Evaluating Deep Learning Methods for Detection of AI Generated Images: A Study on CIFAKE

Isabela Iacob and Emilia-Maria Nuță

Babeș-Bolyai University, Cluj-Napoca, România

Abstract. In a world full of images and data, where artificial intelligence is now more powerful than ever and can generate complex and lifelike digital media, it is essential to maintain a well-defined line between real content and AI-generated creations. Therefore, these authors propose a comparison between two machine learning techniques aimed at classifying images: a custom Convolutional Neural Network (CNN) and a Residual Neural Network (ResNet). (modificam schimbarile la arhitectura)

The custom CNN architecture relies on multiple convolutional layers to extract hierarchical features from the images, pooling layers for dimensionality reduction, and fully connected layers for final classification. On the other hand, the ResNet introduces residual connections, or skip connections, allowing the network to mitigate the vanishing gradient problem and train deeper architectures effectively. ResNet is particularly well-suited for capturing complex patterns in image data by preserving information across layers.

Our attention will be focused on the CIFAKE dataset, a large dataset containing both AI-generated images and real photographs, labeled as 'Fake' and 'Real'. We will present the final results of both architectures alongside the methodology and experiments conducted. Relevant metrics such as accuracy, precision, and F1-score will be used to compare the two approaches.

These experiments are crucial in an era dominated by artificial intelligence to maintain ethical and secure use of media across the internet and to address potential future legal implications.

Keywords: First keyword · Second keyword · Another keyword.

1 Introduction

Recently, the field of synthetic image generation through artificial intelligence (AI) has evolved rapidly, creating a critical need to detect these images to ensure authenticity and veracity. As AI-generated content becomes more sophisticated, detecting synthetic images will become increasingly challenging, posing significant risks in fields such as law, where the authenticity of visual evidence could influence the outcome of a case. This issue is particularly crucial when determining the veracity of images used in legal proceedings, where misidentification of AI-generated content could have serious consequences for justice.

Therefore, our research proposes the application and optimization of well-established deep learning architectures to detect AI-generated images. Additionally, we employ adversarial attacks on these architectures to assess their robustness and effectiveness in real-world scenarios, where attempts to deceive detection systems are becoming more prevalent.

The paper is structured into five main sections (excluding the introduction). Section 2 reviews the related work on detecting AI-generated images, as this area has become essential for ongoing research and development. In Section 3, we describe the dataset used in the study: the CIFAKE dataset, a widely recognized collection of both AI-generated and real images. Section 4 outlines the methodology used in our experiments, detailing the deep learning models applied and the approach for testing their robustness through adversarial attacks. In Section 5, we present the results obtained from our experiments, followed by an in-depth discussion of their implications. Finally, Section 6 concludes the paper, summarizing the key findings and suggesting potential future research directions in this rapidly advancing field.

2 Related work

The rapid advancements in image generation techniques, such as Generative Adversial Networks (GANs), Variational Autoencoders (VAEs) and diffusion models have shifted the attention to the detection of these AI generated images.

3 Dataset

In this study, we will use the CIFAKE dataset, a widely recognized dataset that includes a substantial collection of AI-generated synthetic and authentic images (Wang et al., 2024). The images are pre-labeled and categorized into two classes: 'Fake' and 'Real,' which facilitates the training process by eliminating the need for extensive pre-processing steps. Additionally, the dataset is already split into training and testing sets, comprising 100,000 training images and 20,000 testing images. This large volume of data ensures that the models have sufficient samples for training and testing, making the final results more reliable and representative.

4 Methodology

Hardware and software resources The library employed for the detection of AI synthetic generated images was Keras (maybe link??). For reproductability purposes, we set the seed to 42. All algorithms in this study were run on Google Colab resources.

5 Results and discussion

6 Conclusion

6.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 1 gives a summary of all heading levels.

Table 1. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

$$x + y = z \tag{1}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 1).

Fig. 1. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Proof. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].

Acknowledgments. A bold run-in heading in small font size at the end of the paper is used for general acknowledgments, for example: This study was funded by X (grant number Y).

Disclosure of Interests. It is now necessary to declare any competing interests or to specifically state that the authors have no competing interests. Please place the statement with a bold run-in heading in small font size beneath the (optional) acknowledgments[1], for example: The authors have no competing interests to declare that are relevant to the content of this article. Or: Author A has received research grants from Company W. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.

References

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25

[1] If EquinOCS, our proceedings submission system, is used, then the disclaimer can be provided directly in the system.