Описание функциональных характеристик

программа для ЭВМ
«Искусственный интеллект «АТОМ»

Термины

XML	расширяемый язык разметки
C++	компилируемый, статически типизированный язык программирования общего назначения
Uuid	стандарт идентификации, используемый в создании программного обеспечения, стандартизированный Open Software Foundation (OSF) как часть DCE — среды распределённых вычислений

Аннотация

Настоящий документ содержит описание технических решений на основе существующих компетенций ЗАО «Сервис-Газификация» в области автоматизации распознавания и структурирования информации, реализованных в программе для ЭВМ «Искусственный интеллект «АТОМ» (далее по тексту равнозначны полному наименованию: «ИИ «АТОМ», «ИИ», «приложение «АТОМ»»).

Описание программной платформы ИИ «АТОМ»

Программная платформа представляет собой набор системных инструментов для сканирования и распознания документов с последующим структурированием информации. Основным программным компонентом используется Openframeworks C++, реализующий типовые функции информационной системы таких как чтение изображения, организации потоков, чтение/запись xml. Также используются библиотеки OpenCV, OpenCL, DeepCL, PyTorch C++/CUDA, MuPDF, Tesseract.

Область применения платформы ИИ «АТОМ»

На данный момент платформа используется для построения искусственного интеллекта на языке C++. За основу берется клиент-серверная архитектура. На основе данной архитектуры могут быть построены как приложения для закрытых сетей, так и публичные, доступные пользователям через интернет, так и гибридные варианты с распределением прав доступа к различным частям и данным приложения со стороны различных групп пользователей.

Описание прикладных решений ИИ «АТОМ»

Цель разработки

Автоматизация проверки документации

Оптимизация трудоресурсов при работе с бумажным документооборотом

Цифровизация и типизация документации

Решаемые задачи

Цифровазация неограниченого количества документации

Автоматическая типизация неограниченного количества отцифрованной документации

Извлечение необходимых атрибутов отцированной документации

Автоматизация создания реестров отцифрованной документации

Выявление несоотвествий полученных результатов из реестров

Оптимизация трудозатрат

Область применения

Любая классифицируемая документация

Описание модулей платформы ИИ «АТОМ»

AIDocProcessor – основное приложение осуществляющее извлечение данных.

Основные модули/алгоритмы:

Модуль потоковой загрузки PDF файлов. Основная задача: загрузить пакет pdf, представить в памяти в виде изображений 3 масштабов: изображение исходное, изображение для нейронного классификатора и изображение для нейронного сегментатора. По мимо этого выполняется коррекция ориентации изображения по данным тессеракта и доворот скана если присутствует небольшой угол поворота. Также выполняется инициализация модуля AIRectRefiner в потоке для каждой страницы. На вход берет путь к каталогу PDF на выходе создается массив PageRawData* - страниц с изображениями. В процессе обработки эти структуры «обрастают» найденными полями с текстом.

Модуль AIRectsRefiner. Модуль компьютерного зрения, осуществляющий сегментацию страницы – исходного изображения скана. Сегментация изображения на строки. Этот модуль используется как вспомогательное средство при уточнении координат текстовых областей на различных стадиях обработки документа. Например, найденных текстовых областей при помощи нейронной сегментации и табличного процессора.

Модуль потокового полнотекстового распознавания текста. На вход получает массив PageRawData* и для каждой страницы выполняется полное

распознавание тессеректом. Все найденные tesseract'ом данные записываются в виде иерархии узлов. Эта иерархия узлов хранится внутри PageRawData*. Где каждый узел – это блок текста, строка, слово. Для каждого слова и строки хранятся распознанные текстовые значения слов и прямоугольные координаты.

Модуль нейросетевой классификации документа. Работает в основном потоке на GPU. На вход получает массив PageRawData* и для каждой страницы берется её уменьшенная копия скана. Это уменьшенная копия типизируется нейросетью. По сути, для каждой страницы проставляется индекс - некий номер/тип документа.

Модуль нейросетевой сегментации документа. Работает в основном потоке на GPU. На вход получает массив PageRawData* и для каждой страницы берется её уменьшенная копия скана. Для этой уменьшенной копии нейросеть определяет области интересующего нас текста. Фактически формируются координаты прямоугольной области. Для каждой области проставляется класс текстового поля. Формируется массив AiSegmentationOut который хранит распознанные прямоугольные области.

Модуль получения данных по uuid документа. Обобщенно можно сказать, что это набор классов и функций который обеспечивают получение нужных параметров относительно интересующих нас документов и их текстовых полей согласно их uuid. Данные собираются из различных конфигурационных файлов. Но основные файлы – settings.xml и docTypes.xml. Именно этот модуль позволяет назначать те или иные алгоритмы/модули или скрипты который будут применятся в последствии к документу/странице/текстовому полю.

Модуль потокового распознавания текста согласно прямоугольным областям. На вход получает массив AiSegmentationOut для каждого PageRawData* и для каждого прямоугольного поля выполняется распознавание тессеректом. Все найденные tesseract'ом данные записываются в AiSegmentationOut.

Модуль текстовой верификации типа документа. Работает параллельно, распараллеливание по страницам. Рассматриваются узлы с текстом для каждого PageRawData*. Данный модуль проверяет наличие ключевых текстовых признаков документа согласно выявленному нейросетью типу страницы.

Алгоритм кластеризации страниц в документы. Данный модуль собирает страницы в документ. Работает в основном потоке. Для этого должны быть соблюдены несколько правил. 1 – должен быть найден ключевой признак на первой странице. Как правило, это номер документа. Этот признак должен быть уникален и не должен повторяться в рамках одного документа. 2 – Все

последующие страницы должны пройти текстовую верификацию по признакам текста. Как только перестанет подтверждаться 1 или 2 пункт – формируется документ из отобранных страниц.

Модуль извлечения текста на основе анализа текста по нечетким признакам. Модуль работает потоково. Распараллеливание на уровне каждого документа. То есть берется текст со всех страниц, относящихся к одному документу. Сложный скриптовый алгоритм, который работает иерархически и находит ключевые текстовые признаки в найденных текстовых узлах страницы PageRawData*. Предварительно все иерархические узлы преобразуются в набор линий с сохранением их пространственных координат. После чего он забирает текстовую информацию согласно заданному сценарию. Сценарий определяет то, какой текст относительно признака будет собран. Текст может быть взять между признаками или над/под признаком. Также можно взять ближайший подчеркнутый текст относительно найденного признака. Более того найденный текст можно будет снова прогнать через текстовое извлечение и так неограниченное количества раз. Это нужно для того, чтобы достать конкретное значение в общем предложении или блоке текста. Именно поэтому этот алгоритм мы называем иерархическим методом извлечения текста на основе нечеткого сравнения строк. Результат работы алгоритма – найденные прямоугольные области с интересующим нас текстом.

Модуль табличного процессора. Модуль компьютерного зрения, осуществляющий детекцию и распознавание таблиц. Согласно заданному для данного типа документа скрипту, система ищет таблицу по логическим критериям. А именно размерность таблицы, количество столбцов и строк по принципу задания логических условий. Например, ищем таблицу, где строго 5 столбцов и больше 2 строк**.** Затем для каждой такой найденной таблицы определяется правило сохранения данных. Можно сохранить конкретные ячейки, строки, столбцы или всю таблицу. Можно найти столбец с нужным именем и взять все ячейки или конкретную ячейку. По сути, результат работы этого модуля схож с нейросетевой сегментацией – на вход берется изображение, а возвращаются прямоугольные области с определенным типом, для которых можно применить распознавание текста.

Модуль извлечения машиночитаемых документов. Комбинирует в своей работе подходы, примененные в модуле табличного процессора и в модуле извлечения текста. Ориентирован на работу с документами, которые создавались в машиночитаемом формате – четкая структура, большая часть данных в таблицах фиксированной размерности.

Модуль OcrNet. Модуль, осуществляющий построчное распознавание текста. Был разработан нейросетевой модуль, осуществляющий распознавание текста. Берет на вход прямоугольную область с изображением. Возвращает строку с распознанным текстом.

Модуль улучшения качества изображения aiImageImrpove. Улучшает изображение с текстом. Специально обученная нейросеть на синтетических изображениях. Она убирает шумы и восстанавливает поврежденные символы на изображении. Берет на вход изображение, возвращает изображение той же размерности.

Модуль сохранения данных. Осуществляет сохранения данных согласно требованиям заказчика. Сохраняет данные 3 видов: xml данные с распознанными данными страниц. Каждый xml файл – отдельная страница. Csv файл – где каждая строка распознанные данные с документа/страницы. А также Csv файл на каждый распознанный документ, где извлечены все данные по документу. PDF файлы, которые были сформированы согласно найденным и обработанным входным файлам.

Модуль пост. обработки текста. Применяется к конкретным полям документа. Для каждого поля документа может быть назначено не ограниченное количество различных инструкций по пост. обработки файла. Начиная с регикса и заканчивая различными замещениями фрагментов текста по словарю и любые другие произвольные конверсии текста.

Модуль группировки документов. Осуществляет группировку документов - склейка реестров и pdf документов разного типа в один файл согласно настройкам. Работает в основном потоке.

Внешние вспомогательные утилиты:

1. **TrainUnetTextSegmentation –** приложение, осуществляющее обучение нейросети-сегментатора детекции областей текста на изображении. Где каждое изображение – одна страница. Каркас - Openframeworks. В основе нейросеть построенная на фреймворке PyTorch.
2. **PageClassificator** – приложение, осуществляющее обучение нейросети-классификатора. Обучает нейросеть типизировать страницу документа по ее изображению. Каркас - Openframeworks. В основе нейросеть построенная на фреймворке DeepCL.
3. **Модуль обучения нейросети aiImageImrpove -** приложение осуществляющее обучение нейросети которая улучшает качество изображения скана для определенных текстовых областей.
4. **Модуль обучения нейросети OcrNet** - модуль, осуществляющий обучение нейросети OcrNet.

5. **Модуль DatasetUtils** - утилита, которая позволяет собирать бинарные датасеты для нейросетей из размеченных изображений. Обладает различные функциями редактирования датасетов. (Объединение, конвертация, нарезка и прочее)

6. **Модуль IDP** – модуль осуществляет ограниченный просмотр обработанных AIDocProcessor'ом данных. Также при помощи него осуществляется разметка датасетов для нейросети сегментации и классификации. Из этих сырых данных потом формируется бинарный датасет на которых обучается та или иная нейросеть.

7. **Модуль AtomGUI** – развитие модуля IDP, здесь осуществляется полноценный просмотр обработанных данных, есть возможность их редактировать и экспортировать. Также здесь можно исправить ошибки классификации и кластеризации и повторно извлечь данные из документов с учетом этих исправлений.

Особенности разработки платформы ИИ «АТОМ»

Разрабатываемое решение содержит различные уникальные алгоритмы обработки данных в области компьютерного зрения и искусственного интеллекта.

Используются свои технологические решения для типизирования документов и извлечения данных.

Были выработаны концепции извлечения признаков и типизации документов на основе триплетной функции потери и строкового сходства Дамерау-Левенштейна.

Была выработана концепция непрерывной аугментации при обучении нейронных сетей.