SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Understanding The State of the Art of Publicly-Available Deepfake Detection Tools

Berdiguly Yaylymov

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Understanding The State of the Art of Publicly-Available Deepfake Detection Tools

Der Stand der Technik bei der Erkennung von Deepfakes durch öffentlich zugängliche Tools

Author:	Berdiguly Yaylymov
Supervisor:	Prof. Dr. Jens Großklags
Advisor:	M.A. Severin Engelmann
Submission Date:	15.08.2023

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Munich, 15.08.2023 Berdiguly Yaylymov

Acknowledgments

Abstract

Deepfake technology, a fusion of deep learning and fake media, has rapidly evolved and become a powerful tool for generating highly realistic synthetic content. This advancement brings with it significant challenges in media authentication, cybersecurity, and privacy. As deepfakes become more sophisticated and accessible, the need for effective detection tools has become paramount. This thesis aims to provide a comprehensive understanding of the state of the art of publicly-available deepfake detection tools.

The study begins with a literature review that explores the evolution of deepfake technology, the various methods used for deepfake generation, and the existing approaches for deepfake detection. By analyzing the strengths and limitations of these techniques, this study sets the foundation for evaluating the effectiveness of publicly-available deepfake detection tools.

A robust methodology is employed to collect and analyze data on the available tools. The evaluation criteria include accuracy, efficiency, scalability, versatility, and user-friendliness. The selected deepfake detection tools, encompassing open-source projects, commercial offerings, and academic research projects, are assessed in detail to provide insights into their features, capabilities, and performance.

The findings of this study reveal the strengths and weaknesses of the evaluated deepfake detection tools. Comparative analysis sheds light on their distinctive characteristics and effectiveness in detecting deepfakes across different media types. Additionally, the study identifies gaps and challenges within the current landscape of deepfake detection, offering recommendations for future research, development, and policy-making.

The implications of this research extend to a wide range of domains, including media forensics, journalism, law enforcement, and online platforms. The ability to distinguish between genuine and manipulated content is crucial for safeguarding information integrity, maintaining trust, and combating disinformation campaigns. The insights provided by this thesis contribute to the ongoing efforts to develop effective deepfake detection mechanisms that keep pace with the evolving landscape of deepfake technology.

In conclusion, this thesis provides a comprehensive overview of publicly-available deepfake detection tools, offering an in-depth evaluation and comparison of their features and capabilities. The study highlights the urgent need for ongoing research

and development in the field of deepfake detection to counter the growing threat posed by synthetic media manipulation. By promoting a deeper understanding of the state of the art in deepfake detection, this research aims to contribute to the advancement of techniques and policies that can effectively mitigate the risks associated with deepfakes and uphold the integrity of digital media.

Contents

1 Introduction

The rapid and continuous development of Artificial Intelligence (AI) has given birth to numerous applications that have pushed the boundaries of what we previously believed to be possible. This thesis will delve into one of the most fascinating and alarming developments in this field, deepfakes. This document seeks to provide an exhaustive review of the current state of the art in publicly-available deepfake detection tools.

1.1 Background and Motivation

In an era where digital media forms the cornerstone of communication, the advent of deepfakes, AI-enabled synthetic media, poses an unprecedented challenge to information integrity. Deepfakes, a portmanteau of 'deep learning' and 'fake', is a technology that manipulates or fabricates audio-visual content to make it appear real, often indistinguishable from the original.

The proliferation of deepfake technology was initially sparked by its application in creating misleading celebrity images and videos, before quickly expanding into other sectors. One of the earliest examples that drew significant attention to deepfakes was a video created by an anonymous Reddit user called 'deepfakes' in late 2017. This user began to post digitally altered pornographic videos, realistically swapping the faces of actresses onto the bodies of porn stars. However, it wasn't long before the technology was used outside of pornographic content.

A notable instance that clearly demonstrated the power of deepfakes, and arguably brought it to mainstream attention, was a video of former U.S. President Barack Obama, released in April 2018 by Buzzfeed and Jordan Peele [1], [2]. The video features a deepfake of Obama saying things he never actually said, with Peele providing the voiceover. This deepfake video, viewed by millions, effectively highlighted the potential misuse of this technology in spreading misinformation and propaganda.

In recent years, the sophistication of deepfake technology has reached an unprecedented level. A perfect example of this progression can be seen in the creation of 'Tom Cruise deepfakes' that circulated on social media in early 2021. The videos, created by Belgian visual effects artist Chris Ume in collaboration with actor Miles Fisher, who impersonated Cruise's voice and mannerisms, were shared on TikTok under the

account name @deeptomcruise. These deepfake videos show the synthetic 'Tom Cruise' doing various activities - performing a magic trick, playing golf, or simply telling a story about Mikhail Gorbachev.

The 'Tom Cruise deepfakes' took the internet by storm due to their uncanny resemblance to the real actor, in terms of both appearance and behavior. Unlike the early deepfake videos, which often exhibited glaring imperfections, these deepfakes were so convincing that many viewers initially believed they were watching the actual Tom Cruise. This level of realism underscored the strides made in deepfake technology, while simultaneously highlighting the potential dangers of its misuse.

Driven by advances in machine learning, especially deep learning, deepfake technology has grown significantly in sophistication and accessibility. The potential applications of deepfakes range from benign, such as in film production and entertainment, to malicious uses, including disinformation campaigns, identity theft, and deepfake pornography. As these applications become more widespread, deepfake technology has raised profound questions and challenges for society, especially regarding media authenticity, privacy, and cybersecurity.

However, it is not just the creation of deepfakes that has improved; strides have also been made in detection. There are now more sophisticated, AI-powered tools that can analyze videos and images for signs of manipulation. These tools operate on multiple levels, from detecting inconsistencies in lighting and shadows to looking for signs of digital artifacts and abnormal facial movements. But as detection tools become more sophisticated, so too do the techniques used to create deepfakes. This constantly evolving technological arms race underscores the critical need for ongoing research and development in deepfake detection.

In response to these challenges, there is an increasing need for robust and reliable deepfake detection tools. However, despite the flurry of research and development in this area, a comprehensive understanding and evaluation of the available detection tools remain elusive. This knowledge gap not only impedes the technological advancements in deepfake detection but also complicates the task of policy-making and regulation in this sphere.

This thesis is motivated by the need to bridge this gap and advance our understanding of publicly-available deepfake detection tools. By examining these tools, this study aims to contribute to the ongoing efforts to mitigate the risks associated with deepfakes and uphold the integrity of digital media.

1.2 Objectives of the Study

1.3 Scope and Limitations

1.4 Thesis Structure

2 Literature Review

2.1 History of Deepfakes

2.2 Techniques Used in Deepfakes

2.3 Publicly Available Deepfake Tools

2.4 Ethical and Legal Concerns

2.5 Existing Countermeasures and Detection Methods

3 Methodology

3.1 Research Design

3.2 Selection Criteria for Publicly-Available Tools

3.3 Evaluation Metrics

3.4 Datasets

3.5 Overview of Selected Publicly-Available Deepfake Tools

4 Analysis of Publicly-Available Deepfake Tools

4.1 Sensity

4.2 FaceForensics++

4.3 FaceSwap

4.4 XceptionNet

4.5 Comparative Analysis

5 Case Studies

5.1 Entertainment and Art

5.2 Politics and Media

5.3 Cybersecurity and Privacy

5.4 Deepfake Generation Tools

6 Results

6.1 Dataset Augmentations

6.2 Frequency Analysis

6.3 Final Results

7 Discussion and Recommendations

7.1 Effectiveness and Accessibility of Publicly-Available Tools

7.2 Potential Future Developments

7.3 Recommendations for Policy Makers and Researchers

8 Conclusion

8.1 Summary of Findings

8.2 Future Researcher Directions

9 Test

9.1 Section

Acronyms must be added in `main.tex` and are referenced using macros. The first occurrence is automatically replaced with the long version of the acronym, while all subsequent usages use the abbreviation.

E.g. `\ac{TUM}, \ac{TUM}` ⇒ Technical University of Munich (TUM), TUM

For more details, see the documentation of the `acronym` package[1].

9.1.1 Subsection

See Table 9.1, Figure 9.1, Figure 9.2, Figure 9.3.

Table 9.1: An example for a simple table.

A	B	C	D
1	2	1	2
2	3	2	3

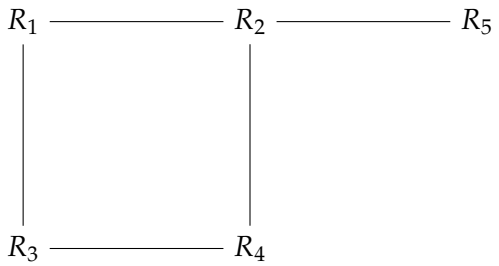

Figure 9.1: An example for a simple drawing.

[1] `https://ctan.org/pkg/acronym`

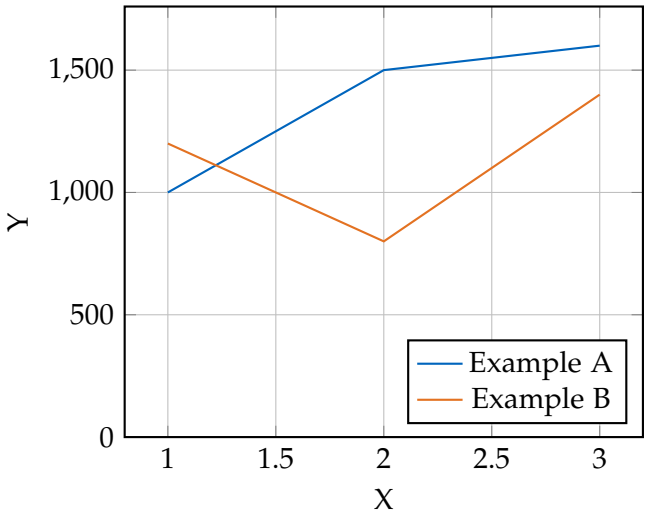

Figure 9.2: An example for a simple plot.

```
SELECT * FROM tbl WHERE tbl.str = "str"
```

Figure 9.3: An example for a source code listing.

Abbreviations

TUM Technical University of Munich

AI Artificial Intelligence

List of Figures

List of Tables

Bibliography

[1] BuzzFeed. *You Won't Believe What Obama Says In This Video!* Accessed: 21.05.2023. 2018. URL: https://www.youtube.com/watch?v=cQ54GDm1eL0.

[2] S. Greengard. "Will Deepfakes Do Deep Damage?" In: *Commun. ACM* 63.1 (Dec. 2019), pp. 17–19. ISSN: 0001-0782. DOI: 10.1145/3371409. URL: https://doi.org/10.1145/3371409.