

Robust Detection of AI-Generated Images

Георгий Валерьевич Килинкаров

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 $H \times W \times C$, $y_i \in \{0, 1\}$.

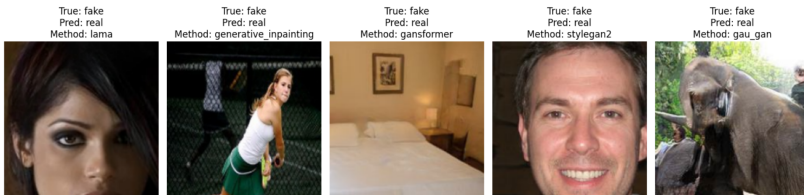
Необходимо построить отображение $\mathbf{F} : \mathbb{N}_0^{H \times W \times C} \rightarrow \{0, 1\}$.

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

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

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.



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:

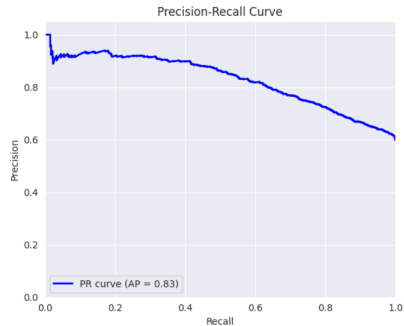
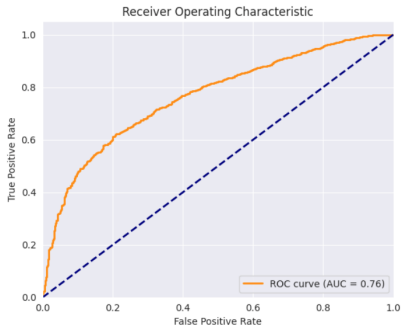
$\mathbf{f} : \mathbb{N}_0^{H \times W \times C} \rightarrow \mathbb{R}^d$ — векторизация изображения

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

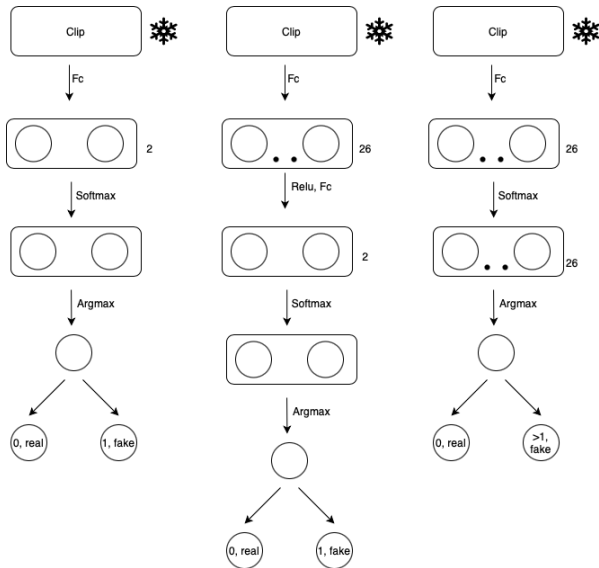
В статье для обучения \mathbf{F} обучается только голова классификатора \mathbf{g} , а \mathbf{f} фиксировано и не обучается. Для векторизатора \mathbf{f} рассматривается Clip от OpenAI.

Рос-Аус, PR-curve и другие

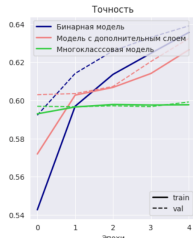
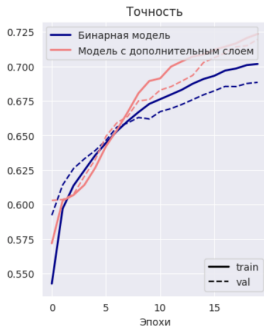
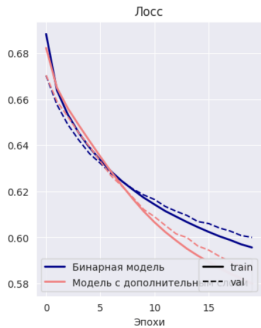
accuracy	precision	recall	f1-score
0.689	0.679	0.655	0.658



Увеличение выхода сети



Графики обучения



Промежуточные результаты

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

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

Ещё планируется сделать:

- ▶ Разобраться с проблемами многоклассовой классификации и попробовать меньшее число классов
- ▶ Побобрать конкретные модели для конкретных методов и протестировать эту модель