ВЕСТНИК ЮГОРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2020 г. Выпуск 4 (59). С. 7–12

DOI: 10.17816/byusu20200407-12
УДК 004.738.5:347.2

В. В. Бурлуцкий, В. А. Балуев

РЕАЛИЗАЦИЯ АГЕНТА АНАЛИЗА ИЗОБРАЖЕНИЙ В РАМКАХ ИНТЕЛЛЕКТУАЛЬНОЙ МУЛЬТИАГЕНТНОЙ СИСТЕМЫ ПОИСКА ПРОТИВОПРАВНЫХ МАТЕРИАЛОВ В СЕТИ ИНТЕРНЕТ

Аннотация. В статье описан процесс разработки агента-анализатора, использующего методы глубокого машинного обучения, и входящего в мультиагентную систему поиска противоправной информации в сети Интернет. Задачей анализатора является извлечение текстовой информации из изображений, где параметры изображения и текста невозможно предопределить.

Ключевые слова: глубокое машинное обучение, OCR, image processing.

V. V. Burlutskiy, V. A. Baluev

IMPLEMENTATION OF AN IMAGE ANALYSIS AGENT WITHIN THE FRAMEWORK OF AN INTELLIGENT MULTI-AGENT SYSTEM FOR SEARCHING FOR ILLEGAL MATERIALS ON THE INTERNET

Abstract. The article describes the development process of an analyzer agent that uses deep machine learning methods and is part of a multi-agent system for searching for illegal information on the Internet. The task of the analyzer is to extract text information from images, where the image and text parameters cannot be predefined.

Key words: deep machine learning, OCR, image processing.

Введение

В настоящее время информационные технологии используются в различных сферах человеческой деятельности, а рост объемов информации носит экспоненциальный характер. Сегодня основной источник информации для многих – всемирная сеть Интернет, предоставляющая мгновенный доступ к огромному количеству источников информации. В связи с этим актуальной задачей является разработка алгоритмов, моделей и программных комплексов, эффективно решающих задачи по обработке больших объемов данных, в том числе и задачи обеспечения информационной общественной безопасности. С целью автоматизации процессов поиска и классификации материалов из сети Интернет противоправного характера (связанной с терроризмом, наркоторговлей, принуждением к суициду и т. д.) в Югорском НИИ информационных технологий разработана автоматизированная информационная система «Поиск» (далее АИС «Поиск»).

Современная версия архитектуры АИС «Поиск» построена на использовании мультиагентного подхода [1]. Данное решение позволяет гибко настраивать механизмы выявления запрещенных материалов. В рамках системы функционируют агенты-анализаторы, которые обрабатывают различные структурные элементы материалов. Так, существуют реализации агентов-анализаторов для классификации текстов, для анализа изображений и для анализа аудиофайлов.

Постановка задачи

В результате экспертизы материалов, обрабатываемых агентом анализа изображений, было выявлено, что в настоящий момент наблюдается тенденция публикации противоправной информации на изображениях с текстом, нанесенным с помощью различных графических редакторов.

Например, на некоторых сайтах в сети Интернет публикуются сообщения о продаже наркотиков без указания объекта торговли в самом сообщении. Однако на прикрепленных к сообщению изображениях содержится текст о продаже наркотических веществ. В данном случае имеющиеся базовые модули анализа текста АИС «Поиск» не смогут обнаружить противоправный контент, к тому же вышеописанный случай имеет систематический характер, что усугубляет проблему.

Для решения этой проблемы и для повышения общей эффективности работы экспертов АИС «Поиск» требуется разработка агента, способного извлекать текстовую информацию из изображений. Результат работы данного агента будет передаваться другому агенту-анализатору, который уже будет анализировать содержимое полученного текста.

Для реализации поставленной цели необходимо решить следующие задачи:

1. Сформировать адекватные обучающие и тестовые наборы данных.
2. Проанализировать существующие методы распознавания и экспериментальным путем выявить наиболее эффективный и адаптировать его к условиям нашей задачи.
3. Разработать агент-анализатор распознавания текста с изображений на основе полученного метода.
4. Интегрировать разработанный агент-анализатор в АИС «Поиск».

Обзор существующих методов и технологий решения

Задача распознавания текста с изображений относится к семейству технологий визуальной идентификации объектов, так как необходимо определить, является ли графический объект текстом, а если является, то к какому классу он принадлежит [2]. Данная технология носит название оптического распознавания символов. Оптическое распознавание символов (англ. Optical Character Recognition – OCR) – это механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные [3].

Системы OCR состоят из следующих основных блоков, предполагающих аппаратную или программную реализацию:

- блок сегментации (локализации и выделения) элементов текста;
- блок предобработки изображения;
- блок выделения признаков;
- блок распознавания символов;
- блок постобработки результатов распознавания.

Блок сегментации отвечает за выделение текстовых областей, строк и разбиение связанных текстовых строк на отдельные знакоместа, каждое из которых соответствует одному текстовому символу.

В блоке предобработки изображения, представленные в виде двумерных матриц пикселов, подвергаются сглаживанию, фильтрации с целью устранения шумов, нормализации размера, а также другим преобразованиям с целью выделения образующих элементов или численных признаков, используемых впоследствии для их распознавания.

Распознавание символов происходит в процессе сравнения выделенных характерных признаков с эталонными наборами и структурами признаков, формируемыми и запоминаемыми в процессе обучения системы на эталонных и/или реальных примерах текстовых символов.

На этапе постобработки смысловая или контекстная информация может быть использована как для разрешения неопределенностей, возникающих при распознавании отдельных

символов, обладающих идентичными размерами, так и для корректировки ошибочно считанных слов и даже фраз в целом [4].

В современных OCR-системах для нахождения сегментов и распознавания символов используются методы машинного обучения. Это обусловлено невысокой эффективностью алгоритмических подходов, так как искомые объекты слишком разнообразны по структуре. По этой причине в работе будут рассматриваться только OCR-системы, использующие методы машинного обучения.

Для оценки работы таких систем была подготовлена валидационная выборка, состоящая из четырех категорий изображений с текстом по 50 изображений на каждую из следующих категорий:

1. Синтетический текст – изображения с нанесенным поверх текстом посредством графических редакторов.
2. Полусинтетический текст – фотографии синтетического текста. Например, фотография афиши, содержащей печатный текст.
3. Полурукописный текст – фотографии рукописного текста печатными буквами. Например, фотография самодельного плаката.
4. Рукописный текст – фотография рукописного текста.

На основе подготовленной валидационной выборки была проведена оценка результативности современных OCR-систем: Tesseract [5], Google Vision [6] и onlineocr [7].

Tesseract – это свободно распространяемая программа для распознавания текстов. Существенный недостаток этой системы заключается в том, что она корректно функционирует только на однородных изображениях, таких как ксерокопии документов или созданные в графических редакторах изображения на одноцветном фоне.

Google Vision и onlineocr являются коммерческими системами распознавания текста с закрытым кодом и требуют абонентскую плату за использование.

Результаты валидации данных моделей приведены в таблице 1.

Таблица 1 – Результаты валидации готовых решений

Модель	**Синтетический текст**	**Полусинтетический текст**	**Рукописный печатный текст**	**Рукописный текст**	**Среднее**	**Без рукописного текста**
Tesseract	0,455	0,306	0,285	0,065	0,28	0,35
Google Cloud Vision	0,69	0,735	0,465	0,062	0,49	0,63
OnlineOCR	0,59	0,735	0,465	0,041	0,46	0,60

Поскольку разрабатываемое решение будет являться частью государственной информационной системы поиска противоправной информации в сети Интернет, то это накладывает дополнительные условия, а именно необходимость обработки огромных объемов контента и открытость программного кода. Эти условия с учетом относительно невысоких результатов валидации существующих решений определяют актуальность разработки модифицированного метода для распознавания русскоязычного текста для агента анализа изображений.

Описание решения

Для первого этапа OCR-сегментации, т. е. нахождения области с распознаваемым текстом, в качестве базовой была выбрана модель CRAFT [8], поскольку она показывает лучшие результаты по критерию точности в классе мультиязычных моделей, а также уже предобучена и не требует существенной доработки.

На втором этапе – распознавании слов – было решено использовать модель MORANv2 [9], так как данная модель построена на перспективном подходе, так называемых «механизмах внимания». Механизмы внимания – это подход в машинном обучении, заключающийся

в выделении части входных данных (регионов изображений, фрагментов текста) для более детальной обработки. Более того, механизмы внимания повышают точность результатов за счет использования контекста. Например, во время распознавания буквы модель будет оценивать не только ее вид и форму, но и также обращаться к результатам распознавания предыдущих букв. Таким образом это снизит вероятность ошибочного распознавания, например, четырех согласных подряд.

Однако для обучения модели распознавания слов необходима обучающая выборка, так как для русского языка публичных выборок такого рода не имеется в открытом доступе. Для решения данной задачи был разработан генератор обучающей выборки.

При реализации генератора было выделено 3 модуля: модуль работы с фоновыми изображениями, модуль генерации битовой маски с текстом, модуль нанесения битовой маски на изображение.

Модуль работы с фоновыми изображениями должен выбирать случайное изображение из предложенных, а также найти все области, пригодные для нанесения текста.

Далее на этапе работы с битовой маской текста случайным образом выбирается произвольное количество областей, пригодных для нанесения текста. На основе размера каждой области строится изображение текста. Наносимый текст определяется набором дополнительных случайных параметров: гарнитурой шрифта, размером шрифта, цветом, углом поворота и межбуквенным интервалом. На этапе формирования текста должна быть сохранена информация о расположении каждой использованной буквы.

На этапе совмещения битовой маски и фонового изображения должны быть реализованы различные методы объединения изображений, а кроме того, данное совмещение должно учитывать выбранные на первом этапе области. На этом этапе также должно проводиться внесение визуальных искажений над текстом, таких как зашумление, размытие и прочие.

На основе этих принципов был разработан генератор, способный сформировать качественные наборы изображений, необходимых для обучения и валидации моделей распознавания.

Эксперимент и анализ

С помощью генератора была подготовлена обучающая выборка для модели MORANv2. Выборка состояла из черно-белых изображений, что обусловлено форматом входных данных модели MORANv2, и содержала одно слово из словаря русского языка. Пример изображения из сгенерированной обучающей выборки представлен на рисунке 1.

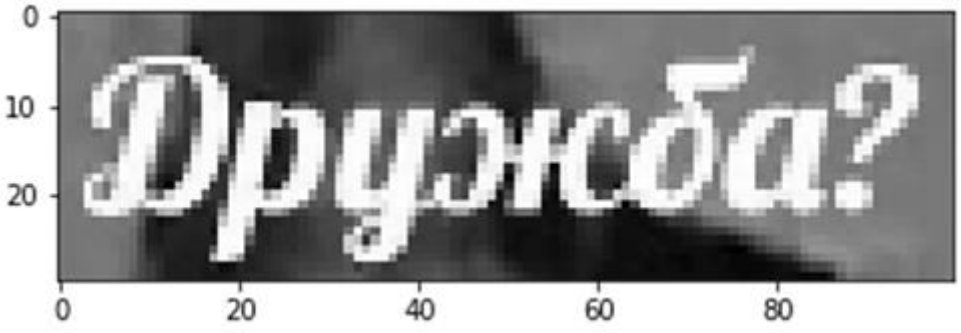

Рисунок 1 – Изображение из сгенерированной обучающей выборки для MORANv2

Модель была обучена на созданной выборке, а также проведена её интеграция с моделью CRAFT. Результаты обучения полученного базового прототипа приведены в таблице 3.

Таблица 3 – Результаты валидации прототипа относительно готовых решений

Модель	Синтетический текст	Полусинтетический текст	Рукописный печатный текст	Рукописный текст	Среднее	Без рукописного
Tesseract only	0,455	0,306	0,285	0,065	0,28	0,35
Google Cloud Vision	0,69	0,735	0,465	0,062	0,488	0,63
OnlineOCR	0,59	0,735	0,465	0,041	0,46	0,60
Moranv2 + CRAFT	0,431	0,199	0,16	0,0189	0,20	0,26

Несмотря на то, что точность распознавания несколько ниже, чем у аналогов, важно учитывать, что это результаты базового прототипа, которые можно существенно улучшить за счет более точного обучения и дополнения механизмами предобработки и постобработки. Что и было сделано.

Для усовершенствования модели была добавлена корректировка возвращаемых моделью слов по словарю с помощью Yaspeller [11], а также дополнительного обучения модели MORANv2. В результате этого удалось добиться на русском тексте результата, близкого к коммерческим англоязычным аналогам. Итоговые результаты валидации для модифицированной модели приведены в таблице 4.

Таблица 4 – Итоговые результаты валидации модели

Модель	Синтетический текст	Полусинтетический текст	Рукописный печатный текст	Рукописный текст	Среднее	Без рукописного
Tesseract	0,455	0,306	0,285	0,065	0,28	0,35
Google Cloud Vision	0,69	0,735	0,465	0,062	0,488	0,63
OnlineOCR	0,59	0,735	0,465	0,041	0,46	0,60
Moranv2 + CRAFT	0,673	0,586	0,445	0,091	0,45	0,568

Заключение

Разработанная модель была реализована в формате микросервиса и интегрирована в АИС «Поиск». В настоящее время эксперты из правоохранительных органов, а также члены кибердружин активно используют этот функционал АИС «Поиск» для выявления и направления на блокировку в Роскомнадзор интернет-страниц с размещенными на них противоправными материалами, особенно связанными с распространением наркотиков. Пример такой страницы, на данный момент заблокированной, приведен на рисунке 2. Разработанный метод позволил реализовать в рамках системы АИС «Поиск» агента-анализатора, распознающего противоправный текст на изображениях, что повысило процент выявления и фильтрации противоправного контента в сети Интернет.

Рисунок 2 – Пример страницы, распространяющей наркотические вещества, выявленной по изображению

Литература

1. Weiss, G. Multiagent systems: a modern approach to distributed artificial intelligence / G. Weiss. – Cambridge : MIT Press, 1999. – 585 p.

2. Optical Character Recognition, Line Eikvi // Semanticscholar. – URL: https://pdfs.semanticscholar.org/9484/96f9d73cab9c7b4fd5c3b656d1e5b1dc50d3.pdf (дата обращения: 20.11.2020).

3. OCR Технологии. – Текст : электронный // Encyclopedia Britannic. – URL: https://www.britannica.com/technology/OCR (дата обращения: 10.11.2020).

4. Обработка и анализ изображений в задачах машинного зрения: курс лекций и практических занятий / Ю. В. Визильтер, С. Ю. Желтов, А. В. Бондаренко [и др.]. – Москва : Физматкнига, 2010. – 672 с. – Текст : энепосредственный.

5. Tesseract Open Source OCR Engine (main repository) // Tesseract-ocr URL: https://tesseract-ocr.github.io/ (дата обращения: 10.11.2020).

6. Google Vision API // Google Cloud's Vision API offers powerful pre-trained machine learning models through REST and RPC APIs. – URL: https://cloud.google.com/vision (дата обращения: 10.11.2020).

7. Online OCR. – URL: https://www.onlineocr.net/ru/ (дата обращения: 10.11.2020).

8. Official implementation of Character Region Awareness for Text Detection (CRAFT) // GitHub. – URL: https://github.com/clovaai/CRAFT-pytorch (дата обращения: 10.11.2020).

9. MORAN // A Multi-Object Rectified Attention Network for Scene Text Recognition. – URL: https://github.com/Canjie-Luo/MORAN_v2. (дата обращения: 10.11.2020).

10. Yaspeller Search tool typos in the text, files and websites // GitHub. – URL: https://github.com/hcodes/yaspeller. (дата обращения: 10.11.2020).