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

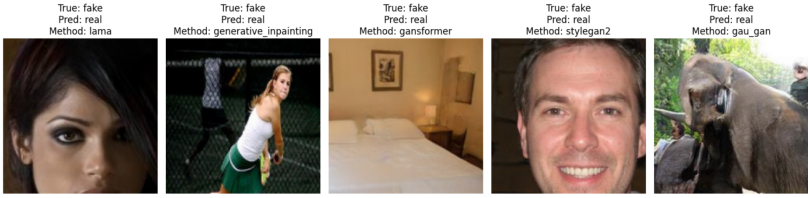
Задана выборка

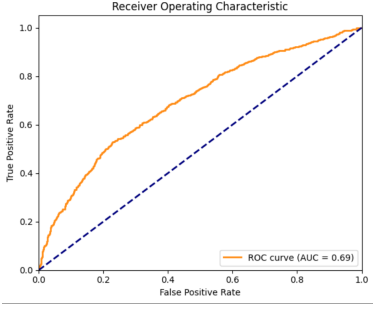
$$\mathfrak{D} = \{\boldsymbol{x_i}, y_i\},\ i = 1, ..., N,$$

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):

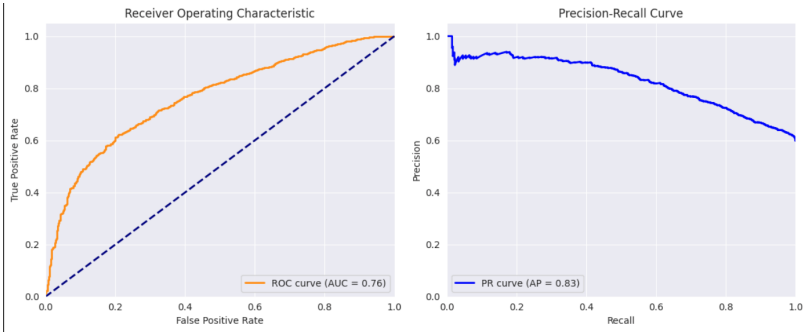
$$\boldsymbol{F}^* = \arg \min_{\boldsymbol{F}^* \in \mathcal{F}} \mathrm{BCE}(F).$$

Artifact и Оценочные метрики

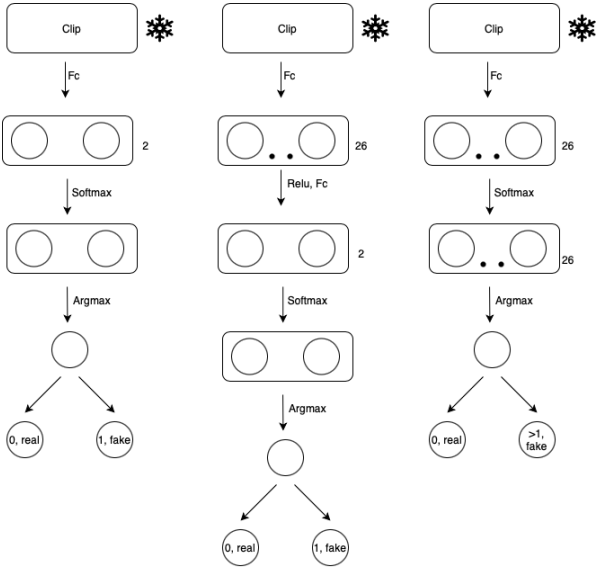




Roc-Auc и PR-curve



Увелечение выхода сети



Графики обучения