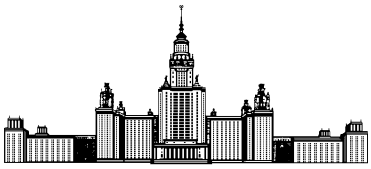
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ОТЧЕТ О ПРЕДДИПЛОМНОЙ ПРАКТИКЕ СТУДЕНТА 417 ГРУППЫ

”Распознавание и генерация изображений рукописных текстов на русском языке”

Выполнил:
студент 4 курса 417 группы
Тыцкий Владислав Игоревич

Москва, 2021

Аннотация

Распознавание текста (Optical Character recognition) одна из известных задач в области машинного обучения. Много лет люди хорошо решают эту задачу для печатных текстов, но в то же время в распознавании рукописных текстов возникает за собой много трудностей, которые сложно решить. В данном отчете будет будет описан сбор датасета, обучение модели распознавания, а также генерация синтетического датасета с помощью GAN (Generative Adversarial Network)

Содержание

1 Введение

Распознавание текста (Optical Character recognition) одна из известных задач в области машинного обучения. Много лет люди хорошо решают эту задачу для печатных текстов, но в то же время распознавание рукописных текстов влечет за собой много трудностей, которые нужно решать. Одна из проблем – это малое количество данных для обучения. Сбор датасета с рукописными текстами времязатратен и дорог, помимо этого получение изображений текста лишь первый этап, за которым следует обработка и разметка данных. Вторая проблема – высокая сложность генерации синтетических данных. В отличии от рукописных данных, печатные тексты несложно сгенерировать: выбираем шрифт, пишем определенный текст и подставляем его на белый фон. Даже такие простые манипуляции могут давать данные, на которых можно обучить работающую модель распознавания.

1.1 Данные

На сегодняший день известно не так много датасетов с рукописными текстами.

Некоторые из них:

- IAM database [6] – датасет английских рукописных текстов
- RIMES database – датасет французских рукописных текстов
- Cyrillic Handwriting Dataset (CyrHD) [4] – датасет русских рукописных текстов *(преимущественно слов)*
- Handwritten Kazakh and Russian (HKR) [5] – датасет русских/казахских рукописных текстов *(не для коммерческого использования)*

1.2 Модели распознавания

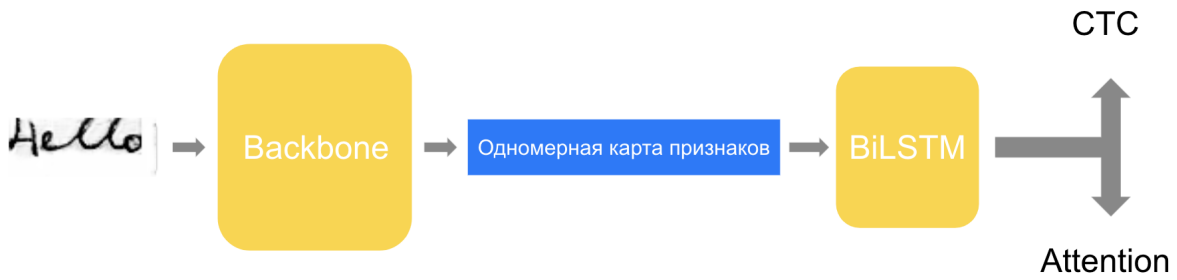


Рис. 1: Стандартная модель распознавания

Современные модели распознавания текстов построены по похожим принципам. Разберем некоторые особенности на примере модели, которую мы будем использовать в качестве основной в данной работе. На Рис.1 изображена стандартная архитектура нейросети для распознавания. Разберем основные особенности:

- На вход подается изображение с заранее неизвестной шириной
- Backbone – это любая нейросеть, умеющая извлекать высокоуровневые признаки из картинки. Обычно используются модели, обученные на ImageNet
- На выходе из backbone получаем одномерную карту признаков – изображение с большим количеством каналов, высотой h=1px и шириной зависящей от входного изображения
- Т.к. мы считаем, что в тексте символы сильно связаны необходимо привнести в полученную карту признаков некоторые нелокальные зависимости. В нашей архитектуре этого можно добиться, используя BiLSTM.
- На выходе из BiLSTM мы получаем одномерную карту признаков того же размера, но с учтенными нелокальными зависимостями.
- Для преобразования карты признаков в текст можно использовать рекуррентую нейросеть с вниманием, либо CTC loss, часто использующийся в OCR

1.3 Метрики

Для понимания качества распознанного текста используются нетривиальные техники подсчета качества. Они делятся на два вида: пословные, посимвольные. Отличие состоит в том, что мы считаем минимальной сущностью (токеном), которая может быть правильной и неправильной при предсказании. На уровне слов ошибка в одном символе приведет к ошибке на всем слове, на уровне символов ошибка – это неправильный символ на некоторой позиции. Ошибки бывают трех типов: ошибка в токене, пропуск токена, лишний токен. Для того, чтобы верно учитывать ошибки, связанные с пропуском/добавлением токена необходимо "выровнять" предсказанный текст до верного. Для этого запустим запустить процедуру поиска расстояния Левенштейна, которая вернет количество вставок, удалений и замен. Выделяют две основных метрики: word error rate (WER), character error rate (CER)[1]

[1]Заметим, что эти метрики могут быть больше 1

$$CER = \frac{S_c + D_c + I_c}{N_c}$$

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

где S_* – количество замен, I_* – количество вставок, D_* – количество удалений

1.4 Генерация синтетических изображений

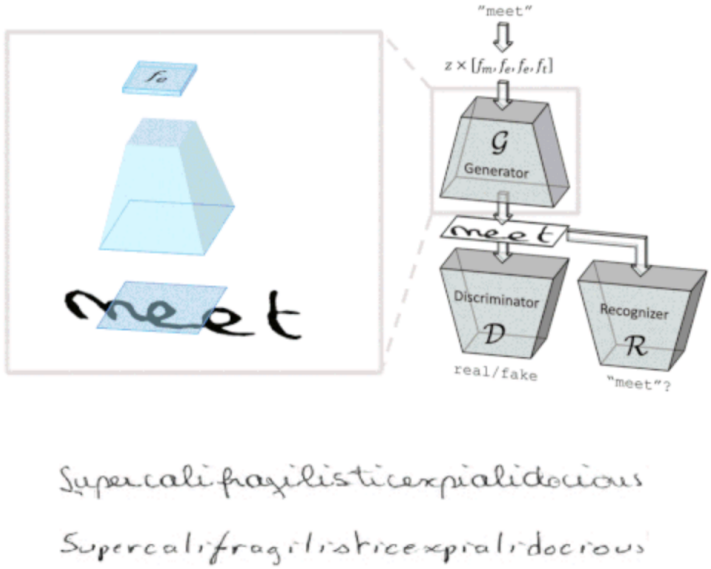


Рис. 2: Scrabble Gan

Для генерации синтетических изображений рукописного текста использовалась архитектура Scrabble Gan [2].

Данная модель состоит из трех частей:

- Генератор – полностью сверточная нейросеть, использующая транспонированные свертки для получения конечного изображения из бинарных векторов соответствующих символам алфавита.
- Дискриминатор, который ”исследует” изображение в целом и относит изображение к ”подделка”, ”реальные данные”
- Распознаватель, который ”читает”, что написано на картинке и сравнивает с эталоном.

2 Этапы работы

Цель всей работы состоит в обучении модели распознавания рукописного текста на русском и английском языке. Для этого необходимы данные для обучения. В случае английского языка можно использовать IAM dataset [6], для русского языка существует датасет CyrHD, но он не подходит под специфику коммерческого использования. Данный датасет преимущественно состоит из слов, не не из линий/строк текста, что автоматически делает его менее полезным для реального использования. Поэтому было решено начать сбор своего датасета.

2.1 Сбор данных

Сбор данных проводился через краудфандинг сервис "Толока".

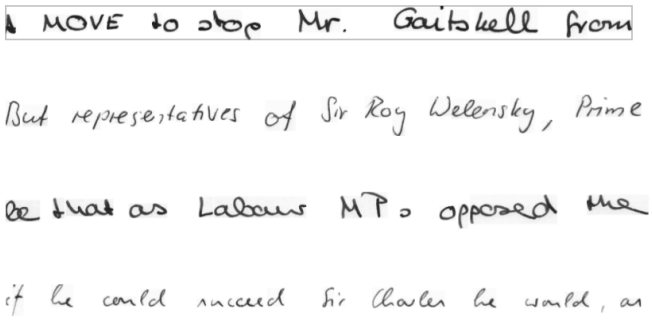


Рис. 3: IAM

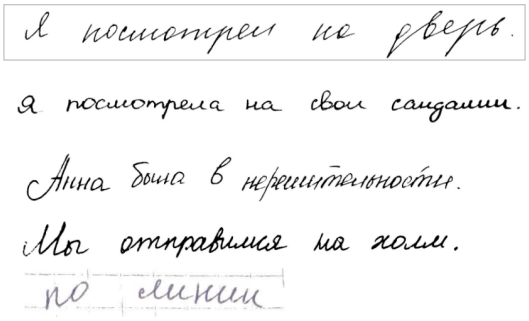


Рис. 4: RHD (ours)

Сбор данных делился на следующие этапы:

- Создание/сбор текстов для задания "написать и сфотографировать текст"
- Люди пишут и фотографируют заранее заданную строку текста, которая выбирается из собранных данных на предыдущем этапе
- Другие люди проверяют изображение на соответствие некоторым критериям
- Строки текста извлекаются с помощью детектора, обученного на печатных текстах. Данный детектор хорошо работает и в случае рукописных текстов.

2.2 Аугментация

В случае рукописных текстов нельзя не воспользоваться искусственным методом увеличения выборки – агументацией.

Были разработаны два вида аугментации (Рис.5):

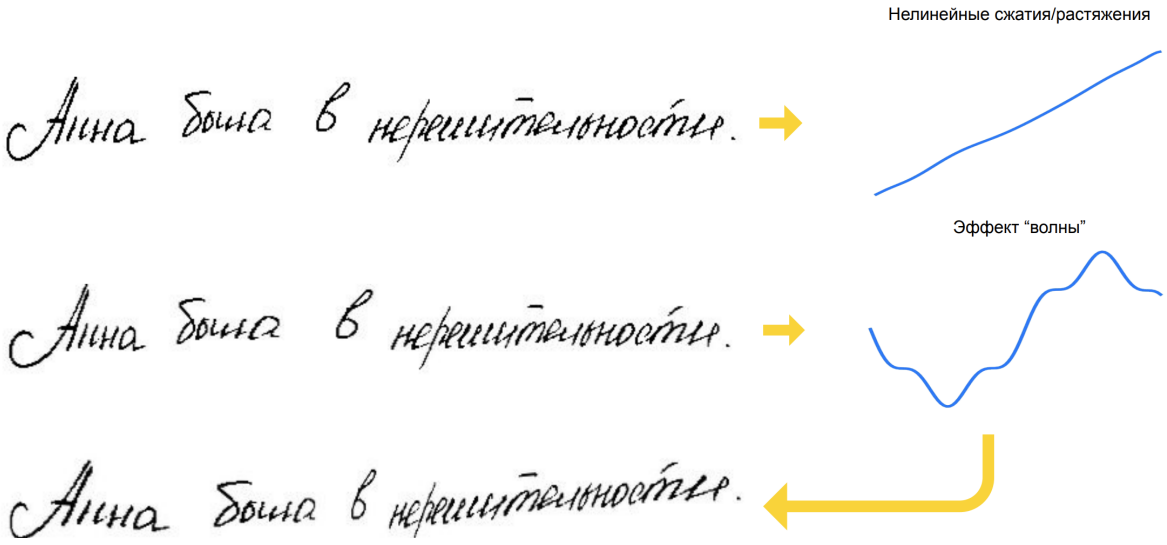


Рис. 5: Augmentation

- Нелинейные сжатия/растяжения – каждая точка изображения сдвигается по координате X на некоторое расстояние от своего изначального положения.

- Эффект ”волны” – каждая точка изображения сдвигается по координате Y на некоторое расстояние.

2.3 ”Рукописные” шрифты

В виду легкой доступности генерации ситнетических печатных текстов было решено попробовать сгенерировать тексты, использующие шрифты, похожие на рукописные (Рис.6)

Рис. 6: ”Рукописные” шрифты

2.4 Эксперименты

2.4.1 IAM

Datasets	WER	CER
Синтетика	0.87	-
IAM	0.32	-
IAM + aug	0.25	-
Синтетика + IAM (10%)	0.27	-
Синтетика + IAM + aug (10%)	**0.21**	**0.084**
Синтетика + IAM + aug (15%)	**0.21**	**0.084**
Синтетика + IAM + aug (5%)	0.27	-

Таблица 1: Сравнение качества нашей модели на разных данных

Methods	IAM		RIMES	
	WER	CER	WER	CER
(Salvador et al. 2011)	22.4	9.8	-	-
(Pham et al. 2014)	35.1	10.8	28.5	6.8
(Bluche 2016)	24.6	7.9	12.6	2.9
(Sueiras et al. 2018)	23.8	8.8	15.9	4.8
(Bhunia et al. 2019)[1]	**17.2**	8.4	10.5	6.4
(Zhang et al. 2019)	22.2	8.5	-	-
DAN	19.6	**6.4**	**8.9**	**2.7**

[1] Word-level recognition, where the words in the original image are cropped out then recognized.

Таблица 2: Предыдущие работы

В Таб.1 представлены результаты экспериментов для разных соотношениях синтетических данных, исходного IAM, аугментированного IAM. Модель, обученная на полностью синтетических данных, как и ожидалось показала плохие результаты. Эксперименты с IAM без аугментаций дают результаты существенно лучше, чем чистая синтетика. Добавление аугментации улучшило качество еще на 7%. Наиболее успешной конфигурацией оказалась синтетика в сочетании с агументированным IAM

В Таб.2 агрегированы результаты некоторых статей. Можно видеть, что в сравнении наша модель занимает 2-ое место[2] по метрике WER, что можно считать очень хорошим результатом, учитывая, что мы не отходили от стандартной архитектуры модели.

2.4.2 Русскоязычные датасеты

В Таб. 3 представлены результаты экспериментов для разных соотношениях исходного RHD, аугментированного RHD. Эксперименты с RHD без аугментаций дают результаты неплохой результат, но добавление аугментации разительно улучшает качество на 19%. Наиболее успешной конфигурацией оказалась синтетика в сочетании с аугментированным IAM

В Таб. 4 в столбце CyrHD можно сравнить качество нашей модели с лучшим решением с платформы Kaggle. Нам удалось улучшить качество по метрике WER на 10%, получив значение 0.39.

[2]Bhunia er al. 2019 в расчет не берется, они сами вырезали строки текста из исходного датасета

Datasets	WER
rhd	0.3
rhd + aug	0.11
Синтетика + rhd (10%)	0.13
Синтетика +rhd + aug (10%)	0.11
Синтетика +rhd + aug (15%)	**0.09**

Таблица 3: Сравнение качества нашей модели на разных данных

	CyrHD		Logs		rhd	
Data	**WER**	**CER**	**WER**	**CER**	**WER**	**CER**
Синт + rhd+aug (10%)	2.60	1.40	0.93	0.46	0.09	0.04
Синт + rhd+aug(10%) + cyrhd(10%)	**0.39**	**0.11**	**0.57**	**0.26**	0.09	0.04
Kaggle	0.50	0.11	-	-	-	-

Таблица 4: Предыдущие работы

2.5 Генерация изображений

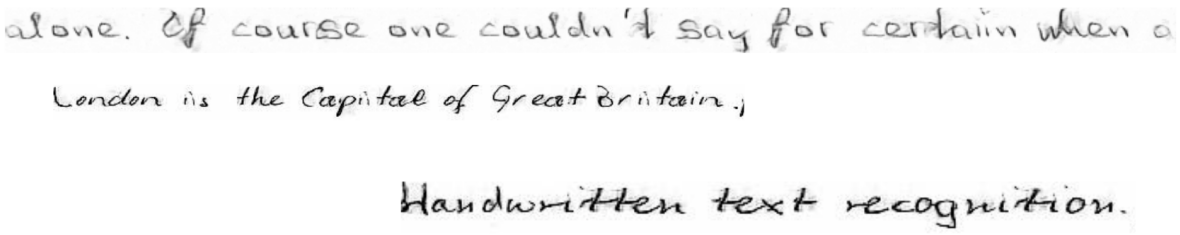


Рис. 7: Примеры сгенерированных изображений на английском языке

Для генерации английского языка был обучен Scabble GAN на датасете IAM. Нам удалось восстановить результаты из статьи и получить рукописные тексты приемлемого качества

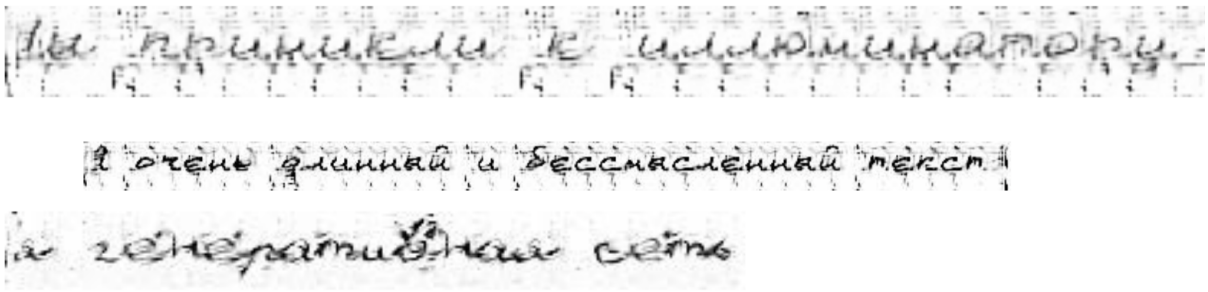


Рис. 8: Примеры сгенерированных изображений на русском языке

Для генерации русского языка был обучен Scabble GAN на датасете RHD. Нам не удалось получить приемлемое качество для русского языка. Одной из причин может служить семантическая простота нашего датасета – текст часто был достаточно прост и предсказуем, поэтому распознаватель мог переобучиться под него. Также русский рукописный текст субъективно более сложен для написания, в нем присутствуют нетривиальные связи между буквами в отличии от английского языка, где таких связей меньше.

3 Заключение

В данной работе был собран большой датасет русского рукописного текста, обучена модель на английском языке показывающая качество, сравнимое с лучшими моделями, обучена модель на русском языке, а также исследовано поведение генеративной сети Scrabble Gan для генерации русского рукописного текста.

Список литературы

[1] Abdallah A., Hamada M., Nurseitov D. Attention-based Fully Gated CNN-BGRU for Russian Handwritten Text //Journal of Imaging. – 2020. – Т. 6. – №. 12. – С. 141.

[2] Fogel S. et al. Scrabblegan: Semi-supervised varying length handwritten text generation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – С. 4324-4333.

[3] Wang T. et al. Decoupled attention network for text recognition //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 07. – С. 12216-12224.

[4] www.kaggle.com/constantinwerner/cyrillic-handwriting-dataset

[5] github.com/abdoelsayed2016/HKR_Dataset

[6] https://fki.tic.heia-fr.ch/databases/iam-handwriting-database