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

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 GenImage 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: Computer Vision · Deep Learning · Generative AI

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 GenImage 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

Existing detection techniques. The accurate detection of AI generated images is paramount, and as such there exists several deep learning approaches for this task. More prevalent are learning based methods. Wang et al. [9] use a ResNet-50 architecture pretrained with the ImageNet as a classifier and train it in a binary classification setting using a ProGAN generated dataset. They find that the CNN model could generalize well in the detection of other GAN generated images. While research suggests that learning based methods are viable for this task, Ojha et al. [7] show that real-vs-fake image classification models trained on a specific generative model have limited generalizability to other generative models, and that the learned features are biased towards recognizing patterns from one class disproportionately better than the other. Other works [4,10] present that AI generated images have unique traces (called fingerprints) that depend on the architecture and training characteristics. Our work tries to train neural networks on data generated by different generative models and assess their performance.

AI generated images. The field of AI generated images has rapidly evolved in the last years, mainly with the use of GANs and DMs. In this study we focus on this type of AI generated images to determine whether different learning based methods categorize them as fake.

3   Dataset

In this study, we use the GenImage dataset [11], a comprehensive resource comprising over 2.6 million images, including 1.35 million AI-generated and 1.33

million real images. The dataset leverages 1,000 distinct labels from the ImageNet database, ensuring a diverse range of image categories that span various subjects such as animals, objects, and scenes. The image generation process for the GenImage dataset employs several state-of-the-art generative models. These include diffusion models such as Midjourney [5], Stable Diffusion V1.4 [8] and V1.5 [8], GLIDE [6], VQDM [3] and ADM [2], as well as GANs like BigGAN [1]. Since the original dataset proposed by the authors is very big in size, we use a tinier version of it available on Kaggle[1]. In the Kaggle version, just 5000 images are kept for 7 of the aforementioned generative models (Stable Diffusion V1.4 is excluded). For each model, data is divided into train (4000 images) and validation (1000 images), divided further into ai (2000 images for train and 500 images for validation) and nature (2000 images for train and 500 images for validation). It is not mentioned how many of the 1,000 labels are kept, but we will continue in a binary classification fashion, either 'nature' or 'ai'.

Fig. 1. Sample images from the GenImage dataset

Augmentation.

4   Methodology

In this section, we introduce the methodology followed when conducting the experiments.

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.

---

[1] https://www.kaggle.com/datasets/yangsangtai/tiny-genimage

5   Results and discussion

6   Conclusion

References

1. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis (2019), https://arxiv.org/abs/1809.11096
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf
3. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10686–10696 (2022). https://doi.org/10.1109/CVPR52688.2022.01043
4. Marra, F., Gragnaniello, D., Verdoliva, L., Poggi, G.: Do gans leave artificial fingerprints? In: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). pp. 506–511 (2019). https://doi.org/10.1109/MIPR.2019.00103
5. Midjourney: (2022), https://www.midjourney.com/home
6. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 16784–16804. PMLR (17-23 Jul 2022), https://proceedings.mlr.press/v162/nichol22a.html
7. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24480–24489 (2023). https://doi.org/10.1109/CVPR52729.2023.02345
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE Computer Society (2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042
9. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot...for now. In: CVPR (2020)
10. Yu, N., Davis, L., Fritz, M.: Attributing fake images to gans: Learning and analyzing gan fingerprints. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7555–7565 (2019). https://doi.org/10.1109/ICCV.2019.00765
11. Zhu, M., Chen, H., YAN, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. In: Advances in Neural Information Processing Systems. vol. 36, pp. 77771–77782 (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf