СРАВНЕНИЕ АРХИТЕКТУР U-NET И FCN ДЛЯ ДЕТЕКТИРОВАНИЯ ТЕКСТОВЫХ НАДПИСЕЙ НА ИЗОБРАЖЕНИЯХ РЕАЛЬНЫХ СЦЕН

Ю.А. Иванова, к.т.н., доц. ОИТ ИШИТР,
Гао Жэньцзе, студент гр. 8ВМ03
Томский политехнический университет
E-mail: zhencze2@tpu.ru

Введение

Детектирование текстовых надписей на изображениях реальных сцен является важной темой компьютерного зрения и часто является необходимым условием для распознавания текста. Использование модели семантической сегментации для детектирования объектов является широко используемым методом.

Целью работы является сравнение применения сверточных нейронных сетей на основе U-net и FCN для детектирования текстовых надписей на изображениях реальных сцен.

Описание работы

Для обучения и тестирования была выбрана база данных KAIST Scene Text [1], из которой было отобрано 1025 изображений, из которых 800 использовались для обучения и 225 – для валидации. Каждому изображению соответствует маска изображения, которая используется для определения положения текста. Входными данными для сверточных нейронных сетей U-net и FCN является предварительно обработанное изображение, а выходными данными является маска изображения с сегментированными текстовыми областями. Так как размерность выходной карты равна размерности входного изображения, то прогноз может быть произведен для каждого пикселя при сохранении пространственной информации в исходном входном изображении, что позволяет классифицировать объекты на изображении с попиксельно. FCN — это так называемая «полностью сверточная нейронная сеть» [2], для лучшей производительности сегментации часть исходного FCN с понижающей дискретизацией заменена глубокой остаточной сетью Res-Net 34 [3]. Часть повышающей дискретизации инициализируется билинейным ядром свертки, и в общей сложности апсемплинг выполняется три раза.

U-net усовершенствован на основе FCN, архитектура состоит из сужающего пути для захвата контекста и симметричного расширяющегося пути, обеспечивающего точную локализацию [4]. Первая половина используется для извлечения признаков, а вторая половина – для повышения частоты дискретизации. Такая архитектура называется структурой кодер-декодер. В сверточных слоях и слоях обратной свертки используется функция активации ReLU. Выходной слой обеих сверточных нейронных сетей использует сигмовидную функцию активации. Чтобы определить производительность сети, по формуле 1 вычисляется точность:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \qquad (1)$$

где TP (true positive) – сумма всех пикселей изображения, которые модель отнесла к положительному классу («текст») и угадала; FP (false positive) — сумма всех пикселей изображения, которые модель отнесла к положительному классу («текст») и ошиблась; TN (true negative) – сумма всех пикселей изображения, которые модель отнесла к отрицательному классу («нетекст») и угадала; FN (false negative) – сумма всех пикселей изображения, которые модель отнесла к отрицательному классу («нетекст») и ошиблась.

Результаты работы

На рисунках 1 и 2 показаны примеры карт масок для результатов предсказания сетей FCN и U-net соответственно (слева-исходное изображение, справа-результат предсказания). Хотя эти две сети в большинстве случаев могут правильно детектировать область текста, из-за сложного и изменчивого фона изображения текста в реальной сцене, в некоторых случаях сеть считает текстом узоры стен, узоры между текстом и вывески.

Рис. 1. Примеры масок для FCN

Рис. 2. Примеры масок для U-net

Заключение

В результате точность обучения сверточной нейронной сети FCN на этом наборе данных составляет 82,0 %, а точность теста — 79,4 %. Точность обучения сверточной нейронной сети U-net составляет 90,1 %, а точность теста — 86,8 %.

Видно, что общая производительность U-net лучше, чем у FCN, а также лучше обработка деталей изображения. В дальнейшем планируется увеличить количество изображений в наборе данных, а также попытаться сравнить и выбрать подходящие методы оптимизации изображений при предварительной обработке изображений.

Список использованных источников

1. KAIST Scene Text Database. [Электронный ресурс]. – URL: http://www.iapr-tc11.org/mediawiki/index.php/KAIST_Scene_Text_Database (дата обращения 01.02.2022).
2. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation (2014), arXiv:1411.4038 [cs.CV].
3. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
4. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.