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

1 Аннотация

В связи с улучшением качества машиносгенерированных изображений становится очень сложно отличать реальное изображение от сгенерированных. Существующие на данный момент решения имеют низкую обощающую способность. В этой статье рассматриваются разные модели, в том числе несвязанные с нейронными сетями. Также используется вся существующая информацию и модели, для подбора наилучшего решения. Дополнительно строится модель, которая сначала проверяет метод генерации, потом уже использует конкретную модель для этого метода генерации. Помимо этого, используются методы графических редакторов, на основе искуственного интеллекта.

Keywords Машинносгенерированные изображения

2 Введение

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 (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 проводится попытка 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 сгенерировано ли изображение уже непосредственно с предположением определенного типа генерации.

Преимущество этого подхода заключается в подборе оптимальной модели для конкретного класса генирации, проблема заключается в высокой цене ошибки: если произойдет ошибка в предсказании класса генерации, то будет использоваться заведомо плохо подходящая модель.

3 Постановка задачи

Задана выборка

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

где $\boldsymbol{x_i} \in \mathbb{N}_0^{H \times W \times C}$ — изображение размера $H \times W \times C$, $y_i \in \{0, 1\}$.

Необходимо построить отображение $\boldsymbol{F} : \mathbb{N}_0^{H \times W \times C} \to \{0, 1\}$.

Для нахождения оптимального отображения $\boldsymbol{F}^*$ в классе моделей $\mathcal{F}$ используется Binary Cross-Entropy Loss (BCE):

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

4 Теория

Отображение $\boldsymbol{F} : \mathbb{N}_0^{H \times W \times C} \to \{0, 1\}$. представляет из себя композицию двух отображений: $\boldsymbol{F} = \boldsymbol{f} \circ \boldsymbol{g}$, где:

$$\boldsymbol{f} : \mathbb{N}_0^{H \times W \times C} \to \mathbb{R}^d \ - \text{векторизация изображения}$$

$$\boldsymbol{g} : \mathbb{R}^d \to \{0, 1\} \ - \text{классификатор}$$

В статье для обучения $\boldsymbol{F}$ обучается только голова классифкатора $\boldsymbol{g}$, а $\boldsymbol{f}$ фиксировано и не обучается. Для векторизатора $\boldsymbol{f}$ рассматривается Clip от OpenAI.

Clip - модель на основе траснформера, представляющая из себя серию из 5 ResNets и 3 Vision Transformers. На выходе получаем размер 512. Эта модель является одним из state-of-art классификаторов, для многих задач достаточно сделать голову классификатора и получится высокая точность для классификатора.

5 Вычислительный эксперимент

Целью вычислительного эксперемента является проверка качества двухступенчатой классификации.

В работе рассматривается датасет данных Artifact [6], для вычислительного эксперемента взято 13200 изображений. Датасет включается в себя реальные изображения и 25 методов генерации изображений, включая 13 GANs, 7 диффузионных, и 5 других методов генерации. Изображения по классам берутся так, чтобы их количество было одинаково для всех классов, кроме real(их очень много) и ddpm(их очень мало). Количество данных для классов на тестовой выборке можно посмотреть в таблице.

В качестве основной модели будет использоваться Clip от OpenAI 4. Выборка была поделена на обучающую и тестовую в соотношении 70 на 30.

В работе рассматривается несколько разных моделей для сравнения и на основе этих эксперементов выбирается лучшая. 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.

Для оценки качества модели используются следующие оценочные метрики: Accuracy, Recall, Precision, ROC-AUC, PR-curve. Эти метрики помогают в проверке качества моделей и сравнения моделей между собой. Для многоклассовой классификацией также рассматривается confusion matrix для понимаю природы ошибок и слабых мест модели. Модели сравниваются по всем выше перечисленным метрикам, а также по графику обучения, включающие в себя 3 графика: график Loss от номера эпохи, Train, Test Accuracy от номера эпохи и время, затраченное на backward и forward от номера эпохи.

Обучение происходит с помощью оптимизатора Adam, с критерием CrossEntropy и 20 эпохами.

6 Результаты

Ниже представлена таблица для базовой модели для разнообразных метрик и разных классов. Для конкретного класса проведен фильтр на то, является ли реальная метка таковой, поэтому precision = 1, а recall и accuracy одинаковы.

	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

Также было проведено сравнение первых двух моделей. Сравниваем по 3 графикам: график Loss от номера эпохи, график Accuracy от номера эпохи и время на проходы от номера эпохи.

Сравним модели и обнаружим, что многоклассовая получилась совсем неудачной, когда как увеличение слоев подняла качество без потери времени.

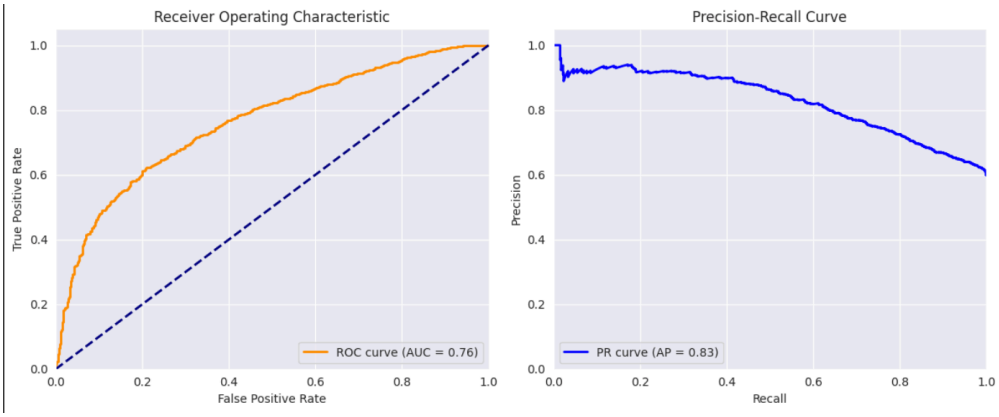

Figure 2: Roc-Auc и PR-curve для базовой модели

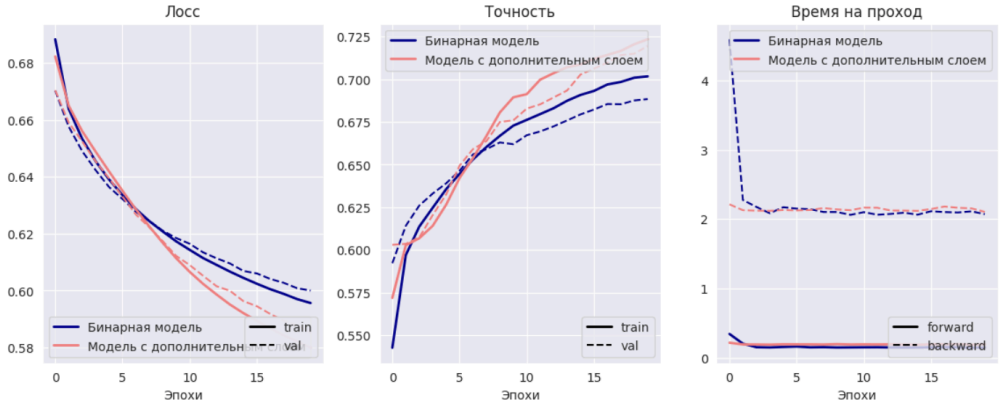

Figure 3: Сравнение первых двух моделей

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.