УДК 004.4

РАЗРАБОТКА СТАНДАРТИЗИРОВАННОЙ БАЗЫ ДАННЫХ СИМВОЛОВ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ НА ОСНОВЕ РАСШИРЕННЫХ КИРИЛЛИЧЕСКИХ АЛФАВИТОВ ТЮРКСКИХ ПИСЬМЕННОСТЕЙ

А. Д. Яблонцева
Научный руководитель – А. А. Голубничий
Хакасский государственный университет им. Н. Ф. Катанова, пр. Ленина, 92/1, 655017, г. Абакан, Россия, 91919919@mail.ru

Данная статья описывает процесс создания стандартизированной базы данных символов расширенных кириллических алфавитов тюркских письменностей, которую можно использовать для построения моделей машинного обучения. В статье приводятся результаты работы отдельных моделей машинного обучения на разработанном наборе данных.

Ключевые слова: *машинное обучение, системы машинного зрения, стандартизованные наборы данных, кириллические письменности.*

DEVELOPMENT OF A STANDARDIZED DATABASE OF SYMBOLS FOR BUILDING MACHINE LEARNING MODELS BASED ON EXTENDED CYRILLIC ALPHABETS OF TURKIC WRITING

A. D. Yablontseva
Scientific supervisor – A. A. Golubnichiy
Katanov Khakass State University, ave. Lenin, 92/1, 655017, Abakan, Russia, 91919919@mail.ru

This article describes the process of creating a standardized database of symbols of extended Cyrillic alphabets of Turkic writing, which can be used to build machine learning models. The article provides the results of the work of individual machine learning models at the developed data set.

Keywords: *machine learning, machine vision systems, standardized data sets, Cyrillic writing writing*

В области искусственного интеллекта и машинного обучения существует множество наборов данных, используемых для тренировки моделей машинного обучения, наиболее популярными из них являются MNIST (набор рукописных символов), CIFAR-10 (набор тематических изображений из 10 классов), ImageNet (набор аннотированных изображений, с несколькими фактами об объекте) и ряд других датасетов, относящихся к конкретной предметной области, таких как Fashion-MNIST (изображения одежды), Caltech-UCSD Birds 200 (изображения птиц) и т. д. При этом в настоящее время отсутствуют достаточно серьезные и массивные датасеты, содержащие сведения о печатных буквах основной и расширенной кириллиц, содержащих множественные искажения.

Актуальность создания такого датасета для детальной настройки моделей машинного обучения обусловлена тем, что в настоящее время культурное наследие многих народов России, в том числе Енисейской Сибири, представлено в виде печатных трудов, вышедших в период 1930–1980 годов. Твердые копии многих произведений находятся в достаточно сомнительном состоянии, использование стандартных средств оцифровки без осуществления распознавания символов в значительной степени усложняет использование исходных текстов, при этом языковые модели для систем распознавания символов разработаны исключительно для популярных языков.

В качестве основы для построения системы оптического распознавания символов была выбрана письменность 23 тюркских языков, таблица показана в правой части слайда, языки имеют кириллическое написание, при этом некоторые из рассмотренных языков, помимо основной кириллицы, содержат дополнительные буквы. На левой части данного слайда представлены всевозможные варианты букв, используемых в системе оптического распознания, также именуемые как «Белые списки». Помимо основных 33 букв кириллицы, в набор дополнительно включены 29 букв из тюркоязычной письменности. Основные сведения об языках, включенных в процесс набора данных, представлены в таблице.

Набор данных для создания датасета

Письменность	Буквы добавлены /исключены	Письменность	Буквы добавлены /исключены
азербайджанская	Ғ Ә Ј Қ Ө Ү Һ Ч ' / Ё Й Ц Щ Ъ Ь Э Ю Я	сойотская	Ғ І Қ Һ Ң Ө Ү Ҷ Ә
алтайская	Ј Н̆ Ӧ Ӱ	татарская	Ә Ж Ң Ө Ү Һ
башкирская	Ғ Ҙ Ҡ Ң Ө Ҫ Һ Ә	тофаларская	Ғ і Ҕ Һ ҥ Ө Ҷ / Ц
гагаузская	Ӓ Ӝ Ӧ Ӱ	тувинская	Ң Ө Ү
казахская	Ә Ғ Қ Ң Ө Ұ Һ І	туркменская	Ж Ң Ө Ү Ә
караимская	Гъ Дж Къ Нъ Ӧ Ӱ Хъ / Ё	узбекская	Ў Қ Ғ Ҳ ' / Щ Ы
каракалпакская	Ә Ғ Қ Ң Ө Ў Ҳ	уйгурская	Ғ Һ Ж Қ Ң Ө Ү / Ё
карачаево-балкарская	Гъ Дж Къ Нг	хакасская	Ғ І Ң Ӧ Ў Ҷ
киргизская	Ң Ө Ү	чувашская	Ӑ Ё Ҫ Ӳ
крымско-татарская	Гъ Къ Нъ Дж	шорская	Ғ Қ Ң Ӧ Ӱ
кумыкская	Гъ Гь Къ Нг Оь Уь	якутская	Һ Ҕ Ө Ү
ногайская	Аь Нъ Оь Уь		

Исходный набор данных EMNIST TCyrillic, построенных на основе ранее разработанного датасет [1], составляет 74 088 изображений, относящихся к одному из 147 классов. Таким образом, количество изображений

на один класс составляет всего 504 экземпляра. Столь малый объем вполне приемлем для печатных наборов данных по причине того, что исполнение букв в печатном виде несет значительно меньший уровень вариаций и искажений в сравнении с рукописным начертанием. Соответственно, отпадает необходимость в использовании нескольких тысяч изображений для каждого класса, как это представлено в наборе MNIST (более 7 000 изображений каждого символа).

Первые результаты для набора данных при использовании линейного классификатора, реализованного посредством персептрона, показали соразмерный уровень предсказания на уровне 92–93 %, что соответствует аналогичным результатам для набора данных MNIST (92 %). Для более сложных моделей, таких как LeNet-5 результат также оказался соразмерным с уровнем предсказания 98 % (рис.).

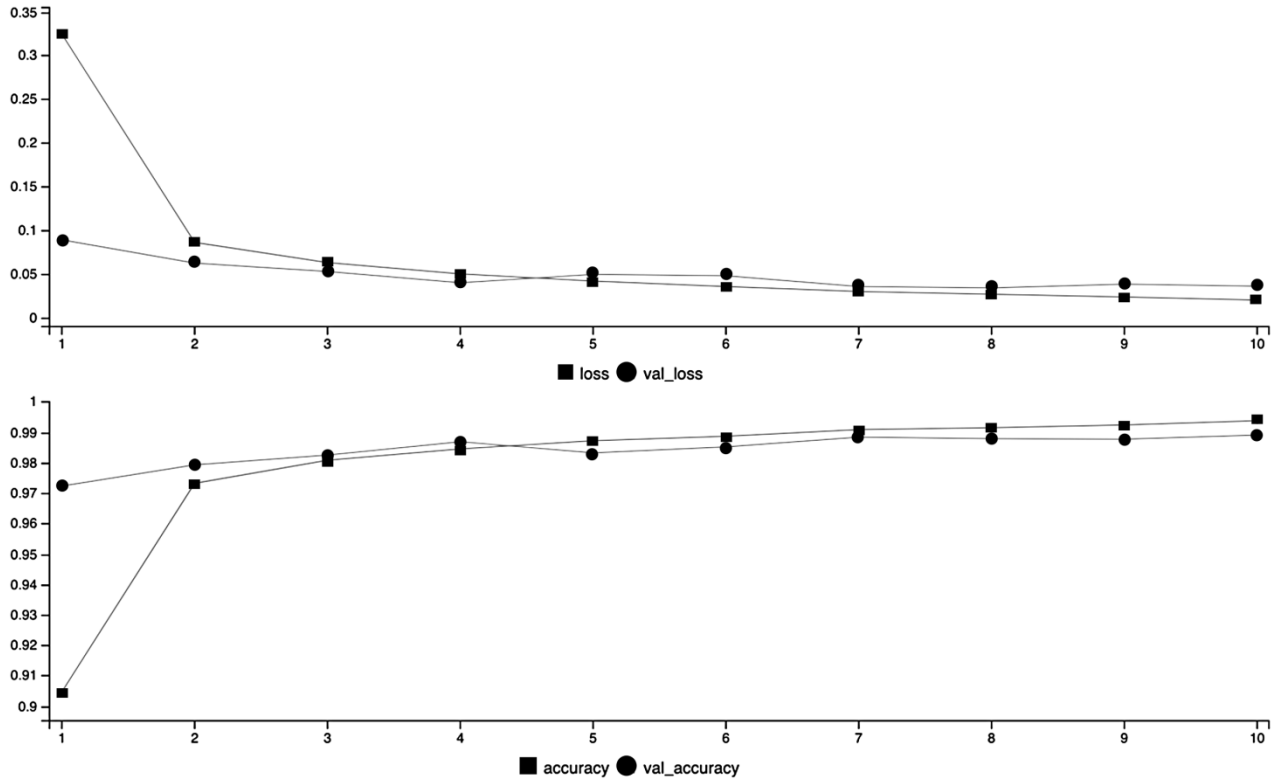

Рис. 2. Результаты предсказания для набора данных EMNIST TCyrillic

Таким образом, результаты показали схожий уровень предсказания с погрешностью не более 2 % для различных моделей машинного обучения, что свидетельствует об успешном построении тестового набора данных.

Библиографический список

1. Golubnichiy A. A., Yablontseva A. D. Outline of a Database of Symbols of Extended Cyrillic Alphabets of Modern Turkic Writing ExTL. Autom. Doc. Math. Linguist. 56, 320–323 (2022). https://doi.org/10.3103/S0005105522060036

© Яблонцева А. Д., 2023