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 оригиальные, устойчивую к методам генерации.

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):

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

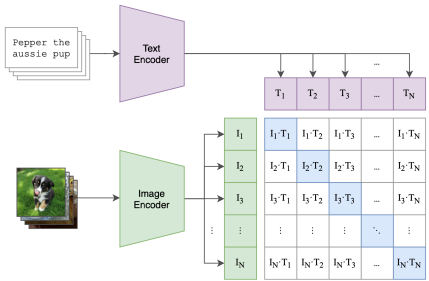
Предлагаемое решение: Clip

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:

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

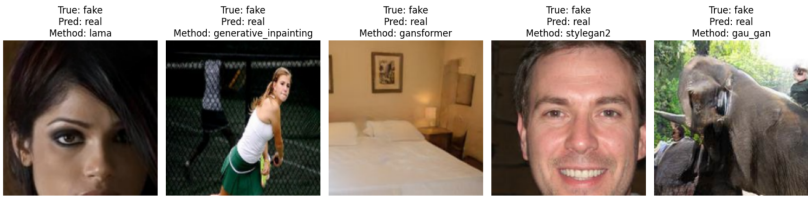
$$\boldsymbol{g} : \mathbb{R}^d \to \{0, 1\} \quad \text{— классификатор}$$

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



датасет Artifact

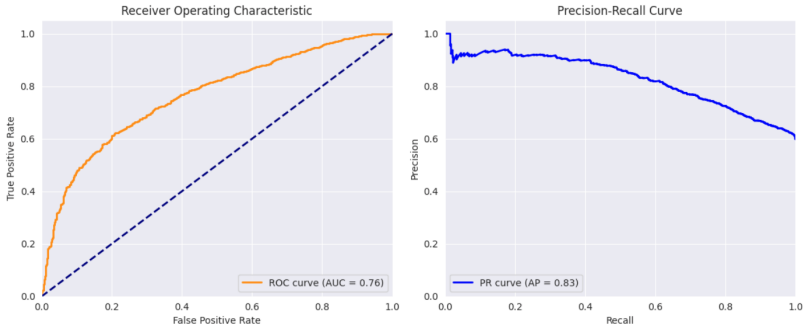
В работе рассматривается датасет данных Artifact. Датасет включается в себя реальные изображения и 25 методов генерации изображений, включая 13 GANs, 7 диффузионных, и 5 других методов генерации.



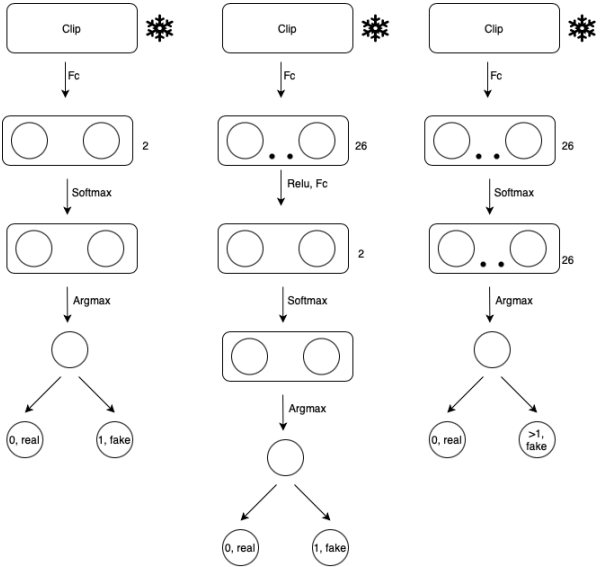
Roc-Auc и PR-curve

В таблице приведены

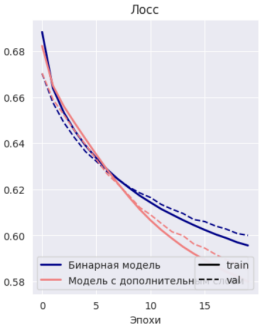
accuracy	precision	recall	f1-score
0.689	0.679	0.655	0.658

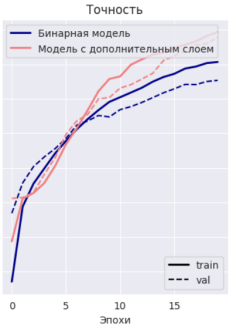


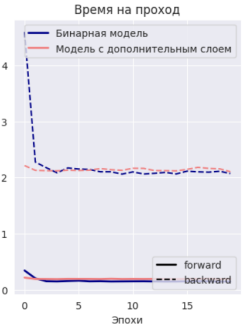
Разные классификаторы

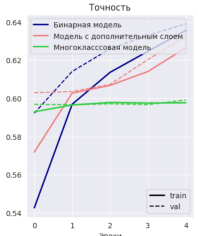


Графики обучения









Промежуточные результаты

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

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

Ещё планируется сделать:

- Обучить модель отдельно на одном классе генерации
- Разобраться с проблемами многоклассовой классификации и попробовать меньшее число классов
- Побобрать конкретные модели для конретных методов и протестировать эту модель