International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

DETECTION OF DEEP FAKE VIDEOS USING LONG DISTANCE ATTENTION

[1]NANDIGUM YESHWANTH CHOWDARY,[2]MEKA MAITHILI,[3]GANGIREDDY AKHILA,[4]G LAKSHMI SAAHITHI,[5]MOHAMMAD HAZERA BEGUM

[1,2,3,4]Students, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100
[5]Associate Professor, Department of computer Science And Engineering, Malla Reddy Engineering College (Autonomous),Hyderabad Telangana, India 500100

ABSTRACT

With the rapid advancement of deepfake technologies, facial video forgeries have become increasingly realistic, posing serious security threats. Detecting such manipulations has become both critical and challenging. While most existing approaches frame deepfake detection as a basic binary classification task, this project addresses it as a fine-grained classification problem due to the subtle distinctions between real and forged faces. It is observed that deepfake videos often introduce artifacts in both spatial and temporal domains—such as generative anomalies within single frames and inconsistencies across consecutive frames. To tackle this, a novel spatial-temporal model is proposed, comprising two key components: one for detecting spatial artifacts within individual frames, and another for identifying temporal inconsistencies across frames. Both components are built using an innovative long-distance attention mechanism that generates attention maps in the form of patches. This mechanism enables a broader contextual understanding while preserving fine-grained local features, allowing the network to focus on the most discriminative facial regions. The proposed method enhances detection accuracy by effectively capturing global and local forgery traces.

Index Terms— Deepfake detection, face manipulation, attention mechanism, spatial and temporal artifacts.

INTRODUCTION

Deepfake videos are created by replacing one person's face with another's, leveraging advanced generative models to produce highly realistic content. The rise of face forgery applications has made it easier for anyone to create convincingly deceptive videos. Consequently, deepfake videos have proliferated across the Internet, posing significant risks, particularly in the context of spreading misinformation and inciting harm in society. The advent of high-quality deepfake videos that are indistinguishable to the human eye has attracted substantial attention from researchers, underscoring the urgent need for effective detection methods.

The process of generating deepfake videos typically involves dividing the video into individual frames, locating and cropping the face in each frame, and using a generative model to replace the original face with the target face. These manipulated faces are then reassembled into a video. During this process, two types of defects are introduced. First, the face generation process often introduces visual artifacts in the spatial domain due to the limitations of the generative model.

Second, inconsistencies between frames arise when assembling the video, as the generative model lacks global constraints across the frame sequence.

Several detection methods have been proposed based on spatial domain defects. Some of these methods focus on the semantic inconsistencies in deepfake videos, as generative models often fail to ensure global coherence in the generated faces. This can lead to abnormal facial features, such as misaligned facial parts, asymmetric faces, or mismatched eye colors. However, these methods are fragile, as their performance significantly drops when deepfake videos do not exhibit these specific semantic defects. Therefore, a more robust detection approach is crucial to counter the growing threat of deepfake technology.

LITERATURE SURVEY

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,

"Generative Adversarial Nets," in Advances in Neural Information Processing Systems, vol. 27, Montreal, CANADA, 2014.

We propose a new framework for estimating generative models via adversarial nets, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1/2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitatively evaluation of the generated samples.

T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in International Conference on Learning Representations, Vancouver, Canada, 2018

We describe a new training methodology for generative adversarial networks. The key idea is to grow both the generator and discriminator progressively: starting from a low resolution, we add new layers that model increasingly fine details as training progresses. This both speeds the training up and greatly stabilizes it, allowing us to produce images of unprecedented quality, e.g., CelebA images at 1024^2. We also propose a simple way to increase the variation in generated images, and achieve a record inception score of 8.80 in unsupervised CIFAR10. Additionally, we describe several implementation details that are important for discouraging unhealthy competition between the generator and discriminator. Finally, we suggest a new metric for evaluating GAN results, both in terms of image quality and variation. As an additional contribution, we construct a higher-quality version of the CelebA dataset.

Q. Duan and L. Zhang, "Look More Into Occlusion: Realistic Face Frontalization and Recognition With BoostGAN," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 214–228, 2021.

Many factors can affect face recognition, such as occlusion, pose, aging, and illumination. First and foremost are occlusion and large-pose problems, which may even lead to more than 10% accuracy degradation. Recently, generative adversarial net (GAN) and its variants have been proved to be effective in processing pose and occlusion. For the former, pose-invariant feature representation and face frontalization based on GAN models have been studied to solve the pose variation problem. For the latter, frontal face completion on occlusions based on GAN models have also been presented, which is much concerned with facial structure and realistic pixel details rather than identity preservation. However, synthesizing and recognizing the occluded but profile faces is still an understudied problem. Therefore, in this article, to address this problem, we contribute an efficient but effective solution on how to synthesize and recognize faces with large-pose variations and simultaneously corrupted regions (e.g., nose and eyes). Specifically, we propose a boosting GAN (BoostGAN) for occluded but profile face frontalization, deocclusion, and recognition, which has two aspects: 1) with the assumption that face occlusion is incomplete and partial, multiple images with patch occlusion are fed into our model for knowledge boosting, i.e., identity and texture information and 2) a new aggregation structure integrated with a deep encoder-decoder network for coarse face synthesis and a boosting network for fine face generation is carefully designed. Exhaustive experiments on benchmark data sets with regular and irregular occlusions demonstrate that the proposed model not only shows clear photorealistic images but also presents powerful recognition performance over state-of-the-art GAN models for occlusive but profile face recognition in both the controlled and uncontrolled environments. To the best of our knowledge, this article proposes to solve face synthesis and recognition under poses and occlusions for the first time.

F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in IEEE Winter Applications of Computer Vision Workshops, Waikoloa, USA, 2019, pp. 83–92.

High quality face editing in videos is a growing concern and spreads distrust in video content. However, upon closer examination, many face editing algorithms exhibit artifacts that resemble classical computer vision issues that stem from face tracking and editing. As a consequence, we wonder how difficult it is to expose artificial faces from current generators? To this end, we review current facial editing methods and several characteristic artifacts from their processing pipelines. We also show that relatively simple visual artifacts can be already quite effective in exposing such manipulations, including Deepfakes and Face2Face. Since the methods are based on visual features, they are easily explicable also to non-technical experts. The methods are easy to implement and offer capabilities for rapid adjustment to new manipulation types with little data available. Despite their simplicity, the

methods are able to achieve AUC values of up to 0.866.

D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a Compact Facial Video Forgery Detection Network," in IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 2018, pp. 1–7.

This paper presents a method to automatically and efficiently detect face tampering in videos, and particularly focuses on two recent techniques used to generate hyper-realistic forged videos: Deepfake and Face2Face. Traditional image forensics techniques are usually not well suited to videos due to the compression that strongly degrades the data. Thus, this paper follows a deep learning approach and presents two networks, both with a low number of layers to focus on the mesoscopic properties of images. We evaluate those fast networks on both an existing dataset and a dataset we have constituted from online videos. The tests demonstrate a very successful detection rate with more than 98% for Deepfake and 95% for Face2Face.

PROPOSED METHODOLOGY

Service Provider

In this module, the Service Provider logs in using a valid username and password. Upon successful login, they gain access to a variety of operations, such as browsing the dataset, viewing trained and tested results, examining predicted video detection types, analyzing the predicted video detection type ratio, and viewing all remote users. This module is designed to manage the service provider's administrative tasks related to video detection.

Train and Test Model

Here, the Service Provider splits the dataset into training and testing sets, with a 70% training and 30% testing ratio. The training data (70%) is used to train the model, while the testing data (30%) is used to evaluate the model's performance. This division ensures that the model is trained effectively and tested for accuracy.

Remote User

This module allows remote users to register before performing any actions. Once registered, their details are stored in the database. After successful registration, users log in using their authorized username and password. Upon logging in, users can predict video detection types and view their profiles. This module manages user registration and login processes while providing users with access to the system's key features.

Classification

In the Classification module, users input data that is then classified using pre-trained machine learning models. This step is crucial for identifying the type of video detection and ensuring the system functions properly by accurately categorizing the video content.

IMPLEMENTATION ALGORITHMS

Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning models that are primarily used for classification and regression tasks. In this system, the SVM algorithm builds a model that assigns new data points to one of two categories, making it a binary linear classifier. SVM is ideal for detecting deepfake videos

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

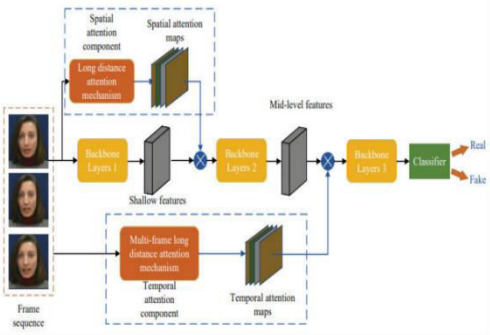
because of its ability to handle high-dimensional data effectively.

Recurrent Neural Network (RNN)

RNNs are a type of neural network where the output from previous steps is fed as input to the current step. This feedback mechanism allows RNNs to "remember" previous inputs, which is crucial when processing sequences, such as video frames or text in deepfake detection. The hidden state in an RNN acts as memory, maintaining information about past inputs to improve prediction accuracy. This is particularly useful for sequential data, like video analysis, where context from previous frames is important.

Gradient Boosting Classifier

Gradient Boosting is an ensemble learning technique that combines weak learners into a strong learner. The algorithm works by training new models to minimize the loss function (e.g., mean squared error or cross-entropy) of previous models. In each iteration, the gradient of the loss function is computed, and a new weak model is trained to reduce this gradient. The predictions of the new model are added to the ensemble, and the process continues until the stopping criterion is met. Gradient Boosting is effective in improving the model's performance and reducing overfitting, making it a powerful tool for deepfake detection.



CONCLUSION

This project addresses deepfake video detection through the lens of fine-grained classification, recognizing the subtle differences between real and fake faces. Given the generation defects of deepfake models in the spatial domain and the inconsistencies in the time domain, we propose a spatial-temporal attention model designed to focus the network on critical local regions. Additionally, a novel long-distance attention mechanism is introduced to capture global semantic inconsistencies present in deepfake videos. To enhance the extraction of texture and statistical information, the image is divided into smaller patches, with their importance recalibrated. Extensive experiments show that our method achieves state-of-the-art performance, demonstrating the effectiveness of the long-distance attention mechanism in guiding the model from a global perspective. Beyond the spatial-temporal model and attention mechanism, a key contribution of this work is the confirmation that not only focusing on pivotal areas is important but also incorporating global semantics. This approach offers a promising strategy for improving current models in deepfake detection.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in Advances in Neural Information Processing Systems, vol. 27, Montreal, CANADA, 2014.

[2] D. P. Kingma and M.Welling, "Auto-Encoding Variational Bayes," 2014.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in International Conference on Learning Representations, Vancouver, Canada, 2018.

[4] Q. Duan and L. Zhang, "Look More Into Occlusion: Realistic Face Frontalization and Recognition With BoostGAN," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 214–228, 2021.

[5]"deepfake," http://www.github.com/deepfakes/ Accessed September 18, 2019.

[6] "fakeapp," http://www.fakeapp.com/ Accessed February 20, 2020.

[7]"faceswap," http://www.github.com/MarekKowalski/ Accessed September 30, 2019.

[8] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in IEEE Winter Applications of Computer Vision Workshops, Waikoloa, USA, 2019, pp. 83–92.

[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a Compact Facial Video Forgery Detection Network," in IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 2018, pp. 1–7.

[10] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-Synthesized Faces Using Landmark Locations," in Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Paris, France, 2019, p. 113–118.

[11] D.-T. Dang-Nguyen, G. Boato, and F. G. De Natale, "Discrimination between computer generated and natural human faces based on asymmetry information," in Proceedings of the 20th European Signal Processing Conference, Bucharest, Romania, 2012, pp. 1234–1238.

[12] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Los Angeles,

USA, June 2019.

[13] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017, pp. 1831–1839.

[14] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, Vigo, Spain, 2016, pp. 5–10.

[15] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,

doi:10.1109/TPAMI.2020.3009287.