Как современные алгоритмы помогают в борьбе с подделками.

Стремительный рост использования искусственного интеллекта для генерации изображений размывает границы между реальностью и подделкой, что подчеркивает насущную необходимость более совершенных инструментов, которые помогут нам определять, где добро, а где зло.

В недавнем исследовании итальянские ученые проанализировали набор моделей ИИ, предназначенных для идентификации поддельных изображений. Результаты, опубликованные в свежем выпуске IEEE Security & Privacy , показали достаточную эффективность текущих методов. Однако они также указывают на разворачивающуюся гонку все более совершенных технологий ИИ, призванную не отставать от стремительно развивающихся генеративных инструментов.

Луиза Вердолива, профессор Неаполитанского университета имени Федерико II, участвовавшая в исследовании, отмечает, что хотя изображения, созданные с помощью ИИ, могут быть развлекательными, их использование в серьезных контекстах несет множество потенциальных рисков.

"К примеру, можно сфабриковать компрометирующую фотографию политического деятеля и использовать ее для дискредитации во время предвыборной кампании, — поясняет Вердолива. — В таких ситуациях крайне важно иметь возможность определять, реальное это изображение или же оно было сгенерировано компьютером".

Существуют два типа признаков, указывающих на то, что изображение создано с помощью ИИ. Первый — это "высокоуровневые" артефакты или дефекты, очевидные невооруженному глазу: странные тени, разводы, асимметрия на лицах. Однако, как отмечает Вердолива, по мере совершенствования генеративных моделей эти явные огрехи будут становиться все менее заметными.

В более глубоких слоях изображения присутствуют артефакты, незаметные для человеческого глаза и выявляемые только путем статистического анализа данных. Они уникальны для каждого отдельного генератора, создавшего изображение.

Концепция похожа на баллистическую экспертизу огнестрельного оружия. Каждая использованная пуля имеет уникальные царапины, соответствующие стволу пистолета, из которого она была выпущена.

По тому же принципу каждое поддельное изображение содержит отличительный

"цифровой отпечаток». Как ни парадоксально, наиболее эффективный способ обнаружить эти отпечатки — разработать новые модели искусственного интеллекта, специально обученные идентифицировать их и устанавливать связь с конкретным ресурсом.

В своем исследовании ученые протестировали 13 моделей ИИ, способных обнаруживать поддельные изображения и/или идентифицировать их происхождение, используя тысячи известных реальных или синтетических изображений. Неудивительно, что модели в целом весьма эффективно выявляли дефекты изображений и генераторы, на которых они были обучены. К примеру, одна модель, обученная на наборе реальных и искусственных фотографий, смогла идентифицировать материалы, созданные DALL-E — это глубокая нейронная сеть, разработанная компанией OpenAI. Она использует архитектуру трансформера для генерации изображений с помощью заданных текстовых описаний. DALL-E является развитием предыдущей модели GPT-3 и специализируется на создании уникальных иллюстраций на основе текстовых запросов.

 Для обучения DALL-E использовалась большая база данных, содержащая миллионы изображений. Нейронная сеть была обучена связывать текстовые описания с соответствующими визуальными представлениями, что позволяет ей генерировать новые изображения, соответствующие заданным текстовым описаниям. С конца 2023 года генерация DALL-E доступна прямо из другого популярного продукта OpenAI — ChatGPT." data-html="true" data-original-title="DALL-E" >DALL-E, с точностью 87%, а изображения, сгенерированные Midjourney — независимая исследовательская лаборатория, выпускающая проприетарную программу искусственного интеллекта под тем же названием, которая создает изображения из текстовых описаний, аналогичную моделям DALL-E и Stable Diffusion. Инструмент в настоящее время находится в стадии открытого бета-тестирования, которое началось 12 июля 2022 года.

 Команду Midjourney возглавляет Дэвид Хольц, соучредитель Leap Motion. Пользователи создают изображения в Midjourney с помощью Discord-бота." data-html="true" data-original-title="Midjourney" >Midjourney, — с точностью 91%.

Что более поразительно, в некоторых случаях модели распознавали изображения, созданные в программах, на которых они конкретно не обучались. Причина в том, что большинство современных ИИ-инструментов используют очень схожие подходы к созданию картинок, что приводит к определенному сходству дефектов.

Однако, как подчеркивает Вердолива, основная проблема заключается в обнаружении новых, ранее не изученных дефектов от генеративных моделей, выпущенных совсем недавно. Эти неизвестные артефакты представляют трудность и для существующих

методов детектирования.

"В конечном счете, нет такого механизма, который работал бы безупречно всегда. По сути, это соревнование между двумя сторонами. Детекторы становятся все более совершенными, но и генераторы также совершенствуются, и обе стороны учатся на своих ошибках", — говорит Вердолива.

Чтобы справиться с этой проблемой в будущем, Вердолива подчеркивает необходимость применять разнообразные модели для выявления поддельных изображений. Это повысит шансы обнаружить необычные дефекты, присущие новым нейросетям.

Однако ключевую роль играет человеческая разборчивость и критическое мышление. Крайне важно, чтобы люди учились не доверять мультимедийному контенту сомнительного происхождения, а искали информацию только из заслуживающих доверия авторитетных источников.

"Это первая и самая важная линия защиты", — отмечает она, добавляя, что никакой искусственный интеллект не сможет полностью обезопасить нас от недоброжелателей в интернете. "Тем временем научное сообщество будет продолжать предоставлять инструменты и методы для участия в этой гонке вооружений".

Таким образом, хотя технологии искусственного интеллекта открывают новые впечатляющие возможности, они также несут серьезные риски и требуют постоянного совершенствования методов обнаружения подделок.