Модель сети с долгой краткосрочной памятью позволяет сразу применять LSTM-блок к векторным представлениям слов после соответствующего слоя. Выход LSTM-слоя подается полносвязному слою, который определяет конечный результат работы сети.

После многократного обучения моделей были получены следующие результаты. Сверточная нейронная сеть достигла точности 92%, а модель с долгой краткосрочной памятью – 88%.

Список использованных источников:

1. Kim, Y. Convolutional Neural Networks for Sentence Classification / Y. Kim // New York University – 2014. – 6 P.
2. Greff, K. LSTM: A Search Space Odyssey / K. Greff, R.Srivastava, J Koutník, B. Steunebrink, J. Schmidhuber// IEEE Transactions on Neural Networks and Learning Systems. A. – 2017. – Vol. 28, № 10. – P. 2222–2321.
3. Заенцев, И. В. Нейронные сети: основные модели/ И. В. Заенцев. – Воронеж, 1999. – 76 с.
4. Mikolov, T. Distributed Representations of Words and Phrases and their Compositionality/ T. Mikolov, I. Sutskever, K.i Chen, G. Corrado, J. Dean// NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems – 2013. – Vol. 2. – P. 3111–3119.

ХРАНИЛИЩЕ ДАННЫХ СИСТЕМЫ КОМПЛЕКСНОГО АНАЛИЗА ДАННЫХ ИНТЕРНЕТ ИСТОЧНИКОВ

Гутковский В.Н.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Пилецкий И.И. – к.ф.-м.н.,доцент

В докладе приводится описание инструмента мониторинга открытых интернет-источников с целью выявления экспертов в некоторой научной области, определения тематик публикаций, оценки популярности публикаций. Описываются принятые решения при построении компонента хранилища аналитического комплекса и полученные результаты его работы.

В настоящее время информация, полученная в результате анализа данных интернет источников, является одной из базовых для принятия решений. Как правило, это неструктурированные текстовые данные, различные мультимедийные данные. Данные могут быть получены как из социальных сетей, так и тематических сайтов (газет, журналов, библиотек, компаний и т. д.), содержащих различные публикации. Есть много работ, которые посвящены принятию решения на основании применения некоторого метода анализа данных. Результатами анализа смогут воспользоваться компании для создания систем поддержки пользователей, социологи для анализа общественного мнения, организаторы мероприятий для получения отклика участников, знаменитости для отслеживания репутации в сети, правительство для контроля настроений в обществе и др.

В данной статье рассматривается проект создания «Системы комплексного анализа данных интернет-источников (СКА)», позволяющей анализировать большие объемы данных из интернет-источников в области научных исследований и предназначенной для сбора информации о научных публикациях, построения графа знаний, что дает возможность определять экспертов предметной области, тематики их работ, их взаимосвязи, а также определять передовые научные направления.

Система должна находить экспертов (авторитетов) в предметной области и выдавать оценку их рейтинга влиятельности. Например, лучше прочитать три книги признанных экспертов в определённой области, чем десять книг дилетантов.

СКА состоит из следующих компонент: сбора данных, фильтрации данных и составление «мешка слов» из N-грамм (векторизации), хранилища данных, библиотеки аналитических модулей, подготовки выдачи результата, клиентского модуля.

В докладе рассмотрен в подробностях компонент хранилища СКА. Компонент хранилища данных – содержит данные из интернет-источников, предварительно обработанные и размеченные данные, необходимые для построения классификатора, «мешок слов», а также служебную информацию, необходимую для работы других модулей системы. В хранилище хранятся сырые данные с сайта, текст, фильтрованный текст, исходные документы, «мешок слов», тематика документов и служебная информация; структура одной из записей для документа приведена ниже:

Структура записи (Hash - primary key of publication, Title - title of publication, Author - author(authors) of publication, Year - publication date, Pages - number of pages, Publisher - publisher of publication, Language - primary language of publication, Topic - topic or topics of publication, Extension - extension of publication file, Tags - array of publication tags, Locator - name of the file).

В настоящее время в хранилище содержится информация более чем о 20 тыс. статей и документов. В дальнейшем планируется получать данные из многих сайтов и при необходимости указывать ссылки к конкретным данным. Документы, статьи читаются с сайта (сайтов), фильтруются,

строится векторное представление («мешок слов») и все данные (сырые данные с сайта, текст, фильтрованный текст, исходные документы и «мешок слов») сохраняются в хранилище.

Выбранная архитектура СКА позволяет её модернизировать и функционально наращивать в процессе эксплуатации. Так, планируется нарастить компонент «хранилище» и дополнить компонент «библиотека модулей» векторными алгоритмами и алгоритмами нейровычислений, что существенно расширит область применения СКА и позволит получать информацию в интересах сотрудников университета, государственных и частных организаций.

Список использованных источников:

1. Data Never Sleeps 6.0 [Электронный ресурс] / Режим доступа: https://www.domo.com/learn/data-neversleeps-6 Дата доступа: 18.01.2019.

2. Пилецкий, И. И. Аналитический комплекс анализа данных из открытых интернет источников / И. И. Пилецкий, В. А. Прытков, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018 – Minsk, BSUIR, 2018. – P. 193 – 199.

3. Прытков, В. А. Анализ репозитория университета с использованием графовой базы данных / В. А. Прытков, И. И. Пилецкий, Н. А. Волорова // BIG DATA Advanced Analytics: collection of materials of the fourth international scientific and practical conference, Minsk, Belarus, May 3 – 4, 2018. – Minsk, BSUIR, 2018. – P. 177 – 183.

СРАВНЕНИЕ ВАРИАЦИЙ ГРАДИЕНТНОГО СПУСКА НА ПРИМЕРЕ ЗАДАЧИ РАСПОЗНАВАНИЯ СИМВОЛОВ

Евжик Д.А., Подвальников Д.С., Тишковский М.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Анисимов В.Я. – к.ф.-м.н.,доцент

В данной работе приведен сравнительный анализ вариаций градиентного спуска, который демонстрирует применимость отдельного метода для задачи распознавания символов. Результаты исследования в данной работе позволяют сделать выбор наиболее подходящего по скорости и качеству обучения метода при решении задач.

В настоящее время можно заметить глубокий интерес к области машинного обучения. Как известно, для построения методов машинного обучения используются средства методов оптимизации для нахождения минимума целевой функции в некоторой области конечномерного векторного пространства.

В ходе исследования были проанализированы три популярных метода нахождения локального минимума функции. Обычный, пакетный и стохастический градиентный спуск.

В данной работе был рассмотрен набор данных *notMNIST* [1], который состоит из изображений размерностью 28×28 первых 10 букв латинского алфавита.

Градиентный спуск – это метод нахождения локального минимума функции путем движения по вектору антиградиента. Этот метод является очень распространенным в машинном обучении для нахождения минимума функции потерь. В начале работы алгоритма задается начальное приближение минимума, затем вычисляется значение градиента функции потерь на всей обучающей выборке, последним шагом является сложение начального приближения с полученным значением градиента со знаком минус. Стохастический градиентный спуск является вариацией обычного градиентного спуска. Его отличие заключается в том, что вычисление значения градиента производится не на всей обучающей выборке, а на одном случайном элементе, засчет этого скорость обучения значительно возрастает. Пакетный градиентный спуск схож со стохастическим, но вместо одного экземпляра из выборки берется какая-то ее часть, несколько экземпляров, что позволяет добиться как хорошей скорости обучения, так и более плавного движения к минимуму функции.

Описанные три вариации градиентного спуска были применены для обучения нейронной сети на наборе данных *notMNIST* размером 10000 элементов. Архитектура нейронной сети состояла из пяти скрытых слоев с 1024, 1024, 512, 256 и 128 нейронами на каждом слое соответственно. На всех скрытых слоях в качестве функции активации была использована *relu* [2] (кусочно-линейная функция), на выходном слое – *softmax* [3] (обобщение логистической функции для многомерного случая). В результате при обучении на 10 эпохах получились следующие графики, показывающие зависимость ошибки обучения от эпохи.

Результаты данного эксперимента в виде зависимости ошибки от порядкового номера эпохи представлены на рисунке 1, рисунке 2 и рисунке 3.