Задача двухклассовой классификации изображений

Гущин Александр
Руководитель: Вадим Стрижов
Консультант: Василий Лексин

Московский физико-технический институт (государственный университет)

28 января 2015 г.

Цели проекта

- Разработать алгоритм двухклассовой классификации изображении на основании того, что различием между этими классами является наличие некоторого запрещенного текста (телефонов, email, url).

Основные идеи алгоритма

- Локализовать текст с помощью Stroke Width Transform
- Распознать текст с помощью Tesseract OCR
- Создать обучающую выборку и обучить классификатор
- С помощью классификатора определить области, соответствующие буквам

Постановка задачи

Вход: цветное изображение. Оно представлено в виде трехмерной матрицы, где первые два индекса отвечают за номер пикселя на изображении, а третий – значения цветов в RGB-системе.

Задача:: построить классификатор и произвести классификацию изображения.

Выход: Действительное число от 0 до 1, представляющее вероятность того, что изображение содержит запрещенную информацию.

Пример входного изображения

Результат работы на изображении

Результат распознавания текста Tesseract OCR :
QM“ Мг Mil/1’s марте Mil/TEE милей mm! тж!

Основные шаги алгоритма

- Локализуем текст с помощью SWT
- Извлекаем признаки из найденных областей
- Распознаем текст с помощью Tesseract OCR
- Применяем TF-IDF к распознанному тексту
- Обучаем классификатор или загружаем уже обученный
- Классифицируем
- Используя результаты классификации выводим ответ

Признаки для классификации

- Области: количество локализованных областей с текстом
- Площадь областей: суммарная площадь локализованных областей
- Feature: результат TF-IDF для распознанного текста

Наличие логотипа на каждом изображении

- **Проблема**: Наличие логотипа авито в правом нижнем углу на каждом изображении
  Решение #1: Закрасить область с с логотипом.
  Решение #2: Вычесть из области изображение, полученное усреднением таких областей. k.
- **Проблема**: Большое время работы SWT и Tesseract (40 тысяч изображений)
  Решение #1: Распараллеливание
  Решение #2: Эффективный код

Результат

Результаты работы классификатора для кросс-валидации с nfold=5 :

LinearSVC на двух первых признаках (количество областей, их суммарная площадь) :
AUC : 0.759 (+/- 0.028)

LinearSVC на трех признаках :
AUC: 0.809 (+/- 0.028)

Сравнение с другими решениями

Алгоритм на нейронных сетях позволяет достичь AUC 0.95

Несмотря на то, что обучение нейронных сетей занимает большее время чем требует рассмотренный алгоритм, судя по результатам конкурса Авито и другим соревнованиям по распознаванию изображений, они являются гораздо более эффективным подходом, чем рассмотренный в этой работе.

Заключение

Был разработан алгоритм классификации изображений, основанный на признаках, полученных при локализации текста и его распознавании.