Распознавание и генерация изображений рукописных текстов на русском языке

Студент: Тыцкий В.И.

Научный руководитель: Майсурадзе А.И.

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

Введение

- Распознавание текста (OCR) важная задача в машинном обучении
- Распознавание **печатного текста** решается достаточно хорошо
- Рукописные тексты обладают большей спецификой
  - Малое количество данных
  - Более сложный домен – каждый почерк уникален

Данные

- Cyrillic Handwriting Dataset (CyrHD) [4]
- Handwritten Kazakh and Russian (HKR) (не для коммерческого использования) [5]
- IAM Handwriting Database [6]

Принято решение собрать свой датасет

Связанные работы

- Attention-based Fully Gated for Russian Handwritten Text [1]
- Scrabblegan: Semi-supervised varying length handwritten text generation [2]

Актуальность

- Модель русского рукописного распознавания может использоваться в индустрии
- Собранные данные могут быть полезны для сообщества
- Генерация синтетических данных GAN'ом может применяться для создания правдоподобных рукописных текстов

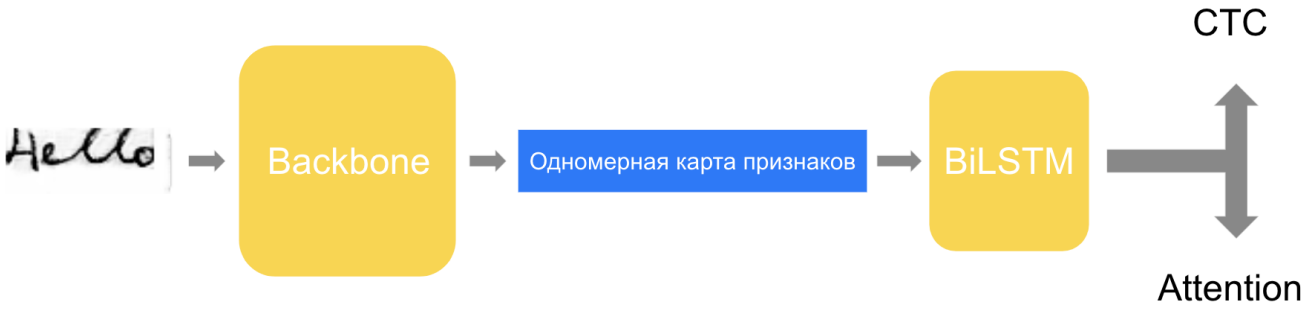
Постановка задачи

- Собрать данные с помощью краудфандинга
- Обучить нейронную сеть для распознавания русских рукописных текстов
- Попробовать использовать синтетические данные, сгенерированные GAN'ом для улучшения качества распознавания

Сбор данных

- Сбор данных проводился в сервисе Толока
- Людей просили написать короткую строчку заранее заданного русского текста
- Проверка корректности внесенных изображений проводилась голосованием с перекрытием
- Данные можно использовать как для обучения распознавания, так и для обучения GAN'a

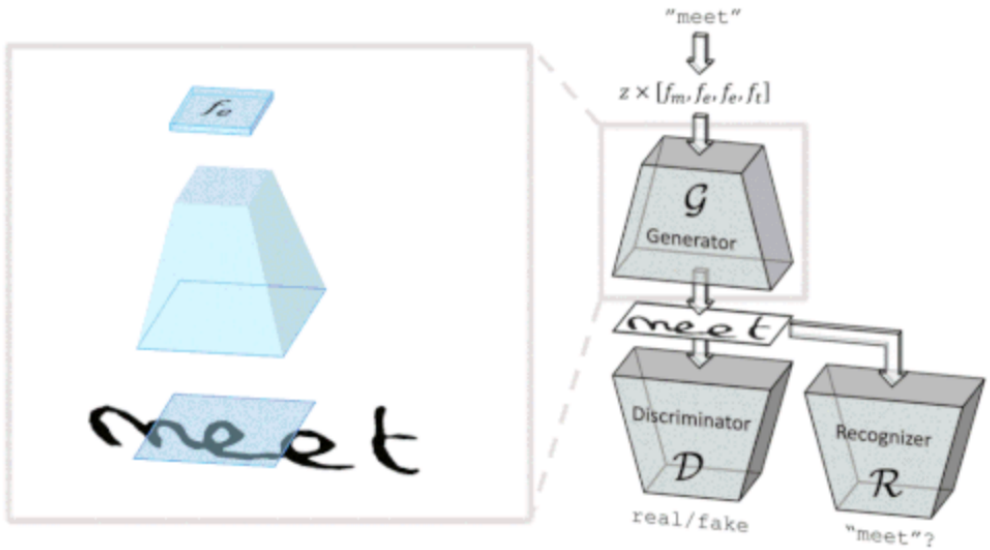
Обучение модели

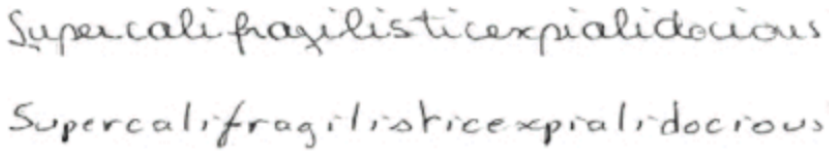


- backbone: Inception
- BiLSTM: 512 hidden size
- Выход: train – Attention + CTC, eval – CTC

Синтетика GAN'ом

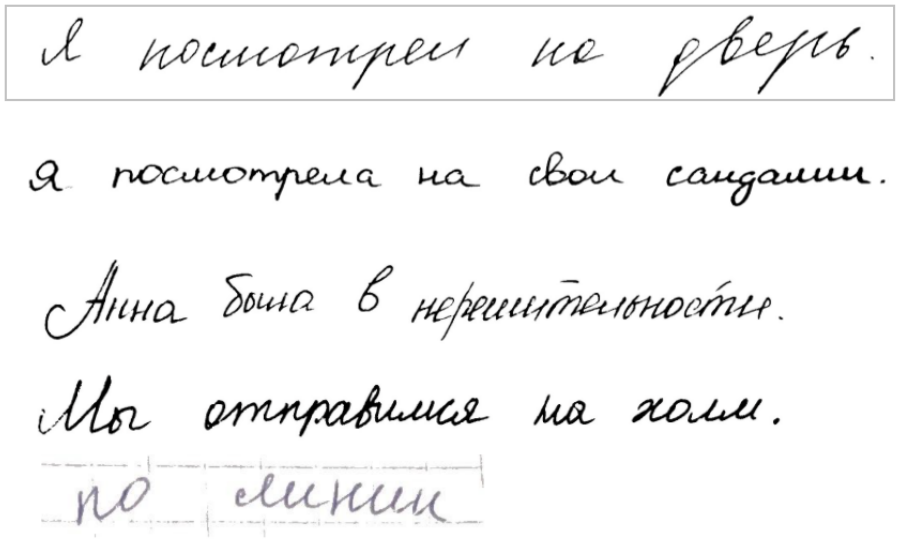
За основу взят Scrabble GAN [2]





Данные

- Легко масштабируемый процесс сбора
- Размер датасета: 13000 строк
- "Некачественных" данных $< 4\%$

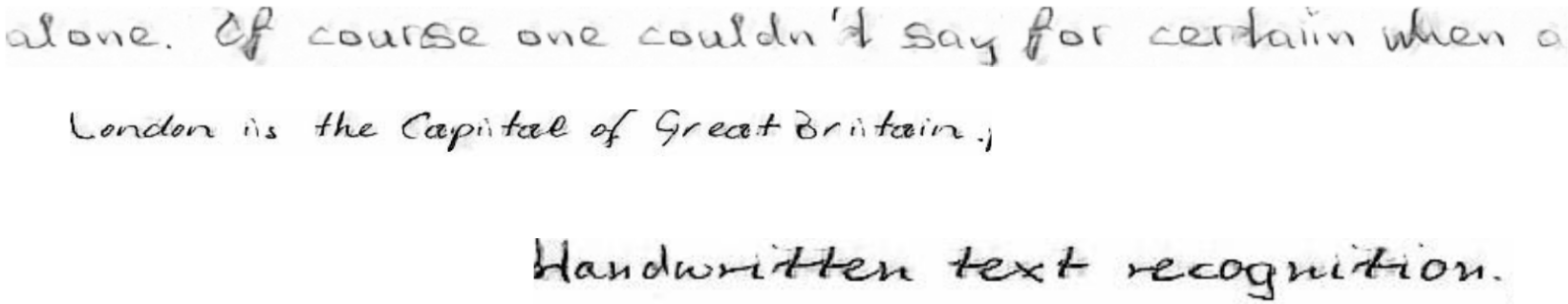


Модель

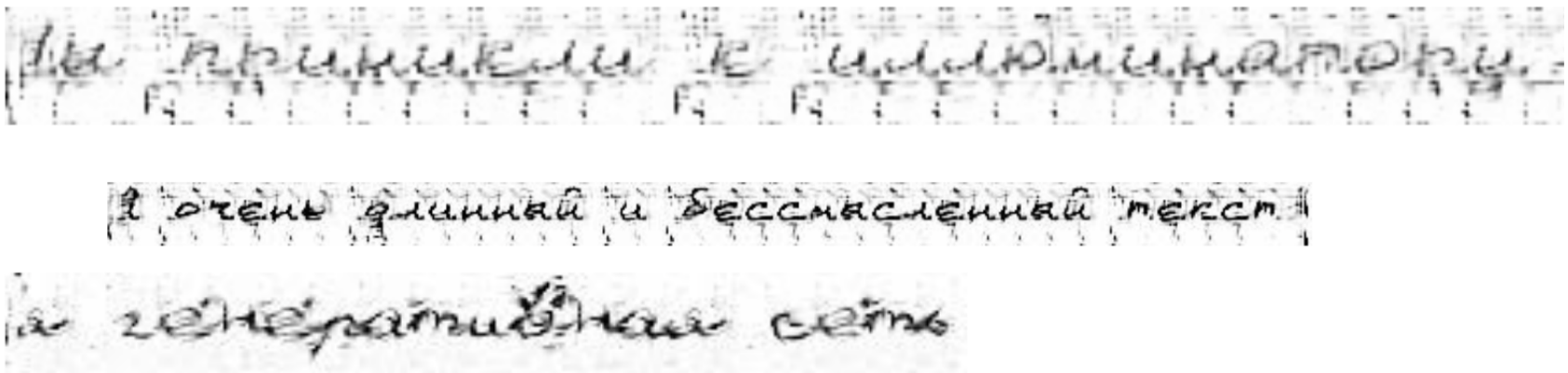
Data	RHD (ours)		CyHD		IAM (EN)	
	WER	CER	WER	CER	WER	CER
Ours	0.09	0.04	**0.39**	**0.11**	**0.21**	**0.08**
CNN-BGRU	-	-	-	-	0.25	0.08
Kaggle	-	-	0.50	0.11	-	-

Синтетика GAN'ом

EN



RU



Заключение

- Собран большой датасет русских рукописных текстов
- Успешно обучена модель распознавания
- Применен GAN для генерации русского рукописного текста

[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