Research Statement: Trustworthy Machine Learning for Information Security and Multimedia "Deepfake" Forensics

Muhammad Ahmad Amin, eeahmadamin@mail.scut.edu.cn

Keywords: Trustworthy Machine Learning, Multimedia Forensics, Generative AI, Multimodel Learning

TL;DR: Contemporary multimedia forensic algorithms rely on the gradient ethos of traditional deep neural networks (DNNs). During my PhD, I developed machine learning (ML) tools to efficiently identify the deepfakes beyond the gradients but with higher-level semantic information like alternative color spaces, frequency statistics, content-aware multimodel learning, and an interpretable temporal coherence analysis mechanism. These insights enable a richer view of the multimedia forensics landscape, quantifying reality and manipulation. During my Postdoc, I want to explore the potential of higher-level information for building more trustworthy ML-based forensic methods and further improve and integrate information sources, like audio, image, and video.

Interested in similar topics? Let's chat!

No time to read on? Don't worry. I'll try to reach out to you ;)

(details below, feel free to skip)

1. Introduction -- Multimedia Forensic

The history of media manipulation traces back to the era of analog photography, yet the digital revolution ushered in unprecedented ease in altering visual content. While digital media offers immense technical advantages, it also amplifies the risks associated with misuse. Today, we witness entire movies synthesized through computer graphics, but alongside this progress, the proliferation of fake media poses a growing threat.

Distinguishing between authentic and synthesized images or videos is paramount, albeit increasingly challenging. Fake media, meticulously crafted to deceive the human eye, utilize sophisticated computational techniques like variational autoencoders (VAEs), generative adversarial neural networks (GANs), and diffusion models (DMs) for enhanced realism and accuracy surpassing human perception.

Emerging image and video manipulation techniques, empowered by deep neural networks, pose formidable challenges to traditional detection methods. By leveraging these networks, manipulators bypass manual editing, yielding results nearly indistinguishable from reality to the unaided eye.

Of particular concern is the phenomenon of face identity swap, one recent example [1], where original facial features in a video are replaced seamlessly with those of another individual while preserving audio, expressions, and movements—a technique commonly known as deepfake. Deepfakes, trained on extensive datasets of source individuals' videos, have garnered significant attention due to their potential ramifications for public stature, privacy, and protection.

Consider the implications in scenarios of geopolitical strain, where these technologies could fabricate statements from world leaders, inciting conflict. The urgency of this research is underscored by the imperative to develop robust mechanisms to detect and combat the proliferation of synthetic media, safeguarding the integrity of visual content and societal trust.

2. Research Specifications and Explainability

In line with generative techniques, several highly efficient approaches for identifying video forgeries rely on deep learning models. While these models offer unmatched accuracy, they often sacrifice explainability and trustworthiness, characteristics that are hallmarks of classic algorithms. The complexity of deep learning models sometimes obscures the precise reasons behind their predictions, leaving users uncertain about their reliability.

An essential quality of a machine learning classifier is its ability to provide correct predictions based on sound reasoning. To achieve this, it is crucial to delve deeper into understanding the rationale behind the model's predictions. This level of explainability is particularly critical in applications that demand a higher level of safety, such as industrial robots that operate alongside humans or autonomous vehicles.

In our particular scenario, offering justifications for why a video is categorized as a deepfake would not only foster greater confidence in the model but also streamline the manual review procedures. For instance, if law enforcement agencies need to use a fake detection model to assess a suspicious video, a straightforward "yes/no" outcome may not suffice as proof. However, giving a detailed explanation for why the video is deemed fake would significantly strengthen the case.

In the same way, more information about predictions could help people who manage big media websites to find and remove false and AI-generated information more easily. Imagine a scenario where trustworthy machine learning is applied to automated facial analysis, as discussed in Krishnan et al [2]. Facial analysis models are tasked with sensitive operations like face recognition and predicting attributes such as gender or age. In such cases, fairness and privacy become paramount concerns for users and contributors to the training data.

Regulators, including governmental agencies and ethics committees, are responsible for formulating and enforcing privacy and fairness requirements. Thus, it is imperative for machine learning regulation to consider these aspects to ensure the responsible and ethical use of AI technologies.

3. Related Work

In recent years, the fields of generative AI and detection have witnessed a surge in contributions. While certain approaches [3, 4, 5] empower models to autonomously detect and learn features within a supervised environment, utilizing methodologies primarily based on DNNs, some augmented with Recurrent Neural Networks (RNNs) and transformer-based models. These combined architectures leverage audio signals and visual information, encompassing spatial, frequency, and temporal data inherent in videos.

While some detectors for deepfakes provide a mask that identifies the areas in an image that are predicted to be fake, there are many detectors that do not offer clear explanations for the predictions they make. The challenge of elucidating the behavior of DNNs has been addressed in various works, adopting both white-box and black-box approaches. As far as I know, there appears to be a significant absence in the research when it comes to studying the concepts of reliability and transparency in the area of multimedia deepfake identification and generative AI.

4. Research Plan and Objectives

The objective of this proposed research is to explore the trustworthiness and explainability of multimedia forensic algorithms, with the following questions in focus:

- Can deepfake predictions be explained in an interpretable manner? We aim to investigate the feasibility of explaining deepfake predictions in a manner that is interpretable across various architectures of deep neural networks, thereby facilitating wide applicability.

- Do seemingly similar ML models employ distinct features to predict deepfakes? Given the multitude of architectures for Deep Neural Networks (DNNs) in detection, particularly in audio and video domains, understanding how architectural changes influence network reasoning is challenging. Our goal is to clarify the distinctive characteristics that each deepfake detector values for its prediction and identify why some models are more effective than others by utilizing explainers on various detectors.

- Is it as effective to use black box techniques as it is to use model-specific or by-design techniques for generating explanations? We seek to evaluate the efficacy of black box explanations in comparison to other explanation types, such as white box or model-specific ones. By comparing graphical explanations produced by detection models with black box alternatives, we aim to determine if black box methods can achieve similar effectiveness.

- Can explanation techniques be applied to video inputs, leveraging temporal and audio information? Given that some forensic deepfake detectors operate at the video level and incorporate temporal and audio data, we intend to explore the potential of utilizing this information encoded in videos to derive more coherent explanations.

- How can this newfound knowledge be leveraged to enhance models or datasets? Insights from explanations offer valuable guidance for model improvement. We will explore indirect utilization through model refinement and direct integration into the training process.

- What level of detail and intuitiveness can we incorporate into our explanations without compromising their completeness? We recognize the trade-off between the completeness and intuitiveness of explanations. Striking an appropriate balance between these attributes is crucial, considering the complexity of deep learning models and the need for comprehensible explanations. Throughout our research, we will aim to achieve a suitable compromise between these properties.

References

[1] Finance worker pays out $25 million after video call with deepfake ‘chief financial officer, 2024.
(https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk)

[2] Anoop Krishnan, Ali Almadan, and Ajita Rattani. “Understanding Fairness of Gender Classification Algorithms Across Gender-Race Groups”. In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2020, pp. 1028–1035. doi: 10.1109/ICMLA51294.2020.00167.

[3] Y. Yu, R. Ni, Y. Zhao, S. Yang, F. Xia, N. Jiang and G. Zhao, MSVT: Multiple spatiotemporal views transformer for DeepFake video detection, 33, 4462–4471, URL https://ieeexplore.ieee.org/document/10138555 , *IEEE Transactions on Circuits and Systems for Video Technology*.

[4] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang and L. Nie, Voice-face homogeneity tells deepfake, 20 , 76:1–76:22, URL https://doi.org/10.1145/3625231 .

[5] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao and K. Ren, AVoiD-DF: Audio35 visual joint learning for detecting deepfake, 18, 2015–2029, URL https://ieeexplore.ieee.org/document/10081373 , *IEEE Transactions on Information Forensics and Security*.