BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF INTELLIGENT SYSTEMS
ÚSTAV INTELIGENTNÍCH SYSTÉMŮ

DEEPFAKE DETECTION FRAMEWORK

FRAMEWORK PRO DETEKCI DEEPFAKES

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. JAN BERNARD**
AUTOR PRÁCE

SUPERVISOR **Mgr. KAMIL MALINKA, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2022

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

140642

Institut:	Department of Intelligent Systems (UITS)
Student:	**Bernard Jan, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Cybersecurity
Title:	**Deepfake Detection Framework**
Category:	Security
Academic year:	2022/23

Assignment:

1. Learn about deepfakes (voice and video). Explore the current state of deepfakes detection methods (voice and video).
2. Learn about the technologies needed to create web extensions and technologies for creating scalable server applications.
3. Learn about existing deepfake detection solutions (e.g. other commercial web browser plug-ins)
4. Design an extensible framework (server-client or client-only) for deepfakes detection (support for at least 3 detection methods (voice and video)). Design a web extension for deepfakes detection that will use this framework. The solution should support multiple browsers and allow the detection of displayed content and uploaded files.
5. Implement the tool according to the design.
6. Test the functionality and reliability of the resulting implementation. Perform testing on at least two independent publicly available deepfakes datasets.
7. Discuss usability, detection efficiency and possible extensions.

Literature:
Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa
Singh, “Facial retouching and alteration detection,” in Handbook of Digital Face Manipulation and Detection, pp. 367–
387. Springer, 2022
FIRC Anton a MALINKA Kamil. The dawn of a text-dependent society: deepfakes as a threat to speech verification systems. In: Brno: Association for Computing Machinery, 2022

Requirements for the semestral defence:
Items 1 to 4.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malinka Kamil, Mgr., Ph.D.**
Consultant:	Ing. Anton Firc
Head of Department:	Hanáček Petr, doc. Dr. Ing.
Beginning of work:	1.11.2022
Submission deadline:	17.5.2023
Approval date:	3.11.2022

Abstract

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v anglickém jazyce.

Abstrakt

Do tohoto odstavce bude zapsán výtah (abstrakt) práce v českém (slovenském) jazyce.

Keywords

Sem budou zapsána jednotlivá klíčová slova v anglickém jazyce, oddělená čárkami.

Klíčová slova

Sem budou zapsána jednotlivá klíčová slova v českém (slovenském) jazyce, oddělená čárkami.

Reference

BERNARD, Jan. *Deepfake Detection Framework*. Brno, 2022. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Mgr. Kamil Malinka, Ph.D.

Deepfake Detection Framework

Declaration

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana X... Další informace mi poskytli... Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....
Jan Bernard
December 29, 2022

Acknowledgements

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant apod.).

Contents

Chapter 1

Introduction

- deepfake is buzzword (no agreed-upon technical definition) - ...

Chapter 2

Deepfakes

The creation of fake media and their detection have been a problem since photography was invented. Digital photography or video with tools such as GIMP, Adobe Photoshop or Adobe After Effects allows more people to create fakes than before, still some experience in this area is needed. Media that have been modified or otherwise manipulated are called synthetic media, and they do not depend on whether it is an analogue or digital medium. Deepfakes also fall under this category [4]. Tools powered by deep learning allow unexperienced users to easily create trusted fakes.

The quality of deepfakes reached a level when a trained person or even an experienced researcher in this field has a problem of spotting them. This fast development allows creating realistically looking assets to art photography or movie production, unfortunately, it can be used for malicious purposes like creating fake porn videos to blackmail people or manipulate public via fake news. There are many use cases where deepfakes can be applied.

It is putting huge pressure on researchers to develop new forensics tools or any technology which will prevent malicious usage of deepfakes. As mentioned before, creating fakes is not new, and a whole field of study engaged in spotting fakes and developing techniques over 15 years. Despite continuous research efforts in the past, the advent of deep learning changed the rules of the game. [10]

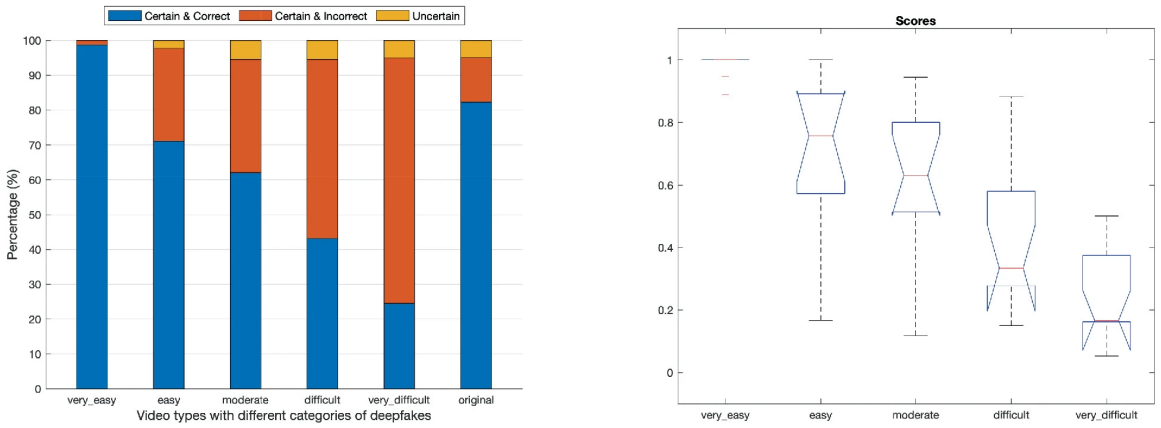
2.1 Human capabilities of deepfake detection

The human ability to recognize fake materials from the originals is in contradiction to their quality. Korshunov and Marcel confirmed this in their research. They created a questionnaire containing several videos, and the subject (interviewee) had to answer after watching the video whether the person in the video was genuine, fake, or they are uncertain. The videos were manually divided into five categories (very easy, easy, moderate, difficult, and very difficult, original).

Videos were split into several categories manually by researchers probably without usage of any metrics but based on their personal feelings, and ANOVA test shows there is an overlap in several categories, so several videos could be moved to different category. However, the categories are still significantly different.

The results of test in Fig. 2.1 certainly demonstrate that people's recognition ability decreases significantly as the quality of deepfakes increases. Also, the audience of this test knows they are looking for fakes, otherwise we can expect worse results if there will be unsuspected audience (for example: deepfakes on social media). It is quite alarming

that the correct answers in the category of "very easy" reach only 71,1 %. The quality of deepfake increases over time, thus it can be expected that human recognition ability will continue to decrease. [6]



(a) Subjective answers

(b) ANOVA test

Figure 2.1: Subjective answers and median values with error bars from ANOVA test for different deepfake categories. Retrieved from [6].

Another research tested only recognition of audio tracks and they were comparing humans versus computer programmes. Attendees had a correct classification between fakes and origins 67 % after the first several rounds. Their accuracy increased while listening and answering to more tracks, but the value stabilizes to 80 %. On average, trained AI performs about 10 % better than human, but this result highly depends on difference of learning and test dataset. Still, it shows that the computer can outperform humans in spotting deepfakes. [8]

2.2 Potentional risks

Humans are not good at recognizing deepfakes, but "why should we be worried?". Almost every technology humankind created could be used for good or bad – deepfake is no exception. There are plenty different deepfake categories, and each has its own attack vector or use case. This section is describing potential risks of those categories and their closer description will be covered in next section 2.3.

Deepfakes are about gaining someone's trust or influence him. For the last couple of years there has been an increasing trend to scam people, mostly via phone or computer. Targeting only one person/victim, for example, to gain their money or information. Those attacks are getting better and more credible and using deepfake to impersonate close friend of victim could be next step how to improve it, if it is not already happening. [2]

Creating "fake news" to influence a large audience is the most common use case of deepfakes because we live in an information era. There are many targets of "fake news" such as rigging elections, demoralizing military units, or manipulating the stock market. In this case, politicians, celebrities, and significant personalities will be used in deepfakes to influence audience. We can only imagine what one person or high quality deepfake can change with enough media reach. For example, after one tweet from Elon Musk about Tesla's stock, sends shares down more than 10 % almost immediately [9]. [4]

A real example of deepfake is famous video with Barak Obama insulting Donald Trump, which should spread awareness regarding the fast developing category of new threads [1]. Several years later another video stating Volodimir Zelenskyj talking about surrendering, it was proved that it is a manipulated video, and its purpose was to demoralize Ukraine army and make them capitulate [2].

Another field where deepfakes could be used is to tricky biometrics systems in which the attacker is a different person to the gain access to the building, to secured equipment, etc. It was proven that biometrics system is not ready to deal with deepfakes, and it will probably require to add a new module to authentication pipeline which will be detecting deepfakes. Face or voice biometrics recognition systems are in greatest danger. The falsification of documents is related to this topic and there was a case of smuggling people across borders with an official passport containing morphed photos of two individuals.

These cases are only the tip of the iceberg, and in the future, everyone should ask if video on social media with film celebrity is real or even worse, if the evidence in courts is trustworthy or not. The solution for this will be easy to use tools capable of detecting deepfakes for unskilled users and also for experts.

2.3 Types of deepfakes and their creation process

There are plenty of methods on how to create deepfakes, and as its name suggests some of them are based on deep neural networks, but not exclusively. This section describes most common types of learning networks used for creating image/video or voice deepfakes. One of the most popular types for face deepfakes is Generative adversarial network (GAN), and it is used to create completely new faces or face manipulations.

Each method leaves traces in the medium that can then be detected. This is one way to recognize deepfakes so understanding process of creation is an advantage. Detecting will be described in more detail in Chapter 3.

2.3.1 Neural networks

All the facts regarding neural networks were retrieved from [7]. Neural networks are composited from neurons arranged in layers, and each layer is connected sequentially via synapses. Synapses are weighed, and the process of finding the proper value of all weights is called a learning neural network. To obtain results from the input of n-dimensional “x” process “forward-propagation” is used to propagate “x” through each layer.

Input to layer is vector “a” of values calculated by previous layer or in case of first layer “x” itself. That means result of each layer is also vector calculated by activation function f(a*W+b), where f is activation function (Sigmod, ReLU, etc.), “a” is input vector, W is matrix of weights between layers i and i+1 and b is dimensional bias. Dimensional bias is a constant offset that helps the network shift the activation function toward the positive or negative side [1].

Now let’s consider the neural network M as a black box and denote its execution as M(x) = y. Supervise learning to train M is using paired samples with from (xi, yi) and loss function L is defined. Loss function is to generate a signal at the output of M and propagate him back to find error of each weight in synapses.

[1] https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed
[2] https://www.youtube.com/watch?v=X17yrEV5sl4

Optimalization algorithms such as gradient descent are then used to calculate new weights of M for the number of epochs. As a result of this process, the network learns the function M(xi) yi and is capable of making prediction on unseen data. More detailed descriptions of this could be found in the work of Y. Mirsky and W. Lee [7].

Next list shows types of neural networks used for generating deepfakes [7]:

- Generative Adversarial Networks (GAN) – Consist of two neural networks working against each other. One layer is generator and second is discriminator. Generator producing fake features trying to fool discriminator, on the other hand, discriminator is learning to distinguish between real sample and fake one.
- Encoder-Decoder networks (ED) – Contains at least two networks, encoder and decoder. It has narrowed layers towards its center. If encoder "En" and decoder "De" are symmetric and they are trained as De(En(x)) = x, then the network is called autoencoder. Generating deepfakes using ED trained with function De(En(x)) = xg, where xg is fake generated features. There is possibility to use multiple different ED chained after each other or using specific variant of ED called variational autoencoder.
- Convolutional Neural Network (CNN) – CNN is learning pattern hierarchies in the data. For deepfakes purposes, it learns filters applied over the input and forming an abstract feature map as the output.
- Recurrent Neural Networks (RNN) – RNN can handle variable length data and it is remembering stat after processing which can be used in next iteration. RNN are mostly used in audio.

Each category has its own subcategories that have small modifications or using some techniques from different category.

2.3.2 Voice deepfakes

Speech synthesis is divided into two categories based on input data. Text to speech (TTS) converts written text to artificial speech and second category is called voice conversion (VC). The voice conversion consumes source voice, and both methods produce synthesis voice saying desired phrases specified by the input. [3]

Voice deepfakes are used independently or with deepfake video (full puppet). Creating synthesis voice is computationally challenging and one of the goals is making real-time voice conversion. There are several projects that are trying to accomplish this [3] [4].

2.3.3 Image or video deepfakes

The list of the following deepfakes is based on the work R. Tolosana, et al. [5]:

- Identity swap – Replacing the face of subject with the face of target as shown in Fig. 2.2. There are two different approaches, classical computer graphics-based technique and deep learning technique. Generally, the process of swap could be described as face detection, cropping, extraction of intermediate representations, synthesis of new face, and blending the generated face.

[3] https://github.com/SolomidHero/real-time-voice-conversion
[4] https://www.resemble.ai/speech-to-speech/

Figure 2.2: XXX. Retrieved from [5].

- Full puppet – Method related to identity swap allows creation of so-called puppet. One person (master) is source of facial expression and body movements that are mapped onto target person as shown in Fig. 2.3. [4]

Figure 2.3: XXX. Retrieved from [6].

- Morphing – It is a type of manipulation that is used to create artificial biometric face samples. Final face contains resemble biometric information of two or more individuals. It should be possible to be successfully verified by biometrics systems for all individuals who were source for given deepfake. Fig. 2.4 shows an example of a morphed image.

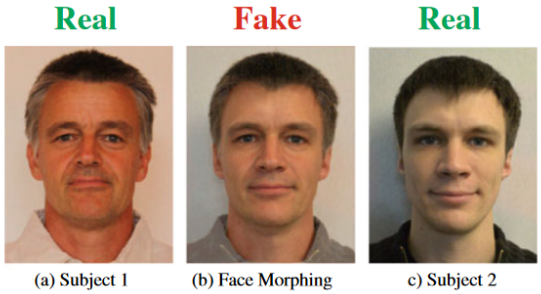


Figure 2.4: XXX. Retrieved from [5].

- Attribute manipulation – Face editing or face retouching technique involves modifying some attributes such as length or color of hair, color of skin, sex, age, adding glasses or other artefacts, and more. Fig. 2.5 shows an example of this technique.

Figure 2.5: XXX. Retrieved from [5].

- Expression swap – Modifying facial expression of the subject as shown in Fig. 2.6. This technique is used as one of part for full puppet.

Figure 2.6: XXX. Retrieved from [5].

- Audio/text to video – This method related to expression swap synthesising facial expression from audio or text. It is also known as lip-sync deepfakes. Diagram in Fig. 2.7 shows how this method works.

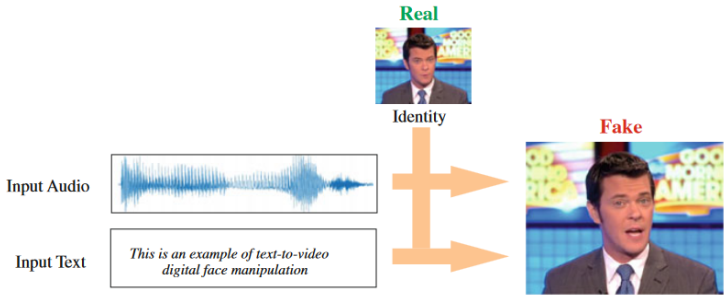


Figure 2.7: XXX. Retrieved from [5].

Creating deepfakes nowadays is complex task and many deepfakes is using more techniques so that they could be included into more than one category. Attackers could create deepfake that will fall under identity swap category and after that use attribute manipulation to tune final results.

Chapter 3

Deepfake detection

Chapter 4

Analysis of existing tools for detecting deepfakes

4.1 Deepware

- https://scanner.deepware.ai/developer/

4.2 Deepstar

- https://www.zerofox.com/deepstar-open-source-toolkit/

Chapter 5

Architecture and technologies analysis

5.1 Requirements

5.2 Containerazation

5.3 Web server

5.4 Web browser plugin

5.5 Selected detection methods

Chapter 6

Framework architecture

6.1 High level architecture

6.2 Containerazation and scaling

6.3 Input layer

6.4 Data preparation layer

6.5 Inividual detection containers

Chapter 7

Client architecture

7.1 Web browser plugin

Chapter 8

Framework implementation

Chapter 9

Client implementation

Chapter 10

Test experiment and results

Chapter 11

Conclusion

Bibliography

[1] *What Is the Necessity of Bias in Neural Networks?* [https://www.turing.com/kb/necessity-of-bias-in-neural-networks]. [cit. 2022-12-29].

[2] Amos, Zachary. *Hybrid Vishing Attacks Skyrocketing: What to Know* [https://itsupplychain.com/hybrid-vishing-attacks-skyrocketing-what-to-know/]. 2022 [cit. 2022-12-28].

[3] Firc, A. *Applicability of Deepfakes in the Field of Cyber Security*. Brno, CZ, 2021. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Available at: https://www.fit.vut.cz/study/thesis/23761/.

[4] Homeland Security. *Increasing Threat of Deepfake Identites* [https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf]. 2021 [cit. 2022-12-16].

[5] Ibsen, M., Rathgeb, C., Fischer, D., Drozdowski, P. and Busch, C. An Introduction to Digital Face Manipulation. In: *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing, 2022, p. 3–26. DOI: 10.1007/978-3-030-87664-7_5. ISBN 978-3-030-87664-7. Available at: https://doi.org/10.1007/978-3-030-87664-7_5.

[6] Korshunov, P. and Marcel, S. The Threat of Deepfakes to Computer and Human Visions. In: *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing, 2022, p. 97–115. DOI: 10.1007/978-3-030-87664-7_5. ISBN 978-3-030-87664-7. Available at: https://doi.org/10.1007/978-3-030-87664-7_5.

[7] Mirsky, Y. and Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* New York, NY, USA: Association for Computing Machinery. 2021, vol. 54, no. 1. DOI: 10.1145/3425780. ISSN 0360-0300. Available at: https://doi.org/10.1145/3425780.

[8] Müller, N. M., Pizzi, K. and Williams, J. Human perception of audio deepfakes. In: *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*. Association for Computing Machinery, 2022, p. 85–91. DOI: 10.1145/3552466.3556531. ISBN 9781450394963. Available at: https://doi.org/10.1145/3552466.3556531.

[9] Shead, S. *Elon Musk's tweets are moving markets — and some investors are worried* [https:

//www.cnbc.com/2021/01/29/elon-musks-tweets-are-moving-markets.html]. 2021 [cit. 2022-12-28].

[10] VERDOLIVA, L. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*. 2020, vol. 14, no. 5, p. 910–932. DOI: 10.1109/JSTSP.2020.3002101.