УДК 004.85

ИССЛЕДОВАНИЕ И АДАПТАЦИЯ МЕТОДОВ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ ДЛЯ ТИПИЗАЦИИ ГРАФИКОВ В ТЕКСТОВЫХ ДОКУМЕНТАХ

Сашин В. А.

(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Научный руководитель – кандидат физико-математических наук, доцент факультета СуиР, доцент ВШ ЦК Бойцев Антон Александрович

(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Данная работа вносит вклад в решение глобальной актуальной проблемы извлечения информации из текстовых документов, не являющихся текстовыми файлами (например, pdf файлов). Работа сфокусирована на решении задачи типизации (классификации) графиков среди всех иллюстраций, которые могут быть распознаны на предшествующем этапе извлечения информации – детектировании объектов на страницах документа (текст, таблицы, иллюстрации и т.д.). Предложенное решение позволяет эффективно типизировать графики, тем самым предоставляя больше информации для извлечения.

Введение. В настоящее время существует и обрабатывается колоссальное количество текстовых документов, не являющихся текстовыми файлами. Такие файлы, например, pdf-файлы, представляют собой набор изображений, каждое из которых отражает фрагмент документа с текстом, иллюстрациями и другими компонентами. Основной проблемой при работе с такими документами является недоступность содержащейся в них информации, невозможность ее автоматического получения, обработки, поиска, анализа и т.д. Поскольку документы такого типа представлены наборами изображений, решение описанной проблемы сводится к извлечению информации из изображений и их компонентов – комплексной задаче, которую целесообразно декомпозировать на распознавание и дальнейший анализ распознанных объектов, таких как текст, заголовки, таблицы, различные виды иллюстраций и другие. Данное исследование посвящено второму этапу – анализу распознанных на документах объектов, и сфокусировано на конкретной задаче типизации (классификации) графиков (гистограмм, круговых, рассеяния и других), среди иллюстраций, распознанных на страницах документа. Как уже было отмечено, решение данной задачи позволит внести вклад в комплексную задачу экстракции информации из текстовоых документов.

Основная часть. Поставленная задача сводится к задаче классификации изображений, представляющих собой иллюстрации текстовых документов, и для ее решения предлагается применить современные архитектуры глубоких сверточных нейронных сетей, которые в настоящее время представляют собой самый популярный метод классификации изображений, являющийся одним из самых эффективных. Более того, для улучшения качества классификации графиков предлагается разделить данную задачу на 2 этапа и использовать две модели: одну для бинарной классифкации иллюстраций текстовых документов с целью отделения графиков от прочих иллюстраций, а вторую – для многоклассовой классификации тех изображений, которые были классифицированы предыдущей моделью как графики. Описанный подход не только положительно влияет на качество классификации, но и позволит в дальнейшем легче масштабировать полученное решение на поддержку большего количества типов иллюстраций.

Основными этапами работы были:

1. поиск, сбор и подготовка наборов данных для обучения моделей;
2. проведение экспериментов по обучению различных архитектур глубоких сверточных нейронных сетей с применением лучших практик глубокого обучения;

3. внесение правок в данные на основе результатов экспериментов и их повторное проведение.

Для решения поставленной задачи из различных источников были сформированы 2 набора данных, которые в ходе проведения экспериментов были модифицированы. Набор данных для бинарной классификации содержал в классе «не график» естественные изображения, таблицы и иллюстрации, вручную собранные из открытых источников, и схемы (не графики) из датасета AI2D, а в классе «график» – подмножество датасета ICPR CHART-Infographics 2020 - UB PMC, также дополненное вручную. Датасет для многоклассовой классификации был основан на ICPR CHART-Infographics 2020 - UB PMC, однако 2 класса из исходных 15 были объединены в один из-за неточностей в разметке и отсутствия практической необходимости различать эти 2 класса; данный набор данных также был дополнен вручную.

Алгоритм экспериментов включал применение лучших практик глубокого обучения и был основан на технике transfer learning:

1. разделение набора данных на обучающую, валидационную и тестовую выборки в соотношении 60/20/20, нормализация параметрами даатсета ImageNet, изменение размера изображений;
2. применение аугментаций;
3. загрузка весов предобученной на ImageNet модели, модификация последнего блока сети, «заморозка» всех слоев модели кроме последнего блока (параметры «замороженных» слоев не оптимизируются в процессе обучения);
4. нахождение оптимального learning rate по методу Cyclical Learning Rates;
5. обучение нескольких эпох (~5) по One Cycle Policy, выбор лучшего состояния модели;
6. «разморозка» всех слоев и подбор оптимального learning rate по методу Cyclical Learning Rates;
7. обучение по методу Discriminative Fine-Tuning с One Cycle Policy;
8. получение предсказаний на тестовом множестве и расчет окончательной метрики классификации (accuracy для бинарной классификации, f-score для многоклассовой).

Выводы. Предложенное решение качественно отличает изображения графиков от прочих иллюстраций текстовых документов, а также производит типизацию изображений графиков на 14 видов, что позволит извлекать больше информации из текстовых документов на этапе анализа объектов (а именно иллюстраций), распознанных и вырезанных из страниц документа. В рамках работы были прозведены сбор и модификация наборов данных, а также проведение экспериментов по обучению популярных архитектур глубоких сверточных нейронных сетей, таких как ResNet, VGG, ResNeXt, DenseNet, Wide ResNet, EfficientNet, RegNet, с использованием современных методов глубокого обучения. В результате лучший результат по бинарной классификации графиков был достигнут с моделью ResNet-152 с accuracy = 0.975 на тестовом множестве. По многоклассовой классификации графиков на данный момент был достигнут показатель f-score, равный 0.89, при помощи модели архитектуры ResNeXt-50, однако планируется провести больше экспериментов с целью улучшения результата многоклассовой классификации.

Сашин В.А. (автор)

Бойцев А.А. (научный руководитель)