Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT:

Detection of Manipulated Visual Content in the Context of Disinformation Using Deep Learning and Multimodal Analysis

SUPERVISOR:

Rafał Kozik PhD DSc, Eng. (main supervisor)
Email: rkozik@pbs.edu.pl
Politechnika Bydgoska im. Jana i Jędrzeja Śniadeckich,
Profesora Sylwestra Kaliskiego 7, 85-796 Bydgoszcz

Ewelina Bartuzi-Trokielewicz PhD, Eng. (auxiliary supervisor).
Email: ewelina.bartuzi@nask.pl
Naukowa i Akademicka Sieć Komputerowa – Państwowy Instytut Badawczy (NASK-PIB),
Kolska 12, 01-045 Warszawa

DESCRIPTION:

Introduction

The development of algorithms for generating visual content, such as Generative Adversarial Networks (GANs)[1] and Stable Diffusion models[2], combined with the increasing computational power of consumer-grade hardware, has led to the emergence of highly realistic manipulated materials and a rapid surge in the volume of content subjected to advanced manipulations. Visual and video materials featuring well-known individuals in contexts crafted to provoke specific social or economic responses, such as false investment or medical advertisements, impersonation of authority figures, or disinformative political speeches, are becoming particularly challenging.

As manipulated media becomes more prevalent online, detecting such manipulations is a growing concern for academic institutions, technology companies, and public sector organizations. Despite ongoing efforts to develop effective detection systems, several key research challenges remain. These include the lack of universal detection methods, difficulty adapting to emerging generative techniques, and limited robustness against obfuscation strategies such as adversarial noise. The development of generative technologies often outpaces the advancement of detection solutions.

This research aims to conduct an in-depth analysis of synthetic visual content and develop methods for detecting manipulated materials. The study will include an examination of generative techniques using data sourced from online platforms, as well as the identification of strategies that hinder detection. A crucial part of the project will involve building a representative experimental dataset encompassing a wide range of generative methods, including both controlled examples and incidental cases and developing and evaluating practical detection tools.

SPECIFIC OBJECTIVES

A) Detectors – classification and evaluation of visual manipulations

- Analysis of deepfake generation techniques – conducting a review and taxonomy of current visual manipulation methods (FaceSwap [3,4], LipSync [5,6], diffusion-based algorithms for images and video [7,8]), as well as identifying their characteristic artifacts and analyzing their impact on detection effectiveness. The outcome will include the development of a reference dataset for research purposes.

- Comparison of classical deep learning methods for visual manipulation detection – evaluating approaches based on convolutional neural networks (CNN)[9], transformers[10], and their variants for detecting deepfakes at the image and video level. The study will also include an analysis of how data augmentation techniques influence model generalization capability.

- Development of evaluation procedures and cross-domain testing – designing comprehensive test scenarios across various domains and datasets (public, commercial, and incidental materials from online platforms) to enable realistic performance comparison of detectors. The evaluation will include cross-dataset testing, robustness assessment against unknown manipulation techniques, and analysis of misclassification cases, including their causes (e.g., quality-related artifacts, false positives/negatives). The goal is to identify the generalization and transferability limitations of detection models.

- Development of a multi-class detector for synthetic content – enabling the classification of manipulation types based on their disinformation potential. The detector will operate on multiple levels of detail and incorporate both low-level features (artifacts) and high-level semantic cues (manipulation type). The model could be implemented in two complementary variants: a universal, identity-independent detector, applicable across a broad spectrum of synthetic content, and a specialized, targeted version, tailored to specific individuals combining a biometric approach.

- Evaluation and advancement of deep learning-based detection methods – comparing the effectiveness of traditional CNN-based approaches, transformers, and multimodal vision-language models in detecting synthetic and manipulated content. The study will include analysis of the impact of selected data augmentation strategies on model generalization, including controlled face occlusion and other masking techniques.

B) Application of VLLM – Vision Language Large Models[11]

- Application of VLLMs to synthetic content analysis – experimental use of VLLMs to support manipulation detection at the semantic level. The research will explore the capabilities of VLLMs in detecting synthetic content, automatically generating descriptions of image and video content (what is visible, what is happening),

identifying inconsistencies and visual artifacts (e.g., distortions, regional blurs, characteristic anomalies), and evaluating the material in terms of applied manipulation techniques.

- Preparation of a training and evaluation dataset for VLLMs – development of a dataset that combines visual components (images/videos) with human-generated semantic descriptions, including detailed annotations of visible elements, indication of potential manipulations and their features, and commentary intended for use in explainability processes (XAI[12]).

- Research on VLLM integration in deepfake detection systems and human–AI interaction – the study will focus on evaluating how vision-language models can support end users (e.g., analysts, moderators) by generating explanations and summaries that aid in assessing the authenticity of visual content.

C) Perturbations – robustness to noise and artifacts
   - Analysis of the impact of perturbations such as noise, compression artifacts, and partial face occlusion on the effectiveness of manipulated visual content detection systems. The study will focus on the vulnerability of preprocessing modules, especially face detection, as these are critical components in most processing pipelines. The goal is to identify algorithmic biases and explore opportunities to improve the robustness of detection systems through: training with noisy and perturbed data, investigating the potential use of denoising modules based on UNet, GAN, and diffusion architectures, as well as creating a synthetic dataset for systematic robustness testing and designing appropriate evaluation schemes.

REQUIREMENTS:

- MSc degree in computer science
- Strong programming skills in Python, with experience in machine learning frameworks
- Experience in knowledge engineering
- Familiarity with current research in generative AI and computer vision techniques
- Knowledge of generative models (e.g., GANs, diffusion models) and multimedia manipulation methods
- Understanding of deep learning architectures for computer vision (e.g., CNNs, transformers)

BIBLIOGRAPHY:

1. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, *Generative Adversarial Networks*, 2014, arXiv:1406.2661, https://arxiv.org/abs/1406.2661

2. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, 2022,

arXiv:2112.10752, https://arxiv.org/abs/2112.10752

3. Pavel Korshunov, Sébastien Marcel, *Deepfakes: a New Threat to Face Recognition? Assessment and Detection*, 2018, arXiv:1812.08685, https://arxiv.org/abs/1812.08685

4. Feifei Wang, *Face Swap via Diffusion Model*, 2024, arXiv:2403.01108, https://arxiv.org/abs/2403.01108

5. K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C. V. Jawahar, *A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild*, 2020, Proceedings of the 28th ACM International Conference on Multimedia (MM '20), https://doi.org/10.1145/3394171.3413532

6. Rithesh Kumar, Jose Sotelo, Kundan Kumar, Alexandre de Brébisson, Yoshua Bengio, ObamaNet: Photo-realistic lip-sync from text, 2017, arXiv:1801.01442, https://arxiv.org/abs/1801.01442

7. OpenAI. (2024). Sora: A Text-to-Video Model. https://openai.com/sora

8. Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, Ying Shan, *VideoCrafter1: Open Diffusion Models for High-Quality Video Generation*, 2023, arXiv:2310.19512, https://arxiv.org/abs/2310.19512

9. Keiron O'Shea, Ryan Nash, An Introduction to Convolutional Neural Networks, 2015, arXiv:1511.08458, https://arxiv.org/abs/1511.08458

10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, Attention Is All You Need, 2017, arXiv:1706.03762, https://arxiv.org/abs/1706.03762

11. Yifan Du, Zikang Liu, Junyi Li, Wayne Xin Zhao, *A Survey of Vision-Language Pre-Trained Models*, 2022, arXiv:2202.10936, https://arxiv.org/abs/2202.10936

12. Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, 2017, arXiv:1708.08296, https://arxiv.org/abs/1708.08296