

# Робастый метод для детекции машинно-сгенерированных изображений

Георгий Валерьевич Килинкар

Научный руководитель: к.ф.-м.н. А. В. Грабовой

Ассистент: Д. Д. Дорин

Анализ данных ФПМИ МФТИ

2025

# Цель и постановка задачи

## Цель работы

Построить модель классификации изображений на машинно-сгенерированные и оригинальные, устойчивую к методам генерации.

## Постановка задачи

Задана выборка


$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, \quad i = 1, \dots, N,$$

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

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

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

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

-  Jordan J. Bird, Ahmad Lotfi (2024). “Image Classification and Explainable Identification of AI-Generated Synthetic Images”. B: URL: <https://ieeexplore.ieee.org/abstract/document/10409290>.
-  Rahman, Md Awsafur и др. (2023). “Artifact: A Large-Scale Dataset With Artificial And Factual Images For Generalizable And Robust Synthetic Image Detection”. B: *2023 IEEE International Conference on Image Processing (ICIP)*, с. 2200—2204. DOI: [10.1109/ICIP49359.2023.10222083](https://doi.org/10.1109/ICIP49359.2023.10222083).
-  Samah S. Baraheem, Tam V. Nguyen (2023). “AI vs. AI: Can AI Detect AI-Generated Images?” B: URL: <https://www.mdpi.com/2313-433X/9/10/199>.

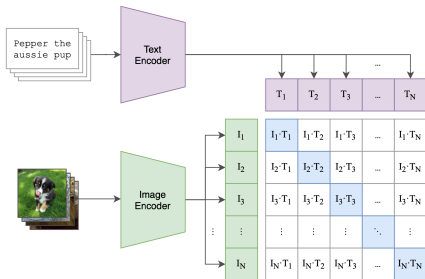
## Предлагаемое решение: Clip

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

$\mathbf{f} : \mathbb{N}_0^{H \times W \times C} \rightarrow \mathbb{R}^d$  — векторизация изображения

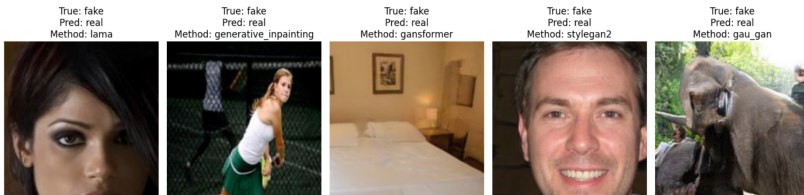
$\mathbf{g} : \mathbb{R}^d \rightarrow \{0, 1\}$  — классификатор

В работе обучается только классификатор  $\mathbf{g}$ . Векторизатор  $\mathbf{f}$  фиксируется и не обучается. В качестве векторизатора  $\mathbf{f}$  рассматривается Clip.



## датасет Artifact

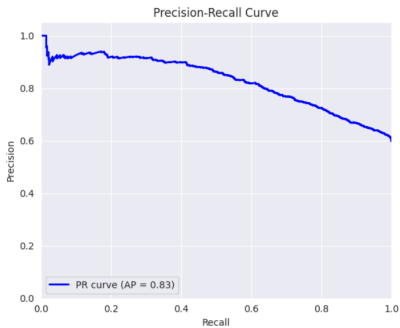
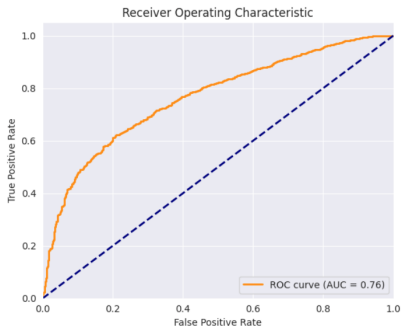
В работе рассматривается датасет данных Artifact. Датасет включает в себя реальные изображения и 25 методов генерации изображений, включая 13 GANs, 7 диффузионных, и 5 других методов генерации.



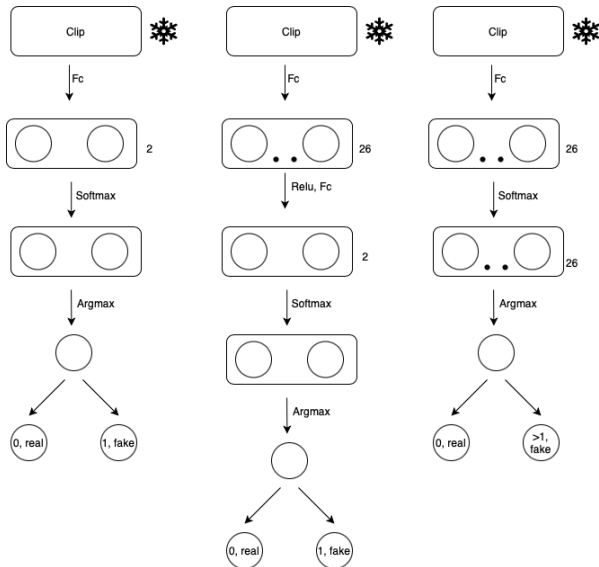
# Рос-Auc и PR-curve

В таблице приведены

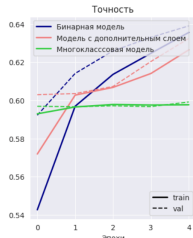
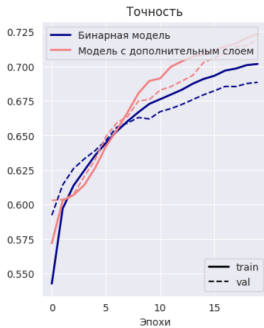
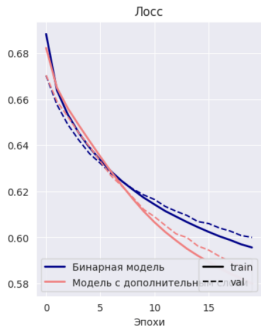
accuracy	precision	recall	f1-score
0.689	0.679	0.655	0.658



# Разные классификаторы



# Графики обучения





В работе были проанализированы разные модели и результаты показали, что:

- ▶ Усложненная модель повысила качество по всем параметрам
- ▶ Многоклассовая классификация себя не оправдала

Что можно ещё сделать в этой работе:

- ▶ Побобрать конкретные модели для конкретных методов и протестировать эту модель