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

УДК 004.855.5

Михлюк
Константин Юрьевич

Классификация предложений на основе сверточных нейронных сетей

АВТОРЕФЕРАТ

на соискание степени магистра информатики и вычислительной техники
по специальности 1-40 81 04 — «Обработка больших объемов информации»

Научный руководитель:
кандидат технических наук, доцент
Егорова Наталья Геннадьевна

Минск — 2020

Работа выполнена на кафедре информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

Научный руководитель: **Егорова Наталья Геннадьевна,**
кандидат технических наук, доцент кафедры информатики учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Рецензент: **Гулякина Наталья Анатольевна**
кандидат физико-математических наук, доцент кафедры интеллектуальных информационных технологий учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Защита диссертации состоится « » июня 2020г. в : часов на заседании Государственной экзаменационной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. Гикало, 9, корп. 4, ауд. 111, тел. 293-85-91, e-mail: inform@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатике и радиоэлектроники».

ВВЕДЕНИЕ

В эпоху больших данных объем информации (изображения, видео, звук и текст) растет в геометрической прогрессии. Исследования, связанные с анализом текста, активно проводятся до настоящего времени. В частности, классификация текста привлекает большое внимание, поскольку текст может иметь такие категориальные метки как настроение, эмоция, пол автора. Анализ эмоциональных состояний обычно включает в себя такие категории как: счастье, радость, удовлетворение, злость. В то же вреям анализ тональности текста определяться по бинарной шкале, либо состоит из нескольких категорий, таких как: положительный, нейтральный и отрицательный. В данной работе упор делается на анализ тональности, который классифицирует текст в одну из категорий тональных оценок. На сайтах с фильмами люди могут публиковать свои комментарии, содержащие эмоции или мнения. Если такое мнение точно прогнозируется, то его можно применить к различным областям, таким как рекомендации фильмов или персонализированная лента новостей. Действительно, многие крупные компании (Netflix, HBO, Disney+) предоставляют пользователям индивидуальные рекомендации для просмотра фильмов по интересующей тематике.

Классификация текста широко применяется во многих областях. С помощью этой техники пользователи могут не тратить время на чтение несвязанных статей. Также анализ тональности может использоваться для прогнозирования рыночных тенденций посредством анализа отзывов клиентов о некоторых конкретных продуктах. Реальный пример такого приложения - проблема с батарями Samsung Note 7 в 2017 году. Количество отрицательных отзывов в Twitter резко возросло после того, как батареи Samsung Note 7 начали возгораться. Наблюдая за изменением количества положительных и отрицательных сообщений, компания Samsung смогла оценить последствия этой проблемы. Другое распространенное использование анализа тональности - сбор и анализ отзывов программных приложений. После каждого обновления приложения в операционной системе iOS или Android разработчики программного обеспечения хотели бы знать, что пользователи думают о недавно разработанной версии. Используя метод анализа тональности, разработчики могут получить обратную связь из комментариев пользователей. В статье [1] авторы представили метод классификации отзывов пользователей, которые были получены из ма-

газинов приложений Apple и Google Play. Отзывы были классифицированы по четырем различным категориям: функциональные запросы, текстовая оценка, впечатления пользователей и сообщения об ошибках. Классификация тональности может применяться и в сфере политики. Когда предлагается новый закон, правительство может собирать мнения людей и материалы для обсуждения из нескольких источников для получения отзывов о новом законе. В таком случае правительство может улучшить или скорректировать принимаемый закон на следующем этапе.

Также ранее было проведено много исследований, в которых использовались методы машинного обучения для классификации текста. Несмотря на то, что методы машинного обучения широко используются и показывают довольно успешную производительность, они сильно зависят от определяемых вручную функций, требующих больших усилий экспертов в предметной области. По этой причине методы глубокого обучения в последнее время привлекают большее внимание, поскольку они могут уменьшить усилия по определению ручных признаков и достичь более высокой производительности. Целью диссератции ставится классификация тональности текстовых данных. Для этого будет предложена архитектура сверточной нейронной сети (CNN), которая является типом модели глубокого обучения. Эффективность предложенной сети будет продемонстрирована путем экспериментального сравнения с другими моделями машинного обучения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследования

Целью диссертационной работы является исследование актуальности применения модели сверточных нейронных сетей для решения задачи классификации текста и разработка модели сети эксперементально подтверждающей результаты исследований.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать существующие решения задачи классификации текста. Найти модели, показывающие наилучший результат в настоящее время, для сравнения с разрабатываемой моделью сверточной нейронной сети.

2. Определить преимущества сверточных моделей в решении задачи классификации текста.
3. Разработать модель сверточной нейронной сети.
4. Сравнить эксперементально результаты производительности разработанной сети с существующими моделями.

Объектом исследования являются процесс классификации текста с помощью нейронных сетей.

Предметом исследования является применение сверточных нейронных для задачи классификации текста.

Основной гипотезой, положенной в основу диссертационной работы является то, что сверточные нейронные сети также как и реккурентные могут захватывать синтаксические и семантические особенности предложений [2] используя преимущества сверточных фильтров. Кроме того, по сравнению с рекуррентными, сверточные сети в основном имеют меньшее количество параметров, что позволяет им обучаться быстрее с гораздо меньшим количеством данных.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

Работа выполнялась в соответствии с научно-техническим заданием и планом работ кафедры «Программное обеспечение информационных технологий» по теме «Разработка моделей, методов, алгоритмов, повышающих показатели проектирования, внедрения и эксплуатации программных средств для перспективных платформ обработки информации, решения интеллектуальных задач, работы с большими массивами данных и внедрение в современные обучающие комплексы» (ГБ № 16-2004, № ГР 20163588, научный руководитель НИР – Н. В. Лапицкая).

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. Г. Егоровой, заключается в формулировке целей и задач исследования.

Структура и объем диссертации

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения и списка использованных источников. В первой главе представлен анализ существующих решений: определены последние исследования

в данной области, найдены современные модели для решения задачи классификации текста. Вторая глава содержит исчерпывающую информацию об алгоритмах классификации тональности текста: подробную концепцию и связанную информацию о нейронных сетях, концепции в технике обработки естественного языка, а также общие метрики оценки для алгоритмов классификации текста. Третья глава содержит описание архитектуры разработанной сверточной нейронной сети, результаты проведенных эксперементов с различными моделями, а также сравнение результата производительности существующих и разработанных в данной работе моделей.

ОСНОВОНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

В **первой главе** приводится анализ существующих решений задачи классификации тональности текста. Рассматриваются как исследования с использованием моделей машинного обучения так и глубокого обучения.

Несмотря на то, что использование моделей машинного обучения с созданием признаков ручным способом показывает неплохие результаты, эти исследования имеют общее ограничение, заключающееся в том, что их производительность варьируется в зависимости от того, насколько хорошо были определены признаки; для разных данных потребуются большие усилия экспертов по предметной области для достижения лучшей производительности. Это ограничение также существует в подходах слияния информации для анализа тональности, в которых сочетаются другие ресурсы (напр. онтология, лексика), поскольку это будет стоить больших затрат времени и усилий экспертов в данной области. Модель глубокого обучения является одним из решений для такого ограничения, поскольку известно, что она автоматически захватывает произвольные шаблоны. Кроме того, как показано в, использование моделей глубокого обучения для анализа текста обеспечит представление характеристик метауровня, которое хорошо обобщает новые области.

Для классификации тональности текста существует две доминирующие техники глубокого обучения: рекуррентные и сверточные нейронные сети. В

данной работе предлагается модель CNN, структура которой была разработана для эффективной классификации тональности текста.

Среди существующих исследований, использующих глубокое обучение для классификации текстов, CNN использует преимущества так называемых сверточных фильтров, которые автоматически изучают функции, подходящие для данной задачи. Например, при использовании CNN для классификации тональности, сверточные фильтры могут захватывать врожденные синтаксические и семантические особенности выражений. Один сверточный слой с комбинацией сверточных фильтров может достичь сопоставимой производительности даже без какой-либо специальной настройки гиперпараметров. Кроме того, CNN не требует экспертных знаний о языковой структуре целевого языка. Благодаря этим преимуществам CNN успешно применяется для анализа различных текстов: семантического анализа, поиска по запросу, моделирования предложений.

Можно утверждать, что для задачи классификации текста лучше применять рекуррентную нейронную сеть (RNN) чем сверточную нейронную сеть (CNN), поскольку она сохраняет порядок последовательности слов. Однако CNN также способна захватывать последовательные шаблоны, что касается локальных шаблонов сверточными фильтрами. Кроме того, по сравнению с RNN, CNN в основном имеет меньшее количество параметров, так что CNN успешно обучается с небольшим количеством данных.

Во **второй главе** представлена информация о нейроне и функциях активации; рассмотрены два типа глубоких нейронных сетей (рекуррентная нейронная сеть и сверточная нейронная сеть); а также, описаны алгоритмы оптимизации, техники обратного распространения ошибки и регуляризации.

В обработке последовательных данных, часто используется рекуррентная нейронная сеть (англ. RNN, Recurrent Neural Network). Модель RNN может быть разделена на четыре типа: «один к одному», «один ко многим», «многие к одному» и «многие ко многим». Эти четыре модели предназначены для решения соответствующих конкретных задач. Задача тегирования предложений в области обработки естественного языка является типичным сценарием для модели RNN. Другим применением модели RNN «многие ко многим» является задача машинного перевода. Задача классификации тональности является репрезентативной задачей для модели RNN «многие к одному». В задаче классификации тональности текста каждое слово в предложении рассматривается как

один вход, единственным выходом является прогнозируемый результат класса тональности для предложения.

При сравнении с моделями CNN разница между RNN и CNN заключается в глубине сети. Глубина сети в модели CNN может быть достаточно большой, в некоторых случаях может превышать 100. Глубина сети в моделях RNN, как правило, мала. Основная причина небольшого количества слоев в сетях RNN заключается в том, что модель RNN разворачивается и рассчитывается вместе с временными шагами, тогда как модель CNN вместо этого рассчитывается в пространстве.

Отсутствие контроля над процессом обучения в глубоких нейронных сетях может привести к проблеме *переобучения*. Переобучение указывает на то, что модели имеют низкую способность к обобщению, что приводит к возможности плохого прогнозирования для тестовых данных, даже если модель достигает высокой производительности в обучающих или валидационных данных. Переобучение происходит, когда разрыв между ошибкой обучения и ошибкой тестирования становится большим. Эта ситуация указывает на то, что модель глубокой нейронной сети обучена научилась "подгонять"результат под данные обучающей выборки вместо изучения реальных шаблонов данных.

Целью обучения модели глубокой нейронной сети является снижение значения функции потерь (англ. loss function). Алгоритмы оптимизации используются для обновления значения параметров модели с целью уменьшения значения целевой функции модели. Когда этап обучения заканчивается, параметры модели в данный момент являются параметрами, которые модель выучила на этапе обучения. Целевая функция также называется функцией потерь (англ. cost function) в глубоких нейронных сетях. Значение целевой функции - это среднее значение рассчитанных потерь, когда данные из набора обучающих данных загружаются для обучения модели DNN. После завершения процесса прямого распространения ошибки, вычисляется среднее значение потерь. Затем применяется алгоритм обратного распространения для вычисления значения градиента в слоях моделей DNN.

Для улучшения способности обобщения глубоких нейронных сетей существует много методов *регуляризации*. К распространенным методам регуляризации относятся: исключение (англ. dropout), пакетная нормализация (англ. batch normalization), механизм ранней остановки (англ. early stopping), методика многозадачного обучения и т.д.

Третья глава посвящена разработке сверточной сети для классификации тональности текста. По экспериментальным результатам показано, что последовательные сверточные слои способствовали повышению производительности на относительно длинном тексте. Предложенные модели CNN достигли около 81% и 68% точности для двоичной классификации и троичной классификации, соответственно. В качестве будущей работы данную нейронную сеть можно применить к другим задачам классификации (к примеру, гендерная классификация). Также можно продолжить поиск лучших структур для классификации текста; например, остаточное соединение для укладки нескольких слоев.

На рисунке 0.1 представлена предлагаемая сеть, которая состоит из слоя внедрения (англ. embedding), двух сверточных слоев, слоя подвыборки (англ. pooling) и полносвязного слоя. Матрица $S \times E$, которая является выходным сигналом слоя внедрения, устанавливается в качестве первого сверточного слоя. Первый сверточный слой представляет собой матрицу $C_1 \times E$, в которой хранится локальная информация, необходимая для классификации тональности в матрице $S \times E$ и передачи информации на следующий сверточный слой. Матрица $C_1 \times E$ "сворачивает"все значения матрицы $S \times E$ с произвольным шагом, вычисляет скалярное произведение и передает результат скалярного произведения на следующий слой. Второй сверточный слой использует матрицу $C_2 \times C_1$ для извлечения признаков из контекстной информации основного слова на основе локальной информации, хранящейся в первом сверточном слое. C_1 и C_2 обозначают размер фильтра каждого сверточного слоя. Два сверточных слоя имеют различные фильтры K_1 и K_2, соответственно, для захвата уникальной контекстной информации. Другими словами, первый сверточный слой используется для просмотра простой контекстной информации при просмотре матрицы $S \times E$, а второй сверточный слой используется для захвата ключевых признаков и их последующего извлечения, которые содержат влияют на классификацию.

После прохождения через слой подвыборки выполняется процесс выравнивания, преобразовывая двумерную карту признаков из выходных данных в одномерный формат и помещая ее в полносвязный слой (англ. FC, Fully-Connected), соединяя все входные и выходные нейроны. Вектор, который проходит через слой FC, формирует выход, который классифицируется как положительный или отрицательный. Функция активации softmax предназначена для классификации нескольких классов. Вывод функции softmax представляет собой значения вероятности для каждого класса.

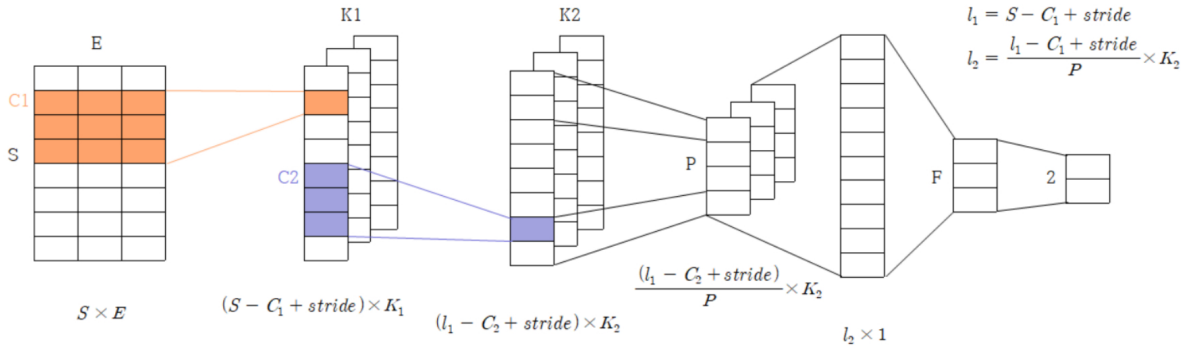

Рис. 0.1 — Графическое представление сети. Выходные размеры каждого слоя представлены внизу соответствующих слоев

С точки зрения F-меры предлагаемая сеть была примерно на 10% больше, чем традиционные модели. Можно предположить, что главная причина - врожденная способность сверточных нейронных сетей захватывать местные модели. Сверточные сети обладают способностью захватывать локальные шаблоны и шаблоны более высокого слоя через сверточные слои и слои подвыборки. Предполагаю, что такая способность модели CNN приводит к разрыву в производительности. Предлагаемая сеть также была лучше, чем современные модели глубокого обучения. Сеть, предлагаемая в диссертации состоит же из последовательных фильтров, которые имеют два преимущества: (1) извлекаются иерархические (более высокие) функции, и (2) фильтры высокого уровня имеют более широкий диапазон, чтобы видеть локальные шаблоны.

ЗАКЛЮЧЕНИЕ

В работе были проанализированы существующие решения задач классификации текста. Были выделены модели, показывающие наилучший результат в настоящее время. Также были определены преимущества использования сверточных нейронных сетей вместо рекуррентных. К примеру, важными преимуществами было то, что сверточные сети также способны захватывать последовательные шаблоны во входном тексте; или то, что по сравнению с рекуррентными сетями, сверточные в основном имеют меньшее количество параметров, так что они успешно обучаются с небольшим количеством данных.

Также в работе была рассмотрена информация об алгоритмах классификации тональности текста. Была описана подробная концепция и связанные с ней теоретические знания о нейронной сети, включая базовый модуль нейронной сети, методику регуляризации и стратегии обучения нейронных сетей; были

описаны концепции обработки естественного языка, включая технику вложения слов и алгоритмы классификации текста; также были представлены общие метрики оценки для алгоритмов классификации текста.

Была разработана сверточная нейронная сеть для классификации тональности текста. По экспериментальным результатам бало показано, что последовательные сверточные слои способствовали повышению производительности на относительно длинном тексте. Предложенные модели CNN достигли около 81% и 68% точности для двоичной классификации и троичной классификации, соответственно. В качестве будущей работы данную нейронную сеть можно применить к другим задачам классификации (к примеру, гендерная классификация). Также можно продолжить поиск лучших структур для классификации текста; например, остаточное соединение для укладки нескольких слоев.