АВТОМАТИЗАЦИЯ ОБУЧЕНИЯ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ОБНАРУЖЕНИЯ ТАБЛИЦ НА ИЗОБРАЖЕНИЯХ ДОКУМЕНТОВ.

И. А. Черепанов, А. А. Михайлов

Изображения документов хранят большое количество ценных табличных данных, которые могут быть востребованы различными научными и практическими приложениями. Для использования в дальнейшей обработке такая информация должна быть представлена в текстовой и теговой форме. Сообщество Document Engineering рассматривает задачу (table extraction) извлечения таблиц из документа и представления их в формате, похожую на электронную таблицу. Для извлечения таблицы необходимо сначала обнаружить её область расположения, а после распознать структуру.

С появлением нейронных сетей, начали развиваться методы распознавания и извлечения, такие как двойная классификация на основе свёрточных нейронных сетей [3], обученные модели на обнаружение объектов [4] и семантическая сегментация [5]. При этом есть и старые методы. Например, метод, основанный на эвристиках [1] и метод машинного обучения [2].

Процесс обучения моделей обнаружения таблиц в PDF документах на основе глубоких нейронных сетей включает в себя множество манипуляций с данными, таких как объединение разных наборов данных, преобразование изображений, аугментация и т.д. Также на качество обученной модели влияют параметры конфигурации. Таким образом для получения лучшей модели (с высокой точностью и полнотой) требуется проводить много экспериментов с разными параметрами обучения, данными, способами преобразования и аугментации. Весь этот процесс затратен по времени и требует большое количество ручных действий, начиная процессом подготовки данных и заканчивая получения оценки модели.

Рассматривается проблема автоматизации всего процесса обучения. Для решения данной проблемы был спроектирован рабочий процесс получения модули с помощью метода глубокого машинного обучения. Данный процесс включает в себя последовательность действий: преобразование датасетов в унифицированный формат PASCAL VOC, преобразование изображений, аугментация данных, преобразование в формат TensorFlow record, обучение и тестирование работы модели. Весь процесс автоматизации был реализован в виде управляющего скрипта на языке программирования Python (https://github.com/tabbydoc/dl4td). Данный скрипт запускает в определённом порядке ряд других скриптов, выполняющих определённый этап: преобразования изображения, аугментации данных и т.д. Управляющий скрипт легко настраивается с помощью конфигурационного файла, в котором указываются пути к датасетам, выходные пути (куда сохранить унифицированный датасет и полученные TF records файлы), определить пути к скриптам (выполняющие преобразование датасетов к унифицированному формату, изображений, аугментацию данных и т.д.) и ещё ряд важных параметров. Помимо лёгкой настройки есть возможность пропустить тот или

иной этап. Например, не выполнять аугментацию данных или преобразование датасета.

С помощью данного решения была обучена модель на наборе датасетов Marmot, ICDAR-2017, UNLV, SciTSR, которая дала оценку полноты – 98.4% и точности – 91.2%. Модель обучалась 200 000 шагов с помощью фреймворка TensorFlow Object Detection API.

В будущем планируется добавить поддержку новых датасетов.

Литература

1. Perez-Arriaga, M., Estrada, T., Abad-Mota, S.: TAO: System for table detection and extraction from pdf documents (2016), https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12916

2. Bansal, A., Harit, G., Roy, S.D.: Table extraction from document images using fixed point model. In: Indian Conf. on Computer Vision Graphics and Image Processing. pp. 67:1–67:8. ICVGIP ’14 (2014). https://doi.org/10.1145/2683483.2683550

3. Hao, L., Gao, L., Yi, X., Tang, Z.: A table detection method for pdf documents based on convolutional neural networks. In: 12thIAPR Workshop on Document Analysis Systems. pp. 287–292 (2016). https://doi.org/10.1109/DAS.2016.23

4. Arif, S., Shafait, F.: Table detection in document images using foreground and background features. In: Digital Image Computing: Techniques and Applications. pp.1–8(2018). https://doi.org/10.1109/DICTA.2018.8615795

5. He, D., Cohen, S., Price, B., Kifer, D., Giles, C.L.: Multi-scale multi-task fcn for semantic page segmentation and table detection. In: 14thIAPR Int. Conf. on Document Analysis and Recognition. vol. 01, pp.254–261 (2017). https://doi.org/10.1109/ICDAR.2017.50