Применение свёрточной нейронной сети для решения проблемы регистрации скан-копий документов в электронном архиве

Екатерина Волгина (РТУ МИРЭА)

Выявлена проблематика, возникающая при регистрации в электронном архиве «ЭЛЬДОКа» дублей скан-копий документов, и приведены варианты решения, описана архитектура свёрточной нейронной сети, а также указаны её преимущества и недостатки.

Введение

Электронный архив документов (ЭЛЬДОКа) – программное обеспечение для каталогизации проектной документации на предприятии и регламентации доступа к ней.

Целевым назначением ИТ-решения «ЭЛЬДОКа» является:

- создание сетевого программного обеспечения в форме веб-приложения для ввода, хранения и обработки информации в рамках задачи информационного обеспечения деятельности Компании заказчика;
- ведение электронного архива Компании;
- структурированное хранение и связь документов;
- обеспечение объектно-ролевого доступа к материалам;
- возможность визуализировать документы различных форматов;
- увеличение эффективности и повышение качества контроля над выполнением работ.

Платформа АТОЛЛ обеспечивает единое управление всей поступающей информацией, включая внутренние сервисы поддержания целостности, и реализует возможность предоставления доступа к данным.

Концепция работы электронного архива систематизирует файловые Хранилища и данные из реляционных систем, собирая их виртуально в единую КАРТОЧКУ ОБЪЕКТА (или ДОКУМЕНТА). Многокритериальный поиск в РЕЕСТРЕ КАРТОЧЕК нужной информации в файлах карточек и ее ВИЗУАЛИЗАЦИЯ обеспечивается атрибутивным бизнес-значимым описанием файлов и структурированием их по разделам карточки (рис. 1).

Функциональность системы заключается в автоматизации перечисленных ниже задач бизнеса в разных бизнес-процессах, главными атрибутами которых являются документы:

- формирование электронного архива документов – создание карточек документов и ведение метаданных (каталогизация данных), регистрация файлов для архивного хранения:
- поиск электронных документов;
- работа с электронными документами – визуализация содержимого электронных документов, работа с реестрами электронных документов, выгрузка данных и массовое редактирование карточек документов;
- управление электронным архивом. Задача организации и настройки хранилища электронных документов, управление сервисами, которые обеспечивают работу с архивными электронными документами, доступ к ним, журналирование работы с архивом;
- администрирование;
- настройка электронного архива;
- Аудит действий пользователей.
- подготовка графической отчётности. Функциональный модуль с набором функций для ведения графических данных, в том числе описывающих объекты предметной области;
- паспортизация объектов. Функциональный модуль с набором функций для представления отраслевых данных по объекту (объекту паспортизации) через настраиваемую иерархи-

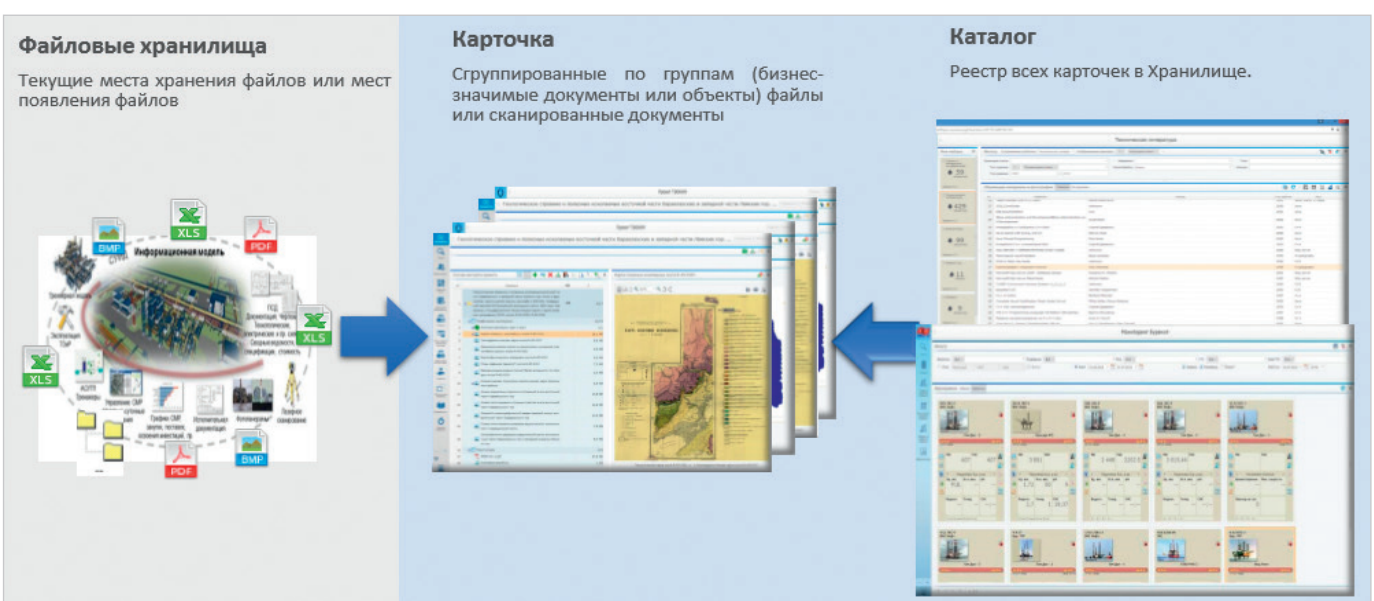

Рис. 1. Концепция работы электронного архива

ческую структуру разделов паспорта и отраслевые визуализаторы.

Цель работы: найти методы разрешения проблематики регистрации дублей скан-копий документов, обучить и протестировать свёрточную нейронную сеть.

Поставленные задачи:

- найти способы обхода недостатков применения метода сравнения скан-копий документов по контрольной сумме и текстовому содержимому;
- провести анализ библиотек на языке Python для найденных способов.

Предпосылки работы

При регистрации документов в электронном архиве разные пользователи зачастую регистрируют одни и те же документы под разными именами и в разных ветках иерархии документов. Это приводит:

- к увеличению расходов на системы хранения документов;
- рассинхронизации разных копий документов при обновлении их версии;
- снижению уровня доверия к электронному хранилищу документов.

Ещё большей проблемой является регистрация разных скан-копий одного и того же документа, так как с точки зрения файловой системы это будут совершенно разные файлы.

Основная идея

Основная задача электронного архива «ЭЛЬДОКа» – каталогизация и классификация документов заказчика в привязке к модели предметной области. При этом каждый документ должен регистрироваться только один раз. Если возникает необходимость привязки того же документа к другой «ветви» или «листу» модели предметной области, то в требуемых местах должна размещаться ссылка на уже загруженный документ.

Требования к регистрации скан-копий документов аналогичны требованиям к размещению документов в офисных форматах. При этом определение дублей скан-копий документов допустимо выполнять в асинхронном режиме, выводя пользователю результат поиска дублей, предоставляя право пользователю подтвердить схожесть скан-копий документов и самому определить место размещения первичной скан-копии.

Методика работы

Существует несколько вариантов решения проблемы поиска дублей скан-копий документов в архиве.

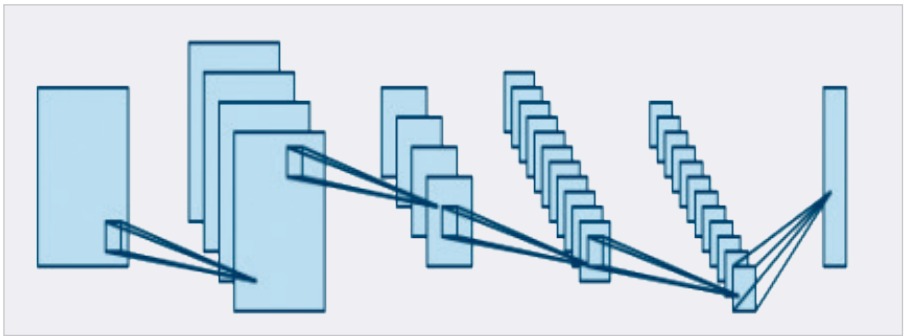

Рис. 2. Структура нейронной сети

1. Поиск дублей документов посредством сравнения по контрольной сумме файлов.

Контрольная сумма файлов — это последовательность цифр и букв, которая используется для проверки данных на наличие ошибок. Если заранее известна контрольная сумма исходного файла, можно использовать служебную программу контрольной суммы, чтобы подтвердить, что сравниваемый файл идентичен исходному.

Чтобы получить контрольную сумму, необходимо запустить программу, которая обработает файл с помощью алгоритма. Типичные алгоритмы, которые используются для этого, включают MD5, SHA-1, SHA-256 и SHA-512.

Алгоритм использует криптографическую хеш-функцию, которая принимает входные данные и создаёт строку (последовательность цифр и букв) фиксированной длины.

Внутри электронного архива реализован алгоритм хеширования MD5. MD5 – в настоящее время является одним из самых распространённых способов защитить информацию в сфере прикладных исследований, а также в области разработки веб-приложений.

Утилита md5sum, предназначенная для хеширования данных заданного файла по алгоритму MD5, возвращает строку. Она состоит из 32 цифр в шестнадцатеричной системе счисления. То есть хеш, полученный от функции, работа которой основана на этом алгоритме, выдает строку в 16 байт (128 бит) . И эта строка включает в себя 16 шестнадцатеричных чисел. При этом изменение хотя бы одного её символа приведёт к последующему бесповоротному изменению значений всех остальных битов строки.

2. Поиск дублей скан-копий документов по схожести содержимого с помощью нейронных сетей.

Для решения задач классификации изображений (распознавания образов) повсеместно используются свёрточные нейронные сети, которые являются разновидностью моделей глубокого обучения. Свёрточные нейронные сети обеспечивают частичную устойчивость к изменениям масштаба, смещениям, поворотам, смене ракурса и прочим искажениям. В основе свёрточной нейросети лежит Deep Learning-алгоритм, который может принимать входное изображение, присваивать важность (усваиваемые веса и смещения) различным областям/объектам в изображении и может отличать одно от другого.

Для достижения цели распознавания скан-копий и вывода ссылки на уже существующую скан-копию необходимо обучить нейронную сеть.

Обучение свёрточной нейронной сети для классификации изображений

Когда работа будет вестись с участием нейронной сети, то возникает необходимость сбора данных для обучения. Обучающий набор данных представляет собой набор наблюдений, для которых указаны значения входных и выходных переменных. В данном случае используются наборы данных, основанных на скан-копиях проектной документации и документов заказчика.

Архитектура свёрточной нейронной сети

Нейронная сеть состоит из пар слоёв – слоёв подвыборки и слоёв свёртки, каждый из которых, в свою очередь, состоит из карт признаков. Нетрудно убедиться в том, что каждая карта признаков в идеале фильтрует изображение, находя какой-то один определённый, специфичный для данной карты, признак (рис. 2).

Исходное изображение подаётся на входной слой. В первом слое подвыборки каждая карта признаков осуществляет поиск определённого, закреплённо-

го только за данной картой, признака. Достигается это за счёт использования общих для всей карты признаков матрицы весов и особой организацией локального рецептивного поля для каждого нейрона такой карты. Каждый нейрон карты признаков получает входные данные от прямоугольной области размера n×m входного изображения.

Смежные нейроны карты признаков получают в качестве входного воздействия смежные прямоугольные области, причём весовые коэффициенты для всех нейронов карты признаков будут одинаковыми.

Для введения инвариантности нейронной сети к смещениям и небольшим деформациям используется слой свёртки. Для каждой карты признаков существует соответствующая ей карта свёртки, которая уменьшает размерность карты признаков с n×m до n/2×m/2 путём усреднения значений по квадрату 2×2 нейронов. После выполнения свёртки сеть теряет часть информации о точном положении найденного признака, но сохраняет информацию относительно взаимного расположения различных признаков. Следующий слой подвыборки осуществляет аналогичную первому слою сегментацию входных данных на прямоугольные области n×m, только входными данными второго слоя служит выход первого слоя. Таким образом, каждая карта признаков второго слоя осуществляет поиск признаков второго порядка одновременно во всех картах признаков первого слоя.

Свёрточной нейронной сети с тремя парами слоёв подвыборки-свёртки вполне достаточно для точного распознавания лиц.

Достоинства СНС1

1. Уменьшение количества обучаемых параметров и повышение скорости обучения по сравнению с полносвязной нейронной сетью.
2. Возможность распараллеливания вычислений и реализации алгоритмов обучения сети на графических процессорах (GPU).
3. Устойчивость к сдвигу позиции объекта во входных данных. При обучении свёрточная нейронная сеть сдвигается по частям объекта. Поэтому обучаемые признаки не зависят от позиции «важных частей». Это свойство свёрточной нейронной сети помогает повышать качество классификации.

Недостатки СНС

1. Высокая сложность архитектуры.
2. Фиксированная площадь окна слоя свёртки.
3. Полносвязность.

Полученные результаты

- Выявлена проблематика регистрации скан-копий внутри электронного архива и изучены варианты решения проблемы.
- Обучена свёрточная нейронная сеть для классификации изображений.
- Проведено тестирование нейронной сети.
- Проведён поиск дублей скан-копий документов.

Заключение

Даны определения свёрточным нейронным сетям (СНС). Проведены этапы разработки нейронной сети и её работы, её реализация кода на языке программирования Python.

Литература

1. *Шолле Ф.* Глубокое обучение на Python. СПб.: Питер, 2019.
2. Свёрточные нейронные сети для распознавания изображений // URL: https://libeldoc.bsuir.by/bitstream/123456789/39033/1/Prokopenya_Svertochnyye.pdf.
3. Архитектура обобщённых свёрточных нейронных сетей // URL: http://www.it-visnyk.kpi.ua/wp-content/uploads/2012/08/54_36.pdf.
4. Свёрточная нейронная сеть для решения задачи классификации // URL: https://mipt.ru/upload/medialibrary/659/91_97.pdf.

НОВОСТИ МИРА

Путин назвал результаты «Ростеха» скромными и раскритиковал развитие 5G и микроэлектроники

Президент России Владимир Путин оценил результаты госкорпорации «Ростех» и направлений, которые она курирует. Российский лидер назвал их скромными, передаёт ТАСС.

«Я знаю, там сейчас коллеги будут говорить, финансирования не хватает, ещё что-то. Но я просто констатирую сам факт того, что происходит», – сказал Путин в ходе заседания Совета по стратегическому развитию и нацпроектам.

Глава государства, в частности, остался неудовлетворён результатом проекта по созданию сетей 5G, развитию производства оборудования для широкого внедрения интернет-вещания. Он отметил, что в прошлом году программу по развитию микроэлектронной отрасли пришлось полностью перезагружать.

В сентябре 2021 года сообщалось, что Минцифры планирует внедрить связь пятого поколения в городах-миллионниках к 2024 году. В министерстве назвали 5G критической технологией, которую в России приняли решение развивать своими силами.

Кроме того, Путин поручил привлечь отечественных инвесторов для развития местных высокотехнологических компаний вроде «Яндекса», ранее финансировавших своё развитие за счёт западного капитала. Об этом он заявил на заседании Совета по стратегическому развитию и нацпроектам, передаёт РБК.

Он признал, что российская финансовая система не в состоянии обеспечивать функционирование компаний, не имеющих активов, но с большими перспективами развития. В связи с этим глава государства поручил сформировать механизмы, которые позволят компаниям привлекать частный капитал на российском рынке.

Какие механизмы имеются в виду, Путин не уточнил. В настоящее время «Ozon», «Яндекс» и другие технологические компании торгуются на Московской бирже. Во всём мире именно такие условия называются возможностью привлекать частный капитал. По всей видимости, речь идёт о стимулах или обязательствах по покупке акций или облигаций компаний.

Российский фондовый рынок рухнул более чем в два раза с октября прошлого года, когда западные страны начали ожидать вхождения российских войск на территорию Украины. По состоянию на 18 июля индекс Мосбиржи находится ниже 2100 пунктов.

russianelectronics.ru

Свобода проектирования

DeltaDesign

В состав Delta Design, обеспечивающей сквозной цикл проектирования печатных плат, входят модули:

- Менеджер библиотек
- Схемотехнический редактор
- Схемотехническое моделирование
- HDL-симулятор
- Редактор правил
- Редактор печатных плат
- Топологический редактор плат TopoR
- Коллективная работа для предприятий

EREMEX +7 (495) 232-18-64 • sales@eremex.ru • www.eremex.ru