РАСПОЗНАВАНИЕ ИЗОБРАЖЕНИЙ, СГЕНЕРИРОВАННЫХ НЕЙРОСЕТЯМИ

recognition of images generated by neural networks

Коптев Д.В.

Федеральное государственное автономное образовательное учреждение высшего образования "Санкт-Петербургский государственный университет аэрокосмического приборостроения"
Коптев Д.В. – студент кафедры “Кафедра прикладной информатики”, научный руководитель старший преподаватель Григорьева Н.Н.

Аннотация

В данной статье рассматриваются способы решения проблемы распознавания сгенерированных изображений. Выделяется наиболее благоприятная реализация и проводится работа по созданию нейронной сети, способной отличить искусственное изображение от настоящего. Рассматриваются тонкости создания обучающей выборки, настройки и оценки полученной модели.

Развитие искусственного интеллекта (ИИ) привнесло в современный мир множество инноваций, но многие из них также сопряжены с рисками и новыми проблемами. [1] В последние пару лет очень остро стала проблема попадания фейковых, сгенерированных ИИ изображений в интернет.

Существует несколько главных подходов к решению данной проблемы: запрет или лицензирование нейросетей, запрет публикации искусственных изображений, маркировка сгенерированных изображений водяными знаками, анализ поисковой выдачи на стороне поисковых движков. Решения, запрещающие сами нейросети или их контент, можно отбросить, так как это потребует сильного государственного регулирования. Наиболее реализуемым кажется решение об обнаружении таких изображений на стороне браузеров, с помощью компьютерного зрения и решение о маркировке водяными знаками.

Маркировка является наиболее надёжным и простым инструментом, так как, поместив водяной знак на изображение на стороне нейросети, можно легко его считать визуально или, с помощью алгоритма. Но данное решение практически невозможно осуществить, так как потребуется обязать все сервисы, предоставляющие услуги генерации образов, наносить стандартизированные метки. Данные действия возможны только при вмешательстве государства и изменении законодательной базы.

Рассмотрим оставшееся решение о распознавании, с помощью нейронных сетей. Предполагаемая система будет с разрешения пользователя сканировать поисковую выдачу и выявлять сгенерированные изображения, поле чего рядом с ними будет появляться маркер. К сожалению, данное решение не обходится без издержек. Самыми главными из них являются: усложнение работы поисковой выдачи, замедление процесса сёрфинга и погрешность обнаружения, от которой невозможно избавиться.

В данной системе наибольший интерес представляет алгоритм распознавания искусственных изображений. Такой алгоритм должен уметь распознавать, является ли визуальный образ искусственным или настоящим. Для данной задачи было выбрано решение, с помощью глубоких нейронных сетей, методом классификации изображений. [2] Самым трудоёмким процессом в данной задаче является формирование выборки данных и последующая их предобработка.

Был сформирован обучающий и валидационный датасет, содержащий порядка 10000 изображений, разделённых на два класса: настоящие и сгенерированные образы. При формировании набора “настоящих” изображений были использованы материалы из сети интернет раньше 2019 года, чтобы исключить возможность попадания ложного образа. Выборка искусственных изображений была отобрана среди специализированных сообществ, посвящённых нейросетям, например в социальной сети reddit. [3] При дальнейшей предобработке были убраны все файлы, имеющие не относящиеся к изображениям расширения.

Так как искомая задача заключается в классификации изображений, то была создана свёрточная нейронная сеть, состоящая из 9 слоёв. Полученная модель показала крайне высокие результаты предсказания на случайных изображениях (погрешность в пределах 5%). Для оценки модели были посчитаны следующие метрики: Precision (0.82), Recall (1), BinaryAccuracy (0.92). Высокие значения подтверждают корректность работы алгоритма.

Реализация алгоритма распознавания показала, что выбранное решение жизнеспособно и требует лишь внедрения в систему браузера. На данном этапе развития нейросетей данной реализации достаточно для решения проблемы попадания искусственных образов в интернет. Учитывая непрекращающиеся темпы развития ИИ, новый подход в решении задачи обнаружения может потребоваться уже совсем скоро, в связи с размытием границы между настоящим и сгенерированным образом.

Список литературы

1. Петерс С.В. Нейросети для генерации изображений: области применения и юридические проблемы эксплуатации // Вестник науки. 2024. №3. Стр. 72

2. Денисенко А.А. Глубокое обучение для классификации изображений в различных цветовых системах // Международный журнал прикладных и фундаментальных исследований. 2020. №3. стр. 42 - 47

3. https://www.reddit.com/r/midjourney/