Оцифровка рукописного текста методами машинного обучения

Глеб Нуждов

Студент

Механико-математический МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: gleb.nuzhdov@gmail.com

Научный руководитель *— Илларионов Егор Александрович*

В период 1883-1938 гг. в Федеральной обсерватории Цюриха (ориг. Eidgenössische Sternwarte) производились измерения солнечных пятен, которые были записаны от руки в таблицы на бумаге. В дальнейшем таблицы были отсканированы и выложены в открытый доступ по ссылке 1. Нашей основной задачей было предложить эффективный способ оцифровки этих таблиц.

Мы разработали подход машинного обучения к оцифровке рукописных таблиц с числовыми данными исторических наблюдений солнечных пятен. В качестве первого шага мы создали алгоритм распознавания табличных сеток и извлечения ячеек сетки. Мы использовали его для сбора набора данных из 15 тысяч изображений содержащих рукописные числа, которые были размечены вручную. На основе этого набора данных мы обучили сверточную рекуррентную модель нейронной сети для распознавания чисел. Модель показывает точность 97.5% для даннных без знака. По итогам работы подготовим и выложим в открытый доступ датасет рукописных чисел из таблицы. Благодаря зависимостям между ячейками в таблице, датасет будет размечен с точностью 100%. Подобные наборы данных пока существуют только для отдельных цифр, и на этих наборах найдены эффективные и простые в реализации решения, тогда как для распознавания чисел готовых моделей пока нет.

Литература

1. Ссылка на данные
   https://www.cloud.aip.de/index.php/s/Ma4DMweJfMpQby3