Анализ графических изображений патентных документов

Д.М. Коробкин, В.С. Щербинин, С.А. Фоменков, А.С. Тозик

Волгоградский государственный технический университет

Аннотация: В настоящее время в патентных документах содержатся графические изображения чертежей устройств, графиков, химических и математических формул, причем формулы зачастую необходимо распознать и привести к унифицированному стандарту. В данной работе осуществляется анализ графических изображений, извлеченных из описаний патентов ФИПС Роспатента. Обеспечивается тематическая фильтрация математических и химических формул, содержащихся в патентных документах, и их распознавание. Теоретическая ценность заключается в разработанных алгоритмах парсинга патентов в системе Яндекс.Патенты; распознавания среди графических патентных изображений химических и математических формул; перевода графических изображений химических формул в формат SMILES; конвертации графических изображений математических формул в формат LaTeX. Практическая значимость работы заключается в разработанном программном модуле анализа графических изображений из патентных документов. Область применения разработанной системы — исследование патентов и приведение графических изображений к унифицированному стандарту для решения задач патентного поиска.
Ключевые слова: патент, изображение, математическая, химическая, формула, LaTeX, SMILES, Яндекс.Патенты, ClickHouse.

Введение

Особенностью патентов является то, что их структура обязана быть строго типизированной. У каждого патента должно быть название, реферат, развернутое текстовое описание, сведения о цитировании, международная классификация и т.д.

За счет строгой типизации патентных документов облегчается создание систем, которые могут осуществлять патентный поиск [1]. Патентный поиск может применяться для различных целей, таких, как проверка степени уникальности изобретения [2], области применения, поиска аналогов [3] и получения сведений о лицах и компаниях, имеющих отношение к патенту или области. Кроме того, патентный поиск может помочь проводить определенные исследования в различных областях знаний. Например, имея определенный объем данных по патентам в какой-либо области, можно создать патентный ландшафт этой области [4], при помощи которого можно

определить уровень технического развития этого направления и предсказать возможный вектор развития этой области [5].

Для рядового пользователя патентный поиск предоставляется крупными компаниями, которые имеют в своем распоряжении значительный объем патентных документов. Например, сервис Яндекс.Патенты [6] по состоянию на 2023 год имеет в своем распоряжении более трех миллионов патентов.

Помимо текстовой информации в патентных документах содержатся также графические изображения чертежей устройств, графиков, химических и математических формул. Именно графические изображения различных формул можно распознать, привести к унифицированному стандарту и затем использовать в патентном поиске.

На данный момент есть несколько способов распознавать графические изображения из патентных документов. Например, можно приобрести подписку на сервис AWS Textract [7], который позволит подключаться к API и распознавать изображения встроенными средствами сервиса. Другим способом является написание собственного программного обеспечения для распознавания патентных изображений. Можно программно получать содержимое ресурса Яндекс.Патенты для дальнейшего анализа и классификации изображений.

Описание автоматизированного процесса

На первом шаге пользователь должен сформировать файл с ссылками на патенты системы Яндекс.Патенты за нужные ему временные промежутки. После этого необходимо запустить подмодуль скачивания графических изображений из патентов, указанных в конфигурационном файле, и поместить результат в общую папку. После этого программный модуль классифицирует изображения по заданным классам и запишет их в унифицированном представлении (SMILES, LaTeX) вместе с текстовыми

полями указанных патентов. После завершения работы модуля пользователь при помощи SQL запросов к базе данных может получить извлеченную информацию и использовать ее в своих исследованиях.

Программный модуль должен обеспечивать возможность выполнения нижеперечисленных функций:

– Создание конфигурационного файла cj ссылками на исследуемые патенты.

– Парсинг текстовых полей патентных документов с сохранением в БД.

– Парсинг графических изображений из патентных документов с сохранением в файловой системе.

– Классификация графических изображений.

– Конвертация графических изображений формул в унифицированные представления SMILES и LaTeX.

Алгоритм парсинга графических изображений

Алгоритм представленный на рис.1 и на рис.2 направлен на формирование набора изображений для дальнейшего преобразования в унифицированный формат. Происходит рекурсивный обход по всем патентам по начальной ссылке из конфигурационного файла и сохранение изображений в созданную директорию в файловой системе.

Для обучения модели используется сверточная нейронная сеть, которая имеет следующие достоинства [8]:

– Автоматическое нахождение признаков на изображениях, что упрощает процесс предобработки данных и минимизирует необходимость вручную определять признаки.

– Высокая точность и производительность в задачах распознавания изображений, которое может быть сравнимо или превосходить точность человеческого восприятия.

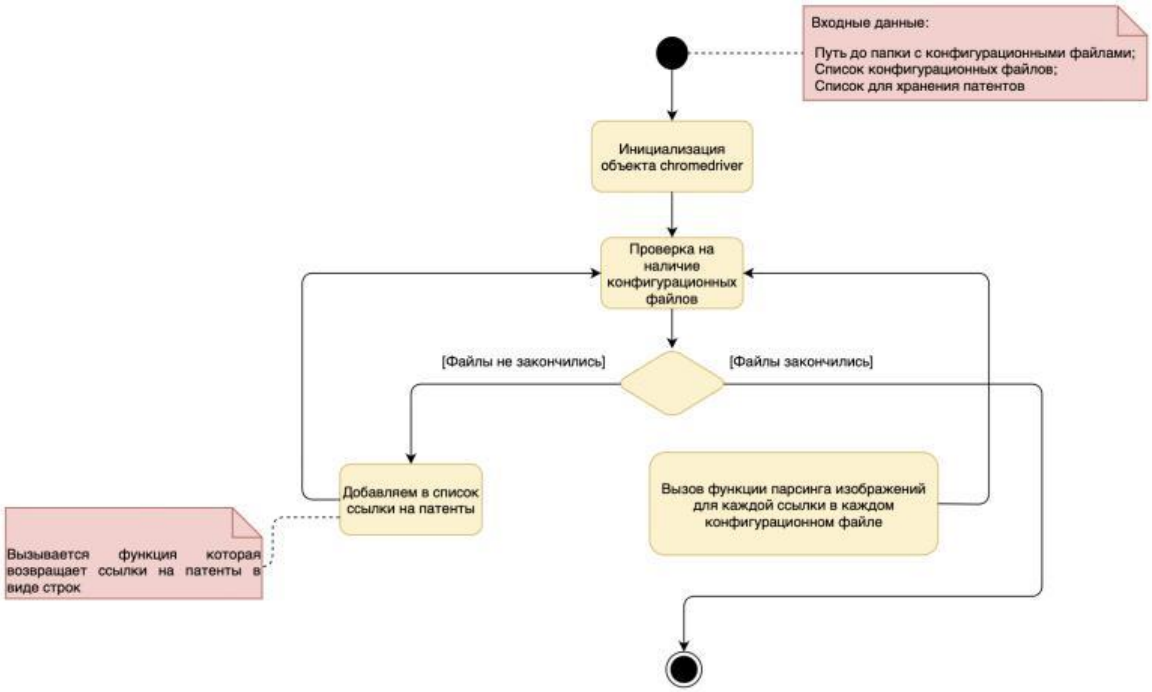

Рис. 1. – Общий алгоритм парсинга графических изображений

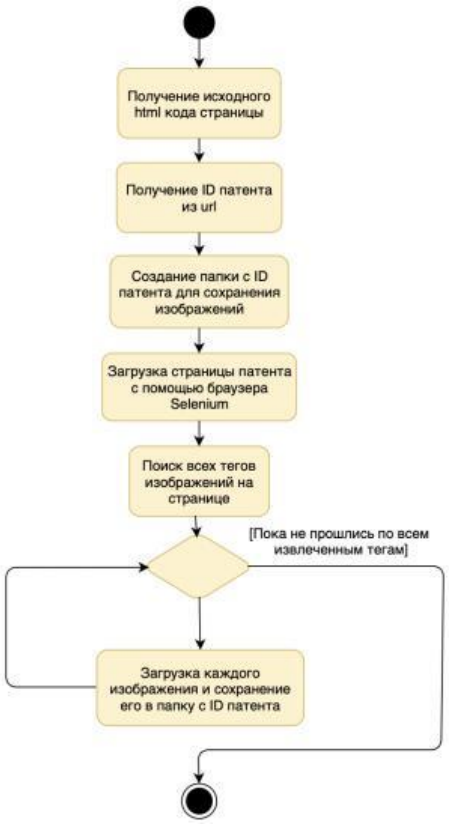

Рис. 2. – Алгоритм «Функция парсинга изображений для каждой ссылки в каждом конфигурационном файле»

Алгоритм классификации

Алгоритм начинается с создания пустой последовательной модели с помощью функции tf.keras.Sequential.

Затем добавляются сверточные слои Conv2D, которые используются для обнаружения признаков на изображении. Каждый слой имеет определенное количество фильтров и размер ядра фильтра. В данном алгоритме используется 4 сверточных слоя с разным количеством фильтров (16, 32, 64, 128) и одинаковым размером ядра фильтра (5, 5). К каждому сверточному слою применяется функция активации ReLU, которая помогает сети ускорять обучение и избегать проблемы затухающих градиентов [9].

После каждого сверточного слоя следует слой подвыборки MaxPooling2D, который используется для уменьшения размерности изображения и извлечения более важных признаков. В данном алгоритме используются слои подвыборки с размером ядра (2, 2).

Далее следует полносвязная часть нейронной сети, которая используется для классификации извлеченных признаков. В данном алгоритме используются два слоя Dense с 1024 и 256 нейронами соответственно. Каждый слой содержит функцию активации ReLU и слой Dropout, который помогает сети избежать переобучения.

Наконец, добавляется выходной слой с функцией активации softmax, который присваивает вероятность каждому из трех классов, на их основе производится классификация изображения.

На обучающую выборку было выделено по 1000 изображений на два исследуемых класса, на тестовую - 500 изображений.

После обучения импортируем модель в проект с целью классифицировать изображения. В данном блоке входными данными будут являться:

– Путь к сохраненной модели;

– Путь к папке с изображениями.

В качестве выходных данных функция возвращает предсказанные метки классов через генератор.

Алгоритм конвертации графических изображений

Данный алгоритм представленный на рис.3 и на рис.4 использует библиотеки pix2tex [10] и deepsmiles [11] для конвертации изображений в текстовый формат. Для работы с данными решениями можно использовать следующие варианты:

– Использование инструмента командной строки.

– Использование решения с графическим интерфейсом.

– Использование API, завернутого в docker контейнер.

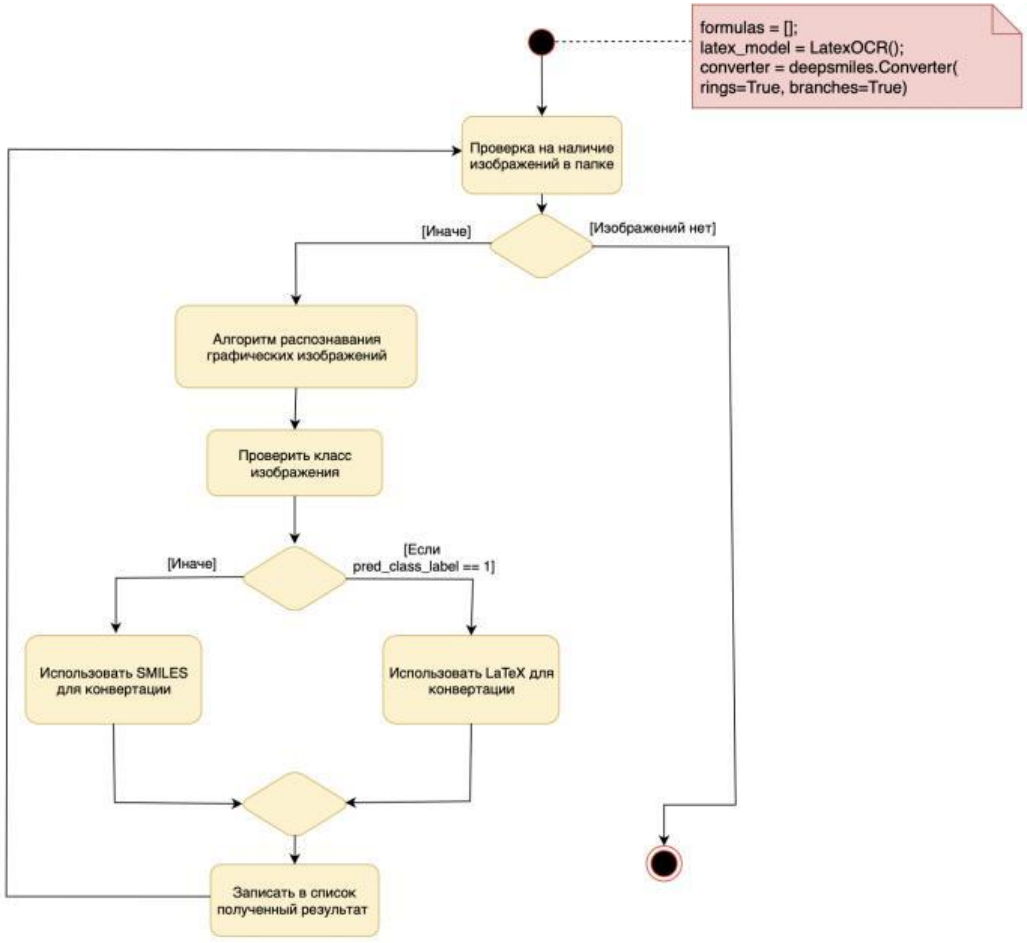

Рис. 3. – Алгоритм конвертации графических изображений в унифицированный формат

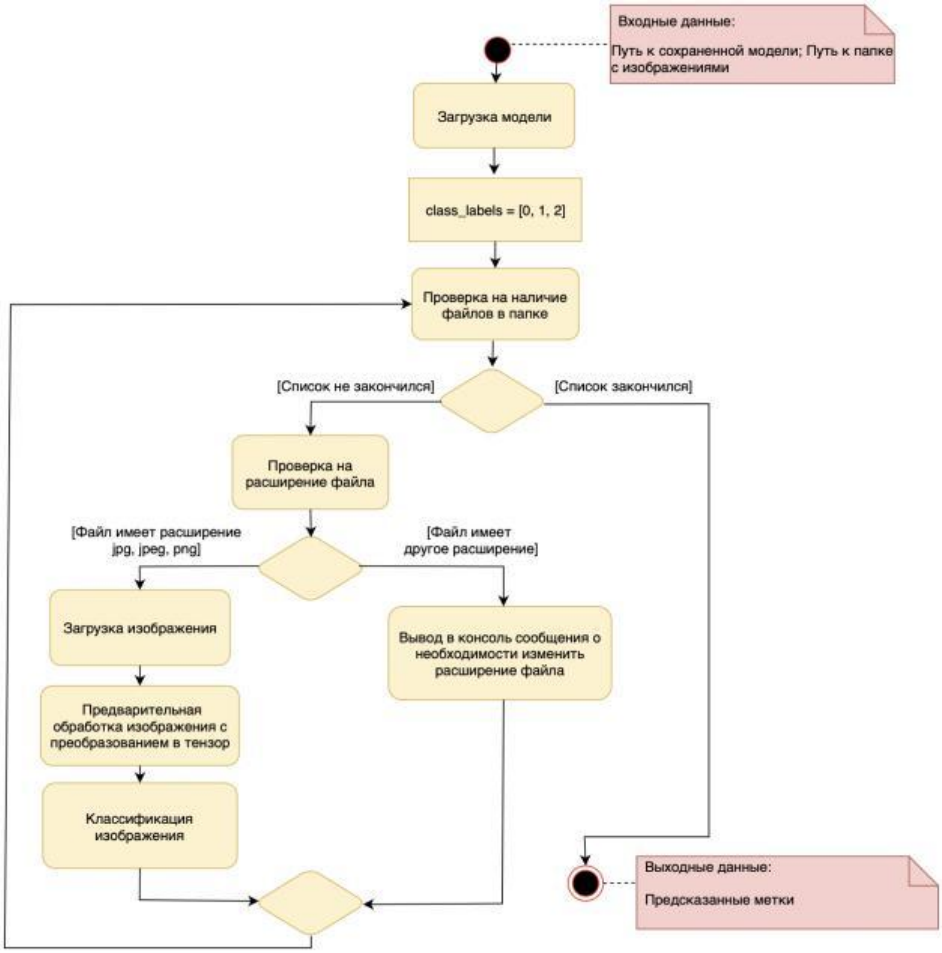

Рис. 4. – Алгоритм распознавания графических изображений

В рамках данной задачи будет использоваться вариант с запуском через инструмент командной строки. Такое решение позволяет использовать возможности библиотеки в связке с остальными подмодулями как единое решение. Входными данными, помимо пути до изображений, является также предобученная модель из работы другого подмодуля. Для работы с изображениями используется библиотека Pillow.

Выходными данными является заполненный унифицированными представлениями графических изображений список, который можно посмотреть в заполненной базе ClickHouse.

Проектирование программного модуля

Для реализации разрабатываемого модуля был выбран язык программирования Python 3.11.

Для разработки парсера использовалась библиотека selenium для автоматизации работы веб-браузера, а также библиотека Beautiful Soup для парсинга текстовых полей.

Для обучения модели на распознавание и классификации изображений использовались библиотеки TensorFlow и Keras. Также использовалась библиотека numpy. Это модуль для python с открытым исходным кодом, который предоставляет общие числовые операции. Для построения визуализаций использовалась библиотека matplotlib.

Для разработки алгоритма конвертации графических изображений в текстовый формат использовались библиотеки pix2tex для преобразования в LaTeX и deepsmiles для конвертации в формат SMILES.

Для хранения информации было выбрано два хранилища: СУБД ClickHouse [12] и файловая система. ClickHouse разрабатывалась как СУБД для хранения больших объемов логов, представляющих небольшие текстовые файлы, что удобно для хранения информации из патентных документов. Для последующего модуля классификации изображений их необходимо хранить в удобно виде, для этого используется файловая система.

Модуль анализа графических изображений из патентных документов представляет собой консольное приложение, предназначенное для обработки и анализа патентной информации и записи обработанных данных в базу данных. Перед запуском модуля необходимо вручную или с помощью скрипта получить конфигурационный файл со ссылками на исследуемые патенты.

Блок парсинга графических изображений является скриптом, который позволяет загружать в файловую систему изображения из патентов с

сохранением их в папку, название которой соответствует текстовому полю id патента. При этом пользователю может выводиться информация об отсутствии изображений. Изображения нумеруются переменной счетчика, начиная с «0.jpg». Каждое новое изображение нумеруется «{i + 1}.jpg», где i – номер предыдущего изображения. После классификации изображений остаются только соответствующие химическим и математическим формулам, как показано на рис.5.

Блок парсинга текстовых полей патента показанный на рис.6 обеспечивает следующий функционал:

– Считывает конфигурационный файл для получения ссылок на исследуемые патенты.

– Выполняет извлечение текстовых полей из патента. К тестовым полям относятся id, title, assignee, authors, publications, link, classifications, abstract, description, claims, cited.

– Предоставляет пользователю доступ к извлеченной текстовой информации при помощи web-интерфейса сервиса Clickhouse cloud.

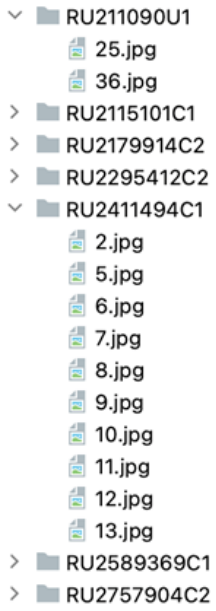

Рис. 5. Блок парсинга

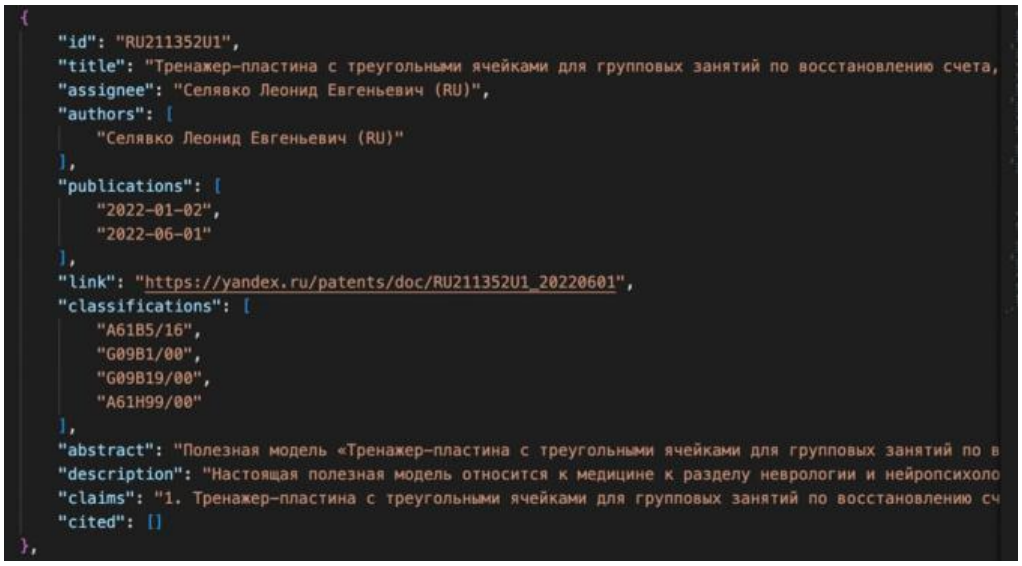

Рис. 6 – Структура извлекаемых текстовых данных

Блок классификации графических изображений, принимая на вход обучающий датасет, состоящий из изображений химических и математических формул, обучает модель для распознавания патентных изображений. На рис.7 представлен процесс обучения модели с метриками accuracy и val_accuracy.

```
Epoch 1/5
21/21 - 150s - loss: 30.3977 - accuracy: 0.6538 - val_loss: 0.5299 - val_accuracy: 0.8000 - 150s/epoch - 7s/step
Epoch 2/5
21/21 - 145s - loss: 0.3188 - accuracy: 0.8825 - val_loss: 0.3594 - val_accuracy: 0.8686 - 145s/epoch - 7s/step
Epoch 3/5
21/21 - 145s - loss: 0.1763 - accuracy: 0.9376 - val_loss: 0.2951 - val_accuracy: 0.9029 - 145s/epoch - 7s/step
Epoch 4/5
21/21 - 146s - loss: 0.1487 - accuracy: 0.9496 - val_loss: 0.2603 - val_accuracy: 0.8857 - 146s/epoch - 7s/step
Epoch 5/5
21/21 - 147s - loss: 0.1297 - accuracy: 0.9533 - val_loss: 0.2814 - val_accuracy: 0.9029 - 147s/epoch - 7s/step
```

Рис. 7. – Процесс обучения подмодуля модели CNN

Блок классификации и конвертации изображений на основе класса изображения использует определенную библиотеку для формирования унифицированного представления графического изображения (SMILES, LaTeX).

Заполненную базу данных можно просмотреть, используя браузер. Пример заполненного текстовыми данными хранилища можно увидеть на рис.8.

#	id	title	assignee	authors	publicatio…	link	classifica…	abstract	description	claims	cited
1	RU2115101C1	Датчик да…	Институт …	["Голод В…	["1997-05-…	https://ya…	["G01L7/08…	Изобретен…	Изобретен…	Датчик да…	["RU234534…
2	RU2179914C2	Способ га…	ОАО "Юрги…	["Федько …	["2000-05-…	https://ya…	["B23K9/16…	Способ мо…	Изобретен…	Способ га…	[]
3	RU2295412C2	Способ по…	Открытое …	["Чебыкин…	["2005-05-…	https://ya…	["B21C37/2…	Изобретен…	Изобретен…	Способ по…	[]
4	RU2411494C1	Способ оп…	Государст…	["Санкин …	["2009-10-…	https://ya…	["G01N3/40…	Изобретен…	Изобретен…	Способ оп…	[]
5	RU2589369C1	Способ оц…	Акционерн…	["Корбут …	["2015-07-…	https://ya…	["B64F5/00…	Изобретен…	Изобретен…	Способ оц…	["RU265556…
6	RU2757904C2	Химически…	СПАГО НАН…	["АКСЕЛЬС…	["2018-01-…	https://ya…	["C07F7/18…	Изобретен…	Область и…	1. Химиче…	[]
7	RU2589369C1	Способ оц…	Акционерн…	["Корбут …	["2015-07-…	https://ya…	["B64F5/00…	Изобретен…	Изобретен…	Способ оц…	["RU265556…
8	RU211090U1	Оптически…	Акционерн…	["Колбас …	["2022-01-…	https://ya…	["H01S3/05…	Полезная …	Полезная …	Оптически…	[]

Рис. 8. – Пример заполнения базы данных текстовыми полями

На рис.9 – 12 представлен результат работы блоков классификации и конвертации графических изображений.

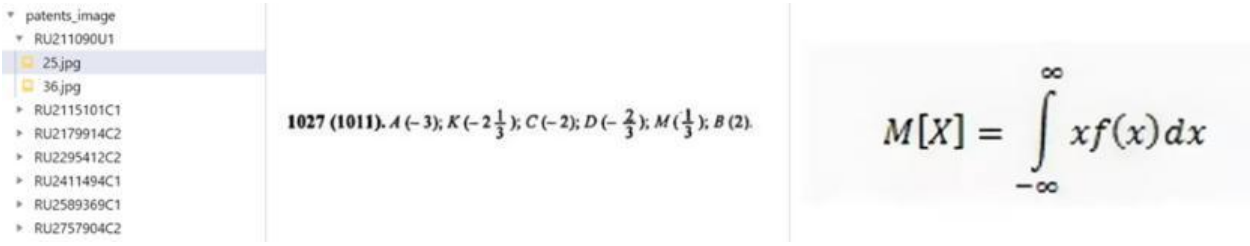

Рис. 9. – Пример изображений патента RU211090U1

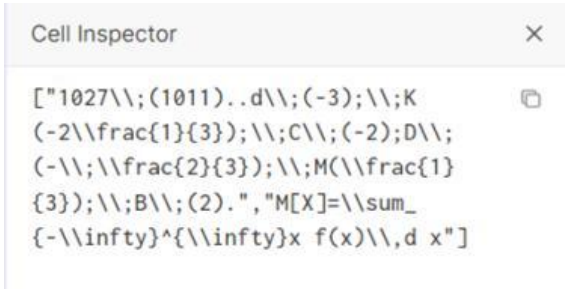

Рис. 10. – Вывод результата конвертации изображений в формат LaTeX

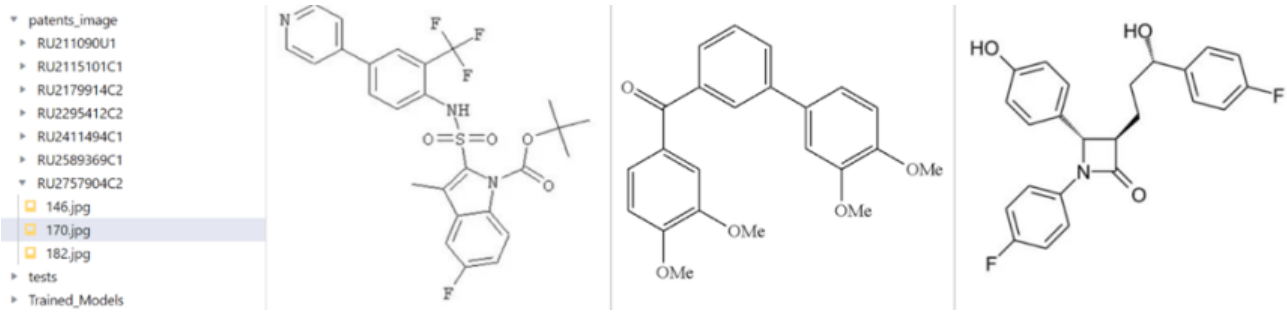

Рис. 11. – Пример изображений патента RU2757904C2

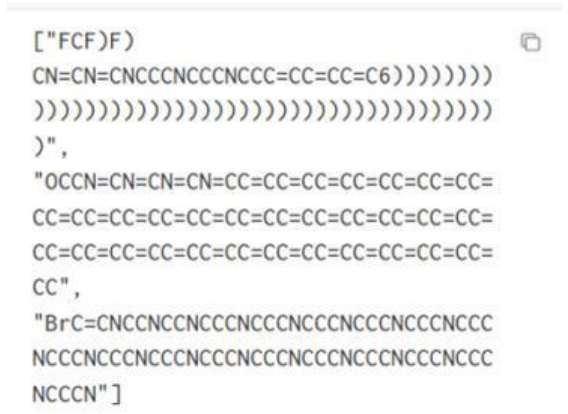

Рис. 12. – Вывод результата конвертации изображений в формат SMILES.

Было проведено тестирование полученной модели машинного обучения, обученной на распознавание химических и математических формул, результаты которого приведены на рис.13 и на рис.14.

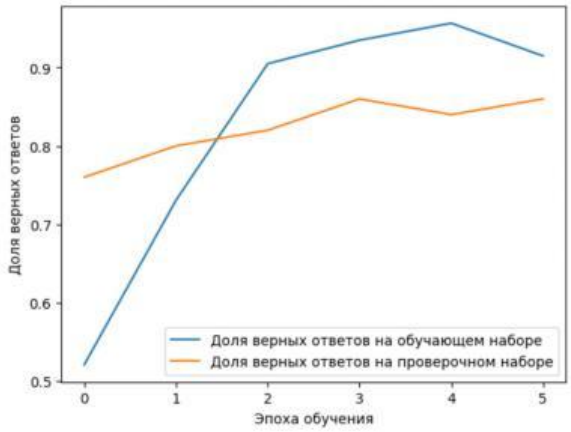

Рис. 13. – Оценка качества верных ответов на обучающем и проверочном наборе данных

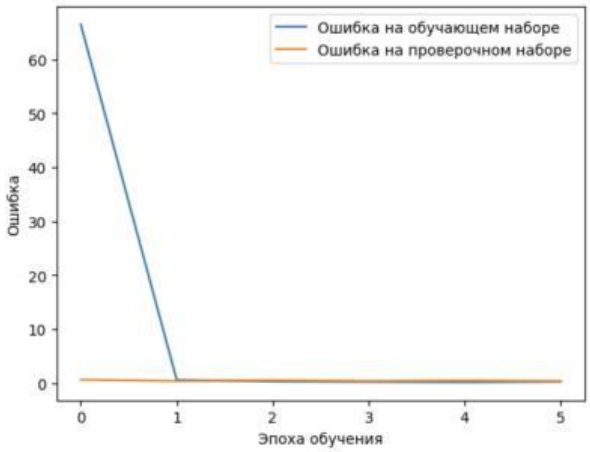

Рис. 14. – Оценка функции ошибки на обучающем и проверочном наборе данных

Заключение

В ходе данной исследовательской работы был успешно разработан, реализован и протестирован модуль для анализа графических изображений, извлеченных из патентных документов. Эта разработка представляет собой важный инструмент в сфере исследования патентов и приведения графических изображений к унифицированному стандарту, что в свою очередь способствует более эффективному информационному поиску. Для дальнейшего усовершенствования и развития этого модуля предлагаются следующие направления:

– Распараллеливание вычислений в блоке парсинга с целью ускорения обработки патентных изображений. Это позволит существенно снизить время анализа и повысить производительность системы;

– Повышение точности распознавания модели, что поможет уменьшить число ошибок при исследовании патентов и повысит надежность результатов;

– Улучшение работы блока классификации путем внедрения дополнительных алгоритмов классификации изображений. Это сделает систему более гибкой и способной к более точному определению типов изображений.

Внедрение данных улучшений позволит значительно повысить эффективность и функциональность разработанного модуля, сделав его еще более ценным инструментом для исследования патентной информации и информационного поиска.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 23-21-00464, https://rscf.ru/project/23-21-00464/.

Литература

1. Бобунов А.В., Коробкин Д.М., Фоменков С.А. Разработка системы информационного поиска для сопоставления с уровнем техники // Моделирование, оптимизация и информационные технологии, 2023. Т. 11, № 3 (42). 15 с. DOI: 10.26102/2310-6018/2023.42.3.023. URL: moitvivt.ru/ru/journal/article?id=1413.

2. Korobkin D.M., Fomenkov S.A., Zlobin A.R., Vereshchak G.A. The Formation of Metrics of Innovation Potential and Prospects // Cyber-Physical Systems Engineering and Control. - Cham (Switzerland): Springer Nature Switzerland AG, 2023. Cc. 17-29.

3. Бобунов А.В., Коробкин Д.М., Фоменков С.А., Васильев С.С. Разработка программного модуля поиска патентов-аналогов // Инженерный вестник Дона, 2022. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/8018

4. Коробкин Д.М., Савельев М.В., Фоменков С.А., Верещак Г.А. Формирование визуализированного представления патентного ландшафта // Инженерный вестник Дона, 2022. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7989.

5. Коробкин Д.М., Фоменков С.А., Верещак Г.А., Сороко М.А. Алгоритм обработки информации для прогнозирования развития технологий на примере теле- и радиовещания // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика, 2023. № 3. Сс. 87-96.

6. Patents // Yandex URL: yandex.ru/patents (дата обращения: 12.02.2023).

7. AWS Textract // Amazon Textract URL: aws.amazon.com/ru/textract (дата обращения: 12.04.2023).

8. Рашка С. Python и машинное обучение. Москва: ДМК Пресс, 2017. 418 с. ISBN 978-5-97060-409-0.

9. Ын А. Теоретический минимум по Big Data. Всё, что нужно знать о больших данных. Санкт-Петербург: Питер, 2019. 208 с. ISBN 978-5-4461-1040-7.

10. Pix2tex – LaTeX OCR // Python Package Index URL: pypi.org/project/pix2tex/ (дата обращения: 12.04.2023).

11. DECIMER-Image-to-SMILES // Github URL: github.com/Kohulan/DECIMER-Image-to-SMILES (дата обращения: 13.04.2023).

12. Чем хорош ClickHouse : главные плюсы быстрой OLAB-СУБД в Big Data // Big Data School URL: bigdataschool.ru/blog/clickhouse-advantages.html (дата обращения: 25.04.2023).

References

1. Bobunov A.V., Korobkin D.M., Fomenkov S.A. Modelirovanie, optimizaciya i informacionny`e texnologii, 2023. V. 11, № 3 (42). 15 p. DOI: 10.26102/2310-6018/2023.42.3.023 URL: moitvivt.ru/ru/journal/article?id=1413.

2. Korobkin D.M., Fomenkov S.A., Zlobin A.R., Vereshchak G.A. Cyber-Physical Systems Engineering and Control. Cham (Switzerland): Springer Nature Switzerland AG, 2023. Pp. 17-29.

3. Bobunov A.V., Korobkin D.M., Fomenkov S.A., Vasil`ev S.S. Inzhenernyj vestnik Dona, 2022. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/8018

4. Korobkin D.M., Savel`ev M.V., Fomenkov S.A., Vereshhak G.A. Inzhenernyj vestnik Dona, 2022. № 11. URL: ivdon.ru/ru/magazine/archive/n11y2022/7989.

5. Korobkin D.M., Fomenkov S.A., Vereshhak G.A., Soroko M.A. Vestnik Astraxanskogo gosudarstvennogo texnicheskogo universiteta. Seriya: Upravlenie, vy`chislitel`naya texnika i informatika, 2023. № 3. Pp. 87-96.

6. Patents. Yandex URL: yandex.ru/patents (accessed: 12.02.2023).

7. AWS Textract. Amazon Textract URL: aws.amazon.com/ru/textract (accessed: 12.04.2023).

8. Rashka S. Python i mashinnoe obuchenie [Python and Machine learning]. Moskva: DMK Press, 2017. 418 p.

9. En A. Teoreticheskij minimum po Big Data. Vsyo, chto nuzhno znat` o bol`shix danny`x [Theoretical minimum for big data. All you need to know about big data]. Sankt-Peterburg: Piter, 2019. 208 p.

10. Pix2tex – LaTeX OCR. Python Package Index URL: pypi.org/project/pix2tex/ (accessed: 12.04.2023).

11. DECIMER-Image-to-SMILES. Github URL: github.com/Kohulan/DECIMER-Image-to-SMILES (accessed: 13.04.2023).

12. Chem xorosh ClickHouse : glavny`e plyusy` by`stroj OLAB-SUBD v Big Data [What is good about ClickHouse: the main advantages of a fast OLAP

database in Big Data]. Big Data School URL: bigdataschool.ru/blog/clickhouse-advantages.html (accessed: 25.04.2023).

Дата поступления: 3.11.2023

Дата публикации: 8.12.2023