"DEEP FAKE VIDEO DETECTION WITH MACHINE LEARNING A LONG DISTANCE ATTENTION APPROACH"

Mrs Mehaboob Karishma, Assistant Professor, MCA, M.Tech (CSE), MCA Department, Sir C R Reddy College,

PG Courses, Eluru.

mehaboob.karishma@gmail.com

ABSTRACT

Deep Learning (DL) has revolutionized numerous fields, including healthcare, industry, and academia, where it has been applied to solve complex challenges such as thyroid diagnosis, lung nodule detection, advanced computer vision tasks, large-scale data analysis, and human-level control. Its adaptability and power have made DL indispensable in modern technology. However, as DL technologies evolve, a darker side has emerged. Deepfake techniques, which can manipulate or fabricate highly convincing video content, are posing severe security threats by enabling facial video forgery. The ability to generate such deceptive videos calls for urgent action in detecting these forgeries. In this paper, the problem is treated as a special fine-grained classification problem since the differences between fake and real faces are very subtle. It is observed that most existing face forgery methods left some common artifacts in the spatial domain and time domain, including generative defects in the spatial domain and inter-frame inconsistencies in the time domain. And a spatial-temporal model is proposed which has two components for capturing spatial and temporal forgery traces in global perspective respectively. The two components are designed using a novel long distance attention mechanism. The one component of the spatial domain is used to capture artifacts in a single frame, and the other component of the time domain is used to capture artifacts in consecutive frames. They generate attention maps in the form of patches. The attention method has a broader vision which contributes to better assembling global information and extracting local statistic information. Finally, the attention maps are used to guide the network to focus on pivotal parts of the face, just like other fine-grained classification methods. The experimental results on different public datasets demonstrate that the proposed method achieves the state-of the- art performance, and the proposed long distance attention method can effectively capture pivotal parts for face forgery.

1. INTRODUCTION

The deep fake videos are designed to replace the face of one person with another's. The advancement of generative models [1]–[4] makes deep fake videos become very realistic. In the meantime, the emergence of some face forgery application [5]–[7] enables everyone to produce highly deceptive forged videos. Now, the deep fake videos are flooding the Internet. In the internet era, such technology can be easily used to spread rumors and hatred, which brings great harm to society. Thus the high quality deep fake videos that cannot be distinguished by human eyes directly have aroused interest among researchers. An effective detection method is urgently needed. The general process of generating deep fake videos is shown in Fig. 1. Firstly, the video is divided into frames and the face in each frame is located and cropped. Then, the original face is converted into the target face by using a generative model and spliced into the corresponding frame. Finally, all frames are serialized to compose the deep fake video. In these processes, two kinds of defects are inevitably introduced. In the process of generating forged faces, the visual artifacts in the spatial domain are introduced by the imperfect generation model. In the process of combining frame sequences into videos, the inconsistencies between frames are caused by the lack of global constraints. Many detection methods are proposed [8]–[10] based on the defects in the spatial domain. Some of the methods take advantage of the defects of face semantics in deep fake videos, because the generative models lack global constraints in the process of fake face generation, which introduces some abnormal face parts and mismatched details in the face from a global perspective. For example, face parts with abnormal positions [10], asymmetric faces [11], and eyes with different colors [8]. However, it's fragile to rely entirely on these semantics. Once the deep fake videos do not contain the specific semantic defects that the method depends on, the performance will be significantly degraded. There are also some "deep" approaches [9], [12], [13], which attempt to excavate spatial defects according to the characteristics of the deep fake generators. However, compared with image contents, the forgery traces in the spatial domain are very weak, and the convolutional networks tend to extract image content features rather than the traces [14]. So blindly utilizing deep learning is not very effective in catching fake contents [15].Since the deep fake video is synthesized frame by frame, and there is no precise constraint between the frame sequences, the inconsistencies in the time domain will be introduced. Some methods exploit these defects of the time domain. The movements of eyes are exploited in [16]. Li et al. [17] use the human blink frequency in the video to detect the deep fake videos. The movement of lip [18] and the heart rate [19] are also exploited as the identification basis between authentic videos and deep fake videos in the time domain. The optical flows and the movement patterns of the real face and fake face are classified in [20] and [21], respectively. All of the methods mentioned above take the deep fake detection as a vanilla binary

classification problem. However, as the counterfeits become more and more realistic, the differences between real and fake ones will become more and more subtle and local which making such global feature-based vanilla solutions work not well [22].

2. LITERATURE SURVEY

1) Deepfake Detection with Spatio-Temporal Consistency and Attention

Abstract: Deepfake videos are causing growing concerns among communities due to their ever-increasing realism. Naturally, automated detection of forged Deepfake videos is attracting a proportional amount of interest of researchers. Current methods for detecting forged videos mainly rely on global frame features and under-utilize the spatio-temporal inconsistencies found in the manipulated videos. Moreover, they fail to attend to manipulation-specific subtle and well-localized pattern variations along both spatial and temporal dimensions. Addressing these gaps, we propose a neural Deepfake detector that focuses on the localized manipulative signatures of the forged videos at individual frame level as well as frame sequence level. Using a ResNet backbone, it strengthens the shallow frame-level feature learning with a spatial attention mechanism. The spatial stream of the model is further helped by fusing texture enhanced shallow features with the deeper features. Simultaneously, the model processes frame sequences with a distance attention mechanism that further allows fusion of temporal attention maps with the learned features at the deeper layers. The overall model is trained to detect forged content as a classifier. We test our method on two popular large data sets, consistently outperforming the related recent methods. Moreover, our technique also provides memory and computational advantages over the competitive techniques.

2) Generalizable Deepfake Detection with Phase-Based Motion Analysis

Abstract: We propose PhaseForensics, a DeepFake (DF) video detection method that leverages a phase-based motion representation of facial temporal dynamics. Existing methods relying on temporal inconsistencies for DF detection present many advantages over the typical frame-based methods. However, they still show limited cross-dataset generalization and robustness to common distortions. These shortcomings are partially due to error-prone motion estimation and landmark tracking, or the susceptibility of the pixel intensity-based features to spatial distortions and the cross-dataset domain shifts. Our key insight to overcome these issues is to leverage the temporal phase variations in the band-pass components of the Complex Steerable Pyramid on face sub-regions. This not only enables a robust estimate of the temporal dynamics in these regions, but is also less prone to cross-dataset variations. Furthermore, the band-pass filters used to compute the local per-frame phase form an effective defense against the perturbations commonly seen in gradient-based adversarial attacks. Overall, with Phase Forensics, we show improved distortion and adversarial robustness, and state-of-the-art cross-dataset generalization, with 91.2% video-level AUC on the challenging CelebDFv2 (a recent state-of-the-art compares at 86.9%).

3) Deepfake Video Detection Using Convolutional Vision Transformer

Abstract: The rapid advancement of deep learning models that can generate and synthesis hyper-realistic videos known as Deepfakes and their ease of access to the general public have raised concern from all concerned bodies to their possible malicious intent use. Deep learning techniques can now generate faces, swap faces between two subjects in a video, alter facial expressions, change gender, and alter facial features, to list a few. These powerful video manipulation methods have potential use in many fields. However, they also pose a looming threat to everyone if used for harmful purposes such as identity theft, phishing, and scam. In this work, we propose a Convolutional Vision Transformer for the detection of Deepfakes. The Convolutional Vision Transformer has two components: Convolutional Neural Network (CNN) and Vision Transformer (ViT). The CNN extracts learnable features while the ViT takes in the learned features as input and categorizes them using an attention mechanism. We trained our model on the DeepFake Detection Challenge Dataset (DFDC) and have achieved 91.5 percent accuracy, an AUC value of 0.91, and a loss value of 0.32. Our contribution is that we have added a CNN module to the ViT architecture and have achieved a competitive result on the DFDC dataset.

4) Local attention and long-distance interaction of rPPG for deepfake detection

Abstract: With the development of generative models, abused Deepfakes have aroused public concerns. As a defense mechanism, face forgery detection methods have been intensively studied. Remote photoplethysmography (rPPG) technology extract heartbeat signal from recorded videos by examining the subtle changes in skin color caused by cardiac activity. Since the face forgery process inevitably disrupts the periodic changes in facial color, rPPG signal proves to be a powerful biological indicator for Deepfake detection. Motivated by the key observation that rPPG signals produce unique rhythmic patterns in terms of different manipulation methods, we regard Deepfake detection also as a source detection task. The Multi-scale Spatial–Temporal PPG map is adopted to further exploit heartbeat signal from multiple facial regions. Moreover, to capture both spatial and temporal inconsistencies, we propose a two-stage network consisting of a Mask-Guided Local Attention module (MLA) to capture unique local patterns of PPG maps, and a Temporal Transformer to interact features of adjacent PPG maps in long distance. Abundant experiments on FaceForensics + + and Celeb-DF datasets prove the superiority of our method over all other

rPPG-based approaches. Visualization also demonstrates the effectiveness of the proposed method.

3. EXISTING SYSTEM

In the past few years, the performance of general image classification tasks has been significantly improved. From the amazing start of Alexnet [31] in Imagenet [32], the method based on deep learning almost dominate the Imagenet competition. However, for fine-grained object recognition [33]–[37], there are still great challenges. The main reason is that the two objects are almost the same from the global and apparent point of visual. Therefore, how to recognize the subtle differences in some key parts is a central theme for fine-grained recognition. Earlier works [38], [39] leverage human-annotated bounding box of key parts and achieve good results. But the disadvantage is that it needs expensive manual annotation, and the location of manual annotation is not always the best distinguishing area [40], [41], which completely depends on the cognitive level of the annotator. Since the key step of fine-grained classification is focusing on more discriminative local areas [42], many weakly supervised learning methods [23], [40], [43] have been proposed. Most of them use kinds of Convolutional attention mechanisms to find the pivotal parts for detection. Fu et al. [43] use a recurrent attention Convolutional neural network (RA-CNN) to learn discriminative region attention. Hu et al. [44] propose a channel-wise attention method to model interdependencies between channels. In [40], a multi-attention Convolutional neural network is adopted and more fine-grained features can be learned. Hu et al. [23] propose a weakly supervised data augmentation network using attention cropping and attention dropping. Deepfake detection and fine-grained classification are similar, that attempt to classify very similar things. Thus we learn from the experience in this field and leverage the attention maps generated with long range information to make the networks focus on pivotal regions.

DISADVANTAGES

- The spatial attention model is not designed to capture the artifacts that existed in the spatial domain with a single frame.
- The system not implemented Effectiveness of spatial-temporal model which leads the system less effective.

4. PROPOSED SYSTEM

The experience of the fine-grained classification field is introduced, and a novel long distance attention mechanism is proposed which can generate guidance by assembling global information.

• It confirms that the attention mechanism with a longer attention span is more effective for assembling global information and highlighting local regions. And in the process of generating attention maps, the non-convolution module is also feasible.

• A spatial-temporal model is proposed to capture the defects in the spatial domain and time domain, according to the characteristics of deepfake videos, the model adopts the long distance attention as the main mechanism to construct a multi-level semantic guidance. The experimental results show that it achieves the state-of-the-art performance.

ADVANTAGES

- High Accuracy to detect the deep fake videos

SYSTEM ARCHITECTURE

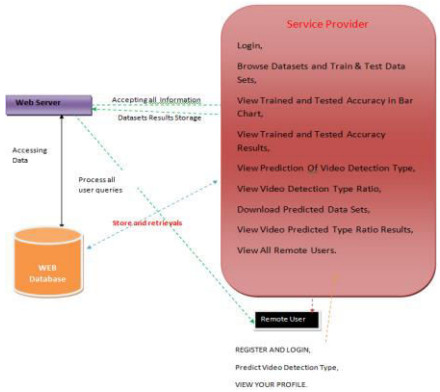

Fig 1: System Architecture

5. ALGORITHMS

5.1 DECISION TREE CLASSIFIERS

Decision tree classifiers are used successfully in many diverse areas. Their most important feature is the capability of capturing descriptive decision making knowledge from the supplied data. Decision tree can be generated from training sets. The procedure for such generation based on the set of objects (S), each belonging to one of the classes C1, C2, …, Ck is as follows:

Step 1. If all the objects in S belong to the same class, for example Ci, the decision tree for S consists of a leaf labeled with this class

Step 2. Otherwise, let T be some test with possible outcomes O1, O2,…, On. Each object in S has one outcome for T so the test partitions S into subsets S1, S2,… Sn where each object in Si has outcome Oi for T. T becomes the root of the decision tree and for each outcome Oi we build a subsidiary decision tree by invoking the same procedure recursively on the set Si.

5.2 LOGISTIC REGRESSION CLASSIFIERS

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar. Logistic regression competes with discriminate analysis as a method for analyzing categorical-

response variables. Many statisticians feel that logistic regression is more versatile and better suited for modeling most situations than is discriminate analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminate analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis.

5.3 SVM

In classification tasks a discriminate machine learning technique aims at finding, based on an independent and identically distributed (iid) training dataset, a discriminate function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, which require computations of conditional probability distributions, a discriminate classification function takes a data point x and assigns it to one of the different classes that are a part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminate approaches require fewer computational resources and less training data, especially for a multidimensional feature space and when only posterior probabilities are needed. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space. SVM is a discriminate technique, and, because it solves the convex optimization

problem analytically, it always returns the same optimal hyper plane parameter—in contrast to genetic algorithms (GAs) or perceptions, both of which are widely used for classification in machine learning. For perceptions, solutions are highly dependent on the initialization and termination criteria. For a specific kernel that transforms the data from the input space to the feature space, training returns uniquely defined SVM model parameters for a given training set, whereas the perception and GA classifier models are different each time training is initialized. The aim of GAs and perceptions is only to minimize error during training, which will translate into several hyper planes' meeting this requirement.

5.4 RANDOM FOREST

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output

of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. The first algorithm for random decision forests was created in 1995 by Tin Kam Ho[1] using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho[1] and later independently by Amit and Geman[13] in order to construct a collection of decision trees with controlled variance. Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

6. RESULTS

6.1 Output Screens

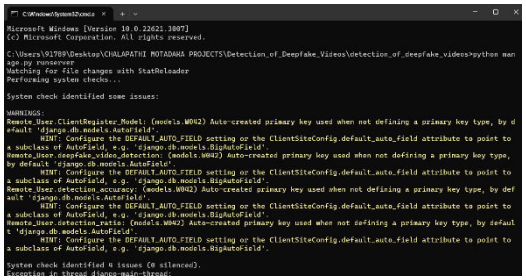

Fig 6.1 Run the manage.py file

In the above screen shows is the execution of manage.py file.

Fig 6.2 Prediction of Deep Fake Video

In above screen shows the prediction of deep fake video

Fig 6.3 Prediction Result

In above screen shows the prediction result of the deep fake video from the dataset.

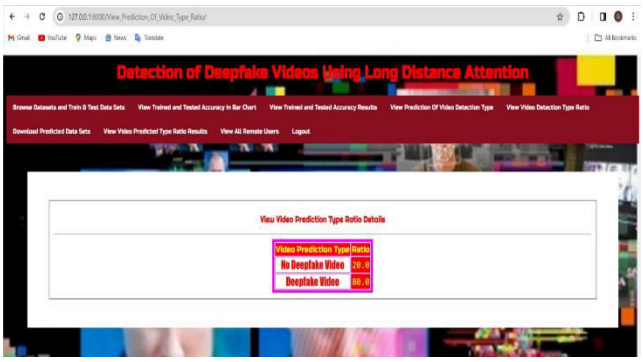

Fig 6.4 Prediction ratio of fake video or no fake video

In above screen shows the prediction ratio for the fake video or no fake video

7. CONCLUSION

In this paper, we detect deep fake video from the perspective of fine-grained classification since the difference between fake and real faces is very subtle. According to the generation defects of the deep fake generation model in the spatial domain and the inconsistencies in the time domain, a spatial temporal attention model is designed to make the network focus on the pivotal local regions. And a novel long distance attention mechanism is proposed to capture the global semantic inconsistency in deep fake. In order to better extract the texture information and statistical information of the image, we divide the image into small patches, and recalibrate the importance between them. Extensive experiments have been performed to demonstrate that our method achieves state-of the- art performance, showing that the proposed long distance attention mechanism is capable of generating guidance from a global perspective. Apart from the spatial-temporal model and the long distance attention mechanism, we think a main contribution of this paper is that we confirm not only focusing on pivotal areas is important, but combining global semantics is also critical. This is a noteworthy point, which can be a strategy to improve current models.

8. 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] Kiran Kumar Kommineni, Ratna Babu Pilli, K. Tejaswi, P. Venkata Siva, Attention-based Bayesian inferential imagery captioning maker, Materials Today: Proceedings, 2023, , ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2023.05.231.

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

[7] “faceswap,” http://www.github.com/MarekKowalski/ Accessed September30, 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] K. K. . Kommineni and A. . Prasad, “A Review on Privacy and Security Improvement Mechanisms in MANETs”, Int J Intell Syst Appl Eng, vol. 12, no. 2, pp. 90–99, Dec. 2023.

[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.
[16] K. K. Kumar, S. G. B. Kumar, S. G. R. Rao and S. S. J. Sydulu, "Safe and high secured ranked keyword searchover an outsourced cloud data," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 2017, pp. 20-25, doi: 10.1109/ICICI.2017.8365348.
[17] M. Li, B. Liu, Y. Hu, and Y. Wang, “Exposing Deepfake Videos by Tracking Eye Movements,” in 25th International Conference on Pattern Recognition, Milan, Italy, 2021, pp. 5184–5189.
[18] Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” in IEEE International Workshop on Information Forensics and Security, Hong Kong, China, 2018, pp. 1–7.
[19] C.-Z. Yang, J. Ma, S. Wang, and A. W.-C. Liew, “Preventing Deepfake Attacks on Speaker Authentication by Dynamic Lip Movement Analysis,” IEEE Transactions on Information Forensics and Security, vol. 16, pp. 1841–1854, 2021, doi:10.1109/TIFS.2020.3045937.
[20] S. Fernandes, S. Raj, E. Ortiz, I. Vintila, M. Salter, G. Urosevic, and S. Jha, “Predicting Heart Rate Variations of Deepfake Videos using Neural ODE,” in IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea (South), 2019, pp. 1721–1729..