УДК 004.85

Генеративные модели для размещения текста на изображении

Ефимова В. А., Университет ИТМО, Санкт-Петербург
Научный руководитель – Фильченков А. А., к.ф.-м.н., доц. ФИТиП университета ИТМО

Введение

С текстом, размещенным на изображении, современный человек сталкивается каждый день в рекламных объявлениях, в журналах или же в социальных сетях. Обычно этот текст написан на натуральном языке. Текст на изображениях используется для коммуникации, передачи и представления информации, распознавание такого текста – сложная и широкоиспользуемая задача, важная часть взаимодействия между человеком и компьютером. Разработано множество техник для извлечения текста из отсканированных документов, но извлечение текста из изображений требует дальнейшего изучения [1-3].

В большинстве случаев текст на изображении размещен вручную дизайнером или другим человеком, мы же рассмотрим автоматическое размещение текста на изображении. В этой области доступных наборов данных крайне мало, существуют наборы данных ICDAR [4, 5] и Street View Text [6], последний, например, содержит 647 слов из 3796 букв на 249 изображениях, полученных из Google Street View. Число таких изображений мало, в сумме порядка пары тысяч, а словарь представленного текста ограничен. Сбор вручную подобного набора данных дорог и требует значительных затрат по времени.

Решением этой проблемы будет автоматическое размещение текста на изображении.

Цель работы

Целью настоящей работы является создание набора данных для дальнейшего использования – автоматического распознавания текста. Кроме огромного объема сгенерированных данных синтезирование изображений с текстом позволяет точно определять bounding box текста, что важно для большинства систем распознавания [6].

Описание предлагаемого подхода

В данном исследовании планируется использовать условную генеративную состязательную сеть (Conditional Generative Adversarial Network, cGAN) [7] с некоторыми модификациями. Ее схема представлена на рис. 1.

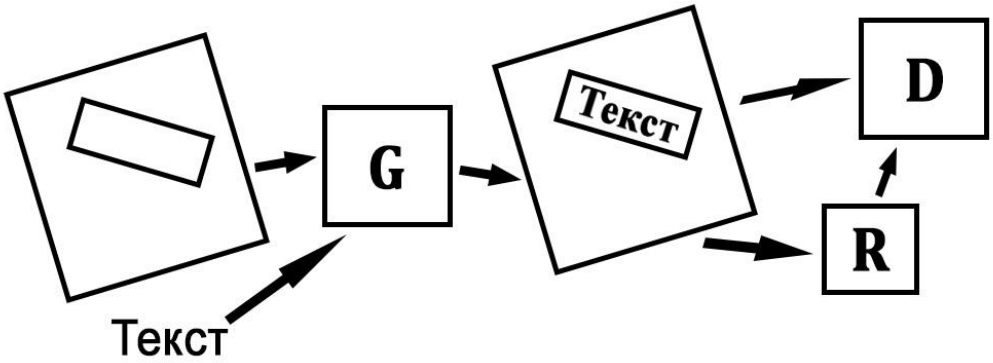

Рисунок 1. Схема предлагаемого метода.

На вход сети подаются три объекта: текст, который нужно будет вписать, область на изображении (bounding box), в которую нужно будет вписать текст, и сам текст. Буквой G обозначен генератор, который синтезирует новое изображение с уже вписанным текстом. Далее синтезированное изображение оценивается с помощью системы распознавания текста (на рисунке R), а результат распознания учитывается при оценке изображения дискриминатором D. Этот шаг добавлен нами в оригинальную архитектуру, так как текст должен оставаться распознаваемым. Все это повторяется заданное число раз.

В результате получаем обученное распределение, которое наилучшим способом повторяет данное распределение.

В дальнейшем планируется не задавать извне bounding box, а предсказывать методами машинного обучения, определяя поверхности, на которых можно расположить текст в реальном мире.

Результаты

Пока что полученные изображения далеки от реальных фотографий, требуется дальнейшая доработка. Полученный набор синтетических изображений будет использован для улучшения обучения системы распознавания текста.

Список литературы

[1] Chen, D., Tsai, S., Chandrasekhar, V., Takacs, G., Chen, H., Vedantham, R., Grzeszczuk, R. and Girod, B. Residual enhanced visual vectors for on-device image matching // Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR). 2011, pp. 850-854.

[2] Epshtein, B., Ofek, E. and Wexler, Y. // Detecting text in natural scenes with stroke width transform // IEEE Computer Society Conference on Computer Vision and Pattern Recognition // 2010, pp. 2963-2970.

[3] Neumann, L. and Matas, J. Real-time scene text localization and recognition // Conference on Computer Vision and Pattern Recognition // 2012, pp. 3538-3545.

[4] Lucas, S.M. ICDAR 2005 text locating competition results // Eighth International Conference on Document Analysis and Recognition (ICDAR'05). 2005, pp. 80-84.

[5] Shahab, A., Shafait, F. and Dengel, A. ICDAR 2011 robust reading competition challenge 2: Reading text in scene images // International conference on document analysis and recognition // 2011, pp. 1491-1496.

[6] Wang, T., Wu, D.J., Coates, A. and Ng, A.Y. End-to-end text recognition with convolutional neural networks // Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012) // 2012, pp. 3304-3308.

[7] Mirza, M. and Osindero, S. Conditional generative adversarial nets // arXiv preprint arXiv:1411.1784, 2014.