РАСПОЗНАТЬ И ТРАНСКРИБИРОВАТЬ

АВТОГРАФЫ ПЕТРА ВЕЛИКОГО И ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Татьяна Базарова,
кандидат исторических наук, заведующая Научно-историческим архивом и группой источниковедения СПбИИ РАН

Денис Димитров,
senior data scientist группы компьютерного зрения ПАО «Сбербанк», механико-математический факультет МГУ

Марк Потанин,
senior data scientist группы компьютерного зрения ПАО «Сбербанк», ФПМИ МФТИ

Мария Проскурякова,
кандидат исторических наук, старший научный сотрудник СПбИИ РАН

В 2022 году исполняется 350 лет со дня рождения первого российского императора — Петра Великого. С конца XVIII века не утихают споры о его личности и характере. Каждая юбилейная дата, связанная с жизнью и деятельностью венценосного реформатора, заставляла глубже всматриваться в прошлое, тщательнее изучать рукописное наследие Петра Великого и его современников

СОБИРАЯ РУКОПИСНОЕ НАСЛЕДИЕ

В 1872 году, когда российское общество торжественно отмечало 200-летний юбилей Петра I, у одного из основателей Императорского Русского Исторического общества, академика Афанасия Федоровича Бычкова (1818–1899) зародилась идея публикации писем и бумаг первого российского императора в качестве единого комплекса источников. Грандиозный замысел получил поддержку министра народного просвещения Дмитрия Андреевича Толстого и был одобрен императором Александром II. В декабре 1872 года была создана Комиссия по изданию писем и бумаг Петра Великого под председательством Д.А. Толстого[1]. На своих заседаниях члены Комиссии выработали принципы отбора и копирования документов. В течение долгих лет велась кропотливая работа по выявлению документов в российских архивах, библиотеках, в зарубежных собраниях. К 1885 году Комиссии удалось собрать около 15000 копий документов. В 1887 году увидел свет первый том «Писем и бумаг императора Петра Великого»[2], в который вошли документы за 1688–1701 годы.

Работа по подготовке томов продолжающегося издания (со значительными остановками) велась и после кончины А.Ф. Бычкова[3]. Второй выпуск тринадцатого тома вышел из печати в 2003 году (материалы за июль – декабрь 1713 года). Значительная часть эпистолярного наследия Петра Великого до сих пор остаётся не введённой в широкий научный оборот. Рукописные копии писем и бумаг Петра Великого (с 1688 по 1725 год), созданные в результате работы Комиссии, в настоящее время хранятся в Научно-историческом архиве Санкт-Петербургского института истории РАН (ф. 270) (Ил. 1). Это собрание известно и востребовано исследователями, занимающимися историей России Петровской эпохи.

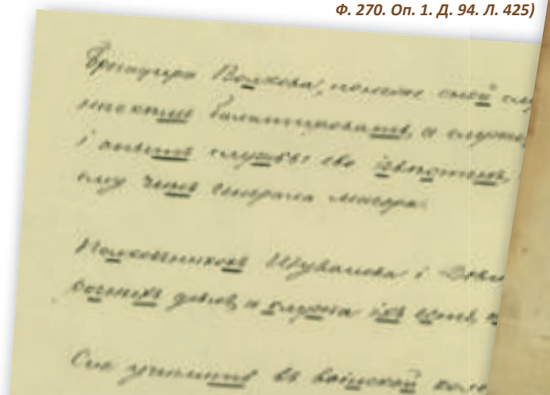
Рукописное наследие Петра обширно и многообразно и включает как собственноручные письма и записки, черновики указов, уставов и инструкций (часто с многочисленной правкой), так и небольшие приписки и подписи под писарским текстом (Ил. 2). Работа по выявлению и публикации бумаг Петра Великого не остановилась после прекращения деятельности Комиссии. Ученые продолжают выявлять в российских и европейских архивных собраниях и вводить в научный оборот автографы Петра Великого. В связи с этим проблема прочтения и транскрибирования текста, написанного его рукой, продолжает оставаться актуальной.

В РУСЛЕ ПРОБЛЕМ МИРОВОЙ НАУКИ

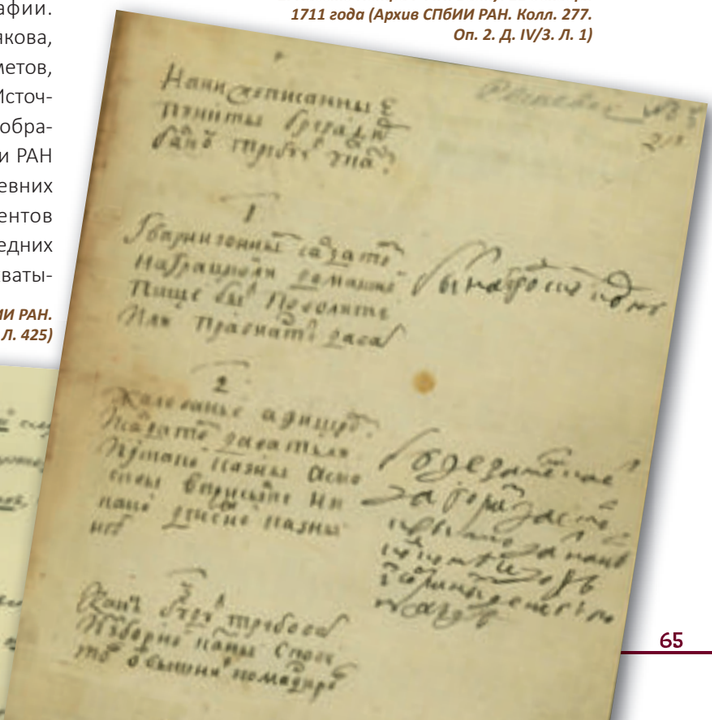
Перспективы применения новых методов при изучении петровских автографов открываются благодаря научно-исследовательскому проекту «Автографы Петра Великого: Чтение технологиями искусственного интеллекта», инициированному Российским историческим обществом и ПАО «Сбербанк».

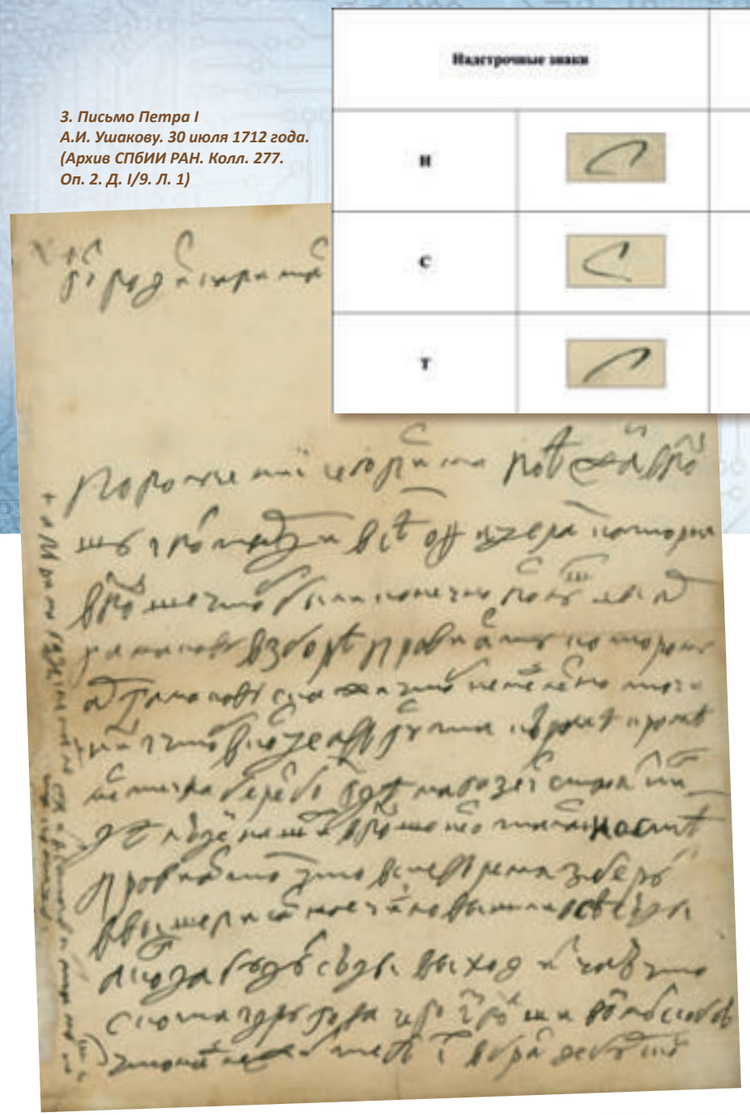
В ходе работы над проектом в Санкт-Петербургском институте истории РАН была сформирована рабочая группа, состоящая из научных сотрудников – специалистов по истории Петровской эпохи, а также палеографии и археографии. В группу вошли Т.А. Базарова, М.Е. Проскурякова, И.А. Поляков, А.А. Калашникова, Н.А. Шереметов, Е.С. Дилигул, Ю.Б. Фомина, М.В. Тихонова. Источниковой базой проекта стали рукописи из собраний Санкт-Петербургского института истории РАН и Российского государственного архива древних актов (РГАДА). Выявление и отбор документов производились на основе материалов последних томов «Писем и бумаг Петра Великого», охватывавших период с 1709 по 1713 год. Основным критерием отбора являлось наличие в документе текста, написанного рукой Петра (от двух-трех слов до нескольких страниц) (Ил. 3). Огромную помощь и поддержку оказали Росархив и РГАДА, которые предоставили рабочей группе цифровые копии автографов.

1. Резолюция Петра I. Копия XIX века. (Архив СПбИИ РАН. Ф. 270. Оп. 1. Д. 94. Л. 425)

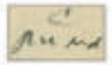


2. Письмо Петра I Ф.Н. Балку. 11 декабря 1711 года (Архив СПбИИ РАН. Колл. 277. Оп. 2. Д. IV/3. Л. 1)





3. Письмо Петра I А.И. Ушакову. 30 июля 1712 года. (Архив СПбИИ РАН. Колл. 277. Оп. 2. Д. I/9. Л. 1)

Надстрочные знаки		Примеры слов	
н		имено	
с		писма	
т		отпиз	

4. Примеры написания Петром I выносных букв

Изучение массива документов позволило выявить характерные черты почерка Петра Великого. Важная особенность петровских автографов – малая вариативность написания: как правило, буква имеет одну, реже две формы (символа). В то же время начертание ряда выносных знаков близко друг к другу, а порой даже идентично. Так, например, схожим надстрочным знаком передаются буквы «н», «с», «т» (Ил. 4). Прочтение и транскрибирование текста затрудняет отсутствие пробелов между словами и использование сокращений. Пётр употреблял два вида сокращений: по типу инициальных сиглей (сокращение до начальной буквы слова или начальных букв нескольких слов, например, «ч» – «ч(еловек)», «г ф» – «г(енерал)-ф(ельдмаршал)») и по типу суспенсии (сокращение слогов в отдельных словах, например, «де» – «де(нь)»).

Как и у многих его современников, в почерке Петра наблюдается вариативность в написании «ер» (ъ) и «ерь» (ь), а также заглавных и строчных букв. Кроме того, в исследованных материалах не было зафиксировано ни одного использования букв «кси» (ѯ) и «пси» (ѱ). Между тем буква «пси», хотя и была упразднена в ходе реформы 1710 года, встречалась в рукописных и печатных текстах на протяжении всей Петровской эпохи. В собственноручных текстах Петра Великого не удалось обнаружить букву «ферт» (ф), вместо которой употреблялась «фита» (Ѳ). Примечательно, что эту же букву царь активно использовал в качестве выносной в конце слов. В таких случаях она служила «заменой» букве «в». Последнее наблюдение, к примеру, позволило выявить несколько неточных прочтений петровских автографов в «Письмах и бумагах Петра Великого».

ПРИНЦИПЫ И МЕТОДИКИ

Учитывая эту специфику, рабочая группа совместно со специалистами по анализу данных Сбербанка Д.В. Димитровым и М.С. Потаниным выработала методику передачи текста для последующей компьютерной обработки. Основным принципом стала максимально точная передача знаков (букв), которые использовал Петр Великий. Тексты документов, опубликованные в томах «Писем и бумаг Петра Великого», данному требованию не отвечали. Помимо букв современного нам алфавита при компьютерном наборе использовались «ять» (Ѣ) и «и десятеричная» (і), мягкий и твёрдый знаки при их отсутствии не восполнялись, сокращения не раскрывались.

Одним из приёмов издания автографов Петра Великого стало выделение курсивом выносных букв[4]. Участники настоящего проекта этому правилу не следовали: надстрочные знаки вносились в строку и не выделялись графически. Компьютерному набору подлежал только собственноручный текст государя в документе (включая зачеркнутые буквы, слова и предложения). Каждой строке присваивался порядковый номер, так было выделено 9238 строк. Компьютерный набор несколько раз сверялся с цифровой копией подлинника, спорный случай подлежал коллегиальному обсуждению.

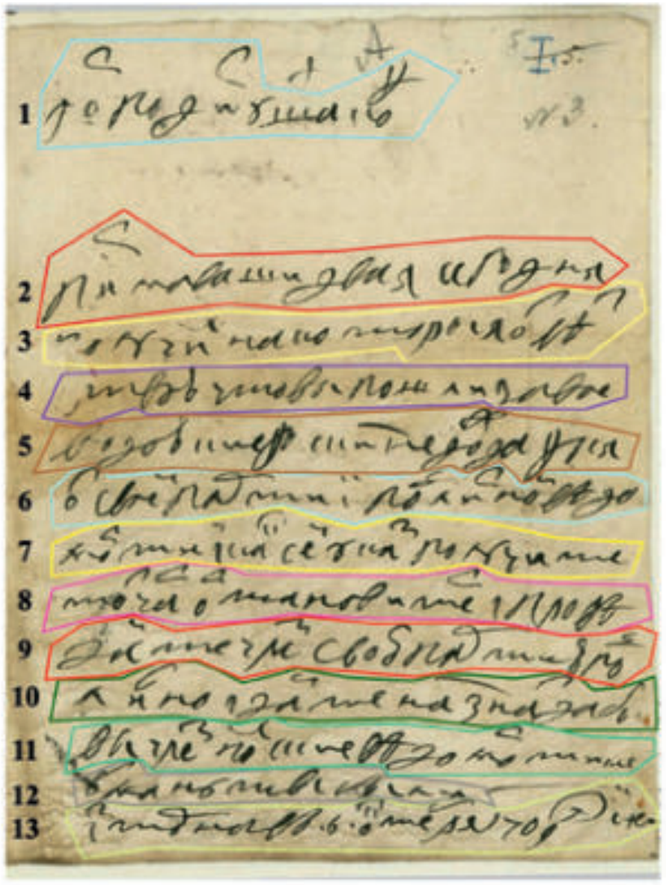
Следующий этап работы начался после загрузки цифровых копий в web-приложение Computer Vision Annotation Tool (далее – CVAT) – необходима была построчная разметка цифровых копий документов. На этом этапе к выполнению проекта подключились имевшие успешный опыт работы со скорописью XVII - XVIII веков аспиранты и магистранты НИУ ВШЭ (М.Д. Аксенова, А.О. Видничук, А.Д. Новикова, М.И. Парфеня, М.С. Петрова, М.А. Всемирнов, А.Д. Гусак, А.М. Новикова, А.И. Репникова). Сотрудники Сбербанка Д.В. Димитров и М.С. Потанин удалённо провели обучение как членов рабочей группы (Т.А. Базаровой, М.Е. Проскуряковой, А.А. Калашниковой), так и аспирантов и магистрантов, а также постоянно держали под контролем процесс разметки текста. Благодаря тесному сотрудничеству и совместной работе специалистов из разных учреждений, подготовку данных для разработки программы удалось выполнить на высоком уровне и в установленный срок.

Реализация проекта находится в русле проблем, стоящих перед мировой наукой. Схожие задачи решают европейские исследователи, реализуя идею построения алгоритма для распознавания древних рукописных текстов, прежде всего на латинице. Целями международного проекта «The Recognition and Enrichment of Archival Documents (READ)» заявлены усовершенствование технологии распознавания рукописных текстов и транскрибирование исторических документов (https://eadh.org/projects/read). Этот проект курирует Европейская ассоциация цифровых гуманитарных наук (European Association for Digital Humanities (EADH). Он объединяет более десяти европейских научных центров, а также архивы. Одним из результатов стало создание платформы Transkribus для машинного чтения древних рукописей (http://transkribus.eu). Платформа включает около 50 моделей для распознавания текстов и позволяет обрабатывать документы, написанные ранними формами кириллического письма – уставом и полууставом.

Эту задачу призваны решить модели, построенные на распознавании текста церковно-богослужебных книг XI и XVI веков. Однако тестовое транскрибирование нескольких образцов полууставного письма выявило многочисленные неточности и ошибки прочтения. Прежде всего следует отметить низкое качество распознавания надстрочных знаков, которые в отдельных случаях «видятся» алгоритмом как самостоятельные строки и неверно идентифицируются как предлоги или окончания слов верхней строки. Кроме того, основной массив русской делопроизводственной документации XV–XVII веков написан скорописью, а её Transkribus «читать» не умеет.

Научно-исследовательский проект «Автографы Петра Великого: Чтение технологиями искусственного интеллекта» призван раскрыть перед российскими историками новые возможности работы с рукописными текстами. Главной задачей стало создание алгоритма, который способен автоматически распознавать и транскрибировать автографы Петра Великого, то есть преобразовывать изображение со скорописью в текст. Кроме того, разработанный алгоритм должен послужить отправной точкой для участников соревнования **«Digital Пётр: распознавание рукописей Петра I»**, проводимого Сбербанком в рамках Artificial Intelligence Journey 2020 (AIJ 2020, https://ai-journey.ru/contest/task01) – ежегодной серии мероприятий, посвящённых искусственному интеллекту.

На данном этапе исследований перед разрабатываемым алгоритмом, или моделью, ставилась задача распознавать не полное изображение документа, а лишь предварительно вырезанные строки. Цель участников соревнования – улучшить текущую



5. Документ, разрезанный на строки, и результат транскрибирования изображения

1 господин ушакоф
2 писма ваши два я сего дня
3 получил на которыя отвѣт
4 твую что вы пошли за вое
5 водою киефъским не дождафъся
6 от своей партіі подлинной вѣде
7 мости і как сей указ получите
8 тотчас остановитес і провѣ
9 дайте чрез свою партию под
10 линно і дайте нам знат дабы
11 вы чрез полские вѣдомости не
12 обмануты были
13 із торна въ 6 октября 1709 Piter

модель или предложить совершенно новый метод, тем самым увеличив качество распознавания автографов Петра Великого. При этом перед авторами статьи стоит не менее амбициозная задача – научить модель автоматически разрезать страницу на строки[5] (которые в некоторых случаях могут быть даже вертикально ориентированными) для последующего распознавания. Такой алгоритм будет представлен позднее. Подобное упрощение сделано сознательно, в том числе для того, чтобы участники соревнования могли сконцентрироваться именно на качестве распознавания текста, а не на качестве автоматического разрезания страницы на строки.

В ПОИСКАХ ОПТИМАЛЬНЫЙ МОДЕЛИ

После изучения литературы по распознаванию рукописного текста[6] был выбран подход, основанный на глубоком обучении и нейронных сетях. Были собраны два датасета[7], состоящие из транскрибированных и разрезанных на строки документов с автографами Петра. Первый датасет необходим для обучения нейронной сети (поэтому его называют обучающим). Обучение подразумевает тонкую настройку параметров нейронной сети для адаптации под конкретную задачу. Второй датасет требуется для подсчёта качества распознавания, то есть тестирования обученной модели (поэтому этот датасет

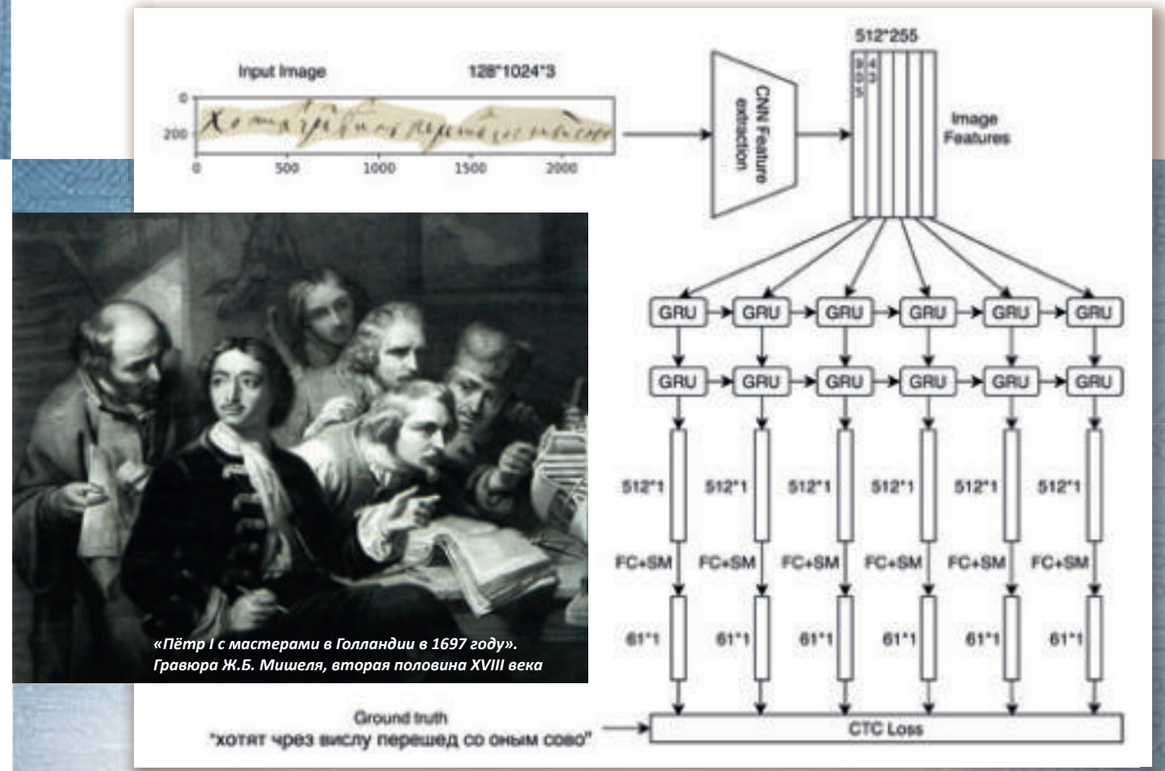
6. Извлечение признаков с помощью сверточной нейронной сети

называют тестовым). Качество обязательно вычисляется на наборе данных, который нейронная сеть «не видела» при обучении, поэтому первый и второй датасеты не пересекаются.

Первичными данными являлись отсканированные письма и бумаги Петра Великого, которые затем были прочтены и транскрибированы рабочей группой СПбИИ РАН. После этого требовалось вручную разделить исходные документы на отдельные строки и сопоставить их с транскрибированным текстом. Пример разметки исходного документа на строки в web-приложении CVAT представлен на Ил. 5. Размер обучающей выборки составил 6196 строк. Выборка, на которой производилось тестирование качества работы модели, состояла из 3042 строк.

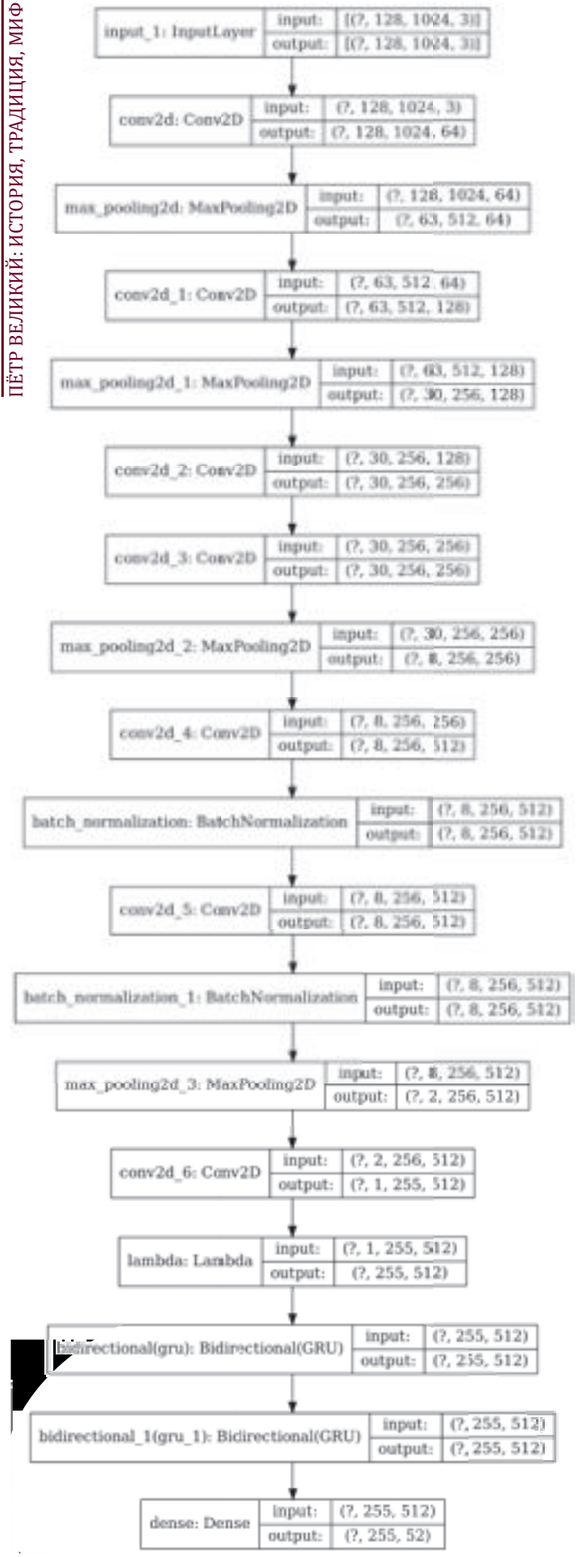
Используемая искусственная нейронная сеть в определённой степени моделирует то, как сам человек воспринимает и анализирует информацию, поступающую в мозг через зрительную сенсорную систему[8]. Изображение со скорописью сначала подается на вход свёрточным слоям (или convolutional layers) для извлечения полезных признаков. Нейронная сеть сама выучивает характерные черты написания тех или иных букв, настраивая нужным образом свои параметры (с помощью метода оптимизации – градиентного спуска). Пример такого извлечения признаков представлен на Ил. 6.

После этого каждый выученный признак (в виде вектора) подается на вход соответствующему GRU-слою (Gated Recurrent Unit). Рекуррентный слой необходим для представления входных признаков в виде последовательности и распознавания сетью символов в порядке, в котором они находятся в обучающем тексте. После обработки рекуррентным слоем каждый вектор передаётся полносвязному слою и активационной функции softmax. Размерность полносвязного слоя равна длине массива всех возможных распознаваемых символов с добавлением blank-маркера. Данный маркер модуль CTC–Loss использует для обозначения отсутствия символа в тексте, так как большая часть всех подаваемых на вход последовательностей символов будет по длине меньше максимально возможной. Softmax-функция представляет входной вектор в виде вектора распределения вероятностей для всех возможных символов, включая и маркер отсутствия символа (blank) (см. Ил. 7).



«Пётр I с мастерами в Голландии в 1697 году». Гравюра Ж.Б. Мишеля, вторая половина XVIII века

7. Принцип работы нейронной сети, которая распознает рукописный текст



8. Архитектура используемой нейронной сети

Таким образом, разработанная нейросеть состоит из нескольких CNN-блоков, за которыми следуют несколько слоёв GRU, и использует CTC–Loss. Такая структура называется CRNN. Количество настраиваемых параметров используемой нейросети составило 7 948 733. Подробная информация об её архитектуре представлена на Ил. 8.

В качестве постобработки, а именно для spell checking, была использована языковая модель, обученная на коллекции текстов XVII века. Коллекция была предложена организаторами соревнования **GramEval2020**[9]. Такая языковая модель может исправлять ошибки, допущенные основной моделью при распознавании, используя знания о словах и выражениях, которые употреблялись в XVII веке.

Для оценки качества распознавания используются следующие метрики:

1. CER – character error rate

$$\mathrm{CER} = \frac{\sum_{i=1}^{n} \mathrm{dist}_c(\mathrm{pred}_i, \mathrm{true}_i)}{\sum_{i=1}^{n} \mathrm{len}_c(\mathrm{true}_i)}$$

2. WER – word error rate

$$\mathrm{WER} = \frac{\sum_{i=1}^{n} \mathrm{dist}_w(\mathrm{pred}_i, \mathrm{true}_i)}{\sum_{i=1}^{n} \mathrm{len}_w(\mathrm{true}_i)}$$

3. String Accuracy – метрика, показывающая сколько полных строк (в процентах) было распознано идеально правильно, то есть полностью без ошибок (включая и пробелы):

$$\mathrm{String\ Accuracy} = \frac{\sum_{i=1}^{n} [\mathrm{pred}_i = \mathrm{true}_i]}{n}$$

В приведённых выше формулах $pred_i$ – строка из символов, распознанная моделью на изображении; $true_i$ – истинный перевод, произведённый экспертом; n = 3042 – количество строк в тестовом датасете. В формулах для CER и WER $dist$ – расстояние Левенштейна[10], то есть минимальное количество токенных операций (а именно вставок, удалений, замен), необходимых для превращения одной последовательности символов в другую. Только для CER токенами для сравнения являются отдельные символы ($dist_c$), а для WER токенами являются целые слова ($dist_w$). По аналогии len_c – длина строки в символах (включая пробелы), len_w – длина строки в словах. В формуле для String Accuracy используется скобка Айверсона:

$$[pred_i = true_i] = \begin{cases} 1, & \mathrm{pred}_i = \mathrm{true}_i, \\ 0, & \mathrm{pred}_i \neq \mathrm{true}_i \end{cases}$$

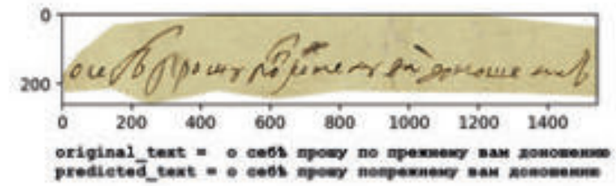
Последняя версия модели имеет следующие значения метрик:

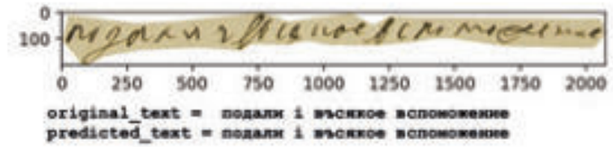
Метрика	Значение
CER	9.94%
WER	42.97%
String Accuracy	23.01%

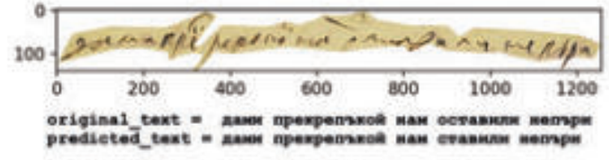
Так как мы считаем частоту ошибок, то показатель CER в 10% означает, что 90% символов распознаны правильно, что является неплохим результатом. По приведённым в Приложении 1 примерам работы модели видно, что зачастую ошибки скрываются в неправильно расставленных пробелах. Это – распространённая проблема для подхода CRNN и CTC–Loss.

Справа представлены примеры распознавания строк из документов Петра, которые попали в тестовую выборку: original_text – перевод эксперта, predicted_text – перевод нейросети, а также исходное изображение.

Преимущества описанного подхода – гибкость и способность модели быстро дообучаться на новом почерке, не требуя большого количества транскрибированных и размеченных данных: достаточно нескольких страниц с новым почерком. Создание отечественного программного обеспечения для распознавания и транскрибирования автографов Петра I откроет возможности для построения аналогичных моделей чтения русской скорописи XVI – начала XVIII веков, а также письма других эпох.







Приложение I. Примеры распознавания строк из тестовой выборки

Реализацию модели на языке Python, а также всю информацию о соревновании **«Digital Пётр: распознавание рукописей Петра I»** можно найти на следующих ресурсах: https://github.com/sberbank-ai/digital_peter_aij2020 и https://ods.ai/competitions/aij-petr

[1] *В состав комиссии вошли петербургские и московские ученые С.М. Соловьев, Н.А. Попов, К.Н. Бестужев-Рюмин, Е.Е. Замысловский, Н.В. Калачов, А.Е. Викторов.*

[2] *Письма и бумаги императора Петра Великого. СПб., 1887. Т. 1.*

[3] *Подробнее см.: Подъяпольская Е.П. Об истории и научном значении издания «Письма и бумаги императора Петра Великого» // Археографический ежегодник за 1972 г. М., 1974. С. 56–70.*

[4] *Этот прием был выработан в самом начале деятельности «Комиссии по изданию писем и бумаг Петра Великого».*

[5] *Подробнее см.: Alberti M. et al. Labeling, Cutting, Grouping: an Efficient Text Line Segmentation Method for Medieval Manuscripts, 2019 // URL: https://arxiv.org/pdf/1906.11894.pdf (Дата обращения 28.09.2020).*

[6] *Shi B., Bai X., Yao C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition, 2015 // URL: https://arxiv.org/pdf/1507.05717.pdf (Дата обращения 28.09.2020); Jaderberg M. et al. Deep structured output learning for unconstrained text recognition. In ICLR, 2015 // URL: https://arxiv.org/pdf/1412.5903.pdf (Дата обращения 28.09.2020); Cilia N. D. et al. A ranking-based feature selection approach for handwritten character recognition, 2019 // URL: https://www.sciencedirect.com/science/article/abs/pii/S0167865518301272 (Дата обращения 28.09.2020); Зеленцов И.А. Методика распознавания древнерусских скорописных текстов: Дисс… канд. техн. наук. М., 2011 // URL: https://rusneb.ru/catalog/000199_000009_005407087 (Дата обращения 28.09.2020).*

[7] *Датасет (англ. Dataset) – набор данных.*

[8] *Mahapattanakul P. From Human Vision to Computer Vision – Convolutional Neural Network (Part 3/4) // URL: https://becominghuman.ai/from-human-vision-to-computer-vision-convolutional-neural-network-part3–4–24b55ffa7045 (Дата обращения 28.09.2020).*

[9] *GramEval-2020 // URL: https://competitions.codalab.org/competitions/22902#learn_the_details (Дата обращения 28.09.2020).*

[10] *Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. М., 1965. Т. 163. № 4. С. 845–848.*