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, такую классификацию я в работе называю дальше двухступенчатая.

2 Problem statement

Задана выборка

$$\mathfrak{D} = \{\mathbf{x_i}, y_i\},\ i = 1, ..., N,$$

где $\mathbf{x_i} \in \mathbb{N}^{m \times n \times r}$ - изображение разрешения $m \times n$ с r каналами, $y_i \in \{0, 1\}$

Строится отображение $\mathbf{F} : \mathbb{N}^{m \times n \times r} \to [0, 1]$ - отображение из изображения в вероятность того, что изображение сгенерированно.

Используется метрика точности(Log Loss). А именно:

$$\text{LogLoss}(\mathbf{F}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\mathbf{F}(y_i)) + (1 - y_i)\log(1 - \mathbf{F}(y_i)),$$

Решается задача нахождения оптимального отображения $\mathbf{F}^*$ в своём классе моделей $\mathcal{F}$, т.е.:

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

3 Experiments

Целью вычислительного эксперемента является проверка качества двухступенчатой классификации.

3.1 Описание данных и работы модели

В качестве данных я взял датасет ”Artifact” [6], а точнее его уменьшенную версию ”Smaller Artifact”, состоящую из 132000 изображений. Датасет включается в себя реальные изоюражения и 25 методов генерации изображений, включая 13 GANs, 7 диффузионных, и 5 других методов генерации.

В качестве основной модели будет использоваться модель на основе трансформеров, а именно Clip от OpenAI.

3.2 Оценночные метрики

Для оценки качества модели будем использовать следующие оценочные метрики:

1. **Accuracy**: Показывает долю правильных ответов классификации
2. **Recall**: Показывает долю объектов положительного класса, предсказанных верно
3. **Precision**: Показывает долю объектов, имебщих положительный класс из объектов, у которых был предсказан положительный класс
4. **ROC-AUC**: Кривая TPR(доля истинно положительных результатов) от FPR(доля ложных положительных результатов) при изменении порога
5. **ROC-AUC**: Кривая Presicion от Recall при изменении порога

Мы будем использовать эти метрики для проверки качества моделей и сравнения моделей между собой

4 Results

За базовую модель была взята модель с Clip бинарным классификаором. Вторая модель к Clip добавляет два линейных слоя с функцией активацией Relu Один линейный слой с выходом 26(именно столько классов

изображений существует в датасете). Третья модель является многоклассовым классификатором, но Test Accuracy считается для возможности сравнения как для бинарной модели.
Ниже представлена таблица для базовой модели

	accuracy	precision	recall	f1	count
real	0.484222	1.000000	0.484222	0.652493	919.0
stylegan1	0.641509	1.000000	0.641509	0.781609	53.0
stylegan2	0.758865	1.000000	0.758865	0.862903	141.0
stylegan3	0.500000	1.000000	0.500000	0.666667	54.0
big_gan	1.000000	1.000000	1.000000	1.000000	53.0
pro_gan	0.784314	1.000000	0.784314	0.879121	51.0
projected_gan	0.962264	1.000000	0.962264	0.980769	53.0
gau_gan	1.000000	1.000000	1.000000	1.000000	53.0
star_gan	0.943396	1.000000	0.943396	0.970874	53.0
gansformer	0.924528	1.000000	0.924528	0.960784	53.0
generative_inpainting	0.754717	1.000000	0.754717	0.860215	53.0
mat	0.740741	1.000000	0.740741	0.851064	54.0
palette	0.660377	1.000000	0.660377	0.795455	53.0
taming_transformer	0.703704	1.000000	0.703704	0.826087	54.0
ddpm	1.000000	1.000000	1.000000	1.000000	6.0
latent_diffusion	0.905660	1.000000	0.905660	0.950495	53.0
stable_diffusion	0.622642	1.000000	0.622642	0.767442	53.0
vq_diffusion	0.924528	1.000000	0.924528	0.960784	53.0
glide	0.886792	1.000000	0.886792	0.940000	53.0
lama	0.870370	1.000000	0.870370	0.930693	54.0
denoising_diffusion_gan	0.547170	1.000000	0.547170	0.707317	53.0
face_synthetics	1.000000	1.000000	1.000000	1.000000	53.0
cips	1.000000	1.000000	1.000000	1.000000	53.0
cycle_gan	0.981132	1.000000	0.981132	0.990476	53.0
sfhq	0.981132	1.000000	0.981132	0.990476	53.0
diffusion_gan	0.849057	1.000000	0.849057	0.918367	53.0
overall	0.689112	0.678583	0.655488	0.658375	2287.0

Figure 1: Результаты базовой модели для разных классов

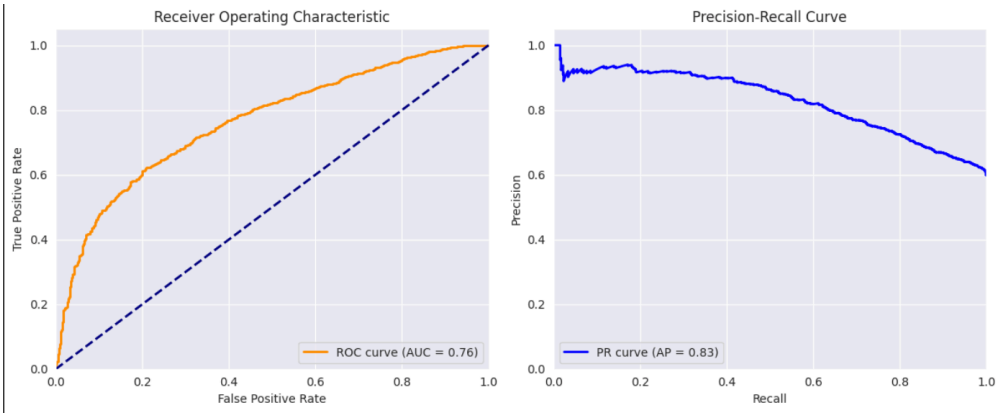

Figure 2: Roc-Auc и PR-curve для базовой модели

Также было проведено сравнение первых двух моделей

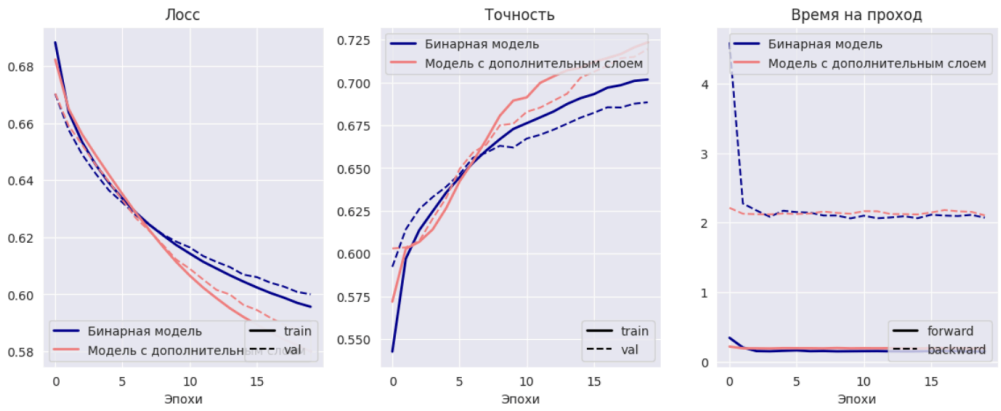

Figure 3: Сравнение первых двух моделей

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.

[6] Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, Zaber Ibn Abdul Hakim, and Shaikh Anowarul Fattah. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2200–2204, 2023. doi:10.1109/ICIP49359.2023.10222083.