Заключение

В данной работе был проведен обзор существующих подходов к распознаванию текста на изображениях. В результате был поставлен эксперимент по сравнению двух моделей на вырезанных словах с товарных ценников: STAR-net и tesseract-ocr. По полученным метрикам видно, что наилучший результат показала модель STAR-net. Анализ ошибок показал, что если модель и ошибается, то делает чаще всего это в одном символе, что может говорить о высоком результате распознавания. Также были замечены изображения, на которых разметка не соответствовала действительности. При этом tesseract-ocr часто путал схожие символы, а также не находил часть текста. В качестве улучшения можно предложить попробовать другие параметры обучения модели STAR-net. Для tesseract-ocr же можно попробовать препроцессинг данных, а также дообучить его на датасете.

Список литературы

1.Baek, Jeonghun, et al. “What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, doi:10.1109/iccv.2019.00481.

2.Graves, Alex, et al. “Connectionist Temporal Classification.” Proceedings of the 23rd International Conference on Machine Learning - ICML '06, 2006, doi:10.1145/1143844.1143891.

3.He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi:10.1109/cvpr.2016.90.

4.Liu, Wei, et al. “STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition.” Procedings of the British Machine Vision Conference 2016, 2016, doi:10.5244/c.30.43.

5.Liu, Xuebo, et al. “FOTS: Fast Oriented Text Spotting with a Unified Network.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, doi:10.1109/cvpr.2018.00595.

6.Stanford cs class cs231n: Convolutional neural networks for visual recognition. http://cs231n.github.io/neural-networks-case-study/.

7.Tesseract OCR, https://tesseract-ocr.github.io/tessdoc/

8.Karatzas, Dimosthenis, et al. “ICDAR 2013 Robust Reading Competition.” 2013 12th International Conference on Document Analysis and Recognition, 2013, doi:10.1109/icdar.2013.221.

9.Karatzas, Dimosthenis, et al. “ICDAR 2015 Competition on Robust Reading.” 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, doi:10.1109/icdar.2015.7333942.

10.Wang, Kai, et al. “End-to-End Scene Text Recognition.” 2011 International Conference on Computer Vision, 2011, doi:10.1109/iccv.2011.6126402.

11.Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In NIPS, pages 2017–2025, 2015.

ПРОВЕДЕНИЕ СРАВНИТЕЛЬНОГО АНАЛИЗА ATTENTION OCR И TESSERACT В ЗАДАЧЕ РАСПОЗНАВАНИЯ СИМВОЛОВ НА ИЗОБРАЖЕНИЯХ ПРЕЙСКУРАНТОВ.

Марков Андрей Владиславович
студент,
Челябинский государственный университет,
Россия, г.Челябинск

АБСТРАКТ

Решения классической задачи распознавания символов является высоко востребованной на практике. В рамках данной работы будет рассматриваться задача распознавания символов с изображений прайс-листов табачной продукции. Для разметки изображений использовался сервис Yandex OCR. Сравнивалась модель Attention OCR и технология Tesseract по качеству распознавания изображений слов, вырезанных с прайс-листов. Attention OCR показала более качественное распознавание символов по сравнению с Tesseract.

Ключевые слова: Tesseract, Attention OCR, прайс-листы.

Введение

Задача распознавания символом является одной из базовых задач компьютерного зрения. Отличительной особенностью данной задачи состоит в разнообразии данных. Текст может быть представлен различными символами, языками, иметь разный шрифт, фон, размер, а также ориентацию в пространстве. Актуальность же данной задачи состоит в большом практическом значение в областях, где решение данной задачи позволяет автоматизировать процесс сбора информации с изображений. В рамках данной работы будет рассматриваться задача распознавания символов с изображений сигаретных прайс-листов. Решение задачи распознавания символом на этих данных полезна для автоматического сбора различной информации, в том числе цен, представленных на прайс-листах табачной продукции. Данная информация может быть полезна для дальнейшего исследования полноценного end-to-end решения для сбора информации с изображений прайс-листов.

Обзор

Одно из первых открытых решений задачи распознавания символов является технология Tesseract[5] (https://tesseract-ocr.github.io/tessdoc/)..

Данная технология способна как находить текст, так и распознавать его. При этом, Tesseract позволяет распознавать больше сотни языков, включая русский и английский. Также, Tesseract способен работать не только с изображениями отдельных слов, но также отдельных абзацев или даже страниц. Tesseract хорошо подходит в качестве базового решения, при этом есть возможность дообучить его на своем датасете.

Решения задачи распознавания символов являются востребованными, в связи с чем существует несколько облачных решений. Такая возможность есть у таких облачных гигантов как Amazon, Google, Yandex. Преимуществом таких решений является высокое качество моделей из за большого количества тренировочных данных. В качестве недостатков стоит заметить дороговизну таких решений, скорость работы, а также, иногда, необходимо разворачивать в облаке отдельные сервисы для поддержания инфраструктуры.

Кроме готовых решений существуют различные модели, которые можно натренировать на своих данных. Один из типов таких моделей заключается в архитектуре, которая представлена в виде пары энкодер-декодер, а также есть слой внимания. В качестве энкодера выступает сверточная нейронная сеть, которая выделяет различные признаки с изображения, а в качестве декодера рекуррентная нейронная сеть, которая, основываясь на полученных признаках, делает предсказания символов. В качестве примера можно привести сеть Attention OCR[1], которая на датасете FSNS[7] показала долю правильных ответов 82.4%.

Также, существуют модели, которые не только решают задачу детекции символов, но и задачу нахождения текста. Одна из таких моделей - CharNet[2]. Основой архитектуры сети являются сети ResNet[3] и HourGlass[4] после которых есть два пайплайна распознавания. Первый пайплайн реализует обнаружения текста на уровне слов, второй же, находит текст на уровне символов, что позволяет модели решать задачу распознавания символов.

В работе проводится сравнение технологии Tesseract и Attention OCR. Attention OCR не требует данных, размеченных посимвольно, Tesseract используется как базовое решение.

Датасет

Датасет был автоматически собран с помощью сервиса Yandex OCR. Изображения представляли собой изображения прайс-листов табачной продукции (рисунок 1) из которых в дальнейшем были вырезаны слова (рисунок 2). Минимальные и максимальные размеры изображений вырезанных слов: высота - (9, 48), ширина - (3, 564). Полученные пары (изображение - текст) в дальнейшем были почищены от пар, текст которых встречался только один раз, так как это было неверное распознавание Yandex OCR. Также, чтобы предотвратить переобучение, были частично удалены пары, у которых текст был “СИГАРЕТЫ”, так как это было самое часто встречающееся слово. В результате, был получен датасет, в котором не было каких либо сильно часто встречающихся слов. Количество пар в обучающей выборке составило 50050, и 5000 в тестовой.

Также был составлен алфавит: “**!"&'()*+,-./0123456789:;ABCDEFGHIJKLMNOPQRSTUVWXYZ||~°ЁАБВГДЕЖЗИЙКЛМНОПРСТУФХЦЧШЩЫЬЭЮЯ№**”, с которым в дальнейшем обучалась модель Attention OCR.

№	Код товара	Название товара	Цена
53	3277866	СИГАРЕТЫ WINSTON SS SILVER 1П.	143руб. 00коп.
54	3277847	СИГАРЕТЫ WINSTON WHITE 1П.	143руб. 00коп.
55	3227660	СИГАРЕТЫ WINSTON XS BLUE 1П.	148руб. 00коп.
56	3227662	СИГАРЕТЫ WINSTON XS SILVER 1П.	148руб. 00коп.
57	3384812	СИГАРЕТЫ WINSTON XSTYL.SIL.ПАЧ	132руб. 00коп.
58	3356406	СИГАРЕТЫ WINSTON XSTYLE BLU1П.	132руб. 00коп.
59	2041790	СИГАРЕТЫ ДОН.ТАБАК ТЕМНЫЙ 1ПАЧ.	90руб. 00коп.
60	3197020	СИГАРЕТЫ ДОНСКОЙ ТАБАК СВ.ПАЧ	90руб. 00коп.
61	3609614	СИГАРЕТЫ ПЕТР I ОСОБ.ЧЕР.ПАЧКА	108руб. 00коп.
62	3670192	СИГАРЕТЫ ПЕТР I ЭТАЛ.КОМ.ПР.ПАЧ	107руб. 00коп.
63	3439972	СИГАРЕТЫ ПЕТР I ЭТАЛ.КОМП.1ПАЧ	107руб. 00коп.
64	3276717	СИГАРЕТЫ ПЕТР I ЭТАЛОН 1П.	125руб. 00коп.
65	3439975	СИГАРЕТЫ ПЕТР I ЭТАЛОН ОС.ПАЧ.	102руб. 00коп.
66	3631558	СИГАРЕТЫ ЯВА 100 ЗОЛ.КЛАС.ПАЧКА	96руб. 00коп.
67	3454111	СИГАРЕТЫ ЯВА БЕЛ.ЗОЛ.КЛАСС.1ПАЧ	90руб. 00коп.
68	3201664	СИГАРЕТЫ ЯВА ЗОЛОТАЯ КЛАСС.1П.	96руб. 00коп.
69	3981551	СТИКИ HEETS FR.PAR.AM.SEL.ПАЧКА	145руб. 00коп.
70	3981554	СТИКИ HEETS FR.PAR.PURP.W.ПАЧКА	145руб. 00коп.
71	3981553	СТИКИ HEETS FR.PAR.TUR.S.ПАЧКА	145руб. 00коп.
72	3981552	СТИКИ HEETS FR.PAR.YEL.S.ПАЧКА	139руб. 00коп.

Рисунок 1. Пример изображения прайс-листа.

Рисунок 2. Пример изображений.

Метрика

В работе были выбраны следующие метрики качества:

Accuracy - доля совпадений истинного текста и предсказанного

Расстояние Левенштейна.

Эксперимент и анализ результатов

В работе сравниваются две модели: Attention OCR и Tesseract. Tesseract использовался со следующими конфигурациями:

lang=rus+eng --eom 1 --psm 7

lang=rus+eng --eom 1 --psm 8

lang=rus+eng --eom 1 --psm 10

lang=rus+eng --eom 1 --psm 13

Модель Attention OCR обучалась на тренировочной выборке. Конфигурация обучения:

Количество эпох epoch = 3500

Начальный шаг обучения lr = 1

Максимальная длина предсказываемого текста была задана в 30 символов

--target-embedding-size=128

Метрики на тестовом датасете показаны в таблице 1.

Таблица 1

Значения метрик на тестовом датасете

	AOCR	Tesseract --psm 7	Tesseract --psm 8	Tesseract --psm 10	Tesseract --psm 13
accuracy	**0.934**	0.526	0.513	0.525	0.512
Levenshtein	**0.574**	2.437	2.46	2.434	2.461

Из полученных метрик лучшей оказалась модель Attention OCR. Полученная модель оказалась более устойчивой к шумам и к размытым изображениям. Также, так как модель была обучена на тренировочном датасете. она способна учитывать различную специфику датасета. При этом некоторые ошибки модели были посчитаны из за неправильного распознавания Yandex OCR, в то время как модель правильно предсказала текст на изображении.

В качестве улучшения результатов для Attention OCR можно предложить увеличить количество скрытых слоев сети, попробовать различные препроцессинги, а также попробовать дообучить модель с меньшим шагом обучения. Для улучшения результатов распознавания Tesseract также можно попробовать различные препроцессинги изображения, такие как эрозия, выравнивание текста или бинаризация.

Заключение

В данной работе была рассмотрена задача распознавания символов с изображений прайс-листов табачной продукции. В ходе эксперимента была обучена модель Attention OCR. Обученная модель сравнивалась с технологией Tesseract. В данном эксперименте модель Attention OCR по качеству распознавания символов оказалась лучше, чем Tesseract. В дальнейших исследованиях, для улучшения качества распознавания символом, можно попробовать расширить датасет изображений прайс-листов синтетическими данными, а также опробовать различную аугментацию данных.

Список литературы

1. Zbigniew Wojna, et al. “Attention-based Extraction of Structured Information from Street View Imagery” arXiv:1704.03549v4 [cs.CV] 20 Aug 2017. https://arxiv.org/pdf/1704.03549.pdf

2. Xing, Linjie, et al. “Convolutional Character Networks.” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, doi:10.1109/iccv.2019.00922.

3. He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, doi:10.1109/cvpr.2016.90.

4. H. Law and J. Deng. Cornernet: Detecting objectsas paired keypoints. InProceedings of the EuropeanConference on Computer Vision (ECCV), pages 734–750, 2018

5. Tesseract OCR, https://opensource.google/projects/tesseract .

6. Stanford cs class cs231n: Convolutional neural networks for visual recognition. http://cs231n.github.io/neural-networks-case-study/.

7. R.Smith,C.Gu,D.-S.Lee,H.Hu,R.Unnikrishnan,J.Ibarz,S.Arnoud, and S. Lin, “End-to-end interpretation of the french street name signs dataset,” in European Conference on Computer Vision. Springer, 2016, pp. 411–426.