EXPOSING DEEP FAKE VIDEOS THROUGH LONG RANGE ATTENTION

Sivasankar Chittoor[1], M Chanikya[2], Agepati Gayathri[2], Vallemma Lavanya[2], Kasthuri Anil Kumar [2], Abhishek Kumar Pal[2]

[1] *Professor, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*
[2] *Research Scholar, Department of Computer Science & Information Technology, Siddharth Institute of Engineering & Technology, Andhra Pradesh, India*

ABSTRACT

With the rapid progress of deepfake techniques in recent years, facial video forgery can generate highly deceptive video content and bring severe security threats. And detection of such forgery videos is much more urgent and challenging. Most existing detection methods treat the problem as a vanilla binary classification problem. In this article, 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 interframe inconsistencies in the time domain. And a spatial-temporal model is proposed which has two components for capturing spatial and temporal forgery traces from a global perspective, respectively. The two components are designed using a novel long-distance attention mechanism. 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.

.

Keyword*: Long, Distance, Traces, Patches.*

1. INTRODUCTION

In recent years, there has been a significant increase in the volume of financial transactions due to the expansion of financial institutions and the popularity of web-based e- commerce. Fraudulent transactions have become a growing problem in online banking, and fraud detection has always been challenging. Along with credit card development, the pattern of credit card fraud has always been updated. Fraudsters do their best to make it look legitimate, and credit card fraud has always been updated. Fraudsters do their best to make it look. legitimate. They try to learn how fraud detection systems work and continue to stimulate these systems, making fraud detection more complicated. Therefore, researchers are constantly trying to find new ways or improve the performance of the existing methods. People who commit fraud usually use security, control, and monitoring weaknesses in commercial applications to achieve their goals. However, technology can be a tool to combat fraud. To prevent further possible fraud, it is important to detect the fraud right away after its occurrence.

There are two mechanisms, fraud prevention and fraud detection, that can be exploited to avoid fraud-related losses. Fraud prevention is a proactive method that stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudster attempts a fraudulent transaction. Fraud detection in banking is considered a binary classification problem in which data is classified as legitimate or fraudulent. Because banking data is large in volume and with datasets containing a large amount of transaction data, manually reviewing and finding patterns for fraudulent transactions is either impossible or takes a long time. Therefore, machine learning-based algorithms play a pivotal role in fraud detection and prediction. Machine learning algorithms and high processing power increase the capability of

handling large datasets and fraud detection in a more efficient manner. Machine learning algorithms and deep learning also provide fast and efficient solutions to real-time problems.

2. RELATED WORKS

2.1 FINE-GRAINED CLASSIFICATION

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.

2.2 VISION TRANSFORMER

Transformer [45], a kind of self-attention architectures, is initially applied in natural language processing (NLP) and shows excellent performance. Its variant in the field of computer vision, Vision Transformer (ViT) [28], is first proposed by the Google team in 2020 and attracts a lot of attention. In vision, attention is usually used as a component of convolutional networks while keeping the overall structure. ViT shows that reliance on CNNs is not necessary. To apply the transformer to images directly, they firstly split the image into patches and project the patches to linear embedding. As a classification model, it generates a final discriminant vector through several stacking layers of self-attention modules. The self-attention modules are used to integrate the features of each patch with the self-attention mechanism. The self-attention mechanism is a stunning mechanism, which draws global dependencies and assembles global information.

It may be a promising candidate to deal with the detection of deepfake videos since the deepfake videos need to be considered from a global perspective and focused on the
critical regions. However, we find that it is not effective

3. ANALYSIS OF DEEP FAKE

The deepfake videos, generated by GANs [1] and VAEs [2], are formidably realistic and difficult for human eyes to discriminate. Since the differences between authentic videos and deepfake videos are subtle, detectors that blindly utilizing deep learning are not effective in catching fake content [15]. Similar problems have been studied in the field of fine-grained classification. A crucial experience is that using an attention mechanism to make the network focus on pivotal local regions can greatly improve the classification performance.
The generative models also have some inherent defects, which make deepfake detection possible. Whether it's GANs or VAEs, the generative networks will have an up-sampling process in the generation process to generate high-resolution images from latent coding [1], [2]. This allows the network to fill in details into the rough image. Deconvolution allows the model to draw a larger square from a point in the small graph. However, deconvolution is

prone to uneven overlap, especially when the kernel size cannot be divided by the step size. In theory, the neural network can learn the weight parameters carefully to avoid this kind of defect, but in fact, the neural network cannot completely avoid this kind of defect [46]. This overlapping style is reflected in two dimensions. The uneven overlapping multiplication of two coordinate axes results in the image block similar to chessboard [46], and resulting in a loss of facial texture details. Liu et al. [47] observe that the upsampling is a necessary step of most face forgery techniques and utilize phase spectrum to capture the up-sampling defects of face forgery. Since the up-sampling occurs between adjacent pixels, it is advantageous to capture the local information and collect statistics by using small blocks of appropriate size [48]. On the other hand, deepfake often generates abnormal face semantics. For example, unconvincing specular reflections in the eyes, either missing or represented as white blobs, or roughly modeled teeth, which appear as a single white blob [26]. The semantics and textures of the human face also appear in the form of the region [49]. Therefore, the processing of facial features in the form of patches is conducive to extracting local statistical information and capturing forgery traces. In the long distance attention, the input image is divided into manynon-overlapping small patches to collect local information.

However, some face semantics are normal from the local perspective but abnormal from the global perspective. That's because the GANs lack global constraints which introduce abnormal facial parts and mismatched details. It is observed that the density distributions of normalized face landmark locations on real and GAN-synthesized fake faces are different [10], because there is no coordination mechanism in the generation process of face components. This also leads to the asymmetry of the face [11]. In addition to the global structure as clues, the difference of details between facial components is also a key to the detection. For example, human eyes are always separated by a certain distance and have the same color, but the eyes of the fake face sometimes show different color [26]. An example of defects in the spatial domain is shown in Fig. 3. The first row reflects defects in a local region, and the next two rows reflect defects from a wider vision. It is also observed that biological signals are not coherently preserved in different synthetic facial parts [15].

In addition to the generative defects in the spatial domain, temporal defects are also existed in deepfake videos. In [17], the temporal inconsistencies are caught by the frequency of eye blinking. The inconsistency is also reflected in the face motion. The face motion patterns of real videos and deepfake videos have some differences, and can be used for classification [21]. Furthermore, there is a strong correlation between facial expression and head movement [25]. Changing the former without modifying the latter may expose a manipulation. It is also observed that temporal consistencies of human biological signals are not well preserved in GAN-erated content [15]. Thus, it is beneficial to modeling the continuity of face in the videos for deepfake detection.We exploit these inconsistencies in the time domain with a temporal model.

4. THE EXISTING SYSTEM

4.1 OVERVIEW

In Existing system, In the past few years, the performance of general image classification tasks has been significantly improved. From the amazing start of Alexnet in Imagenet, the method based on deep learning almost dominate the Imagenet competition. However, for fine- grained object recognition, 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 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, 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, many weakly supervised learning methods have been proposed. Most of them use kinds of convolutional attention mechanisms to find the pivotal parts for detection. Use a recurrent attention convolutional neural network (RA-CNN) to learn discriminative region attention. a channel-wise attention method to model interdependencies between channels. In a multi-attention convolution neural network is adopted and more fine-grained features can be learned. Propose a weakly supervised data augmentation network using attention cropping and attention dropping.

4.2 DISADVANTAGES OF EXISTING SYSTEM

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

5. THE PROPOSED SYSTEM

5.1 OVERVIEW

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.

5.2 ADVANTAGES OF PROPOSED METHODOLOGY

- Accuracy
- Adaptability
- Improved regulatory compilance

6. SYSTEM DESIGN

6.1 INTRODUCTION

It is a process of planning a new business system or replacing an existing system by defining its components or modules to satisfy the specific requirements. Before planning, you need to understand the old system thoroughly and determine how computers can best be used in order to operate efficiently.

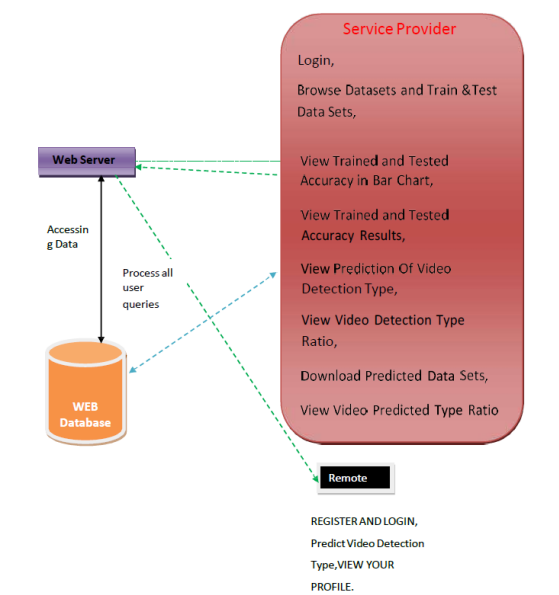

Fig -1 System Architecture

6.2 MODULES DESCRIPTION

- Service Provider:
  In this module, the functionalities are follows :
  1. Login
  2. Browse & Train & test Data set
  3. View trained & tested Accuracy in Bar chart
  4. View trained & Tested Accuracy Results
  5. View prediction of Video Detection type
  6. View Video Detection type ratio
  7. Download prediction data set
  8. View Video Prediction type ratio results

9. View all Remote users
10. Log Out

- Remote User:
  In this module, the functionalities are follows :
  1. Register
  2. Login
  3. Predict Video Detection type
  4. View Your Profile
  5. Logout.

7. RESULT ANALYSYS

7.1 USE CASE DIAGRAM FOR REMOTE USER

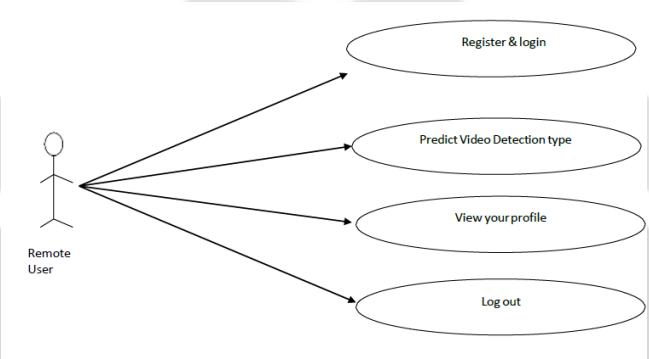

Fig-2 Use Case Diagram For Remote User

7.2 USE CASE DIAGRAM FOR SERVICE PROVIDER

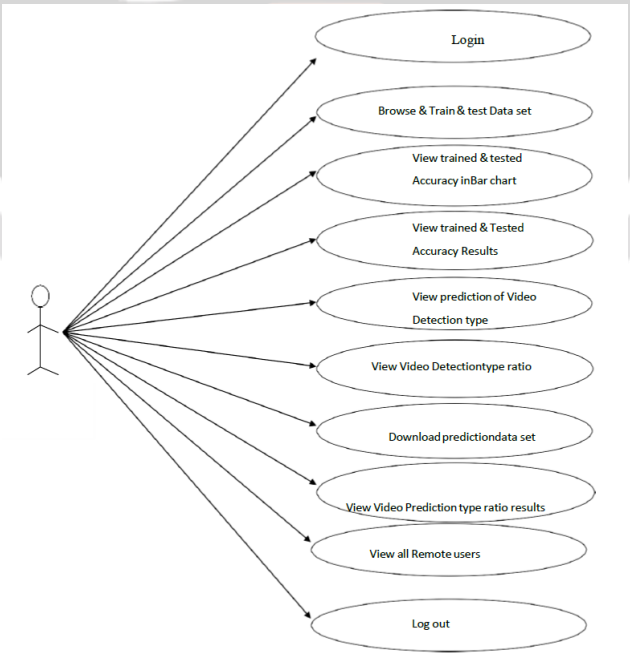

Fig -3 Use Case Diagram For Service Provider

7.3 EXECUTION PROCEDURE

The Execution procedure is as follows :
1. In this research work with data with attributes are observable and then all of them are floating data. And there's a decision class/class variable. This data was collected from Kaggle machine learning repository.
2.In this research 70% data use for train model and 30% data use for testing purpose. 3.RNN is used as Classifier .
4. In the classification report we were able to find out the desired result
5. In this analysis the result depends on some part of this research. However, which algorithm gives the best true positive, false positive, true negative, and false negative are the best algorithms in this analysis.

8. CONCLUSION

In this project, we detect deep fake video from the perspective of fine-grained classification sincethe 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.

9. FUTURE SCOPE

Develop more advanced models that can better capture long-range dependencies and subtle inconsistencies in Deepfake videos. Create larger and more diverse datasets to train detection models, including different types of Deepfake videos and more real videos for comparison.

11. 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 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] 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.