth region with the target head to achieve a more realistic lip motion. This approach outperformed classic methods such as the Active Appearance Model)[23] and recent counterparts like Face2Face [20] in terms of lip sync clarity and smoothness. Chen and Li 2018.[5] proposed an approach that combines audio and image embedding to generate lip images that match audio speech. They used an identity framework to encode visual information, which was fused with audio information using duplications and concatenation. The model addressed the challenge of cross-modality by using audio-video correlation loss and three-stream adversarial learning loss, achieving significant improvement over existing methods on multiple datasets. This was the first research to consider speech-lip movement correlations for generating lip images. Song and Zhu et al. 2018 [8] presented a method to generate talking face videos with accurate lip-sync by integrating image and audio features. They used a conditional recurrent generation network with spatial-temporal and lip-reading discriminators. Audio features were extracted using MFCCs and fed into an auto-encoder, while image features were obtained from multiple frames and concatenated to generate a hybrid feature. Their approach included two discriminators to enhance lip movement accuracy and image quality. Park et al. 2022 [19] proposed SyncTalkFace, which uses an Audio-Lip memory to achieve fine-grained audio-visual coherence in lip-sync videos. They used an auto-encoder model, an audio-key memory, and a lip-value memory to generate high-quality videos that best match the ground truth. The loss function combines various losses to improve the visual and lip-sync quality of the generated videos. This approach outperformed all other state-of-the-art models in lip-sync, visual quality, and realness according to a human evaluation survey.

These above discussed approaches [[4],[8]] do not explicitly model emotions or predict the sentiment of the input speech, which can result in facial expressions that do not match the speech's tone. The mouth texture synthesis assumes that the mouth texture can be fully determined by positions of lip landmarks[4] [19], which may not be true for some sounds that require the use of the tongue.

4.1.2 Speech-driven talking head deep fakes

Speech-driven talking head deep fakes refers to using machine learning to generate a video of a person's head that appears to be speaking in sync with an audio clip. The model learns to synthesize a realistic-looking image of the person's head while also predicting the visual features that correspond to the audio input. This process involves the extraction of facial landmarks, the mapping of audio features to visual features, and the synthesis of realistic-looking images.

Various approaches to generate deepfakes using Generative Adversarial Networks(GAN) have been proposed. Zakharov et al. 2019 [6] suggests a system to generate personalized talking head models by meta-learning on a large dataset of videos using three networks: an embedder, a generator, and a discriminator. Chen et al. [17] proposed a temporal GAN with a multi-modal generator and a regression-based discriminator that evaluates generated videos using both sequence and

frame-level information. Das et al. 2020[24] proposed a strategy using two GAN networks to learn motion and texture separately for lip-sync. Chen et al, 2020[25] aims to generate talking heads with rhythmic head motion and proper lip sync. These approaches try to solve the problem of generating facial animation, head motions, and lip deformations.

Li et al,2021 [26] proposes a text-based talking head generation framework that synthesizes facial expressions and head motions according to the contextual sentiments and speech rhythm and pauses. The framework has speaker-independent stage and speaker-specific stage designed to capture generic relationships between texts and visual appearances. This generates holistic facial expressions compared to other works([27], [4], [28]).

Papantoniou et al. 2022 [21] propose a hybrid method called Neural Emotion Director (NED), which uses a parametric 3D face representation to manipulate the emotional state of actors in "in-the-wild" videos. The method translates the 3D representation to different domains and uses a video-based neural renderer to synthesize the target face. NED achieves photo-realistic manipulation of facial performances to any of the 6 basic emotions plus neutral, using only the semantic label as input, while retaining the original mouth motion. The method also enables the attachment of a specific style to the target actor, without requiring person-specific training, by extracting the reference style from any given video during testing.

With these approaches, there is a noticeable personality mismatch when using landmarks from a different person[6], making it necessary to adapt the landmarks to the specific individual in order to create more convincing deep fakes. The mapping[24] between audio and lip motion is not a one-to-one mapping. Different people have different lip motions for the same audio. This will lead to different final results compared to the ground truth. Also, There is an absence of nuances in human facial expressions and movements signifying emotions in these approaches except [21].

4.2 Deepfake Detection

The detection of deepfakes, which refers to manipulated or synthetic media created using deep learning algorithms, is a challenging task as the generated media can be quite realistic. For deepfake videos involving humans, several methods have been developed to aid in detection. Mirsky et al. (2021) [18] survey categorizes these techniques into two groups: those that rely on artifacts and those that utilize undirected approaches. The former approach involves the use of deep learning and machine learning models to identify specific artifacts that may not be perceivable by humans but can help detect deepfakes. In contrast, the latter technique does not have any predetermined features and instead relies on the deep learning model to identify relevant characteristics. We will focus on the artifacts in this literature.

Artifacts may occur when the generated content is blended back into the frame. Researchers have proposed several techniques to identify such artifacts for the observer's benefit, including quality measures [29] and frequency analysis [30]. Many times, the content of a fake face may seem anomalous compared to the rest of the frame, and residuals from face warping processes [11], changes in lighting, and variations in fidelity can provide clues to the presence of generated content.

We will be studying emotional aspect which involves analyzing the emotional content of an image or video to identify whether it is real or manipulated. Emotions are a critical component of human communication and expression, and therefore, deepfakes often lack the emotional nuance and complexity present in real media.

Mittal et al, 2020 [31] proposed a deepfake detection method that combines audio and video modalities with perceived emotion features extracted from both to detect falsification or alteration

in input videos. A Siamese network-based architecture is used to model these multimodal features and perceived emotions, offering a novel way to detect deepfakes by utilizing multiple modalities and perceived emotions simultaneously. With the DFDC[32] dataset, it gives an accuracy of 85% and 95% for DF-TIMIT[33] dataset.

Hosler et al [34] propose a deepfake detection method that focuses on detecting unnatural and inconsistent emotions conveyed by audio and video, utilizing the valence-arousal model of emotion. The approach estimates changes in valence and arousal over time by analyzing audio and facial LLDs, and inputting these into a supervised classifier to detect deepfakes by identifying emotional inconsistencies that may indicate falsification. This dataset gives an accuracy of 99.5% for the DFDC dataset.

5 Discussion

There has been considerable effort and success in generating deep fakes which are as real as possible to the human eye in parts. Though these papers have every time improved on the accuracy of best state-of-the-art approaches to generate deep fakes, many have limitations they do not consider like emotion modeling, and do not predict the sentiment of the input speech. Many studies [[8],[17]] consider MFCC features as input to the network for audio files. MFCC features are not very robust against noise signals as they can change if one frequency band is skewed. All of the papers mentioned above do not consider taking the raw audio file. Park et al. 2022 [19] involves the audio-lip memory to store lip motion features but struggles to synthesize the outputs with editable emotions.

It is also necessary to incorporate personalized speaking styles when generating deep fakes to ensure authenticity to someone familiar with the person being synthesized. While [4] proposed a solution for Barack Obama, this model may not work for other individuals, and generating such deep fakes using this technique would require a significant amount of data which may be difficult to find. Moreover, some models do not account for extreme poses, camera movements, light conditions, and audio noise.

Many frameworks[6] which have been discussed in the literature are not flexible enough to generate arbitrary length sequences, with much control over how the face moves and acts. Also, tongue modeling is not considered as well. There is not much work done to align the face as a whole to generate the deep fakes with expressions, face, head movements, and emotions.

There are only two papers [[25], [21]] that attempt to solve these problems as a whole. This approach synthesizes video and displays emotional full facial expressions, rhythmic head motions, and so on. But, this still has several limitations. One of them is to not consider emotion[25] in generated lip and head animations.

Despite the numerous approaches developed for detecting deepfakes, there are several limitations to their effectiveness. These limitations include a lack of transferability between datasets, making them less reliable in real-world scenarios. Additionally, deepfake detection[18] models are vulnerable to adversarial attacks that add noise or perturbations to the data, making detection less reliable. Furthermore, deepfake generation is rapidly evolving, leading to a concept drift scenario where detection models can become less effective as deepfake techniques and methods continue to advance.

6 Appendix

ChatGPT has been used for mostly to correct and rephrase sentences.

- Check for grammatical errors and correct any spelling mistakes.
- Make the paragraph in present continuous form.
- Define the term MFCC?
- What does this paragraph mean?

References

[1] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[2] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional lstm," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4884–4888.

[3] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7-8, pp. 857–876, 2013.

[4] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[5] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 520–535.

[6] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9459–9468.

[7] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9299–9306.

[8] Y. Song, J. Zhu, D. Li, X. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," *arXiv preprint arXiv:1804.04786*, 2018.

[9] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3661–3670.

[10] S. Battiato, O. Giudice, and A. Paratore, "Multimedia forensics: discovering the history of multimedia contents," in *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, 2016, pp. 5–16.

[11] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[12] J. Vanschoren, "Meta-learning: A survey," *arXiv preprint arXiv:1810.03548*, 2018.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[15] I. Goodfellow, "Nips 2016 tutorial: Generative adversarial networks," *arXiv preprint arXiv:1701.00160*, 2016.

[16] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[17] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss," *arXiv preprint arXiv:1905.03820*, 2019.

[18] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[19] S. J. Park, M. Kim, J. Hong, J. Choi, and Y. M. Ro, "Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2062–2070.

[20] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.

[21] F. P. Papantoniou, P. P. Filntisis, P. Maragos, and A. Roussos, "Neural emotion director: Speech-preserving semantic control of facial expressions in" in-the-wild" videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 781–18 790.

[22] C. Yang and S.-N. Lim, "One-shot domain adaptation for face generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5921–5930.

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[24] D. Das, S. Biswas, S. Sinha, and B. Bhowmick, "Speech-driven facial animation using cascaded gans for learning of motion and texture," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 408–424.

[25] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*. Springer, 2020, pp. 35–51.

[26] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan, "Write-a-speaker: Text-based emotional and rhythmic talking-head generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 1911–1920.

[27] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Transactions On Graphics (TOG)*, vol. 36, no. 4, pp. 1–11, 2017.

[28] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-based editing of talking-head video," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–14, 2019.

[29] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 659–665.

[30] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," *arXiv preprint arXiv:1911.00686*, 2019.

[31] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2823–2832.

[32] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[33] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[34] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm, "Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1013–1022.