Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.912

Давыдовский
Сергей Васильевич

Использование алгоритмов глубокого обучения для анализа тональности текста

АВТОРЕФЕРАТ
на соискание степени магистра информатики и вычислительной техники

по специальности 1-40 81 04 Обработка больших объемов информации

_____

Научный руководитель

доцент, канд. физ.-мат. наук
Сиротко Сергей Иванович

_____

Минск 2018

ВВЕДЕНИЕ

Современный мир насыщен огромным объемом текстовых данных в электронном виде, в них – человеческие знания, эмоции и опыт. А еще – спам, который необходимо отличать от полезной информации. Люди хотят общаться с теми, кто не знает их родной язык. А еще – управлять своим телефоном/телевизором/умным домом с помощью голоса. Всё это обеспечивает востребованность и бурное развитие методов обработки естественного языка.

Обработка естественного языка (англ. *Natural Language Processing, NLP*) – общее направление искусственного интеллекта и математической лингвистики, которое изучает задачи компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез – генерацию грамотного текста. Решение этих задач будет означать создание более удобной формы взаимодействия компьютера и человека.

Данная работа исследует и разрабатывает систему по анализу тональности текстов на английском языке. Анализ тональности текста – это область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах.

В работе исследуется подход, основанный на машинном обучении с учителем. Его суть состоит в том, чтобы обучить машинный классификатор на заранее размеченных текстах, а затем использовать полученную модель для анализа новых документов. В работе используется набор технологий и алгоритмов машинного обучения, относящихся к так называемому глубокому обучению (англ. *Deep Learning*).

В настоящее время область глубокого обучения – крайне популярное направление, алгоритмы которого превзошли многие свои аналоги. Как часть концепции искусственного интеллекта, глубокое обучение лежит в основе различных инноваций: беспилотные автомобили, распознавание голоса, изображения и т.д.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В данной работе исследуется применимость методов глубокого обучения к задаче анализа тональности текста; исследуется и разрабатывается система по анализу тональности текстов на английском языке. Для этого используются различные техники и приемы из глубокого обучения, в частности рекуррентные нейронные сети с LSTM-ячейками, дропаут и др. В работе также сравниваются различные методы получения векторных представлений слов, описываются их достоинства и недостатки, а также обосновывается выбор метода, который используется в работе.

СОДЕРЖАНИЕ РАБОТЫ

Методы глубоко обучения работают только с векторами чисел, поэтому в начале необходимо найти способ представления слов в виде векторов. Все нижеописанные семейства методов основаны на дистрибутивной гипотезе, которая гласит, что лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.

Методы, основанные на сингулярном разложении, для нахождения векторных представлений слов, итерируют по большому набору текстов и считают, сколько раз каждое слово находилось в контексте других слов. Строится матрица смежности слов X, над которой производится сингулярное разложение (англ. *Singular Value Decomposition, SVD*). Из полученного разложения USV^T векторные представления слов берутся как строки матрицы U. В целом, эти методы плохо масштабируются под новые слова и документы. Вычислительная сложность обучения для матрицы $m \times n$ равна $O(mn^2)$. Однако, несомненным преимуществом данного семейства методов является эффективное использование статистики.

Итерационный класс методов использует модель, которая обучается на каждой итерации, и которая учится вычислять вероятность слова по его контексту. Параметрами этой модели являются векторные представления слов. Идея состоит в том, чтобы тренировать модель на определенной целевой функции, измерять ошибки и распространять их обратно, тем самым, изменяя параметры модели. Существуют несколько методов по вычислению векторных представлений слов, основанных на данной идее. В работе рассматривается один из них, получивший широкую известность, – *word2vec*. Метод работает следующим образом: он принимает большой текстовый корпус в качестве входных данных и затем сопоставляет каждому слову вектор, выдавая координаты слов на выходе. Векторное представление основывается на контекстной близости: слова, встречающиеся в тексте рядом с одинаковыми словами (а, следовательно, имеющие схожий смысл), в векторном представлении будут иметь близкие координаты векторов.

В word2vec существуют два основных алгоритма обучения:

- *CBOW* (Continuous Bag of Words) предсказывает текущее слово, исходя из окружающего его контекста;
- *Skip-gram* действует наоборот – он использует текущее слово, чтобы предугадывать окружающие его слова.

Сравнивая эти два алгоритма, можно сказать, что CBOW работает быстрее, зато Skip-Gram работает лучше, особенно для относительно редких слов. CBOW лучше подходит для больших корпусов текстов (больше ста миллионов слов), так как быстрее и чуть лучше работает с более частотными словами, а Skip-Gram

лучше учитывает редкие слова и работает медленнее. Оптимальный размер окна для Skip-Gram – около 10, для CBOW – в районе 5.

Существуют также два метода обучения:

- негативное сэмплирование (англ. *negative sampling*) считает не все возможные контексты для каждого а слова, а случайным образом выбирает лишь несколько;
- иерархический софтмакс (англ. *hierarchical softmax*) использует древовидную структуру, которая эффективно вычисляет вероятности для всего словаря.

На практике, иерархический софтмакс как правило лучше работает со словами, которые встречаются нечасто, в то время как негативное сэмплирование лучше работает для часто встречающихся слов.

GloVe (Global Vectors for Word Representations) – это технология, которая объединяет в себе преимущества двух упомянутых семейств методов. Она эффективно использует статистическую информацию, а также использует различные приемы из итерационных методов для повышения качества получаемых векторных представлений слов. Как и в методах, основанных на сингулярном разложении, GloVe использует матрицу смежности слов. Как и в итерационных методах GloVe предсказывает появление слова j при контексте i, однако, в отличие от итерационных методов, где для каждого слова осуществлялся проход по всему словарю, GloVe, используя матрицу смежности слов, может вычислить вероятность за одну итерацию. В целом, GloVe превосходит word2vec на задачах словарной аналогии. GloVe достигает лучших результатов быстрее и получает лучшие векторные представления. По этим причинам в данной работе именно GloVe используется как источник векторных представлений слов.

Глубокое обучение – это общее направление машинного обучения, характеризующее качественный прогресс, возникший после 2006 года в связи с нарастанием вычислительных мощностей и накоплением опыта. Хотя термин “глубокое обучение” можно понимать и в более широком смысле, в большинстве случаев он применяется в области (искусственных) нейронных сетей.

По определению любая нейронная сеть с более, чем одним скрытым слоем, считается глубокой. Перемещение в глубину позволяет подавать на вход нейронной сети необработанные входные данные: в прошлом однослойным сетям на вход подавались ключевые признаки, которые выделяли из входных данных с помощью специальных функций. Это значило, что для различных классов задач, например, компьютерного зрения, распознавания речи или обработки естественных языков, требовались разные подходы, что препятствовало научному сотрудничеству между этими областями. Но когда сеть содержит несколько скрытых слоев, она приобретает способность сама обучаться выделять ключевые признаки, которые наилучшим образом

описывают входные данные, таким образом находя применение *end-to-end learning* (т.е. без традиционных программируемых обработок между входом и выходом), а также позволяя использовать одну и ту же сеть для широкого спектра задач, так как больше нет необходимости выводить функции для получения ключевых признаков.

Существует такой тип практических задач, при решении которых необходимо работать не с отдельными объектами, но с их последовательностями, т.е. порядок следования объектов играет существенную роль в задаче. Задачи обработки естественного языка, где мы имеем дело с последовательностями слов, относятся именно к этому типу. Нейронные сети прямого распространения не имеют доступа к предыдущей информации – в каждый момент времени сеть обучается лишь на текущем примере. Рекуррентные нейронные сети (англ. *recurrent neural network,* RNN) решают эту проблему. Они содержат в себе обратные связи, позволяющие сохранять информацию. Однако, классические RNN имеют трудности с долговременными зависимостями, поэтому на практике применяют их модифицированные версии. В работе используется LSTM-сеть – особый тип RNN, способный обучаться долговременным зависимостям.

LSTM-сеть представляет собой цепочку повторяющихся ячеек, где каждая ячейка имеет структуру, показанную на рисунке 1.1.

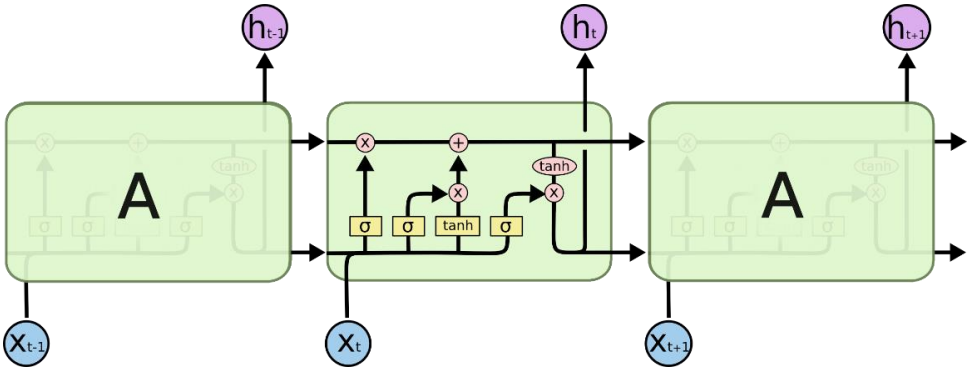

Рисунок 1.1 – Схема LSTM-ячейки

Переобучение (англ. overfitting) – одна из проблем глубоких нейронных сетей (англ. *Deep Neural Networks*, DNN), состоящая в следующем: модель хорошо объясняет только примеры из обучающей выборки, адаптируясь к обучающим примерам, вместо того чтобы учиться классифицировать примеры, не участвовавшие в обучении (теряя способность к обобщению). За последние годы было предложено множество решений проблемы переобучения, но одно из них превзошло все остальные, благодаря своей простоте и прекрасным практическим результатам; это решение – Dropout (в русскоязычных источниках – “метод прореживания”, “метод исключения” или просто “дропаут”). Главная идея Dropout – вместо обучения одной DNN обучить ансамбль нескольких DNN,

а затем усреднить полученные результаты. Сети для обучения получаются с помощью исключения из сети (англ. *dropping out*) нейронов с вероятностью p, таким образом, вероятность того, что нейрон останется в сети, составляет $q = 1 - p$. “Исключение” нейрона означает, что при любых входных данных или параметрах он возвращает 0.

Для программной реализации модели по анализу тональности текста в работе используется библиотека TensorFlow.

Задача анализа тональности текста заключается в определении эмоциональной окраски (позитивная, негативная или нейтральная) последовательности слов.

Для обучения сети в работе используется датасет, содержащий рецензии фильмов, написанные пользователями сайта Imdb. Он состоит из тренировочного и тестового наборов, каждый из которых содержит 25 тыс. обзоров, половина из которых положительные, а другая половина – отрицательные. Тональность обзора определялась оценкой, которую поставил пользователь – обзоры с оценкой ≤ 4 считались отрицательными, ≥ 7 – положительными. Нейтральные или противоречивые обзоры не были включены в датасет.

Поскольку данная работа имеет дело с последовательностями слов, естественной архитектурой будет рекуррентная нейронная сеть. Для решения задачи используется сеть, состоящая из двух слоев LSTM-ячеек и выходным softmax-слоем. Финальный слой использует softmax-функцию, поэтому естественным выбором для функции потерь является кросс-энтропия. В качестве алгоритма минимизации используется оптимизатор Adam (Adaptive Moment Estimation) – алгоритм, схожий со стохастическим градиентным спуском, имеющий, однако “эффект импульса”, который помогает избегать “застревание” в локальных минимумах.

На тестовых данных обученная модель имеет точность $\approx 85\%$, т.е. из 25 тыс. обзоров сеть правильно классифицировала больше 21 тысячи. Модель научилась анализировать тональность текстов, видя исключительно примеры, которые ей предоставили. Она не была специально запрограммирована анализировать различные грамматические и синтаксические структуры языка. Тем не менее, обучаясь на примерах, веса и смещения сети настроились таким образом, который позволяет модели анализировать тональность текста и в большинстве случаев давать хорошие результаты.

ЗАКЛЮЧЕНИЕ

В данной работе была исследована применимость методов глубокого обучения к задаче анализа тональности текста. Было показано, что алгоритмы глубокого обучения, достигают впечатляющих результатов. Все это позволяет предположить их дальнейшее распространение в этой сфере.

Была разработана система, оценивающая тональность произвольного текста. Для разработки были применены различные приемы и алгоритмы глубокого обучения, в частности рекуррентная нейронная сеть с LSTM-ячейками, дропаут и оптимизатор Adam. В результате получилась модель, которая корректно классифицирует около 85% тестовых примеров. Ее можно использовать в сфере маркетинга, автоматически анализируя отзывы покупателей.

Были исследованы способы получения векторных представлений слов, которые способны инкапсулировать синтаксическое и семантическое значения. Применение этих векторных представлений не ограничивается задачей анализа тональности – они используются во многих задачах по обработке естественного языка.

Разработанная модель научилась анализировать тональность, обучаясь только на примерах. Это было достигнуто эффективным использованием статистики на большом количестве данных. В будущем, с увеличением доступных данных, модель можно было бы продолжать улучшать.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Модель Skip-Gram технологии word2vec / Давыдовский Сергей, 2017 – режим доступа: https://libeldoc.bsuir.by/handle/123456789/13101.