ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue.2 April 2024

DETECTION OF DEEPFAKE VIDEOS USING LONG DISTANCE ATTENTION

Aditya Singh [1] Ragalla Kalyan Ram [2],
Konduri Adarsh [3], Ms. V. Sumathi[4]
[1,2,3] UG Student, Dept. of ECE, CMR Institute of Technology, Hyderabad
[4] Assistant Professor, Dept. of ECE,
CMR Institute of Technology, Hyderabad

ABSTRACT

With the rapid progress of deepfake techniques in recent years, facial video forgery can generate highly deceptive video contents and bring severe security threats. And detection of such forgery videos is much more urgent and challenging. Mos t existing detection methods treat the problem as a vanilla binary classification problem. 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-ofthe-art performance, and the proposed long distance attention method can effectively capture pivotal parts for face forgery.

INTRODUCTION

The deepfake videos are designed to replace the face of one person with another's. The advancement of generative model s [1]–[4] makes deepfake videos

become very realistic. In the meantime, the emergence of some face forgery applications[5]–[7] enables everyone to produce highly deceptive forge d videos. Now, the deepfake 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 deepfake 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 deepfake 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 an d spliced into the corresponding frame. Finally, all frames are serialized to compose the deepfake 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 deepfake videos, because the generative models lack global constraints in th e 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 deepfake 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 deepfake 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 deepfake 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 deepfake videos. The movement of lip [18] and the heart rate [19] are also exploited as the identification basis between authentic videos and deepfake 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 deepfake 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]. Similar problems have been studied in the field of finegrained classification. Fine-grained classification aims to classify very similar categories, such as species of the bird, models of the car, and types of the aircraft [23]. Since the deepfake detection and fine-grained classification share the same spirit, that learning subtle and discriminative features, in [22], the deepfake detection is reformulated as a fine-grained classification task. And a convolutional attention module with 1 × 1 is adopted to make a network focus on the subtle but critical regions.

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 an adversarial process, 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 quantitative evaluation of the generated samples. The promise of deep learning is to discover rich, hierarchical

models [2] that represent probability distributions over the kinds of data encountered in artificial intelligence applications, such as natural images, audio waveforms containing speech, and symbols in natural language corpora. So far, the most striking successes in deep learning have involved discriminative models, usually those that map a high-dimensional, rich sensory input to a class label [14, 20].

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

How can we perform efficient inference and learning in directed probabilistic models, in the presence of continuous latent variables with intractable posterior distributions, and large datasets? We introduce a stochastic variational inference and learning algorithm that scales to large datasets and, under some mild differentiability conditions, even works in the intractable case. Our contributions are two-fold. First, we show that a reparameterization of the variational lower bound yields a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods. Second, we show that for i.i.d. datasets with continuous latent variables per datapoint, posterior inference can be made especially efficient by fitting an approximate inference model (also called a recognition model) to the intractable posterior using the proposed lower bound estimator. Theoretical advantages are reflected in experimental results. How can we perform efficient approximate inference and learning with directed probabilistic models whose continuous latent variables and/or parameters have intractable posterior distributions? The variational Bayesian (VB) approach involves the optimization of an approximation to the intractable posterior. Unfortunately, the common mean-field approach requires analytical solutions of expectations w.r.t. the approximate posterior, which are also intractable in the general case. We show how a reparameterization of the variational lower bound yields a simple differentiable unbiased estimator of the lower bound; this SGVB (Stochastic Gradient Variational Bayes) estimator can be used for efficient approximate posterior inference in almost any model with continuous latent variables and/or parameters, and is straightforward to optimize using standard stochastic gradient ascent techniques. For the case of an i.i.d. dataset and continuous latent variables per datapoint, we propose the AutoEncoding VB (AEVB) algorithm. In the AEVB algorithm we make inference and learning especially efficient by using the SGVB

estimator to optimize a recognition model that allows us to perform very efficient approximate posterior inference using simple ancestral sampling, which in turn allows us to efficiently learn the model parameters, without the need of expensive iterative inference schemes (such as MCMC) per datapoint. The learned approximate posterior inference model can also be used for a host of tasks such as recognition, denoising, representation and visualization purposes. When a neural network is used for the recognition model, we arrive at the variational auto-encoder.

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 10242 . 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. Generative methods that produce novel samples from high-dimensional data distributions, such as images, are finding widespread use, for example in speech synthesis (van den Oord et al., 2016a), image-to-image translation (Zhu et al., 2017; Liu et al., 2017; Wang et al., 2017), and image inpainting (Iizuka et al., 2017). Currently the most prominent approaches are autoregressive models (van den Oord et al., 2016b;c), variational autoencoders (VAE) (Kingma & Welling, 2014), and generative adversarial networks (GAN) (Goodfellow et al., 2014). Currently they all have significant strengths and weaknesses. Autoregressive models – such as PixelCNN – produce sharp images but are slow to evaluate and do not have a latent representation as they directly model the

conditional distribution over pixels, potentially limiting their applicability.

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.

Occlusion in facial photos poses a significant challenge for machine detection and recognition. Consequently, occluded face recognition for camera-captured images has emerged as a prominent and widely discussed topic in computer vision. The present standard face recognition methods have achieved remarkable performance in unoccluded face recognition but performed poorly when directly applied to occluded face datasets. The main reason lies in the absence of identity cues caused by occlusions. Therefore, a direct idea of recovering the occluded areas through an inpainting model has been proposed. However, existing inpainting models based on an encoder-decoder structure are limited in preserving inherent identity information. To solve the problem, we propose ID-Inpainter, an identity-guided face inpainting model, which preserves the identity information to the greatest extent through a more accurate identity sampling strategy and a GAN-like fusing network. We conduct recognition experiments on the occluded face photographs from the LFW, CFP-FP, and AgeDB-30 datasets, and the results indicate that our method achieves state-of-the-art performance in identity-preserving inpainting, and dramatically improves the accuracy of normal recognizers in occluded face recognition. In recent years, occluded face recognition has become a research hotspot in computer vision. Unlike unoccluded faces, occluded faces suffer from incomplete visual components and insufficient identity cues, which lead to degradation in recognition accuracy by normal recognizors [1,2,3,4]. Inspired by the recovery mechanism of the nervous system, researchers have proposed two types of approach, i.e., occlusion-robust and occlusion-recovery.

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

With the rapid progress of deepfake techniques in recent years, facial video forgery can generate highly deceptive video contents and bring severe security threats. And detection of such forgery videos is much more urgent and challenging. Mos t existing detection methods treat the

problem as a vanilla binary classification problem. 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-ofthe-art performance, and the proposed long distance attention method can effectively capture pivotal parts for face forgery. The deepfake videos are designed to replace the face of one person with another's.

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.

Disadvantages

ISSN2454-9940

www.ijasem.org

Vol 18, Issue.2 April 2024

- 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.

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

- In the proposed system, the motivation to use long distance attention is given first and then the proposed model is described briefly. As aforementioned, there is no precise global constraint in the deepfake generation model, which always introduces disharmony between local regions in the face forgery from a global perspective.

- In addition to the artifacts that exist in each forgery frame itself, there are also inconsistencies (e.g., unsmooth lip movement) between frame sequences because the deepfake videos are generated frame by frame. To capture these defects, a spatial-temporal model is proposed, which has two components for capturing spatial and temporal defects respectively. Each component has a novel long distance attention mechanism which can be used to assembling the global information to highlight local regions.

ISSN2454-9940

www.ijasem.org

Vol 18, Issue.2 April 2024

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.

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.

[9] Karne, R. K. ., & Sreeja, T. K. . (2023). PMLC- Predictions of Mobility and Transmission in a Lane-Based Cluster VANET Validated on Machine Learning. International Journal on Recent and Innovation Trends in Computing and Communication, 11(5s), 477–483. https://doi.org/10.17762/ijritcc.v11i5s.7109

[10] Radha Krishna Karne and Dr. T. K. Sreeja (2022), A Novel Approach for Dynamic Stable Clustering in VANET Using Deep Learning (LSTM) Model. IJEER 10(4), 1092-1098. DOI: 10.37391/IJEER.100454.

[11] Reddy, Kallem Niranjan, and Pappu Venkata Yasoda Jayasree. "Low Power Strain and Dimension Aware SRAM Cell Design Using a New Tunnel FET and Domino Independent Logic." International Journal of Intelligent Engineering & Systems 11, no. 4 (2018).

[12] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Design of a Dual Doping Less Double Gate Tfet and Its Material Optimization Analysis on a 6t Sram Cells."

[13] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Low power process, voltage, and temperature (PVT) variations aware improved tunnel FET on 6T SRAM cells." Sustainable Computing: Informatics and Systems 21 (2019): 143-153.

[14] Reddy, K. Niranjan, and P. V. Y. Jayasree. "Survey on improvement of PVT aware variations in tunnel FET on SRAM cells." In 2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), pp. 703-705. IEEE, 2017

[17] 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.

[18] 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.

[19] 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.