СИСТЕМА ДЛЯ ПРОГНОЗИРОВАНИЯ ПОПУЛЯРНОСТИ ПУБЛИКАЦИЙ

Калоша А. Л., Хоронеко М. П., Медунецкий М. М.
Кафедра информатики, Факультет компьютерных систем и сетей, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: andreikalosha@mail.ru

Цель данной работы заключается в создании системы для прогнозирования популярности публикаций. В данной системе используется нейронная сеть, которая обучена на наборе метрик, описывающих качество и популярность публикаций. В качестве набора метрик используется количество лайков, просмотров и репостов.

Введение

Объем информации, доступной в сети Интернет, растет с каждым годом. Причем большая часть этой информации представляет собой тексты на естественном языке. В зависимости от области знаний, информация может быть представлена в виде статей, комментариев или сообщений на публичном форуме. Информация в сети Интернет дублируется, уточняется и пополняется ежедневно. Нетрудно понять, что имеющиеся в данный момент доступные ресурсы всемирной сети представляют собой колоссальную базу знаний, представленных в форме, сложно поддающейся компьютерной обработке – в виде текста [1].

Как правило, изучить весь контент (текст) не представляется возможным даже в отдельных областях, поэтому приходится фильтровать получаемую информацию и выбирать лучшую.

Назначение разрабатываемой системы заключается в предсказании популярности статей через определенный промежуток времени. Статья считается популярной при высоком количестве лайков, репостов или просмотров. Данные метрики зависят от множества факторов, таких как название, авторов, время публикации и содержание статьи. Эти параметры наилучшим образом отражают популярность (качество) статьи. Правильно обученная нейронная сеть позволяет с высокой точностью предсказать значения метрик популярности для неопубликованного контента.

Для обучения нейронной сети была выбрана библиотека TensorFlow как один из лучших инструментов машинного обучения. TensorFlow – это библиотека программного обеспечения с открытым исходным кодом для численного расчета с использованием графиков потока данных [2].

Существует прямая зависимость между скоростью обучения нейронной сети и точностью предсказания. Для ускорения процесса обучения используется вычислительная мощность видеокарты, а именно технология CUDA. CUDA – это архитектура параллельных вычислений от NVIDIA, позволяющая существенно увеличить вычислительную производительность благодаря использованию GPU (графических процессоров) [3].

Нейронная сеть – это громадный распределенный параллельный процессор, состоящий из элементарных единиц обработки информации, накапливающих экспериментальные знания и предоставляющих их для последующей обработки [4].

Нейронная сеть сходна с мозгом с двух точек зрения:

1. Знания поступают в нейронную сеть из окружающей среды и используются в процессе обучения;
2. Для накопления знаний применяются связи между нейронами, называемые синоптическими весами [4].

I. Архитектура нейронной сети

Нейронная сеть состоит из 4 слоев (входной, два промежуточных и выходной слой) (см. рис. 1). На промежуточных слоях используется функция активации LeakyReLU, на выходном слое применяется функция softmax. Между всеми слоями, кроме последнего, используется нормализация данных.

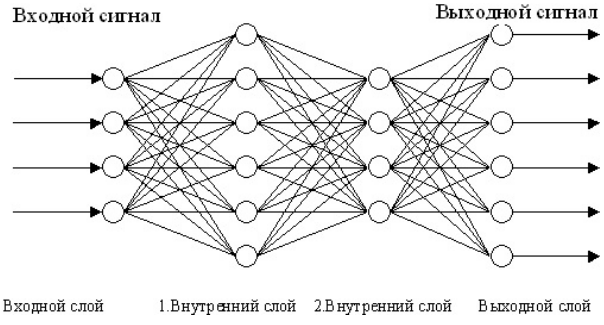

Рис. 1 – Архитектура нейронной сети

II. Обучение нейронной сети

Для обучения нейронной сети необходимо большое количество статей и метаданных, таких как автор, дата создания, ключевые слова и другие.

Перед обучением данные делятся на 2 части: для тестирования и для обучения.

Опишем процедуру обучения нейронной сети. На вход нейронной сети подается матрица векторов MV, каждый вектор V которой содержит информацию о конкретном атрибуте публикации (например, авторе). Для формирования отдельного вектора V перед обучением необходимо получить словарь D всех значений атрибута публикации. Словарь D сортируется по убыванию и отбрасываются последние N значений, чтобы нейронная сеть не обучалась на редко встречающихся элементах, и тем самым не ухудшалась точность классификации. Указанная выше процедура выполняется для каждого атрибута. Для каждого автора публикации, производится поиск в словаре D, если данный автор найден, то под индексом найденного автора в вектор V ставится единица, иначе – ноль. Таким образом, заполняются все векторы матрицы MV [5].

Выходной вектор R описывает количество просмотров через заданный промежуток времени и состоит из единственного дробного числа, находящегося в диапазоне от ноля до единицы. Единица означает максимальное количество просмотров, в данном исследовании выбрано 50 миллионов [5].

Промежуток времени, на который нейронная сеть способна предсказать популярность публикации, является статическим и определяется до обучения нейронной сети. Т.е что бы изменить этот параметр нужно обучить нейронную сеть заново. Для предсказания популярности публикации через несколько временных отрезков, например, неделя, месяц и год можно использовать два варианта:

1. Обучить несколько нейронных сетей;
2. Изменить архитектуру нейронной сети таким образом, что бы на выходном слое был вектор, содержащий значения популярности для нескольких временных интервалов.

У каждого из способов есть достоинства и недостатки и выбирать нужно, исходя из постановки задачи. Плюсом при использовании первого варианта, является проста реализации и тестирования приложения. Минусом является необходимость поддержания нескольких копий приложения, по одному на каждый из временной интервал.

Плюсом при использовании второго варианта является необходимость поддержания только одного экземпляра приложения вместо нескольких, как в первом варианте. Минусом является сложность создания архитектуры, создания приложения и оценки результата, т.к. нейронная сеть может обучится предсказывать некоторые временные участки лучше других, хотя в среднем результат будет оптимальным.

После обучения нейронной сети загружаются тестовые данные, и выполняется процедура тестирования. Далее на основании полученных векторов нейронная сеть предсказывает популярность статей через заданный промежуток времени. Данные обрабатываются и сохраняются в excel для анализа.

Заключение

Обучение сети производилось на более чем 100 000 текстов, что занимает от 4 до 16 часов, в зависимости от глубины обучения и точности результата. В результате нейронная сеть способна предсказать количество просмотров с точностью в 75%. Верным считается ответ, находящийся в диапазоне +/-200 000 просмотров от ответа. Максимальное количество просмотров при обучении составляло 48 миллионов. Коэффициент корреляции для массивов ответов и предсказанных значений составляет 0,3. Это означает, что между входными и выходными данными есть зависимость. Подобрав более точно входные данные или параметры нейронной сети, можно увеличить точность системы [5].

Список литературы

1. Степанов, П. А. Системы анализа текстов естественного языка / П. А. Степанов. – Тамбов: Издательство: Грамота, 2013. – С. 159–161.
2. Library for numerical computation using data flow graphs [Электронный ресурс] / Официальный сайт фрайемворка «TensorFlow». – Режим доступа: https://www.tensorflow.org. – Дата доступа: 11.09.2019.
3. Параллельные вычисления CUDA [Электронный ресурс] / Официальная страница архитектуры «CUDA». – Режим доступа: http://www.nvidia.ru/object/cuda-parallel-computing-ru.html. – Дата доступа: 01.08.2019.
4. Хайкин, С. Нейронные сети: полный курс, 2-е издание / С. Хайкин. – М.: Издательский дом «Вильямс», 2006. – 1104 с.
5. Калоша, А. Л. Система предиктивного анализа для классификации документов текстовых коллекций / А. Л. Калоша, М. А. Медунецкий, М. П. Хоронеко // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 / editorial board: M. Batura [etc.]. – Minsk, BSUIR, 2018. – P. 467–468.