1 слайд

С развитием цифровых технологий и ростом интернет-активности существенно возросла потребность в защите web-ресурсов от автоматизированного взаимодействия. Одним из ключевых инструментов такой защиты являются системы CAPTCHA.

Современные системы CAPTCHA предлагают множество форматов защиты, но одновременно с этим появляются возможности для их автоматического распознаванию, в том числе с использованием методов машинного обучения и нейросетевых архитектур.

2 слайд

Целью данной работы является разработка и анализ комплексного подхода к автоматизации решения CAPTCHA в различных форматах с использованием современных нейросетевых инструментов и API для распознавания.

3 слайд

Для достижения поставленной цели были сформулированы следующие задачи, представленные на данном слайде.

4 слайд

Исторически распространенный тип CAPTCHA был впервые изобретен в 1997 году двумя группами, работающими параллельно. Эта форма CAPTCHA требует ввода последовательности букв или цифр из искаженного изображения.

Набравшая популярность технология reCAPTCHA, была приобретена Google в 2009 году. В дополнение к предотвращению мошенничества с ботами для пользователей, Google использовал технологию reCAPTCHA для оцифровки архивов The New York Times и книг из Google Books в 2011 году.

5 слайд

Данный формат CAPTCHA представлен на текущем слайде. На подобных CAPTCHA, зачастую использовались искажения, среди которых:

1. геометрические искажения;
2. перекрытие символов;

3. добавление шума;
4. нелинейные искажения.

6 слайд

На сегодняшний день CAPTCHA является важной мерой безопасности, так как предотвращает автоматические атаки. Современные системы CAPTCHA используют не только текст, но и аудио, изображения, поведенческие анализы и другие инновационные подходы, чтобы сделать тесты удобными для людей, но сложными для программ.

7 слайд

CAPTCHA в формате изображений, на сегодняшний день, широко используется для защиты ресурсов от автоматизированных ботов и может быть реализована несколькими способами. Наиболее распространены два варианта реализации, которые представлены на слайде:

1. цельное изображение, содержащее несколько объектов, частично размытых или искаженных, при этом изображение разбито на сетку 3×3 или 4×4.;
2. составное изображение, сформированное из 9 или 12 отдельных фрагментов (изображений), каждый из которых представляет собой независимое изображение – зачастую низкого качества, с наложением артефактов или шумов.

8 слайд

В ходе тестирования различных подходов для автоматизации решения CAPTCHA в различных форматах были выбраны наиболее подходящие для каждого конкретного формата. Подходы, которые использовались в данной работе представлены на слайде.

9 слайд

Описать блок-схему.

10 слайд

Описать блок-схему.

11 слайд

Текстовые CAPTCHA на сегодняшний день уже не являются настолько же широкоиспользуемыми как CAPTCHA с изображениями, в связи с чем получение достаточного количества изображений для формирования датасета является трудоемкой задачей.

Качество используемого датасета оказывает существенное влияние на итоговую точность работы модели. Для эффективного обучения необходимо, чтобы набор данных соответствовал следующим требованиям:

1. достаточное количество изображений для каждого символа;
2. разнообразие данных;
3. переменная длина последовательностей символов.

Поскольку в открытом доступе отсутствует достаточное количество данных для формирования сбалансированного датасета, необходимо использовать другие способы для получения разнообразных примеров. Среди таких способов наиболее удобным и подходящим для данной задачи является генерация синтетических изображений с использованием специализированных библиотек. В качестве основного инструмента для решения данной задачи, зачастую, используется библиотека captcha на языке Python, обладающая необходимым функционалом для создания изображений CAPTCHA с заданными параметрами. Данная библиотека поддерживает генерацию изображений с пользовательскими шрифтами и различными эффектами искажений, что исключает необходимость привлечения дополнительных инструментов.

После создания изображений все они прошли этапы предобработки, направленные на улучшение качества данных и повышение эффективности обучения модели. Предобработка включала следующие этапы:

1. преобразование изображений в градации серого для уменьшения количества каналов и снижения вычислительной нагрузки;
2. бинаризация изображений с целью получения контрастного представления символов (белый текст на черном фоне);
3. удаление шумов и фона с использованием морфологических операций, в частности, дилатации.

Результат предобработки показан на данном слайде.

12 слайд

Для проведения экспериментов исходный набор данных, содержащий 100 000 изображений, был случайным образом перемешан и разделен на три подмножества: обучающее, тестовое и валидационное в соотношении 80:10:10. Обучающая выборка использовалась непосредственно для обучения модели, валидационная – для контроля качества процесса обучения на каждой эпохе, а тестовая – для окончательной оценки модели на данных, с которыми она ранее не сталкивалась.

В процессе многократного обучения были экспериментально определены оптимальное количество эпох и значения гиперпараметров, обеспечивающие эффективное снижение функции потерь до приемлемых значений. График изменения функции потерь представлен на слайде.

Для предотвращения переобучения использовался механизм ранней остановки, согласно которому обучение прекращалось при отсутствии уменьшения значения функции потерь на валидационной выборке в течение трех последовательных эпох. В данном эксперименте обучение завершилось на 18-й эпохе. На графике видно, что функция потерь стабилизировалась после 10 эпохе, поэтому 10 эпоха является балансом между точностью распознавания последовательностей и скоростью обучения модели.

Анализ графика сходимости функции потерь показывает наличие резкого увеличения ее значения на 9-й эпохе, что может быть обусловлено следующими факторами:

1. перемешивание данных перед каждой эпохой могло привести к образованию несбалансированной выборки, содержащей значительное число сложных примеров.
2. динамическое изменение скорости обучения, осуществляемое с помощью механизма регулирования скорости обучения, могло повлиять на изменение функции потерь.

13 слайд

Также была построена матрица ошибок, позволяющая проанализировать частоту и характер ошибок модели при классификации различных классов. Данная матрица приведена на слайде.

14 слайд

Окончательная точность распознавания отдельных символов составила 0.9263.

После подбора оптимальных значений гиперпараметров модель была сохранена и протестирована на тестовой выборке. Точность распознавания последовательностей различной длины представлена в таблице.

15 слайд

16 слайд

17 слайд

18 слайд

19 слайд

20 слайд