Список использованных источников:

1. Collberg, C. A Taxonomy of Obfuscating Transformations / C. Collberg, C. Thomborson, D. Low – Auckland: Department of Computer Science, 1997. – 36 p.

2. Сергейчик В. В. Методы лексической обфускации VHDL-описаний / В. В. Сергейчик, А. А. Иванюк // Информационные технологии и системы 2013 (ИТС 2013) : материалы международной научной конференции, БГУИР, Минск, Беларусь, 23 октября 2013 г. = Information Technologies and Systems 2013 (ITS 2013) : Proceeding of The International Conference, BSUIR, Minsk, 24th October 2013 / редкол.: Л. Ю. Шилин [и др.]. - Минск : БГУИР, 2013. – С. 198-199.

3. Garg S., Gentry C., Halevi S., Raykova M., Sahai A., and Waters B. «Candidate indistinguishability obfuscation and functional encryption for all circuits.» FOCS 2013.

СРАВНЕНИЕ СВЕРТОЧНЫХ И РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТА

Витковский А.В.

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

Жвакина А.В. – к.т.н.,доцент

В обработке естественных языков существует задача определения тональности текста. В настоящее время для решения этой задачи используются искусственные нейронные сети. Для определения тональности могут быть использованы сети с различной архитектурой. В данном исследовании сравниваются сверточная нейронная сеть и сеть с долгой краткосрочной памятью.

Обработка естественного языка – направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Анализ тональности текста является одной из задач, решаемых в рамках обработки языка. Суть задачи заключается в автоматическом распознавании в тексте его эмоциональной окраски. Классификация бывает бинарной, при которой определяется позитивную и негативную окраску имеет текст. Также существует классификация на несколько классов, при которой окраска может быть нейтральной или разной степени позитивности или негативности.

Две различные архитектуры нейронной сети будут сравниваться при решении задачи бинарной классификации. В качестве выборки используется выборка отзывов с сайта IMDB. Входными данными для обоих сетей будут являться тексты, где каждой слово закодировано целым числом, при этом значение зависит от частоты появления слова в выборке данных.

Сверточная нейронная сеть (Convolutional Neural Network, CNN) является специальной архитектурой, в основном применяемой для распознавания образов на изображениях. Название архитектура сети получила из-за наличия операции свёртки, суть которой в том, что каждый фрагмент изображения умножается на матрицу (ядро) свёртки поэлементно, а результат суммируется и записывается в аналогичную позицию выходного изображения. Несмотря на основное применение, ее также можно использовать для работы с текстом. Если в случае изображений, фильтр сверточного слоя применяется для нескольких соседних пикселей, то в случае текста фильтр можно применять для нескольких соседних слов.

Долгая краткосрочная память (Long short-term memory, LSTM) – один из видов рекуррентных нейронных сетей. В рекуррентных нейронных сетях связи между нейронами имеют направленную последовательность. Такие сети могут использовать свою внутреннюю память для обработки последовательностей произвольной длины. LSTM-сеть содержит LSTM-модули. LSTM-модуль – это рекуррентный модуль сети, способный запоминать значения как на короткие, так и на длинные промежутки времени. Такое поведение обусловлено тем, что LSTM-модуль не использует функцию активации внутри своих компонентов.

Для обоих нейронных сетей первый слой одинаков и является слоем, осуществляющим замену целых чисел, обозначающих слова, на векторное представление слов. Этот слой обучается в ходе тренировки модели в обоих случаях. В обоих случаях функцией потерь является перекрестная энтропия, обучения происходит при помощи метода обратного распространения ошибки с алгоритмом Adam. Объем тестовых и валидационных данных также одинаков для обоих моделей. Различными являются только сами модели нейронных сетей.

Модель сверточной нейронной сети состоит из трех каналов, каждый канал состоит из сверточного слоя, слоя активации, слоя подвыборки и полносвязного слоя. каналы в модели параллельны. Данная модель позволяет задать различные размеры фильтра сверточного слоя в каждом канале. В реализованной моедели использовались фильтры по 3, 4 и 5 слов. Каналы объединяются при помощи полносвязного слоя.

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,