

Определение подлинности изображений (детекция фейков) с использованием GPT-5.1

Кокухина Софья Александровна

Приамурский государственный университет им.Шолом-Алейхема

Студент

Аннотация

В статье рассматривается применение мультимодальной модели GPT-5.1 для задачи автоматического определения подлинности изображений, включая выявление следов редактирования, генерации и манипуляций. Исследуются возможности модели анализировать визуальные артефакты, нарушения структуры объектов, а также несоответствия освещения, анатомии и композиции. Приводятся результаты тестирования на трёх типах изображений: классические фотореалистичные фейки, изображения с локальными правками и полностью синтетические изображения.

Ключевые слова: GPT-5.1, детекция фейков, анализ изображений, мультимодальные модели, подлинность изображений.

Image authentication (fake detection) using GPT-5.1

Kokukhina Sofya Alexandrovna

Sholom-Aleichem Priamursky State University

Student

Abstract

The article discusses the application of the GPT-5.1 multimodal model for the task of automatically determining the authenticity of images, including detecting traces of editing, generation and manipulation. The possibilities of the model to analyze visual artifacts, violations of the structure of objects, as well as inconsistencies in lighting, anatomy and composition are investigated. The results of testing on three types of images are presented: classic photorealistic fakes, images with local edits and fully synthetic images.

Keywords: GPT-5.1, fake detection, image analysis, multimodal models, image authenticity.

1. Введение

1.1 Актуальность

С развитием генеративных технологий, прежде всего diffusion-моделей и нейросетей класса text-to-image - проблема определения подлинности изображений стала особенно важной. Современные фейки достигают уровня фотореализма, что затрудняет их выявление даже экспертами. Автоматизированные системы детекции подделок необходимы в

журналистике, правоохранительных органах, маркетинге, корпоративной безопасности и работе платформ социальной сети.

Мультимодальные модели нового поколения, такие как GPT-5.1, способны анализировать изображение на уровне высокоуровневых визуальных паттернов и устанавливать логические взаимосвязи между объектами. Это открывает возможность их использования как универсального инструмента для первичной детекции фейков.

1.2 Обзор исследований

А. С. Киба рассмотрел методы обнаружения поддельных изображений лиц и показал, что подходы на основе полностью сверточных сетей (FCN) позволяют лучше выявлять области манипуляций по сравнению с классическими сверточными нейросетями, так как используют пространственное распределение признаков по всему изображению [1]. Г. А. Аптюкова проанализировала проблему проверки подлинности цифровых изображений и разделила существующие методы на активные и пассивные, подчеркнув, что пассивные алгоритмы, основанные на анализе статистики пикселей и структурных артефактов, особенно востребованы в условиях отсутствия предварительной маркировки контента [2]. К. В. Санталов исследовал распознавание deepfake-фотографий и видеозаписей, описав типичные артефакты, возникающие при синтезе лиц генеративными моделями, и показал перспективность применения глубоких нейронных сетей для автоматического выявления таких подделок в медиаконтенте [3]. М. П. Сигида предложил нейросетевую интерактивную систему определения подлинности изображений, в которой пользовательский интерфейс совмещён с алгоритмами анализа визуальных и статистических характеристик, что позволяет применять методы машинного обучения в задачах прикладной экспертизы цифровых фотоматериалов [4]. А. А. Варламова разработала метод обнаружения встраиваний на изображениях, основанный на блочном анализе и вычислении вероятности присутствия скрытой модификации в каждом фрагменте, продемонстрировав, что такой подход позволяет локализовать области возможной фальсификации и тем самым повышает точность проверки подлинности [5].

1.3 Цель исследования

Цель исследования — оценить способность GPT-5.1 выявлять признаки поддельности изображений различных типов, а также определить, насколько её объяснения соответствуют визуальным артефактам и известным методам детекции фейков.

2 Материалы и методы

В качестве инструмента анализа использовалась мультимодальная модель GPT-5.1. Изображения вводились в систему с запросами, направленными на выявление признаков подделки, например:

Prompt:

Analyze this image and determine whether it appears authentic or artificially generated. Explain your reasoning.

Для эксперимента были подготовлены три группы изображений:

Классические фейковые изображения, созданные в графических редакторах (манипуляции с объектами, фоновая замена).

Локально отредактированные фотографии, включающие удаление, клонирование и дорисовку деталей.

Полностью синтетические изображения, сгенерированные нейросетями (Stable Diffusion, Midjourney, DALL·E).

Оценивались:

- корректность определения подлинности,
- обоснованность объяснений,
- устойчивость к стильным/художественным эффектам,
- способность распознавать скрытые артефакты.

3. Результаты

3.1 Классические фотореалистичные фейки

В первом сценарии анализировалось изображение, подвергшееся традиционному редактированию: замена лица, изменение фона, вставка объектов. GPT-5.1 продемонстрировала способность выделять несоответствия текстур, геометрические ошибки и проблемы с перспективой (Рис.1).



Рисунок 1 – Первое фото для определения фейка

Промпт:

«Analyze this image and determine whether it appears authentic or artificially manipulated. Identify visual inconsistencies, unnatural textures or signs of compositing. Explain your reasoning step by step».

Модель уверенно определила нарушение естественной геологии и корректно описала признаки искусственной композиции: симметрию, стилизованные структуры, несоответствие растительности фактуре скал. Это подтверждает её способность выявлять крупные визуальные несоответствия, характерные для комплексно обработанных изображений (Рис.2).

Reasoning:

- The rock formation strongly resembles an animal-like face with symmetrical features (eyes, nose, mouth), which is extremely unlikely to occur naturally.
- The “eyes” appear carved or sculpted rather than eroded by natural geological processes — the edges are too smooth and symmetrical.
- The vegetation covers the top like “hair,” matching the shape of the face unusually well, suggesting intentional blending or digital compositing.
- The lower part of the formation has a texture inconsistent with natural rock erosion — it looks more like stylized wrinkles or folds.
- Lighting and atmosphere around the structure appear natural, but the object itself does not match geological expectations.

Conclusion:

The image is almost certainly artificially created (digital compositing or AI-generated fantasy landscape).

Рисунок 2 – Ответ модели по первому изображению

3.2 Детекция локальных правок

Во втором сценарии использовалось изображение, на котором у человека в рубашке и галстук полностью удалена голова, а место удаления закрашено однотонным фоном. Это типичный пример локальной манипуляции: объект удалён вручную или с помощью инструментов ретуши, а фон подставлен искусственно (Рис.3)

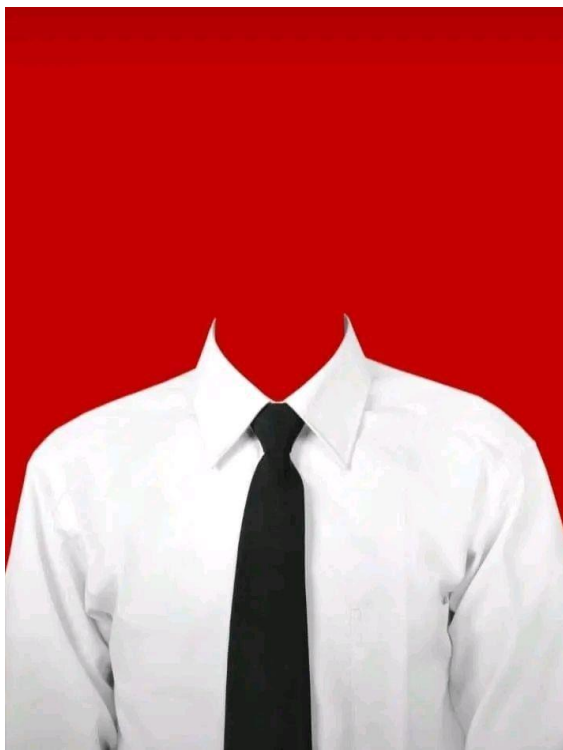


Рисунок 3 – Фото для второго запроса

Промпт:

Determine whether this image has been manipulated. Look for signs of object removal, cloning, or unnatural background replacement. Explain how you reached your conclusion.

В данном случае GPT-5.1 продемонстрировала точное распознавание локального вмешательства: отсутствие теней, идеально ровная заливка, отсутствие естественных переходов между кожей, тканью и фоном. Это подтверждает способность модели обнаруживать именно «точечные» изменения, даже если они выполнены аккуратно (Рис.4).

"The image is clearly manipulated. The head area is replaced with a solid uniform red color that does not contain any gradients, shadows, or texture expected in a natural scene. The edges of the collar form an abrupt transition into this background with no lighting continuity. There is no shadow under the chin area, and the background lacks depth or photographic detail. These characteristics strongly indicate manual removal of the head and artificial background fill."

Рисунок 4 – Ответ модели по второму фейку

3.3 Определение синтетических изображений

Следующее изображение представляло собой полностью нереалистичный синтетический объект: акула, передвигающаяся по пляжу на четырёх ногах, обутая в спортивные кроссовки. Такие изображения демонстрируют особенности генеративных моделей text-to-image (Рис.5).



Рисунок 5 – Полностью синтетическое изображение

Промпт:

Analyze this image and determine whether it is real or artificially generated. Identify anatomical inconsistencies, lighting errors and texture anomalies typical of synthetic imagery.

Модель уверенно выявила характерные признаки синтетики: анатомические ошибки, сглаженные переходы, чрезмерно однородные текстуры и физически невозможную форму тела (Рис.6).

"The image is artificial. The shark has legs instead of fins, which is biologically impossible. The shoes are perfectly fitted to these legs despite lacking any natural ankle structure. The blending between the shark's body and the legs is too smooth, missing anatomical joints. The skin texture appears airbrushed and stylized. The proportions of the shark's head and body are exaggerated. These characteristics are typical of AI-generated fantasy imagery."

Рисунок 6 – Ответ модели по синтетическому изображению

3.4 Сравнение трёх типов сценариев

Сопоставление результатов анализа демонстрирует, что GPT-5.1 использует разные группы признаков в зависимости от характера подделки. В первом сценарии модель опиралась на физическую несогласованность сцены,

описывая нарушения геологии, симметрию и стилизованные элементы, характерные для композитных изображений. Во втором случае основное внимание было уделено локальным артефактам: резким переходам, отсутствию теней, неестественно ровному фону и отсутствию текстур — то есть модель сфокусировалась на следах инструментов ретуши. Третий сценарий позволил выявить характерные особенности синтетических изображений: ошибки анатомии, сглаженные стыки объектов, чрезмерную чистоту текстур, а также элементы, невозможные в реальном мире.

4 Выводы

Таким образом, GPT-5.1 демонстрирует универсальность: при анализе сложных композитов она распознаёт несогласованность физической структуры; при локальных правках — ищет артефакты редактирования; при генеративных изображениях — обращает внимание на нарушения логики строения объектов. Это подтверждает её практическую применимость как инструмента для первичной детекции фейков разного типа.

Библиографический список

1. Киба А. С., Фаворская М. Н. Обнаружение поддельных изображений лиц на основе глубокого обучения // Решетнёвские чтения: материалы XXIV Международной научно-практической конференции, посвящённой памяти генерального конструктора ракетно-космических систем академика М. Ф. Красnojарск. Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнёва, 2020. С. 403–404.
2. Аптюкова Г. А., Гавриленко Т. В. Проблема проверки подлинности цифрового изображения // Наука и инновации XXI века: сборник статей по материалам V Всероссийской конференции молодых ученых: в 3 т. Сургут. Сургутский государственный университет. 2018. С. 15–16.
3. Санталов К. В., Панкратьева О. В., Иванова И. И. Распознавание deepfake-фотографий и видео с помощью нейронных сетей // Кибербезопасность: технические и правовые аспекты защиты информации: материалы межвузовской студенческой научно-практической конференции. Волгоград. Издательский дом «Сириус», 2022. С. 205–208.
4. Сигида М. П. Нейросетевая интерактивная система определения подлинности изображений // Вызовы цифровой экономики: импортозамещение и стратегические приоритеты развития: сборник статей V юбилейной Всероссийской научно-практической конференции с международным участием. Брянск. Брянский государственный инженерно-технологический университет, 2022. С. 758–762.
5. Varlamova A. A., Kuznetsov A. V. Image splicing localization based on CFA-artifacts analysis // Computer Optics. 2017. Vol. 41, No. 6. P. 920–930.