

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):

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

Ошибки на генерациях, качество классификатора

