Robust Detection of AI-Generated Images

Даниил Дмитриевич Дорин
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 сгенерированные и реальные, устойчивую к методам генерации.

Постановка задачи

Задана выборка

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

где $x_i \in \mathbb{N}^{m \times n \times r}$ - изображение разрешения $m \times n$ с r каналами, $y_i \in \{0, 1\}$
Строится отображение $F : \mathbb{N}^{m \times n \times r} \to [0, 1]$ - отображение из изображения в вероятность того, что изображение сгенерированно.
Решается задача нахождения оптимального отображения F^* в своём классе моделей $\mathcal{F}$, т.е.:

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

Ошибки на моделях, качество классификатора