https://doi.org/10.26160/2474-5901-2020-18-50-54

ОБЩИЙ ПОДХОД К СОЗДАНИЮ НАБОРА ДАННЫХ НА ПРИМЕРЕ ФОРМИРОВАНИЯ НАБОРА ИЗОБРАЖЕНИЙ ЛИНЕЙНЫХ ШТРИХ-КОДОВ

Венцов Н.Н., Подколзина Л.А.

Ключевые слова: компьютерное зрение, распознавание образов, подготовка данных, набор данных, аугментация, линейный штрих-код.

Аннотация. Эффективность моделей машинного обучения зависит от качества исходного набора данных. При отсутствии их предварительной обработки может возникнуть ситуация, когда модель обучается на искаженных и ненадежных данных, приводящих работу системы к получению неверных результатов. В научных работах основной упор уделяется процессу улучшения моделей обучения, тогда как работе с самой выборкой не всегда оказывается должное внимание. Таким образом, возникает необходимость первоначального изучения данных с целью улучшения существующего набора. Целью работы является создание набора данных линейных штрих-кодов для проведения дальнейших научных изысканий. Задачей работы является выработка общих рекомендаций при подготовке изображений, включаемых в обучающую и тестовую выборки. Результатом работы является созданный набор данных, включающий в себя 8000 изображений линейных штрих-кодов (4000 изображений и 4000 масок к ним).

THE GENERAL APPROACH TO CREATING A DATASET USING AN EXAMPLE OF BARCODE IMAGES

Ventsov N.N., Podkolzina L.A.

Keywords: computer vision, pattern recognition, data preparation, data set, augmentation, lD barcode.

Abstract. The effectiveness of machine learning models depends on the quality of the original data set. In the absence of their preliminary processing, a situation may arise when the model is trained on distorted and unreliable data, leading the system to obtain incorrect results. In scientific papers, the main emphasis is on the process of improving learning models, while working with the sample itself is not always given due attention. Thus, there is a need for an initial study of the data in order to improve the existing set. The aim of this work is to create a linear barcode data set for further scientific research. The objective of the work is to develop general recommendations for the preparation of images included in the training and test samples. The result of the work is the created data set, which includes 8000 images of linear barcodes (4000 images and 4000 mask for this images).

Введение. Специалистам в области машинного обучения зачастую приходится работать в условиях недостаточности существующего набора данных, причинами которой может являться неполнота информации, наличие ошибочных данных или их несогласованность. В начальной выборке могут вообще отсутствовать необходимые для формирования предикторов данные. В виду того, что эффективность моделей машинного обучения зависит от качества исходного набора данных, требуется время на его подготовку к использованию, а также проведение дополнительных действий по очистке, нормализации и генерации. При отсутствии обработки может возникнуть ситуация, когда модель обучается на искаженных и ненадежных данных, в

которых могут быть пропущены значения. Их использование может привести к неверным результатам. Таким образом, возникает необходимость первоначального изучения набора данных, а также дальнейшего определения степени предварительной обработки или необходимости создания и/или пополнения датасетов релевантной информацией. Важным становится умение проводить сбор, предварительную очистку и обработку данных в целях создания новой выборки, удовлетворяющей условиям, необходимым в контексте решаемой задачи, а в дальнейшем для обучения модели.

Научные работы зачастую посвящены новым и улучшенным моделям, в то время как наборам данных не уделяется должного внимания (для проверки выдвигаемых гипотез применяются наборы общедоступных архивов). Самым большим препятствием для использования глубокого обучения является получение достаточно высокой точности при работе с реальными данными. Улучшение тренировочного набора помогает повысить точность распознавания [1].

Штриховое кодирование широко применяется в розничной торговле, почтовых и грузовых перевозках, для идентификации электронных компонентов и медицинских препаратов. Символика штрихового кода может быть распечатана на достаточно простом оборудовании, её можно масштабировать довольно в широких пределах. Задача распознавания штрих-кодов, отсканированных или сфотографированных в неизвестных заранее условиях, актуальна во многих промышленных системах [2]. Применение мобильных устройств для поиска информации и считывания штрих-кодов стало повсеместным. Однако, различия в параметрах камер, условиях и методах съемки, применяемых при считывании, влияют на дальнейшую точность обнаружения штрих-кодов.

Во многих приложениях используется распознавание линейных штрих-кодов на изображениях [3-5]. Однако, при наличии деформаций и изменений освещенности могут возникать проблемы для традиционных методов их локализации [6]. В этой статье описывается способ создания набора данных для последующего обучения нейронных сетей. **Целью** работы является создание набора изображений линейных штрих-кодов для его дальнейшего применения при распознавании образов. Решаемой задачей является рассмотрение основных этапов создания датасета, процессов внесения необходимых изменений и доработок для формирования качественного набора.

Предложенный подход. Для формирования исходного набора данных студентам, обучающимся в ДГТУ в качестве дополнительного задания, выполняемого на добровольной основе, было выдано техническое задание выдавалось задание на выполнение фотографий штрих-кодов книг. Всего было получено 850 фотографий, пример фотографий представлен на рис.1. После проведения предварительного просмотра было исключено 50 файлов в связи с проблемами файлов (повреждение данных).

Рис. 1. Пример изображений из собранного набора

Далее для выполнения обработки изображений был создан программный модуль на языке Python. Путем поворота и внесения шумов проводилась аугментация для каждой из имеющихся фотографий (рис. 2). Исходное неаугментированное изображение подвергалось различного рода изменениям: поворот на случайный угол, обрезка изображения и увеличение при изменении соотношения сторон не более чем на 20%, добавление шума. В случае, если на обучающей выборке штрих-код занимает все изображение, то при работе с тестовой выборкой найти штрих-код, занимающий 1/3 или 1/5 изображения становится проблематично. Поэтому выполняется масштабирование, чтобы область штрих-кода занимала на все изображение, а лишь его часть. На этом этапе набор данных был увеличен до 4000 фотографий.

Рис. 2. Изображение после аугментации (вращение и обрезка)

Был создан модуль, позволяющий создавать маски штрих-кодов на исходных фотографиях. Под маской подразумеваем монохромное изображение с двумя группами цветов. Первая группа применяется для выделения фона изображения и кодируется 0 (черный). Вторая используется для выделения искомой области – штрих-кода – цвет 1 (белый).

Для проверки точности обнаруженных масок был использован дополнительный проход по выборке, содержащей 50 случайных изображений и 50 масок к ним, когда студенты вручную проверяли маску, созданную модулем, на соответствие ожидаемому результату. В случае выявления несоответствия проводилась ручная корректировка. Полученный набор данных был разделен на 2 директории: в первую определены маски изображений, во вторую – сами изображения. Для упрощения дальнейшей работы постфикс «_M» (рис. 3).

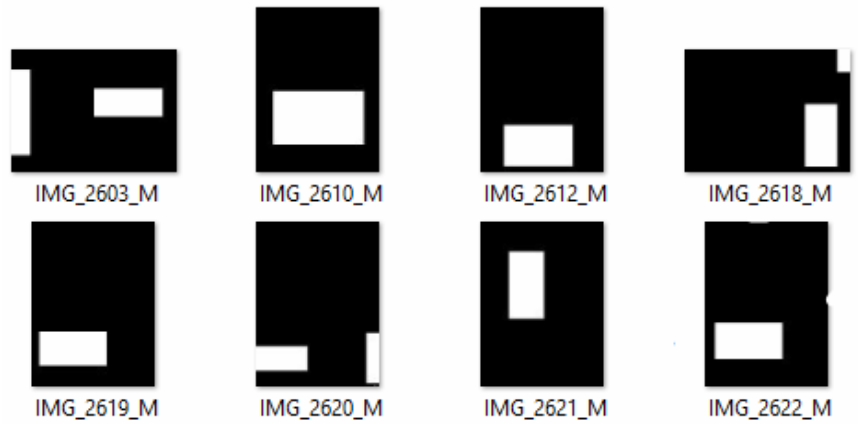

Рис. 3. Маски

Заключение. Результатом работы стало создание набора из 8000 изображений штрих-кодов. Половина набора включает варианты с наличием на изображениях одного или нескольких штрих-кодов, различных масштабов, различных углов поворота и различных коэффициентов кадрирования для повышения обобщающей возможности. Вторая половина представляет собой маски к первой части набора данных.

Работа выполнена при поддержке грантов РФФИ №19-08-00074_а, №19-01-00357_а.

Список литературы

1. Warden P. (2018). Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. ArXiv, abs/1804.0320.
2. Безматерных П.В. и др. Генеративное распознавание двумерных штрихкодов // Институт системного анализа РАН. – 2010. – №4. – С. 63-69.
3. Shen X., Lin Z., Brandt J., Wu Y. Mobile product image search by automatic query object extraction // Proc. ECCV’12. 2012. P. 114-127.
4. Siva Sai Srikar D., Prasanth Reddy E., Aravindh Raju K. Image Based Multiple Bar-Code Detection System? // National Conference on Science, Engineering and Technology (NCSET – 2016). 2016. Vol. 4. Iss. 6. P. 21-23.
5. Zamberletti, A., Gallo, I., Alvertini, S., Noce, L. Neural 1D barcode detection using the Hough Transform // IPSJ Transactions on Computer Vision and Applications. – 2015. – Vol. 7. – P. 1-9.

6. Подколзина Л.А. Аналитический обзор методов распознавания двумерных штрихкодов // Молодой исследователь Дона. – 2018. – №1(10). – С. 55-58.

References

1. Warden P. (2018). Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv, abs/1804.0320.*
2. Bezmaterny P.V. Generative recognition of two-dimensional barcodes // Institute for System Analysis of the Russian Academy of Sciences. 2010. №4. P. 63-69.
3. Shen X., Lin Z., Brandt J., Wu Y. Mobile product image search by automatic query object extraction // Proc. ECCV'12. 2012. P. 114-127.
4. Siva Sai Srikar D., Prasanth Reddy E., Aravindh Raju K. Image Based Multiple Bar-Code Detection System? // National Conference on Science, Engineering and Technology (NCSET – 2016). 2016. Vol. 4. Iss. 6. P. 21-23.
5. Zamberletti, A., Gallo, I., Alvertini, S., Noce, L. Neural 1D barcode detection using the Hough Transform // IPSJ Transactions on Computer Vision and Applications. – 2015. – Vol. 7. – P. 1-9.
6. Podkolzina L.A. An analytical review of two-dimensional barcode recognition methods // Young researcher Don. 2018. No. 1. Vol. 10. P. 55-58.

Венцов Николай Николаевич – кандидат технических наук, доцент кафедры «Информационные технологии», myvnn@list.ru	**Ventsov Nikolay Nikolaevich** – candidate of technical sciences, associate professor of Department of Information technology, myvnn@list.ru
Подколзина Любовь Александровна – аспирант кафедры «Информационные технологии», podkolzinalu@gmail.com	**Podkolzina Lubov Aleksandrobna** – post graduate student of the Information Technologies Department, podkolzinalu@gmail.com
Донской государственный технический университет, Ростов-на-Дону, Россия	Don State Technical University, Rostov-on-Don, Russia

Received 04.02.2020