УДК 004.93

ГЕНЕРАЦИЯ СИНТЕТИЧЕСКОГО ДАТАСЕТА ДЛЯ РАСПОЗНАВАНИЯ ТЕКСТА С ИЗОБРАЖЕНИЯ

В. Б. Четвериков

Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, valerysoftwaredev@gmail.com
Научный руководитель: И. М. Галкин, кандидат физико-математических наук, доцент

В данной работе описана разработка библиотеки для генерации обучающего датасета. Результатом является набор функций, который позволяет автоматически создавать наборы данных для последующей тренировки моделей машинного обучения.

Ключевые слова: язык Python; синтетические данные; библиотека scikit-image; текстовая капча.

Введение

В современном мире текст остается ключевым средством передачи информации. Искусственный интеллект позволяет нам автоматизировать обработку текстовой информации. Специалисты в области анализа данных разрабатывают сложные модели, способные обеспечить высокую точность, но распознавание текста в реальных условиях – это все еще сложная задача.

Эта сложность проистекает из многогранности реального мира. Так, визуальный текст может быть незаметен на фоне других элементов изображений. Фон зачастую имеет сложное цветовое сочетание. Форма и расположение текста могут быть искажены. Условия освещения и наложение элементов также вносят свои трудности.

Дополнительной проблемой для обучения нейронных сетей является отсутствие достаточного количества обучающих данных. Качественный обучающий набор должен содержать обширную коллекцию примеров, отражающих реальность.

Цель работы

Основной целью данной работы является изучение подходов и разработка инструмента для решения проблемы нехватки обучающих данных, используя язык программирования Python.

Задачи работы

1. Выбор исходных данных-примеров.
2. Изучение подходов и существующих решений.
3. Реализация библиотеки генерации датасета.

Выбор исходных данных-примеров

В первую очередь необходимо определиться с реальными данными для изучения походов автоматической генерации синтетического датасета. За основу был взят образец капчи, используемой компанией Яндекс.

Капча представляет из себя изображение с двумя словами, разделенными разрывом, и расположенными по кривой или изогнутой линии, при этом порядок или лексическое значение слова не важны.

Изначально было выявлено три способа получения синтетических данных:

1. Использование генеративных моделей.
2. Специальные инструменты и ПО.
3. Покупка данных у сторонних сервисов.

Рис. 1. Примеры текстовой капчи Яндекса

Изучение подходов и существующих решений

Использование открытых генеративных моделей, таких как DALL·E 2 или Stable Diffusion, в их текущем состоянии их развития не дало достаточной точности и качества, а покупка данных у сторонних сервисов оказалась крайне затратной.

Поэтому было решено искать инструменты и ПО для генерации синтетических датасетов. Одним из таких был рассмотрен инструмент SynthText-Russian. Но из-за специфики искривления текста, было решено, что он не подходит.

Рис. 2. Примеры результатов SynthText-Russian

Подходящим вариантом оказалось использование функции библиотеки scikit-image – PiecewiseAffineTransform, которая делит изображение на множество более мелких и применяет аффинное преобразование.

Аффинное преобразование в данном контексте является термином из области компьютерного зрения. Означает следующие операции, примененные одновременно:

Трансляция: Перемещение всех пикселей изображения в одном направлении.

Масштабирование: Изменение размеров изображения в определенном направлении.

Поворот: Вращение изображения на определенный угол.

Сдвиг (или наклон): Искажение или "наклон" изображения в определенном направлении.

Реализация библиотеки генерации датасета

Библиотека, кроме кода формирования исходного изображения и кода, отвечающего за аугментацию, должна содержать файл списка слов русского языка. А также код функций, таких как синусоида, сохраненных в список, со всеми необходимыми параметрами.

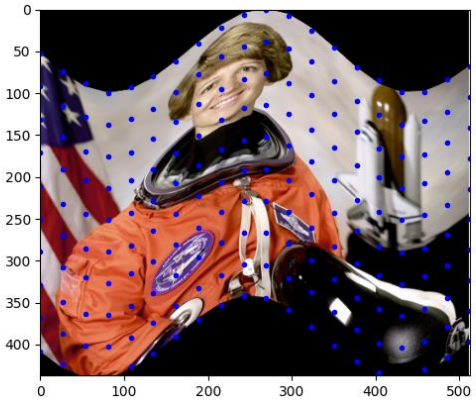

Рис. 3. Пример применения синусоиды к координатам строк изображения через PiecewiseAffineTransform

Формирование исходного изображения представляет собой нанесение букв определенного шрифта и размера на белый фон. Далее к изображению применяется PiecewiseAffineTransform с случайно выбранными из списка функциями.

Опционально, к аугментированному изображению может применяться шум или изменяться яркость.

Результаты работы

1. Реализована библиотека генерации синтетического датасета
2. Изучены подходы и инструменты для генерации синтетических данных

Библиографические ссылки

1. RusTitW: Russian Language Text Dataset for Visual Text in-the-Wild Recognition [Электронный ресурс] / Markov I. [et al.] // arXiv.org. 2023. URL: https://arxiv.org/abs/2303.16531 (дата обращения: 29.03.2023).