Robust Detection of AI-Generated Images

A Preprint

Kilinkarov Georgii
Chair of Data Analysis
Moscow Institute of Physics and Technology
Moscow, Russia
kilinkarov.gv@phystech.edu

Daniil Dorin
Affiliation
Address
email

Andrey Grabovoy
Affiliation
Address
email

Abstract

В связи с улучшением качества машиносгенерированных изображений становится очень сложно отличать реальное изоюражение от сгенерированного. Существующие на данный момент решения имеют низкую обощающую способность. В этой статье тестируем модели, несвязанные с нейронными сетями и изучаем распределение цветов на сгенерированных изоюражениях. Также используем всю существующую информацию и модели, для подбора наилучшего решения, делая валидацию на то, с какой именно моделью работаем. Помимо этого, используем методы графических редакторов, на основе искуственного интеллекта.

Keywords AI-Generated Image

1 Introduction

В современном мире в связи с развитием генераторов изображений человеческому глазу стало уже слишком сложно отличать настоящие изображение и машиносгенерированное. Ещё сложнее человеку отличить реальное изображение от реального, но с использованием графического редактора.[1] В связи с доступностью этих сервисов стали очень распространены разные виды мошенничества, использующие машиногенерацию. Таким образом задача детекции машинносгенерированных изображений стала очень важна.

На данный момент не существует общего подхода к решению этой задачи, устойчивого относительно появления новых моделей. Например, появление диффузионных моделей генерации изображений свело сущесвтующие на тот момент методы к точности около 60 процентов[2]. Таким образом, существующие на данный момент методы имеют низкую обощающую способность. Актуальные научные статьи на эту тему можно поделить на три типа: построение устойчивой модели с помощью добавления новых типов генерации в фазу обучения[3, 1], решение задачи с помощью методов, не использующих AI-методы (с помощью классических методов и рассмотрения спектра света)[4], создание новых более мощных датасетов для данный задачи[2, 5].

AI-модели обучается на всё более новых и новых датасетах, включая в себя новые способы генерации, создаются способы онлайн-обучения [1], что улучшает постепенно качество, но концептуально не отличается от предыдущих методов и не обеспечивает устойчивость в случае, если появится более иновационный метод генерации. До появления диффузионных моделей высокое качество показывал метод, рассматривающий спектр по Фурье [4]. Но на диффузионных моделях не показывает уже высокого качества.

Таким образом, в этой статье мы попробовали объеденить существующие методы и найти новый способ детекции машиносгенерированных изображений. Новизна заключается в объединении методов и построении модели, предполагающей сначала тип генерации, а потом проверяющей на генерацию уже непосредственно с предположением определенного типа генерации.

2 Problem statement

На вход мы получаем $X = [X_1, ..., X_n]$ - выборка, где $X_i \in N_{255}^{d \cdot d \cdot 3}$ - картинка.

Нужно построить отображение $F : N_{255}^{d \cdot d \cdot 3} \rightarrow \{0, 1\}$ - отображение из картинки в её тип (реальная или сгенерированная)

Используется метрика точности(accuracy). А именно:

$$\text{Accuracy}(f) = \frac{1}{N} \sum_{i=1}^{N} I(y_i = F(x_i)),$$

где y_i — истинное значение класса, а $F(x_i)$ — предсказанное значение.

Теперь нам нужно выбрать наилушчую по этой метрике модель f^* в своём классе моделей $\mathcal{F}$, т.е.:

$$f^* = \arg\min_{F \in \mathcal{F}} \text{MSE}(F).$$

References

[1] Oliver Wang Richard Zhang David C. Epstein, Ishan Jain. Online detection of ai-generated imagesonline detection of ai-generated images. 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/DFAD/html/Epstein_Online_Detection_of_AI-Generated_Images__ICCVW_2023_paper.html.

[2] Qiangyu Yan Xudong Huang Guanyu Lin Wei Li Zhijun Tu Hailin Hu Jie Hu Yunhe Wang Mingjian Zhu, Hanting Chen. Genimage: A million-scale benchmark for detecting ai-generated image. 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f4d4a021f9051a6c18183b059117e8b5-Paper-Datasets_and_Benchmarks.pdf.

[3] Tam V. Nguyen Samah S. Baraheem. Ai vs. ai: Can ai detect ai-generated images? 2023. URL https://www.mdpi.com/2313-433X/9/10/199.

[4] Matthias Nießner Luisa Verdoliva Davide Cozzolino, Giovanni Poggi. Zero-shot detection of ai-generated images. 2024. URL https://arxiv.org/abs/2409.15875.

[5] Ahmad Lotfi Jordan J. Bird. Image classification and explainable identification of ai-generated synthetic images. 2024. URL https://ieeexplore.ieee.org/abstract/document/10409290.