ЭКСПЕРИМЕНТАЛЬНОЕ СРАВНЕНИЕ ДВУХ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ EAST И CRAFT В ЗАДАЧЕ ДЕТЕКЦИИ ТЕКСТА НА ИЗОБРАЖЕНИЯХ ЦЕННИКОВ

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

АБСТРАКТ

В работе сравниваются два подхода к обнаружению текста на изображениях магазинных ценников: EAST и CRAFT. Модели сравнивались по F-score. Текст был представлен символами русского и английского алфавитов. В результате наилучшую метрику показал CRAFT. Была произведена аналитика результатов.

Ключевые слова: EAST, CRAFT, детекция текста, ценники.

Введение

Задача обнаружения областей текста на изображении является классической задачей компьютерного зрения. Связано это с большим количеством приложений таких моделей. Одним из таких приложений является обнаружение текста на ценниках в продуктовых магазинах для того, чтобы результаты работы детектора подать в OCR модель. Такая система позволит собирать информацию с ценников в автоматическом режиме. Парсинг информации с ценников с помощью моделей глубокого обучения актуальный еще тем, что не на всех ценниках есть QR-код, по которому, в целом, легко получить необходимую информацию. Здесь, решение задачи обнаружения текста на изображении играет ключевую роль, так как от качества такой модели будет зависеть качество всего пайплайна.

Обзор

Задача детекции текста является важной задачей компьютерного зрения, так как имеет огромное количество приложений. Решать данную задачу с помощью алгоритмов Deep Learning начали сравнительно недавно. Отличительной особенностью такой задачи от обычной задачи обнаружения объектов является то, что текст может быть представлен в очень большом разнообразии, в отличие от конкретных объектов. Текст может находится на различном фоне, иметь разнообразную форму и геометрию. Все это накладывает некоторые ограничения для использования “чистых” методов обнаружения объектов.

Ключевую роль в развитие решений данной задачи является появление таких наборов данных как ICDAR-2013, ICDAR-2015, ICDAR-2017, COCO-TEXT, MSRA-TD500. Эти наборы данных представляют собой разнородные изображения с текстом в естественной среде. Разметка представлена в виде боксов, каждый бокс обрамляет слово на изображении.

Одним из ключевых алгоритмов, показавших в 2017 году SOTA результат является EAST[7]. Модель представляет собой Fully Convolutional Network[6] на базе PVANet[8] с пробросом признаков между слоями, с дальнейшей “головой”, которая может быть представлена двумя способами. Первый способ представляет собой решение задачи регрессии на (x1, y1, x2, y2, alpha), где (x1, y1) и (x2, y2) координаты бокса, а alpha - угол поворота. Второй тип решает задачу регрессии на координаты углов бокса (x1, y1), (x2, y2), (x3, y3) и (x4, y4). В дальнейшем, найденные области отсекаются по порогу и проходят процедуру NMS. Таким образом, добавление угла поворота и предсказывание 8ми координат вместо 4х является попыткой преодолеть ограничения стандартных алгоритмов детектирования объектов. Данная модель смогла достичь F-score равным 0.8072 на датасете ICDAR 2015.

Еще одна модель на базе подхода детекции слов является модель SSTD[4], которая показала значение F-score на ICDAR-2015 равное 0.77. Модель представляет собой исправленную модель SSD для нахождения повернутых боксов, а также модуль внимания. Архитектура такой сети основана на VGG-16[9], а для боксов предсказываются пять значений - координаты бокса и его ориентация. Модуль внимания же позволил снизить количество ложных срабатываний, повысить количество обнаружений специфического текста и повысить точность в нахождении слов.

В дальнейшем, появились алгоритмы, целью которых является не только обнаружить текст, но и прочитать его. Такие модели, как правило состоят из двух частей: детектора и модель, реализующая Optical Character Recognition. Для обучения таких моделей требуется больше данных, а также, иногда, отдельной работы по генерацию разметки на уровне отдельных символов. Также на таких моделях применяют так называемые weakly supervised методы обучения. Но с другой стороны, комбинация двух моделей повышает точность друг друга.

Одной из моделей, которая способна решать задачи обнаружения и распознавания текста является алгоритм FOTS[5]. Основой решения является сеть ResNet-50[3]. Дальше признаки идут на два пайплайна. Признаки в первом пайплайне проходят через слой RoIRotate, который как раз занимается обнаружением областей текста. Дальше, найденные области подаются на второй пайплайн распознавания текста, где используется сеть LSTM с функционалом качества CTC[2]. Такой подход позволил авторам достичь на

датасете ICDAR-2015 значение F-score равное 0.8799, что больше чем у предшествующих моделей.

Еще одним алгоритмом, который способен как находить текст, так и распознавать его, является алгоритм CRAFT[1]. CRAFT по сути, является детектором на уровне отдельных символов. Основа архитектуры состоит из сети, похожую на Unet на базе VGG-16. Голова сети предсказывает как расположение символа на изображении, так и его положение в плоскости. Модель CRAFT показала на датасете ICDAR-2015 значение F-score равное 0.8696.

В данной работе решено было сравнить на датасете ценников значение F-score двух разных подходов. В качестве подхода детекции слов была взята модель EAST, а в качестве подхода детекции отдельных символов модель CRAFT.

Описание данных

Датасет представляет собой изображения ценников крупных продуктовых ритейлеров России. Размер тестовой выборки составляет 1700 изображений. Разметка получена в автоматическом режиме с помощью Yandex OCR API и доразмечена вручную. Ширина изображений W находится в промежутке (43, 1736), высота Y в промежутке (30, 1234). В датасете представлены ценники разных сетей, с разным фоном и разных товаров. Разметка для одного изображения представляет собой набор боксов. Для каждого бокса известны его координаты (x, y, w, h). Пример изображения ценника представлен на рисунке 1.

Рисунок 1. Пример изображения ценника

Метрика

В качестве метрики качества использовались стандартные метрики качества для данной задачи: Precision, Recall и F-score[10]. В качестве определения True Positive служит мера пересечение боксов Intersection Over Union IOU. Если значение IOU больше заданного порога T, то считается что область найдена правильно. Если положить, что Ngt - количество боксов из разметки, а Nalg - количество предсказанных боксов, то тогда:

- Precision = True Positive / Nalg
- Recall = True Positive / Ngt
- F-score = 2 * Precision * Recall / (Precision + Recall)

В рамках эксперимента было положено T = 0.5.

Эксперимент и результаты

В рамках данной работы сравнивались CRAFT (https://github.com/backtime92/CRAFT-Reimplementation), предобученный на синтетических данных и EAST (https://github.com/argman/EAST), предобученный на ICDAR-2013 и ICDAR-2015.

Параметры обучения модели EAST:

- input_size=512
- text_scale=512
- learning_rate=0.0001
- использовалась “голова”, предсказывающая 5 чисел: координаты бокса и его угол поворота

Параметры обучения модели CRAFT:

- Использовался оптимизатор Adam с параметрами learning_rate=3.2768e-5 и weight_decay=5e-4

В таблице 1 представлены метрики моделей EAST и CRAFT. На рисунке 2 представлена визуализация разметки, а на рисунках 3 и 4 результаты работы моделей EAST и CRAFT соответственно.

Таблица 1.

Метрики моделей EAST и CRAFT на тестовом датасете

	Precision	Recall	F-score
EAST	0.7212	0.6859	0.7083
CRAFT	0.6641	0.7899	0.7174

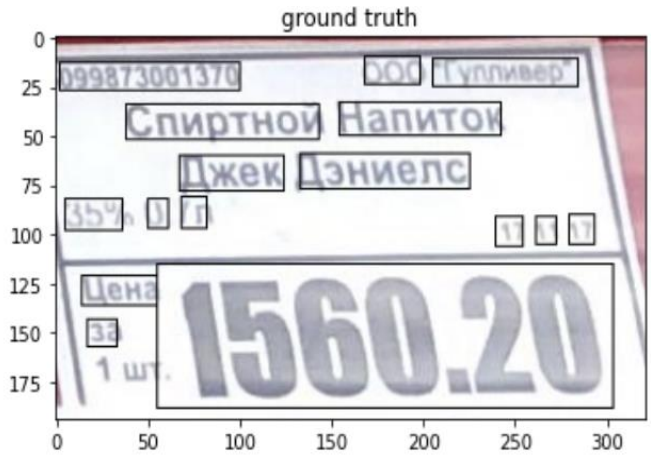

Рисунок 2. Визуализация разметки

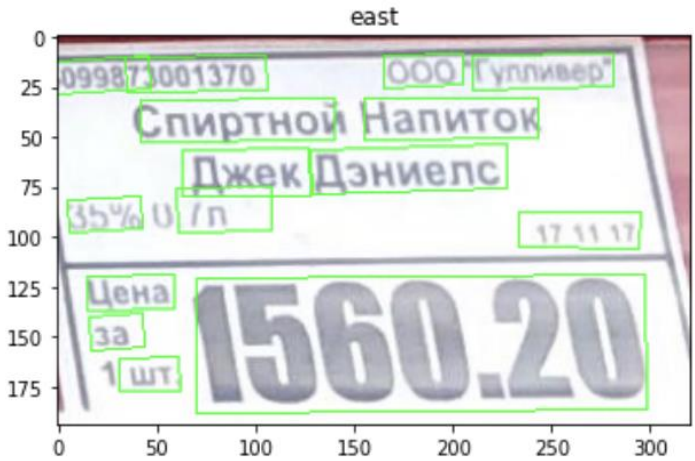

Рисунок 3. Визуализация работы модели EAST

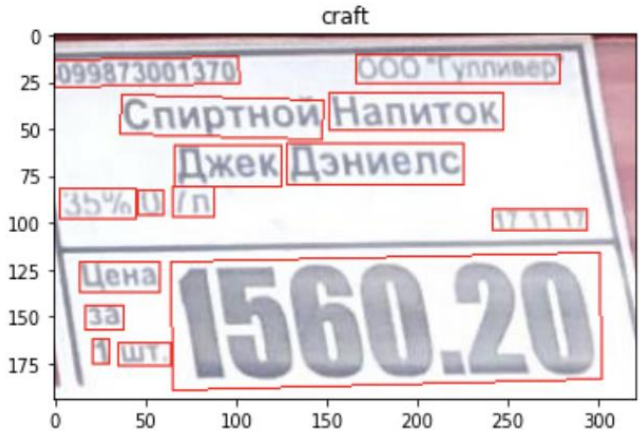

Рисунок 4. Визуализация работы модели CRAFT

Также был произведен замер скорости работы обеих моделей. Результаты приведены в таблице 2. Характеристики сервера:

1.CPU: Intel(R) Core(TM) i5-6600 CPU @ 3.30GHz,

2.GPU: GeForce RTX 2070.

Таблица 2.

Скорость работы на CPU и GPU

	сек./изображение (CPU)	сек./изображение (GPU)
CRAFT	96	**0.08**
EAST	**0.23**	0.22

Обе модели показали близкие результаты. CRAFT способен находить отдельные символы, в то время как EAST скорее всего пропустит их. Также CRAFT находит слова более аккуратно, чем EAST. Но в то же время ориентация боксов отдельных слов текста у CRAFT порой менее стабильна, если слова имеют короткую длину. Также CRAFT лучше себя показывает на менее качественных фотографиях, в то время как EAST не находит на них текст, или находит только крупный. Если же говорить с точки зрения практического применения, то EAST показывает лучшее время на CPU, а CRAFT на GPU. При этом, скорость работы EAST на CPU соизмерима со скоростью работы CRAFT на GPU. Таким образом, решение с EAST будет дешевле с точки зрения эксплуатации модели.

Заключение

Таким образом, был проведен эксперимент по сравнению двух методов детекции текста: EAST и CRAFT. В результате эксперимента была выявлено, что CRAFT показал более качественный и устойчивый результат, чем EAST, но с другой стороны EAST более дешевый в эксплуатации. В качестве улучшения можно обучить модели на данном датасете, а также попробовать обучить модели с другими параметрами обучения, учитывая особенности датасета.

Список литературы

1.Baek Y, Lee B, Han D, et al (2019) Character Region Awareness for Text Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2019.00959

2.Graves A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification. Proceedings of the 23rd international conference on Machine learning - ICML '06. doi: 10.1145/1143844.1143891

3.He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2016.90

4.He P, Huang W, He T, et al (2017) Single Shot Text Detector with Regional Attention. 2017 IEEE International Conference on Computer Vision (ICCV). doi: 10.1109/iccv.2017.331

5.Liu X, Liang D, Yan S, et al (2018) FOTS: Fast Oriented Text Spotting with a Unified Network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. doi: 10.1109/cvpr.2018.00595

6.Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2015.7298965

7.Zhou X, Yao C, Wen H, et al (2017) EAST: An Efficient and Accurate Scene Text Detector. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi: 10.1109/cvpr.2017.283

8.Kye-Hyeon Kim, Sanghoon Hong, Byungseok Roh, Yeongjae Cheon, Minje Park (2016) PVENet: Deep but Lightweight Neural Networks for Real-time Object Detection.

9.Karen Simonyan, Andrew Zisserman (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition

10.Afzal Godil Patrick Grother Mei Ngan The Text Recognition Algorithm Independent Evaluation (TRAIT)

ПРОВЕДЕНИЕ СРАВНИТЕЛЬНОГО АНАЛИЗА ДВУХ НЕЙРОННЫХ СЕТЕЙ EAST И PSENET В ЗАДАЧЕ ДЕТЕКЦИИ ТЕКСТА НА ИЗОБРАЖЕНИЯХ ПРЕЙСКУРАНТОВ.

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

АННОТАЦИЯ

Большое разнообразие текстовых шаблонов и сильно загроможденный фон создают основную проблему точной текстовой локализации. В данном исследовании будет проведен сравнительный анализ существующих решений в области обнаружения текста на изображении, таких как EAST и PSENet. Обе модели показали схожие результаты в обнаружении текста. При этом модель EAST имеет преимущество в скорости, а модель PSEnet - лучше определяет области текста на низкокачественных изображениях.

Ключевые слова: EAST, PSENet, обнаружение текста

Введение

В настоящее время решение задачи чтение текста привлекает всё больше исследователей в области компьютерного зрения. Во многом, это связанно с многочисленным практическим применением и ростом требований бизнеса в сокращении издержек посредством внедрения технологий компьютерного зрения. Эта задача включает в себя две подзадачи: обнаружение и распознавания текста. Данная работа фокусируется на задачах обнаружения которая является более сложной, чем задача распознавания, выполненная на обрезанной части изображения, содержащая слова. Большое разнообразие текстовых шаблонов и сильно загроможденный фон создают основную проблему точной текстовой локализации.

Обзор

Во многом большой прорыв в решении задачи детекции текста заключается в том, что появились большие наборы размеченных данных, на которых исследователи могут строить свои модели и сравнивать их между собой. Таким набором данных