ТЕХНИЧЕСКИЕ НАУКИ — МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ МАШИН, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ (05.13.11)

05.13.11

А.А. Голубничий, А.Д. Яблонцева, В.А. Мясоедова

Хакасский государственный университет имени Н.Ф. Катанова,
инженерно-технологический институт,
кафедра программного обеспечения вычислительной техники и автоматизированных систем,
Абакан, artem@golubnichij.ru

РАЗРАБОТКА СИСТЕМЫ ОПТИЧЕСКОГО РАСПОЗНАВАНИЯ СИМВОЛОВ ТУВИНСКОГО АЛФАВИТА

В работе описываются особенности применения нейронных сетей, для построения системы распознавания символов. Приводится обоснование выбора структуры нейронной сети типа LSTM для решения задачи построения модели распознавания символов тувинского алфавита. Производится описание исходного набора данных, для обучения нейронной сети, а также приводится алгоритм генерации дополнительных данных, посредством программной генерации изображений. Дается итоговая оценка эффективности системы распознавания

Ключевые слова: *системы распознавания символов, малые языки, искусственные нейронные сети, прикладное программное обеспечение, LSTM.*

Введение

Национальные языки, имеющие малое количество носителей, являются основой культурного наследия и достояния многонационального народа Российской Федерации. Тувинский язык относится к саянской группе тюркских языков, общее количество носителей которого составляет порядка 250 тысяч человек. В настоящий момент основой письменности тувинского языка является модифицированный кириллический алфавит, содержащий дополнительные, в сравнении с русским языком, буквы. Большая часть литературы, изданной на тувинском языке, имеется исключительно в печатном варианте и сложна для оцифровки и размещения в открытом доступе в связи с особенностью шрифта. Таким образом, создание системы оптического распознавания тувинского алфавита принимает особую актуальность.

Нейронные сети в системе построения OCR

Искусственные нейронные сети, как известно, являются одной из наиболее распространенных интеллектуальных систем, архитектура которых моделируется, имитируя человеческий мозг [1]. В зависимости от построения, выделяют большое количество разновидностей нейронных сетей, наиболее популярными, с точки зрения решения прикладных задач, являются нейронные сети прямого распространения, рекуррентные нейронные сети, радиально-базисные функции, нейронные сети Кохонена.

Одной из популярных моделей нейронных сетей, используемых в решении задачи NLP (обработке естественного языка), является рекуррентная нейронная сеть. Рекуррентные сети (RNN) – это нейронные сети с обратными связями, данные структуры являются динамическими системами с представлениями временного состояния.

RNN продемонстрировали большой успех во многих задачах обработки естественного языка. Наиболее часто используемым типом таких сетей являются сети долгой краткосрочной памяти (LSTM), которые намного лучше захватывают долгосрочные зависимости, чем RNN. Рекуррентные нейронные сети с долгой краткосрочной памятью используют механизмы стробирования для смягчения взрывающихся и исчезающих градиентов при изучении долгосрочных зависимостей [2].

Рекуррентная нейронная сеть использует полученную ранее информацию для решения последующих задач. Иногда для решения задачи требуется «просмотреть» только последнюю информацию. В тех случаях, когда разрыв между предыдущей информацией и местом, в котором она нужна, невелик, RNN легко справится с задачей. Но по мере увеличения этого разрыва RNN теряют связь между информацией. LSTM специально разработаны для устранения проблемы долгосрочной зависимости. Их специализация – запоминание информации в течение длительных периодов времени. Принцип работы сетей типа LSTM представлен на рисунке 1.

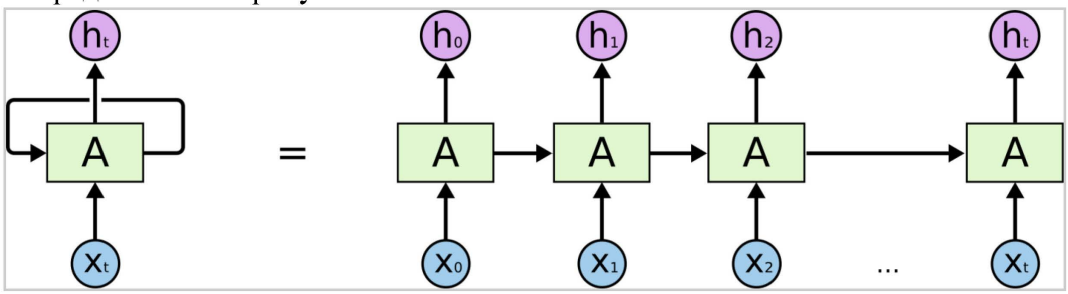

Рис. 1 – Принципиальная структура рекуррентных нейронных сетей с долгосрочной краткосрочной памятью [2]

Сети с долгосрочно-краткосрочной памятью уменьшают или увеличивают количество информации в состоянии ячейки, в зависимости от потребностей. Для этого используются тщательно настраиваемые структуры, называемые гейтами. Гейт – это «ворота», пропускающие или не пропускающие информацию. Гейты состоят из сигмовидного слоя нейронной сети и операции поточечного умножения [3].

Особенности структуры тувинского алфавита

Тувинский язык до 1930 года не имел своей письменности, для переписки до этого времени тувинцы пользовались литературным монгольским языком, письменность которого была основана на традиционном старомонгольском письме. В 1930 году для тувинского языка был введен алфавит, базирующийся на символах единого тюркского алфавита, выполненного на основе латинского, однако уже в 1941 году был разработан, а в 1943 году введен новый алфавит, используемый в качестве основы кириллицу, с дополнениями трех букв (рис. 2).

А а	Б б	В в	Г г	Д д	Е е	Ё ё	Ж ж
З з	И и	Й й	К к	Л л	М м	Н н	Ң ң
О о	Ө ө	П п	Р р	С с	Т т	У у	Ү ү
Ф ф	Х х	Ц ц	Ч ч	Ш ш	Щ щ	Ъ ъ	Ы ы
Ь ь	Э э	Ю ю	Я я				

Рис. 2 – Современный тувинский алфавит, основанный на расширенном кириллическом

Таким образом, в сравнении с русским алфавитом, тувинский алфавит имеет три дополнительные буквы: Ңң (латинское ng, в МФА – [ŋ]); Өө (латинское ö, в МФА – [ø]); Үү (латинское ü, в МФА – [y]).

Обработка исходных текстов

Для построения исходной базы, для обучения модели распознавания, использовались тексты, находящиеся в свободном доступе. В качестве основы для базы послужил журнал «Улуг-Хем», издаваемый на тувинском языке, на базе текстов данного журнала был собран исходный набор данных. Исходные данные для обучения формировались путем выбора страниц с разным качеством текста и наличия деформаций изображения. Для лучшего обучения и расширения базы также были сформированы дополнительные изображения с искусственно введенными искажениями. Для формирования новых изображений были взяты

фрагменты оригинального документа с применением подобранных алгоритмов преобразования, для генерации дополнительных наборов данных для обучения, реализуемых программными методами, имитирующими частые проблемы оцифровки документов. Пример исходного текста, а также одного из алгоритмов искажения представлены на рисунках 3 и 4, соответственно.

Оглунуң канчаар өзүп-доругуп келгенин болгаш авашкыларның канчаар чурттап чорааныңдан бээр ол билир болган. Тыныш, тын чок удуп чытса-даа, шупту чүвени көрүп-билип чыткан бооп-тур. Туруп кээр дээш оттуп чадап каанындан бээр чугаалап турган иргин.

Ам чүү боор, оглу биле ынак кызын эдертип алгаш, төрээн аалынга чанып кээп, орус чондан шиңгээдип алганы ажыктыг мергежилдерин ам кочулуг, хоруглуг эвес апарганынга өөрүп, улай сайзырадып, ада-иезинге, төрел чонунга бараалгап, оюн оя, чигин чире чурттай берип-тир оо!

Рис. 3 – Фрагмент исходного текста без искажений

Оглуиуп каичаар өзүп-доругуп келгенин болгаш авашкыларның канчаар чурттап чорааныңдан бээр ол билир болган. Тыныш, тын чок удуп чытса-даа, шупту чүвени көрүп-билип чыткан бооп-тур. Туруп кээр дээш оттуп чадап каанындан бээр чугаалап турган иргин.

Ам чүү боор, оглу биле ынак кызын эдертип алгаш, төрээн аалынга чанып кээп, орус чондан шиңгээдип алганы ажыктыг мергежилдерин ам кочулуг, хоруглуг эвес апарганынга өөрүп, улай сайзырадып, ада-иезинге, төрел чонунга бараалгап, оюн оя, чигин чире чурттай берип-тир оо!

Рис. 4 – Фрагмент исходного текста,
имитирующий обычное начертание с эффектом нарушения печати

Обучение нейронной сети типа LSTM для тувинского алфавита

Для распознавания в рамках исследования использована одномерная двунаправленная архитектура LSTM. Обнаружено, что одномерная архитектура превосходит своих двухмерных или более высокоразмерных собратьев. В работе использовалась модифицированная версия библиотеки LSTM. Эта библиотека предоставляет одномерные и многомерные сети LSTM, а также выравнивание по основанию с использованием прямого-обратного алгоритма. Библиотека также предоставляет механизм эвристического декодирования для отображения покадрового сетевого вывода на последовательность символов.

На этапе обучения случайно выбранные входные изображения текстовых строк представляются в виде одномерных последовательностей для шага прямого распространения через ячейки LSTM, а затем выполняется прямое-обратное выравнивание выходных данных, далее выполняется обратное распространение для обновления весов, и процесс затем повторяется для следующего случайно выбранного изображения текстовой строки.

Результаты тестирования итоговой нейронной сети на данных, используемых в качестве итогового датасета, показали степень эффективности распознавания в среднем на уровне 96%, что свидетельствует о высоком качестве распознавания.

Список литературы

1. Techradar. Best OCR software of 2021: scan and archive your documents to PDF. Available online: https://www.techradar.com/best/best-ocr-software (дата обращения: 20.11.2021).
2. *Landi, Federico & Baraldi, Lorenzo & Cornia, Marcella & Cucchiara, Rita.* (2021). Working Memory Connections for LSTM. Neural Networks. 144. 10.1016/j.neunet.2021.08.030.
3. *Hochreiter S. and Schmidhuber J.* "Long Short-TermMemory," Nueral Computation, vol. 9, no. 8, pp.1735–1780, 1997.