Интеллектуальный анализ данных и распознавание образов

Нейросетевой метод генерации последовательности символов для синтеза обучающей выборки изображений текста*

П.К. Злобин[I,II], Ю.С. Чернышова[I,II], А.В. Шешкус[I,II], В.В. Арлазаров[I,II]

[I] Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия
[II] ООО «Smart Engines Service», г. Москва, Россия

Аннотация. Объем обучающей выборки – важный фактор при решении задачи оптического распознавания текста, при этом большинство исследований направлено на увеличение разнообразия искажений, которым подвергаются картинки. Однако внутренняя структура текстовой информации также влияет на точность результирующей модели. В статье расмотрен основанный на искусственной нейронной сети метод генерации текста для создания синтетической обучающей выборки изображений, в котором возможно оперировать группами символов – алфавитными кластерами – и использовать последовательность кластеров для предсказания нового символа. Предложенный кластерный подход позволяет создавать неслучайные последовательности, сохраняющие основные свойства целевого языка, но при этом не реализуют полноценную языковую модель. Поскольку предложенный метод работает на небольшом числе кластеров, можно использовать небольшую обучающую выборку и легкую нейронную сеть. Результаты экспериментов с тремя открытыми наборами изображений документов, удостоверяющих личность, демонстрируют эффективность предложенного метода и возможность улучшения современных результатов для целевых полей.

Ключевые слова: *обучающие данные, искусственные нейронные сети, оптическое распознавание символов, генерация текста, синтез данных.*

DOI: 10.14357/20790279230204

Введение

Распознавание документов и, прежде всего, текста в них все шире применяется в повседневной жизни, так как помогает значительно сэкономить время и повысить качество обслуживания клиентов. Технологии распознавания активно используются при обработке паспортов, банковских карт и медицинских документов.

При распознавании текстовой строки изображение одного или нескольких символов поступают

* Исследование выполнено за счёт гранта Российского научного фонда (проект номер 19-29-09066).

на вход классификатора. При раздельном распознавании символов их контекст не имеет значения. Однако если на классификатор подается часть изображения, содержащая несколько символов, соседние символы начинают влиять на точность распознавания. Таким образом, при обучении классификатора важно учитывать окружение символа и обеспечивать его разнообразие.

Поскольку разные поля документов содержат разные символы и их сочетания относительно друг друга, обучающие данные должны создаваться с учетом этих особенностей. Например, один из вариантов написания даты представляет собой последовательность цифр, включающих день, месяц и год, разделенных знаками препинания. Этот вариант исключает наличие букв в его составе. Существуют поля документа, такие как имена людей, содержащие только буквенную последовательность символов за редкими исключениями. Машиночитаемая зона (МЧЗ, MRZ) имеет строгую структуру, так как она специально разработана для распознавания машинами [1]. Эта особенность поля важна при создании данных для обучения классификатора. На рис. 1 показан пример МЧЗ в документе из открытого набора данных MIDV-500 [2]. Таким образом, в документах существуют текстовые строки, которые обладают своей внутренней структурой, не отражающие в полной мере языковую модель. Для таких полей распознаватель с языковой моделью может привести к ухудшению качества распознавания.

Целью нашей работы является разработка метода генерации последовательности символов, которая будет использована для создания обучающей выборки изображений. Разрабатываемый метод учитывает N-граммы заранее заданных алфавитных кластеров. Этот метод позволяет создавать репрезентативные обучающие данные для нейронной сети, что повышает точность распознавания.

Рис. 1. Пример документа с полем МЧЗ из открытого набора данных MIDV-500

1. Обзор литературы

Классическое распознавание изображения текстовой строки включает в себя несколько этапов[3]: предварительная обработка, сегментация, извлечение признаков, классификация. Однако в настоящее время для распознавания текста используются методы глубокого обучения, которые позволяют избавиться от извлечения признаков и существенно уменьшить (или вообще убрать) предварительную обработку изображений, например, оставив только перевод цветного изображения в градации серого [4-6], а иногда и избежать посимвольной сегментации строки.

Распознающая нейронная сеть (распознаватель) обучается на изображениях, содержащих определенную последовательность символов. Существует два различных способа создания обучающих изображений: использование реальных базовых изображений с текстом и создание таких изображений [7]. В первом случае, на изображениях будет содержаться текст, передающий, как минимум, статистику распределения встречаемости букв и N-грамм в языке и содержащий данные в определенном формате, например, в России даты практически всегда пишут в порядке день-месяц-год. Во втором случае символьная последовательность, которая будет отпечатана на обучающих изображениях, создается алгоритмом генерации данных. Создание этой последовательности является важной частью генерации обучающих данных.

При создании текстового заполнения есть два принципиально разных метода и множество промежуточных вариантов. Итак, первый метод – это создание абсолютно случайной последовательности символов [8-10]. В таком случае в данных точно нет никакой языковой модели, которую могла бы выучить распознающая сеть и обоснованно или необоснованно использовать при своей работе. Этот подход прост в реализации и легко адаптируется к различным языкам. При работе со случайными последовательностями можно дополнительно модифицировать распределение исходных символов на основе эвристик и экспертных знаний. Например, можно вместо равномерного распределения по всем символам использовать распределение с повышенными вероятностями знаков препинания. Такая эвристика базируется на том факте, что, например, в китайском языке средняя длина слова – два символа, однако если распределение символов оставить равномерным, то вероятность знака препинания будет меньше 10^{-4}, ведь иероглифов в китайском десятки тысяч. Также возможной эвристикой является ограничение на последова-

тельное написание символов определенных видов, например, запрет на чередование заглавных и прописных букв (строк вида «аАбБвВ») или запрет на длинные последовательности знаков препинания. Однако всевозможные эвристики существенно осложняют систему генерации и могут меняться от языка к языку.

Второй подход – печать реальных текстов или слов, которые были бы на настоящих собранных данных для решаемой задачи [11-15]. В этом случае в данных обязательно присутствует языковая модель, которая будет выучена распознавателем. Использование языковой модели в сети имеет положительные и отрицательные эффекты. К положительным относится способность сети распознавать плохо видимые символы исходя из их соседей. В случае распознавания искаженных изображений такое свойство может быть полезным и соответствовать человеческому восприятию, ведь мы многие слова можем додумать. К отрицательным последствиям относится очень схожий эффект – сеть может «исправить» результат распознавания в редком или написанном с ошибкой слове. Также «исправления» могут происходить при перенесении сети между языками, использующими общую письменность [16]. Такое поведение особенно опасно при распознавании документов, удостоверяющих личность, ведь существует много, например, имен и фамилий, отличающихся одной буквой, но с существенно разной частотой встречи. Также такие сети плохо проявляют себя на строках, существенно отличающихся по своей структуре от обучающей выборке. Например, популярная открытая система распознавания Tesseract OCR 4 [17] использует LSTM-сеть и обучена на данных, взятых из языка (тексты книг, статей и т.д.) и показывает высокие результаты на аналогичных данных. Однако если попытаться ею распознать МЧЗ, снятые в аналогичных условиях и отличающиеся именно внутренней структурой текста, точность распознавания уменьшится [6].

Таким образом, для генерации хорошей последовательности символов система, с одной стороны, должна обладать гибкостью и возможностью создания почти случайных последовательностей, а с другой – сохранять отдельные свойства распределений символов и их кластеров. Стоит заметить, что в области обработки естественного языка (natural language processing) активно исследуются методы генерации текста на основе нейросетевых моделей [18]. Такие модели, умеющие создавать неотличимые от реальных тексты, обычно содержат десятки миллионов параметров [19] и требуют гигантских объемов данных для каждого языка. Однако в нашей задаче нам не требуется большинство возможностей сетей, генерирующих текст, которые приводят к такому количеству параметров, т.е. нам не нужно сохранять семантические связи между полученными словами и не нужно гарантировать какое-либо верное написание слов и предложений. Фактически, нам нужен метод предсказания символа (или его кластера) на основе уже построенной последовательности.

2. Предсказывающая система

2.1 Описание системы

Предсказывающая система включает в себя несколько этапов, при этом ее основным элементом является полносверточная нейронная сеть (подробно описана в следующем разделе).

Перед началом описания системы введем следующие понятия и обозначения:

1) Класс – элемент целевого алфавита при генерации текста, например, буква «А».
2) Кластер – объединение классов, используемое в предсказывающей ИНС, например, «гласные буквы».
3) C – обозначение общего числа кластеров.
4) N – длина последовательности, подаваемой на вход предсказывающей ИНС.

Используемая ИНС принимает в качестве входных данных последовательность из N элементов, представленных в виде индексов кластеров. Соответственно, чтобы предсказать первые N-1 кластер, необходимо задать начальное состояние. Лучший вариант – это случайный выбор кластеров, независимо от их реалистичности. Такой подход позволяет нам генерировать более разнообразные данные. Случайно выбранные кластеры не будут включены в обучающие данные для распознавателя.

Далее ИНС на основе входа вычисляет распределение вероятностей кластеров. На основе вероятностей, вычисленных нейронной сетью, выбирается следующий кластер. Важно отметить, что мы не выбираем всегда кластер с наибольшей вероятностью, так как иначе мы получим много похожих и повторяющихся последовательностей символов.

Следующим шагом после выбора кластера является выбор класса (символа) из этого кластера, чтобы добавить его в результирующую строку. Выбор совершенно случаен, не зависит ни от каких ограничений.

Этапы алгоритма системы для генерации символьной последовательности продемонстриро-

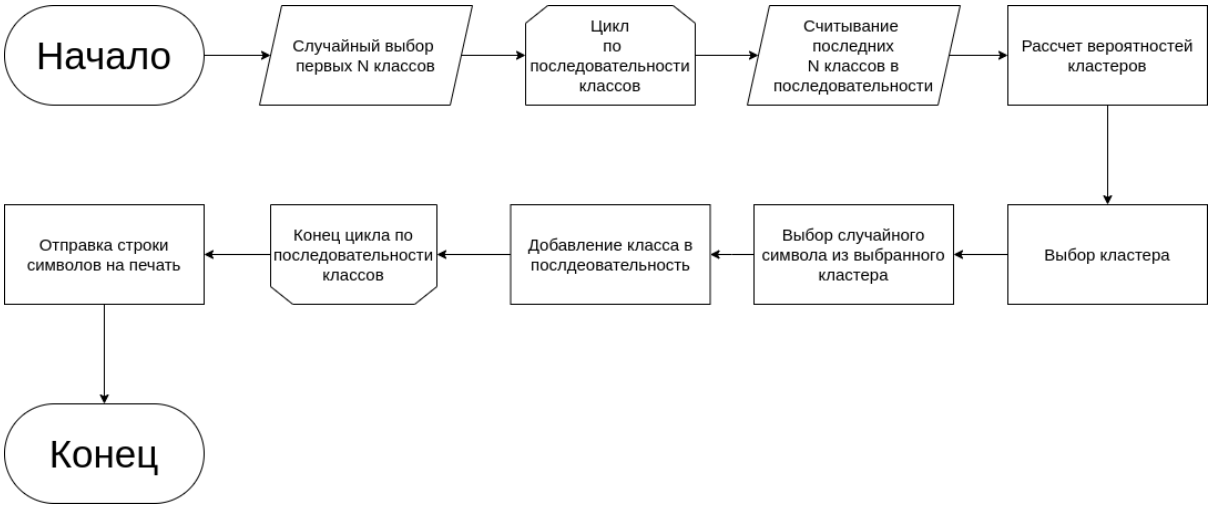

Рис. 2. Этапы работы предсказывающей системы

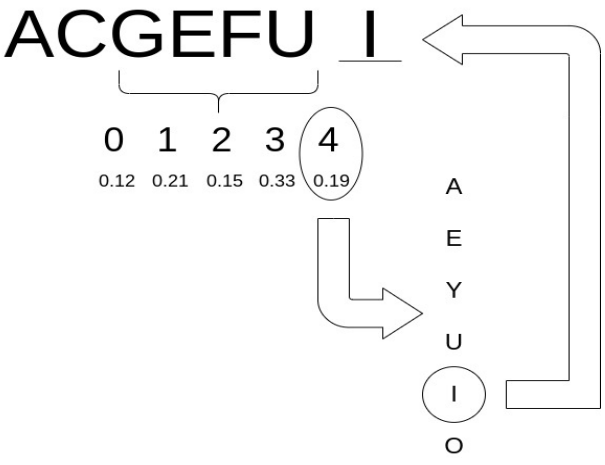

Рис. 3. Процесс предсказания символа (в случае C = 5, N = 4)

ваны на рис. 2. На рис. 3 показан процесс предсказания следующего класса на основе уже созданной строки.

2.2 Предсказывающая ИНС

Архитектура нейронной сети приведена в табл. 1. Архитектура нейронной сети состоит из сверточных слоев с функциями активации ReLU и rf [20]. Архитектура имеет вес $6{,}06 \times 10^3$.

Предсказывающая ИНС обучается на патчах. Каждый патч представляет собой слово из N символов вида $w = (s_1, .., s_{N-1}, s_N)$, где $s_{[1;N-1]}$ – вход нейронной сети, а s_N – предсказанный класс. Каждый символ в строке кодируется соответствующим кластером.

3. Эксперименты

3.1 Постановка экспериментов

В этой статье мы решили использовать поля МЧЗ [21] документов в качестве примеров по нескольким причинам: распознавание полей МЧЗ достигло более высоких значений по сравнению с распознаванием большинства других полей документа; в полях МЧЗ используются слова и цифры, но они не имеют сильных грамматических или семантических связей; МЧЗ имеет строго определенную структуру. Структура МЧЗ имеет некоторые особенности: в этих полях не допускаются знаки препинания, за исключением символа заполнения “<“; поля МЧЗ содержат только заглавные буквы; символы машиночитаемой зоны записываются с помощью шрифта OCR-B (по стандарту) или похожими шрифтами.

Чтобы сделать выводы о результатах распознавания, мы обучали три распознавателя с одина-

Табл. 1

Архитектура предсказывающей сети

Слои			
#	**Тип**	**Функция активации**	**Параметры**
1	Conv	ReLU	32 фильтра 4*1, шаг 1*1, без отступов
2	Conv	ReLU	32 фильтра 1*1, шаг 1*1, без отступов
3	Conv	ReLU	64 фильтра 4*1, шаг 1*1, без отступов
4	Conv	rf[1,1]	32 фильтра 1*1, шаг 1*1, без отступов
5	Conv	rf[1,1]	16 фильтров 1*1, шаг 1*1, без отступов
6		SoftMax	C выходов

ковой архитектурой ИНС, но на разных видах синтезированных данных:

1) Распознаватель NN обучался на данных, созданных с помощью предлагаемой предсказывающей системы, C = 5.
2) Распознаватель RC обучался на данных с полностью случайными последовательностями символов.
3) Распознаватель FA обучался на данных, созданных с помощью предсказывающей системы, где каждому символу, используемому в МЧЗ, соответствовал отдельный кластер, т.е. C = 38.

Также для анализа качества распознавания мы добавили к сравнению Tesseract 4.1.1 [17]. Это необходимо для сравнения наших результатов с базовым уровнем.

Чтобы оценить полученные результаты, мы вычисляем точность посимвольного распознавания по формуле:

$$PCR = \left(1 - \frac{\sum_{i=1}^{L_{total}} \min(lev\left(l_{i_{ideal}}, l_{i_{recog}}\right), len(l_{i_{ideal}}))}{\sum_{i=1}^{L_{total}} \min(len(l_{i_{ideal}}))}\right) * 100\%$$

где L_{total} – общее число строк в тестовой выборке; $len(l_{i_{ideal}})$ – длина i-ой строки из тестовой выборки; $lev(l_{i_{ideal}}, l_{i_{recog}})$ – расстояние Левенштейна между распознанным текстом и аннотацией на i-строке.

3.2 Обучающие и тестовые данные

Для обучения предсказывающей ИНС использовались текстовые данные из открытого набора данных полей МЧЗ [22]. Общий размер набора данных составил 11 400 уникальных файлов. Количество строк МЧЗ в документе варьируется от 2 до 3. Общий объем данных для предсказывающей ИНС составляет примерно 10^6 обучающих примеров. Данные были разделены на обучающую и тестовую выборки в соотношении 90:10. Из-за очень небольшого размера каждого пакета обучение сети на них занимает короткий промежуток времени, несмотря на их большое количество. Благодаря этому не требуется использования больших вычислительных и временных ресурсов. Размер мини-батчей составляет 1024. Во время обучения аугментация данных не применялась. Чтобы получить наибольший объем обучающих данных, мы используем каждые (N-1) символа из результирующей строки со смещением 1 вместо использования непересекающихся уникальных подстрок. Принцип генерации данных для обучения показан на рис. 4. Объединение всех полей в одно устраняет необходимость объявлять начальные символы, чтобы начать генерацию для каждого поля. В первые N-позиций объединенной строки мы записываем пробелы. В наших экспериментах N = 4. Так как наша предсказывающая система работает с кластерами символов, для проведения эксперимента необходимо выделить кластеры для МЧЗ. В наших экспериментах C = 5 и выделены следующие кластеры:

- цифры (“0“,“1“,...,“9“);
- знаки препинания (“<“);
- пробел (“ “);
- заглавные согласные буквы (“B“, “C“,.., “Z“);
- заглавные гласные буквы (“A“, “E“,..., “Y“).

Несмотря на то, что в полях МЧЗ нет пробелов, их присутствие в обучающих данных для сети распознавания необходимо. Поскольку текстовая строка для распознавания обрезана только сверху и снизу, слева и справа от поля есть пустые места. Эти пустые места должны быть обнаружены и отнесены к классу, известному распознавателю. Если сеть не была обучена обнаруживать пробел, то она определит пустое пространство как какой-то другой класс. Это приведет к неправильному распознаванию поля или использованию дополнительной системы обрезки.

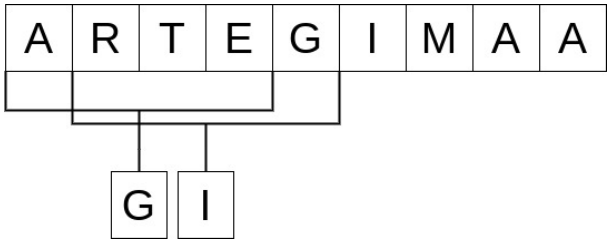

Рис. 4. Формирование обучающих данных (N = 4)

Для обучения распознавателя были созданы изображения с напечатанными на них последовательностями. На рис. 5 показано сравнение строк, созданных с помощью случайного выбора, с использованием предсказывающей системы с 5 кластерами и с предсказывающей системой с полным алфавитом (38 кластеров).

Рис. 5. Сравнение сгенерированных данных без предсказывающей системы и на ее основе с 5 и 38 кластерами

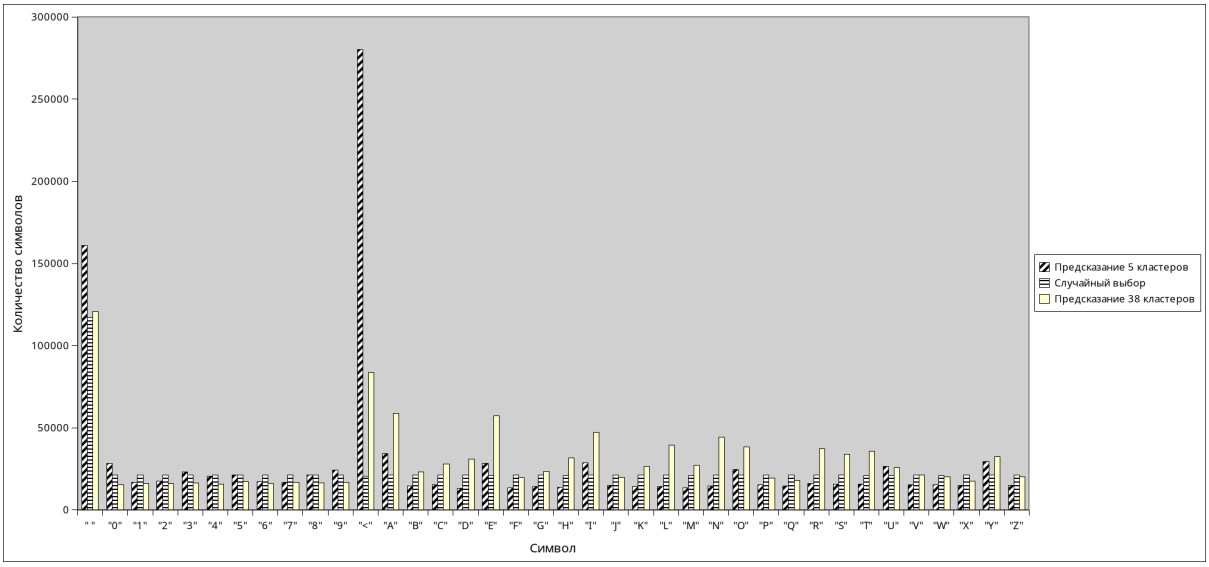

Рис. 6. Количество использований каждого символа в обучающих данных сети распознавания

Количество символов, создаваемых при генерации данных с использованием предсказывающей ИНС, значительно отличается от класса к классу. Количество сгенерированных символов в данных для обучения распознавателя показано на рис. 6.

Использование сети для генерации данных сократило количество пробелов на 60% по сравнению со случайным выбором символов. Напротив, количество символов-заполнителей "<" увеличилось почти в 13 раз. Каждая гласная встречается в строках почти в два раза чаще, чем согласная.

Для эксперимента мы выбрали три тестовых набора данных для оценки влияния предсказывающей нейронной сети на распознавание полей МЧЗ MIDV-500 [2], MIDV-2019 [23] и MIDV-LAIT [24]. Набор данных MLDV-500 включает цветные изображения документов, содержащих 5096 полей МЧЗ. Набор данных MIDV-2019 содержит в общей сложности 3600 строк МЧЗ. MIDV-LAIT включает в себя 72409 изображений документов, из которых можно извлечь 1400 уникальных строк МЧЗ.

3.3 Результаты эксперимента

Результаты распознавания для полей МЧЗ приведены в табл. 2, где даны значения посимвольного качества распознавания:

- обученного на данных, созданных с использованием предсказывающей системы с C = 5 – PCR_{NN};
- обученного на данных, созданных с использованием случайного выбора символов – PCR_{RC};
- обученного на данных, созданных с использованием предсказывающей системы с C = 38, т.е. основанной на полном алфавите – PCR_{FA};
- Tesseract OCR 4.1.1.

Как мы можем видеть, PCR_{NN} – это наилучший результат для каждого тестового набора данных. Это указывает на положительное влияние использования предсказывающей системы на финальное распознавание.

Качество распознавания для PCR_{NN} выше в каждом из протестированных наборов данных. На MIDV-500 разница с PCR_{RC} составляет 0,69%, на MIDV-2019 – 2,27%, на MIDV-LAIT – 0,38%. Можно предположить, что в MIDV-2019 чаще не распознаются не последовательные символы, а отдельные, в то время как соседние распознаются правильно. Сеть, обученная на данных, сгенерированных случайным выбором символов, не может воспользоваться контекстом, окружающим

Табл.2

Результаты распознавания для МЧЗ

Набор изображений	Кол-во полей	Tesseract 4.1.1	PCR_{RC}	PCR_{FA}	PCR_{NN}
MIDV-500	5096	56.58	93.16	92.45	**93.85**
MIDV-2019	3600	45.35	88.08	88.34	**90.35**
MIDV-LAIT	1400	63.69	91.66	91.89	**92.04**

символ, поскольку она была обучена на случайно сгенерированном контексте и, следовательно, символ распознается неправильно. Нейронная сеть, обученная на данных, созданных с использованием предсказывающей ИНС, дополнительно использует преимущество обучения в аналогичном контексте при выдаче результата распознавания. Если сеть не уверена в выборе ответа, этот фактор может оказать решающее влияние. Такое низкое качество распознавания Tesseract OCR 4.1.1 можно объяснить тем фактом, что он обучался на данных, которые не содержали информации из полей МЧЗ. Это подчеркивает важность создания подходящих обучающих данных.

Заключение

Предложенный метод генерации последовательности символов для создания текстового заполнения синтезированных обучающих изображений позволяет увеличить репрезентативность обучающих данных, тем самым повышая качество итогового классификатора. Замеры качества распознавания МЧЗ на актуальных открытых наборах изображений подтверждает эффективность предложенной системы создания данных, показав рост качества на 5-20% в зависимости от используемого набора по сравнению с классификатором, обученным на случайных последовательностях символов. Таким образом, предложенная система предсказания символов показала свою полезность для текстовых полей с нестандартной моделью.

В будущих работах мы планируем применить данный подход для создания нейронной сети, способной генерировать последовательности символов для распознавания других полей документов, требующих иных правил генерации и расширенного алфавита.

Литература

1. *Николаев Д.П., Полевой Д.В., Тарасова Н.А.* Синтез обучающей выборки в задаче распознавания текста в трехмерном пространстве // ИТиВС. 2014. № 3. С. 82-88.
2. *Arlazarov V.V., Bulatov K., Chernov T. and Arlazarov V. L.* “MIDV-500: A dataset for identity document analysis and recognition on mobile devices in video stream,” Computer Optics 43(5), 818–824 (2019). DOI: 10.18287/2412-6179-2019-43-5-818-824.
3. *Naiemi F., Ghods V., Khalesi H.* An efficient character recognition method using enhanced HOG for spam image detection, Soft Computing. 23 (2019)
4. *Bulatov K., Arlazarov V. V., Chernov T., Slavin O., Nikolaev D.* Smart IDReader: Document Recognition in Video Stream // ICDAR 2017 / Manhattan, New York, U.S.: Institute of Electrical and Electronics Engineers Inc. (IEEE). 2017. Т. 6. С. 39-44. DOI: 10.1109/ICDAR.2017.347.
5. *Arlazarov V.L., Arlazarov V.V., Bulatov K.B., Chernov T.S., Nikolaev D.P., Polevoy D.V., Sheshkus A.V., Skoryukina N.S., Slavin O.A., Usilin S.A.* Mobile ID Document Recognition-Coarse-to-Fine Approach // Pattern Recognit. Image Anal. 2022. Т. 32. № 1. С. 89-108. DOI: 10.1134/S1054661822010023.
6. *Chernyshova Y.S., Sheshkus A.V., Arlazarov V.V.* Two-step CNN framework for text line recognition in camera-captured images // IEEE Access. 2020. Т. 8. С. 32587-32600. DOI: 10.1109/ACCESS.2020.2974051.
7. *Jaderberg M., Simonyan K., Vedaldi A. and Zisserman A.* “Synthetic data and artificial neural networks for natural scene text recognition,” in Workshop on Deep Learning, NIPS. 2014.
8. *Hula J.,* Mojž´ısek *D.*, Adamczyk D. and Cech R. “Acquiring Custom OCR System with Minimal Manual Annota- ˇ tion,” in 2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP). 2020. P. 231–236.
9. *Ren X., Chen K. and Sun J.* “A CNN Based Scene Chinese Text Recognition Algorithm With Synthetic Data Engine,” CoRR abs/1604.01891. 2016.
10. *Chernyshova Y.S., Gayer A.V. and Sheshkus A.V.* “Generation method of synthetic training data for mobile OCR system,” in ICMV 2017, A. Verikas, P. Radeva, D. Nikolaev, and J. Zhou, eds., 10696, 1–7, SPIE (Apr. 2018). DOI: 10.1117/12.2310119.
11. *Krishnan P. and Jawahar C.V.* “Generating Synthetic Data for Text Recognition,” CoRR abs/1608.04224. 2016.
12. *Liu Y., Wang Z., Jin H. and Wassell I.* “Synthetically supervised feature learning for scene text recognition,” in Proceedings of the European Conference on Computer Vision (ECCV). 2018. P. 435–451.
13. *Schwarcz S., Gorban A., Serra X.G. and Lee D.-S.* “Adapting Style and Content for Attended Text Sequence Recognition,” in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020. 1586–1595 p.
14. *Namysl M. and Konya I.* “Efficient, Lexicon-Free OCR using Deep Learning,” 2019 International Conference on Document Analysis and Recognition (ICDAR). 2019. P. 295-301. DOI: 10.1109/ICDAR.2019.00055.
15. *Jaderberg M., Simonyan K., Vedaldi A. and Zisserman A.* “Synthetic Data and Artificial Neu-

ral Networks for Natural Scene Text Recognition", Workshop on Deep Learning, NIPS. 2014.
16. *Adnan Ul-Hasan and Thomas M. Breuel.* 2013. Can we build language-independent OCR using LSTM networks? In Proceedings of the 4th International Workshop on Multilingual OCR (MOCR '13). Association for Computing Machinery, New York, NY, USA, Article 9, 1–5. https://doi.org/10.1145/2505377.2505394
17. "Tesseract OCR." https://github.com/tesseract-ocr/tesseract. Online, Accessed: 11.08.2021.
18. *Touseef Iqbal, Shaima Qureshi.* The survey: Text generation models in deep learning, Journal of King Saud University – Computer and Information Sciences, Volume 34, Issue 6, Part A. 2022. P. 2515-2528. https://doi.org/10.1016/j.jksuci.2020.04.001.
19. *Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I.* 2019. Language models are unsupervised multitask learners. OpenAI blog. 1(8). P. 9.
20. *Gayer A.V., Sheshkus A.V., Nikolaev D.P. and Arlazarov V.V.* "Improvement of U-Net Architecture for Image Binarization with Activation Functions Replacement," in ICMV 2020, 11605, SPIE (Jan. 2021). DOI: 10.1117/12.2587027.
21. ICAO Doc 9303 Part 3: Specifications Common to all MRTDs, Machine Readable Travel Documents – International Civil Aviation Organization. 2015.
22. *A. Hartl, C. Arth, and D. Schmalstieg.* "Real-time Detection and Recognition of Machine-Readable Zones with Mobile Devices," VISAPP 2015 – 10th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings 3. 2015. P. 79–87.
23. *Bulatov K., Matalov D. and Arlazarov V.V.* "MIDV-2019: Challenges of the Modern Mobile-Based Document OCR," in ICMV 2019, W. Osten, D. Nikolaev, and J. Zhou, eds., 11433, 1–6, SPIE (Jan. 2020). DOI: 10.1117/12.2558438.
24. *Chernyshova Y.S., Emelianova E.V., Sheshkus A.V. and Arlazarov V.V.* "MIDV-LAIT: a challenging dataset for recognition of IDs with Perso-Arabic, Thai, and Indian scripts," in ICDAR. 2021. P. 1–15.

Злобин Павел Константинович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Аспирант. Количество печатных работ: 1. Область научных интересов: информационные технологии, машинное обучение, анализ данных. E-mail: p.zlobin@ smartengines.com

Чернышова Юлия Сергеевна. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Математик. Количество опубликованных работ: 15. Область научных интересов: синтез данных, оптическое распознавание символов и глубокие нейронные сети. E-mail: chernyshova@smartengines.com (Ответственный за переписку)

Шешкус Александр Владимирович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Научный сотрудник. Количество печатных работ: более 50. Область научных интересов: глубокие нейронные сети, компьютерное зрение, проективно-инвариантная сегментация изображений.
E-mail: asheshkus@ smartengines.com

Арлазаров Владимир Викторович. Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Заведующий отделом. Кандидат технических наук. Количество печатных работ: 130. Область научных интересов: искусственный интеллект, машинное обучение, системы распознавания, информационные технологии. E-mail: vva777@gmail.com

Neural network method for character sequence generation for text images training dataset synthesis

P.K. Zlobin[I,II], Y.S. Chernyshova[I,II], A.V. Sheshkus[I,II], V.V. Arlazarov[I,II]
[I] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia
[II] Smart Engines Service, LLC, Moscow, Russia

Abstract. The size of the training sample is an important factor in solving optical character recognition tasks. Notably, the majority of the research focuses on increasing the variety of distributions that are applied to the images. Yet the internal structure of the textual information also affects the accuracy of the resulting model. We propose a neural network based text generation method for the creation of a synthetic training dataset of images with annotations, in which we propose to operate with groups of characters – alphabetic clusters, and use the sequence of clusters to predict the following character. The proposed cluster approach allows us to create specific sequences that retain the main properties of the target language, but do not contain a full language model. Since the proposed method works on a small number of clusters, we can use a small training set and a light neural network to generate text. The results of experiments with three open datasets of identity document images demonstrate the effectiveness of our method and the possibility of improving modern results for target fields.
Keywords: *training data, neural network, OCR, text generation, data synthesis*
DOI: 10.14357/20790279230204

References

1. *Nikolaev D.P., Polevoy D.V. and Tarasova N.A.* "Sintez obuchayuschey vyborki v zadache raspoznavaniya teksta v trekhmernom prostranstve," ITiVS (3), 82–88 (2014).
2. *Arlazarov V.V., Bulatov K., Chernov T. and Arlazarov V. L.* "MIDV-500: A dataset for identity document analysis and recognition on mobile devices in video stream," Computer Optics 43(5), 818–824 (2019). DOI: 10.18287/2412-6179-2019-43-5-818-824.
3. *Naiemi F., Ghods V., Khalesi H.* An efficient character recognition method using enhanced HOG for spam image detection, Soft Computing. 23 (2019)
4. *Bulatov K., Arlazarov V. V., Chernov T., Slavin O., Nikolaev D.* Smart IDReader: Document Recognition in Video Stream // ICDAR 2017 / Manhattan, New York, U.S.: Institute of Electrical and Electronics Engineers Inc. (IEEE). 2017. T. 6. C. 39-44. DOI: 10.1109/ICDAR.2017.347.
5. *Arlazarov V.L., Arlazarov V.V., Bulatov K.B., Chernov T.S., Nikolaev D.P., Polevoy D.V., Sheshkus A.V., Skoryukina N.S., Slavin O.A., Usilin S.A.* Mobile ID Document Recognition-Coarse-to-Fine Approach // Pattern Recognit. Image Anal. 2022. T. 32. № 1. C. 89-108. DOI: 10.1134/S1054661822010023.
6. *Chernyshova Y.S., Sheshkus A.V., Arlazarov V.V.* Two-step CNN framework for text line recognition in camera-captured images // IEEE Access. 2020. T. 8. C. 32587-32600. DOI: 10.1109/ACCESS.2020.2974051.
7. *Jaderberg M., Simonyan K., Vedaldi A. and Zisserman A.* "Synthetic data and artificial neural networks for natural scene text recognition," in Workshop on Deep Learning, NIPS. 2014.
8. *Hula J.,* Mojž´ısek *D.*, Adamczyk D. and Cech R. "Acquiring Custom OCR System with Minimal Manual Annota- ˇ tion," in 2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP). 2020. P. 231–236.
9. *Ren X., Chen K. and Sun J.* "A CNN Based Scene Chinese Text Recognition Algorithm With Synthetic Data Engine," CoRR abs/1604.01891. 2016.
10. *Chernyshova Y.S., Gayer A.V. and Sheshkus A.V.* "Generation method of synthetic training data for mobile OCR system," in ICMV 2017, A. Verikas, P. Radeva, D. Nikolaev, and J. Zhou, eds., 10696, 1–7, SPIE (Apr. 2018). DOI: 10.1117/12.2310119.
11. *Krishnan P. and Jawahar C.V.* "Generating Synthetic Data for Text Recognition," CoRR abs/1608.04224. 2016.
12. *Liu Y., Wang Z., Jin H. and Wassell I.* "Synthetically supervised feature learning for scene text recognition," in Proceedings of the European Conference on Computer Vision (ECCV). 2018. P. 435–451.
13. *Schwarcz S., Gorban A., Serra X.G. and Lee D.-S.* "Adapting Style and Content for Attended Text Sequence Recognition," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020. 1586–1595 p.
14. *Namysl M. and Konya I.* "Efficient, Lexicon-Free OCR using Deep Learning," 2019 International Conference on Document Analysis and Recognition (ICDAR). 2019. P. 295-301. DOI: 10.1109/ICDAR.2019.00055.

15. *Jaderberg M., Simonyan K., Vedaldi A. and Zisserman A.* “Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition”, Workshop on Deep Learning, NIPS. 2014.
16. *Adnan Ul-Hasan and Thomas M. Breuel.* 2013. Can we build language-independent OCR using LSTM networks? In Proceedings of the 4th International Workshop on Multilingual OCR (MOCR ‘13). Association for Computing Machinery, New York, NY, USA, Article 9, 1–5. https://doi.org/10.1145/2505377.2505394
17. “Tesseract OCR.” https://github.com/tesseract-ocr/tesseract. Online, Accessed: 11.08.2021.
18. *Touseef Iqbal, Shaima Qureshi.* The survey: Text generation models in deep learning, Journal of King Saud University – Computer and Information Sciences, Volume 34, Issue 6, Part A. 2022. P. 2515-2528. https://doi.org/10.1016/j.jksuci.2020.04.001.
19. *Radford A., Wu J., Child R., Luan D., Amodei D. and Sutskever I.* 2019. Language models are unsupervised multitask learners. OpenAI blog. 1(8). P. 9.
20. *Gayer A.V., Sheshkus A.V., Nikolaev D.P. and Arlazarov V.V.* “Improvement of U-Net Architecture for Image Binarization with Activation Functions Replacement,” in ICMV 2020, 11605, SPIE (Jan. 2021). DOI: 10.1117/12.2587027.
21. ICAO Doc 9303 Part 3: Specifications Common to all MRTDs, Machine Readable Travel Documents – International Civil Aviation Organization. 2015.
22. *A. Hartl, C. Arth, and D. Schmalstieg.* “Real-time Detection and Recognition of Machine-Readable Zones with Mobile Devices,” VISAPP 2015 – 10th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings 3. 2015. P. 79–87.
23. *Bulatov K., Matalov D. and Arlazarov V.V.* “MIDV-2019: Challenges of the Modern Mobile-Based Document OCR,” in ICMV 2019, W. Osten, D. Nikolaev, and J. Zhou, eds., 11433, 1–6, SPIE (Jan. 2020). DOI: 10.1117/12.2558438.
24. *Chernyshova Y.S., Emelianova E.V., Sheshkus A.V. and Arlazarov V.V.* “MIDV-LAIT: a challenging dataset for recognition of IDs with Perso-Arabic, Thai, and Indian scripts,” in ICDAR. 2021. P. 1–15.

P.K. Zlobin. PhD student, Institute for Systems Analysis Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. he has published 1 paper. His research interests are IT, machine learning, and data analysis. E-mail: p.zlobin@ smartengines.com

Y.S. Chernyshova. Mathematician, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. Number of papers: 15. Her research interests include training data synthesis, optical character recognition and deep neural networks. E-mail: chernyshova@ smartengines.com (Corresponding author)

A.V. Sheshkus. Researcher, Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, 44/2 Vavilova str., Moscow, 119333, Russia. He has more than 50 published papers. His research interests include deep neural networks, computer vision, and projective invariant image segmentation. E-mail: asheshkus@ smartengines.com

V.V. Arlazarov. Head of the Department for the Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, Moscow, Russia. PhD. Number of articles: 130. Research interests: artificial intelligence, machine learning, recognition systems, information technology, queueing theory.
E-mail: vva777@gmail.com