ПРИМЕНЕНИЕ СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ U-NET ДЛЯ СЕГМЕНТАЦИИ ТЕКСТОВЫХ ОБЛАСТЕЙ НА ИЗОБРАЖЕНИЯХ РЕАЛЬНЫХ СЦЕН

Ю.А. Иванова, к.т.н., доц. ОИТ ИШИТР,
В.А. Лобанова, студент гр. 8ВМ92
Томский политехнический университет
E-mail: val17@tpu.ru

Введение

В настоящее время существует огромное количество информации, хранящейся в виде изображений, содержание которых представляет собой определенную ценность. Детектирование и последующее распознавание текста на изображениях может быть применено в таких областях как перевод фотографий документов в текстовую форму [1], автоматическое определение номерных знаков автомобилей [2], геолокация объекта по названиям улиц, улучшение качества детектирования и распознавания объектов на изображениях. Однако объемы информации, хранящейся в виде изображений, велики, что делает невозможным ее обработку вручную. Тем не менее, автоматизированные методы обработки изображений позволяют успешно справляться с этой задачей.

Целью работы является разработка нейросетевого алгоритма для детектирования текстовых областей на изображениях реальных сцен на основе сверточной нейронной сети архитектуры UNet в комплексе с методами предварительной обработки изображений.

Описание инструментов

Для реализации алгоритма был выбран язык программирования Python и нейросетевая открытая библиотека Keras. Написание программного кода производилось в интегрированной среде разработки PyCharm.

В качестве базы изображений была выбрана база KAIST Scene Text Database, из которой было выбрано 1215 фотографий различных вывесок [3]. Для определения расположения текста для каждого изображения предоставляется изображение-маска. Ввиду малого объема исходной выборки было проведено искусственное увеличение базы изображений в 7 раз за счёт поворота и обрезки изображений. Повороты производились в обе стороны от -18° до 18° с шагом в 6°.

В предыдущих исследованиях было выяснено, что при детектировании текстовых областей часто возникают ложные срабатывания алгоритма в ответ на высокочастотные сигналы, не содержащие текстовых областей. В связи с этим в настоящей работе было решено применять предварительную фильтрацию изображений. Было проведено сравнение таких фильтров, как фильтры сглаживания, фильтр увеличения резкости и фильтры выделения краёв. В результате наилучшие результаты детектирования были получены с применением двустороннего фильтра сглаживания, позволяющего убрать шумы, но сохранить различимость объектов на изображении. Были выбраны параметры фильтра: диаметр пикселя окрестности - 5 пикселей, размер ядер сигма фильтров в цветовом пространстве и в координатном пространстве - 50 пикселей.

В качестве сверточной нейронной сети была применена архитектура U-Net, на вход которой подаётся обрабатываемое изображение, а выходом является изображение-маска, выделяющая текстовые области. Сеть состоит из сужающегося пути, производящего выделение карт признаков и уменьшающего размеры изображения, и расширяющегося пути, увеличивающего размеры изображения и производящего его конкатенацию с соответствующей картой признаков [4, 5]. В сверточных слоях и слоях обратной свертки используется функция активации ReLU. На последнем уровне свертки используется сигмовидная функция активации в форме гиперболического тангенса. Для определения ошибки сети, по формуле 1 вычисляется коэффициент Дайса (dice coefficient), который показывает меру сходства изображений:

$$dice\ coefficient = \frac{2 * intersection(y_{true}, y_{pred}) + smooth}{y_{true} + y_{pred} + smooth}, \quad (1)$$

где *ytrue* - истинное значение пикселей изображений, ypred - предсказанное значение пикселей изображений, intersection - пересечение/умножение значений пикселей изображений, smooth - коэффициент сглаживания.

Чем выше значение коэффициента Дайса, тем большее количество истинных и предсказанных значений пикселей изображений совпадает, соответственно тем лучше детектируются текстовые области.

Для обучения сети был использован набор облачных служб Google Cloud Platform, в котором была создана виртуальная машина.

Результаты работы

Было проведено обучение сверточной нейронной сети U-Net в 20 эпох с предобработкой фильтром сглаживания и, для сравнения, без фильтрации. Значения коэффициента Дайса, полученные на обучающей и валидационной выборке представлены в таблице 1.

Таблица 1. Коэффициент Дайса для обучения и валидации для каждой эпохи

Эпоха	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Обучение без фильтра	12,1	17,9	43,2	47,0	51,6	53,0	55,0	54,4	59,5	62,2	61,0	66,5	70,2	73,5	75,2	75,9	78,8	80,6	81,1	82,2
Валидация без фильтра	11,6	14,0	37,1	45,2	49,5	52,7	54,4	56,6	58,4	61,0	65,4	68,5	73,7	72,0	73,0	76,4	79,6	80,6	81,9	81,8
Обучение с фильтром	25,0	47,3	53,2	56,3	60,5	70,0	76,2	79,7	82,1	83,2	84,6	85,4	86,7	87,0	88,1	89,0	89,8	90,4	90,7	90,9
Валидация с фильтром	44,3	51,4	55,6	59,1	66,0	75,2	77,4	79,8	80,3	84,2	85,0	85,4	85,2	86,1	86,8	87,9	87,9	88,6	88,8	88,2

Из таблицы 1 видно, что точность обучения сверточной нейронной сети с фильтром сглаживания для последних 5 эпох превышала точность без предварительной фильтрации на 11,6% для обучающей выборки и на 9,3% для валидационной.

На рисунке 1 представлены примеры масок для тестовых изображений, полученные в результате предсказания обученной сетью (сверху - оригинал изображения, снизу – результат предсказания). Несмотря на то, что на большей части тестовых изображений текстовые области детектируются правильно, в некоторых случаях сеть выделяет такие области как, окна зданий, просветы между деревьями, вывески с рисунками. Некоторые текстовые надписи с тонким шрифтом плохо определяются за счёт сжатия изображений. Неоднородные текстовые надписи (градиент, контурные шрифты) также представляют собой проблему.

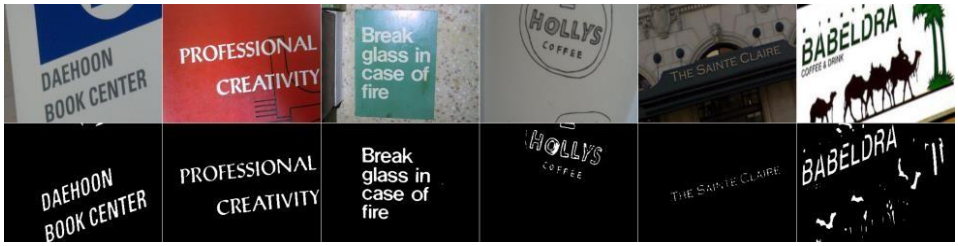

Рис. 1. Примеры масок для тестовых изображений

Заключение

В результате на предварительно обработанной базе изображений двусторонним фильтром сглаживания была обучена сверточная нейронная сеть U-Net и достигнута точность обучения 90,86%, точность валидации – 88,20% и точность тестирования – 87,90%. Для тестовой выборки были получены предсказания в виде изображений-масок высокого качества.

В дальнейшем планируется увеличение обучающей выборки путём добавления новых баз изображений, подачи на вход сверточной нейронной сети фрагментов изображений. Также планируется рассмотреть возможность применения частотного анализа к изображениям в качестве предобработки.

Список использованных источников

1. Mechi O., Mehri M., Ingold R., and Essoukri Ben Amara N. Text line segmentation in historical document images using an adaptive unet architecture. 2019 ICDAR, 2019, pp. 369–374.
2. Chowdhury P. N., Shivakumara P., et al. A new U-Net based license plate enhancement model in night day images. Proc. ACPR, 2019.
3. Jung J., Lee S., et al. Scene Text Extractor Using Touch Screen Interface. ETRI Journal, 2011.
4. Badrinarayanan V., Kendall A., Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 2017, vol. 39, no. 12, pp. 2481-2495.
5. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. Lecture Notes in Computer Science, 2015, vol. 9351, pp. 234-241.