

# Автоматизация решения CAPTCHA в текстовом формате

Студент 5.306М группы: Лаптев А. В.  
Научный руководитель: Калачев А. В.

27 апреля 2025 г.

CAPTCHA давно является стандартным инструментом для защиты веб-ресурсов от спама, автоматизированных ботов и нежелательного извлечения данных.

Автоматизированное распознавание текстовых CAPTCHA позволяет значительно снизить необходимость ручного тестирования веб-приложений, что, в свою очередь, повышает скорость и эффективность тестирования. Кроме того, подобные методы могут использоваться для анализа надёжности внедрённых CAPTCHA, выявления их слабых мест и повышения безопасности веб-приложений, например, за счёт комбинирования нескольких методов защиты.

# Современная реализация текстовых CAPTCHA

Текстовые CAPTCHA обычно состоят из букв и цифр. Зачастую используются символы латинского алфавита (как прописные, так и строчные) и цифры от 0 до 9. Рекомендуемый набор символов в генераторах на некоторых CRM платформах выглядит следующим образом: ABCDEFGHJKLMNPQRSTWXYZ23456789. Длина последовательности символов обычно составляет от 4 до 8 символов.

Для усложнения автоматического распознавания текстовые CAPTCHA подвергаются различным искажениям:

- 1 геометрические искажения;
- 2 перекрытие символов;
- 3 добавление шума;
- 4 сложные фоны;
- 5 нелинейные искажения.

Для распознавания текста с переменной длиной последовательности в задачах САРТСНА наиболее часто применяются следующие архитектуры нейронных сетей:

- 1 оптическое распознавание символов (OCR);
- 2 рекуррентные сверточные нейронные сети (CRNN);
- 3 архитектуры последовательного обучения (Seq-to-Seq).

# Выбор эффективной модели

На начальных этапах экспериментов Seq-to-Seq-модель показала наилучшие результаты среди рассмотренных вариантов. В отличие от OCR- и CRNN-моделей, данная архитектура смогла достичь более высокой точности распознавания последовательностей символов, что обусловлено применением механизма внимания. Дальнейшая работа с моделью была сосредоточена на её оптимизации и улучшении параметров обучения.

# Seq-to-Seq модель. Функция потерь

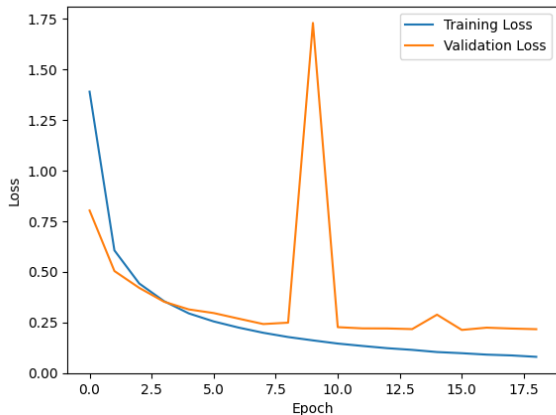


Рис. 1: График изменения значений функции потерь.

# Seq-to-Seq модель. Точность предсказаний

Таблица 1: Точность предсказаний для последовательностей от 4 до 7 символов.

Длина последовательности	Точность распознавания
4 символа	0.9305
5 символов	0.7450
6 символов	0.4575
7 символов	0.1915

# Seq-to-Seq модель. Матрица ошибок

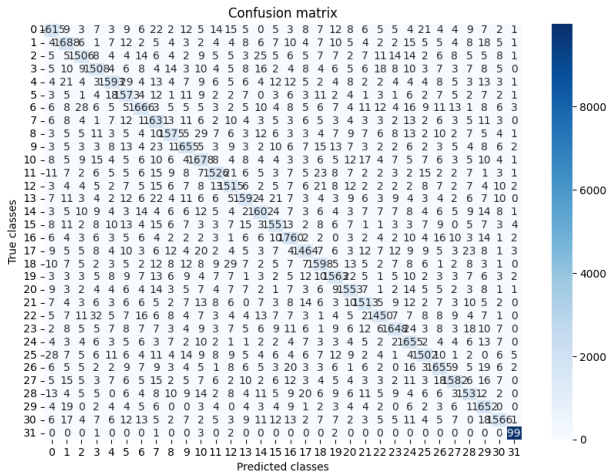


Рис. 2: Матрица ошибок для обученной модели.