Сравнение нейросетевых и непрерывно-морфологических методов в задаче детекции текста (Text Detection)

Гайдученко Н.Е., Труш Н.А., Торлак А.В., **Миронова Л.Р., Акимов К.М**., Гончар Д.А.
October 20, 2018

Аннотация

Данная статья посвящена анализу и сравнению методов распознавания текста на изображениях. Текст - это один из наиболее распространенных способов коммуникации, который для передачи информации может быть представлен в виде документа или являться частью картинки. Несмотря на достигнутые результаты в этой области, задача детекции текста на изображениях требует дальнейшего исследования, особенно для изображений с сложным фоном. В данной статье мы будем рассматривать использование как моделей глубокого обучения, которые активно применяются в машинном зрении, так и непрерывно-морфологических методов обработки изображений. Сравнение будет производиться на датасетах, представляющих разные стороны задачи детекции текстов, с использованием ряда метрик (таких как F-score, etc.) для определения качества работы сравниваемых методов.

Ключевые слова: *нейронные сети, непрерывно-морфологические методы, распознавание текста, обучение без учителя.*

Введение

В последнее время задача детекции текста на изображениях документов и на реальных фотографиях, содержащих надписи, привлекла немалое внимание в научном сообществе. Существуют методы и технологии распознавания строк и букв в сканированных документах, однако они мало применимы к реальным изображениям. Актуальность задачи детекци текста связана со следующими факторами. Во-первых, увеличилось число возможных приложений данной технологии. Например, она может применяться на мобильных устройствах, в роботизированных системах, при поиске картинок в интернете. Во-вторых, текст - это один из наиболее распространенных способов коммуникации, который для передачи информации может быть представлен в виде документа или являться частью картинки. Автоматическое нахождение текста на изображении позволяет увеличить объем получаемой информации. В-третьих, успехи в области компьютерного зрения позволяют применять более продвинутые методы для решения задачи детекции текста на изображениях. В то время как задача оптического распознавания букв на изображениях сканированных документов может считаться изученной, задача детекции текста показывает невысокое качество распознавания, порядка 80%. Данная задача требует учета большого количества факторов, таких как ориентация текста, шрифт, цвет, освещение, фон изображения.

В данной статье проводится сравнительный обзор методов детекции текста на изображениях, рассматриваются как непрерывно-морфологические методы, так и нейросетевые, и выявляется, какие из них в каких случаях показывают наиболее точный результат.

Первый из рассмотренных нами морфологических методов [1] использует геометрический скелет для изображения текста, построенный по бинаризованной картинке, с последующем построением метрики близости и кластеризацией полученного графа и позволяет работать с рукописными документами, так как не опирается на знание о почерке, стиле, языке или о структуре текста.

Во втором [3] используется гауссовский фильтр второго порядка для получения локального представления об ориентации и размере текста. Далее полученные значения собираются в гистограммы для определения общего значений. Полученные распределения бинаризуются для выделения линий текста из фона. Гауссовский фильтр можно эффективно посчитать и получающееся качество близко к более сложным алгоритмам обучения с учителем.

В третьем подходе [4] применяется модификация метода наиболее стабильных экстремальных регионов, посчитанная по 4 цветовым каналам, для выделения букв на картинке. Далее идет анализ связных компонент для выделения регионов с текстом и последующее формирование строк текста. Данный подход может бороться с яркостными эффектами.

Среди нейронных методов мы рассмотрели сверточные сети [5], метод с боксами [6]. А также CTPN [7], который хорошо работает на крупномасштабных и на разноязыковых текстах без дальнейшей обработки, в отличие от многих других методов. На вход он принимает изображение любого размера и обнаруживает text line путем плотного продвижения маленького окна. Такой подход позволяет детектировать символы различного размера за счет того, что размер окна меняется.

Постановка задачи

В данной работе рассмотрена задача определения областей, содержащие отдельные слова, на реальных изображениях и отсканированных документах. Эта задача является частью более общей проблемы детекции текста (строк), которые впоследствии разбиваются на слова.

Исходными данными задачи являются выборка изображений — D, выходными данными являются также изображения, на которых выделены области относящиеся к разным словам внутри строки — ɣ.

Вычисление ошибки

Наша задача - минимизация функционала потерь. Одной из метрик оценки ошибки задачи классификации является Intersection over Union.

Назовем множество рамок с текстом, которые мы ожидаем получить на выходе алгоритма ожидаемым. После работы алгоритма мы получили другое множество рамок, заданных координатами: $x1_i$, $x2_i$, $y1_i$, $y2_i$, по которым мы можем вычислить площади: ожидаемых - $S1_i$, полученных - $S2_i$ и площади их пересечений. Функция вычисляется как отношение площади перекрытия рамок ожидаемой и полученной к площади их объединения:

$$\text{IoU} = \sum((S1_i \cap S2_i) / (S1_i \cup S2_i)). \quad (1)$$

Две другие метрики - precision и recall. Мы знаем правильные ответы для задачи классификации изображений на текст (положительный класс) и не текст (отрицательный класс). Есть алгоритм, который определяет, является ли изображение текстом, то есть относит его или к положительному, или к отрицательному классу. Тогда определим следующие величины:

tp- количество изображений, правильно отнесенных к тексту;

tn- количество правильно не отнесенных к тексту;

fp- количество неправильно отнесенных к тексту;

fn- количество неправильно не отнесенных к тексту.

$$P(\text{precision}) = \frac{\text{tp}}{\text{tp} + \text{fp}}, \quad R(\text{recall}) = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (2)$$

Мера точности характеризует, сколько полученных от классификатора положительных ответов являются правильными. Чем больше точность, тем меньше число ложных попаданий. Мера полноты - способность классификатора «угадывать» как можно большее число положительных ответов из ожидаемых.

Precision и Recall дают характеристику классификатора с разных сторон, и увеличение одной из них приводит к уменьшению другой. Поэтому удобно использовать метрику F1, их среднее гармоническое:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (3)$$

Список литературы

[1] Е. О. Захаров, Л. М. Местецкий. *Сегментация текстовых блоков в изображениях рукописных документов.* 2016.

[2] Q. Ye, D. Doermann. *Text Detection and Recognition in Imagery: A Survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 37, Issue: 7, July 1 2015).

[3] D. Aldavert, M. Rusiñol. *Manuscript Text Line Detection and Segmentation Using Second-Order Derivatives*. 2018 13th IAPR International Workshop on Document Analysis Systems (DAS).

[4] Zh. Liu, Y. Li, X. Qi, Y. Yang, M. Nian, H. Zhang, R. Xiamixiding. *Method for unconstrained text detection in natural scene image.* IET Computer Vision (Volume: 11, Issue: 7, 10 2017).

[5] T. Wang, David J. Wu, A. Coates, A. Y. Ng. *End-to-End Text Recognition with Convolutional Neural Networks.*

[6] M. Liao, B. Shi, X. Bai. *A Single-Shot Oriented Scene Text Detector.* arXiv:1801.02765v3 [cs.CV] 27 Apr 2018.

[7] Zh. Tian, W. Huang, T. He, P. He Y. Qiao. *Detecting Text in Natural Image with Connectionist Text Proposal Network.* arXiv:1609.03605v1 [cs.CV] 12 Sep 2016.