МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
им. Н.Э. Баумана

Факультет «Информатика и системы управления»
Кафедра «Систем обработки информации и управления»

ОТЧЕТ

Домашнее задание
по дисциплине «Методы машинного обучения»

Задача: «Graph Models»

ИСПОЛНИТЕЛЬ: ___Очеретная С.В.__
ФИО
группа ИУ5-25М _____
подпись
"__"_____2024 г.

ПРЕПОДАВАТЕЛЬ: _____Гапанюк Ю.Е.____
ФИО
_____
подпись
"__"_____2024 г.

Москва - 2024

Задание

Домашнее задание включает три основных этапа:

- выбор задачи;
- теоретический этап;
- практический этап.

Этап выбора задачи предполагает анализ ресурса paperswithcode. Данный ресурс включает описание нескольких тысяч современных задач в области машинного обучения. Каждое описание задачи содержит ссылки на наиболее современные и актуальные научные статьи, предназначенные для решения задачи (список статей регулярно обновляется авторами ресурса). Каждое описание статьи содержит ссылку на репозиторий с открытым исходным кодом, реализующим представленные в статье эксперименты. На этапе выбора задачи обучающийся выбирает одну из задач машинного обучения, описание которой содержит ссылки на статьи и репозитории с исходным кодом.

Теоретический этап включает проработку как минимум двух статей, относящихся к выбранной задаче. Результаты проработки обучающийся излагает в теоретической части отчета по домашнему заданию, которая может включать:

- описание общих подходов к решению задачи;
- конкретные топологии нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения, предназначенных для решения задачи;
- математическое описание, алгоритмы функционирования, особенности обучения используемых для решения задачи нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения;
- описание наборов данных, используемых для обучения моделей;
- оценка качества решения задачи, описание метрик качества и их значений;
- предложения обучающегося по улучшению качества решения задачи.

Практический этап включает повторение экспериментов авторов статей на основе представленных авторами репозиториев с исходным кодом и возможное улучшение обучающимися полученных результатов. Результаты проработки обучающийся излагает в практической части отчета по домашнему заданию, которая может включать:

– исходные коды программ, представленные авторами статей, результаты документирования программ обучающимися с использованием диаграмм UML, путем визуализации топологий нейронных сетей и другими способами;

– результаты выполнения программ, вычисление значений для описанных в статьях метрик качества, выводы обучающегося о воспроизводимости экспериментов авторов статей и соответствии практических экспериментов теоретическим материалам статей;

– предложения обучающегося по возможным улучшениям решения задачи, результаты практических экспериментов (исходные коды, документация) по возможному улучшению решения задачи.

Ход работы

Постановка задачи

Для изучения взяли задачу Graph Models.

Описание задачи:

Графовые методы включают в себя архитектуры нейронных сетей для обучения на графах с предварительной структурной информацией, обычно называемые графовыми нейронными сетями (GNN).

В последнее время подходы глубокого обучения расширяются для работы с данными, структурированными на графах, что приводит к появлению серии нейронных сетей на графах, решающих различные задачи. Графовые нейронные сети особенно полезны в приложениях, где данные генерируются из неевклидовых областей и представляются в виде графиков со сложными отношениями.

Некоторые задачи, в которых широко используются GNN, включают классификацию узлов, классификацию графов, прогнозирование связей и многое другое.

Теоретическая часть

Исследования проводили по статьям про сверточные нейронные сети для классификации текста.

Классификация текста — важная и классическая проблема обработки естественного языка. Был проведен ряд исследований, в которых для классификации применялись сверточные нейронные сети (свертка на регулярной сетке, например, последовательности). Однако лишь ограниченное количество исследований изучало более гибкие сверточные нейронные сети на графах (свертка на несеточном, например, произвольном графе) для этой задачи.

В данной работе мы используем сверточные сети графов для классификации текста. Мы строим единый текстовый граф на основе

совпадения слов и документируем отношения слов, а затем изучаем сверточную сеть текстовых графов (Text GCN) для корпуса. Наш текстовый GCN инициализируется с помощью единого представления для слова и документа, затем он совместно изучает вложения как для слов, так и для документов, как это контролируется известными метками классов для документов.

Наши экспериментальные результаты на нескольких эталонных наборах данных показывают, что стандартный текстовый GCN без каких-либо внешних вложений слов или знаний превосходит современные методы классификации текста. С другой стороны, Text GCN также изучает прогнозируемые встраивания слов и документов. Кроме того, экспериментальные результаты показывают, что улучшение Text GCN по сравнению с современными методами сравнения становится более заметным по мере того, как мы снижаем процент обучающих данных, что предполагает устойчивость Text GCN к меньшему количеству обучающих данных при классификации текста.

Методы решения задачи

Традиционная классификация текстов

Традиционные исследования классификации текста в основном сосредоточены на разработке признаков и алгоритмах классификации. Наиболее часто используемой функцией является функция «мешок слов». Кроме того, были разработаны некоторые более сложные функции, такие как n-граммы и сущности в онтологиях. Также существуют исследования по преобразованию текстов в графики и проектированию признаков графов и подграфов. В отличие от этих методов, наш метод может автоматически изучать текстовые представления в виде вложений узлов.

Глубокое обучение для классификации текста.

Исследования по классификации текста глубокого обучения можно разделить на две группы. Одна группа исследований была сосредоточена на моделях, основанных на встраивании слов. Несколько недавних исследований

показали, что успех глубокого обучения классификации текста во многом зависит от эффективности встраивания слов. Некоторые авторы агрегировали неконтролируемые встраивания слов как встраивания документов, а затем помещали эти вложения документов в классификатор. Другие совместно изучали встраивание слов/документов и меток документов.

Наша работа связана с этими методами, основное отличие состоит в том, что эти методы создают текстовые представления после изучения вложений слов, в то время как мы одновременно изучаем вложения слов и документов для классификации текста. Другая группа исследований использовала глубокие нейронные сети. Двумя репрезентативными глубокими сетями являются CNN и RNN. Архитектура представляет собой прямое применение CNN, используемых в компьютерном зрении, но с одномерными извилинами. Также существует LSTM, особый тип RNN, для изучения представления текста. Чтобы еще больше повысить гибкость представления таких моделей, механизмы внимания были введены как неотъемлемая часть моделей, используемых для классификации текста. Хотя эти методы эффективны и широко используются, они в основном сосредоточены на локальных последовательных последовательностях слов, но не используют явным образом глобальную информацию о совместном появлении слов в корпусе.

Графовые нейронные сети

В последнее время тема графовых нейронных сетей привлекает все больше внимания. Ряд авторов обобщили хорошо зарекомендовавшие себя модели нейронных сетей, такие как CNN, которые применяются к регулярной сетке (2-мерная сетка или 1-мерная последовательность) для работы с произвольно структурированными графами. Графовая сверточная сеть (GCN) позволила достичь современных результатов классификации на ряде эталонных наборов графовых данных. GCN также использовался в нескольких задачах NLP, таких как маркировка семантических ролей, классификация отношений и машинный перевод, где GCN используется для кодирования синтаксической структуры. предложений. В некоторых недавних

исследованиях нейронные сети на графах использовались для классификации текста. Однако они либо рассматривали документ или предложение как граф узлов слов, либо полагались на нестандартно доступное отношение цитирования документа. для построения графа. Напротив, при построении корпусного графа мы рассматриваем документы и слова как узлы (следовательно, гетерогенный граф) и не требуем отношений между документами.

Архитектура нашей нейронной сети Text GCN

Мы строим большой и гетерогенный текстовый граф, который содержит узлы слов и узлы документов, чтобы можно было явно смоделировать глобальное совместное появление слов и легко адаптировать свертку графа, как показано на рисунке 1. Количество узлов в текстовом графе — это количество документов (размер корпуса) + количество уникальных слов (размер словаря).

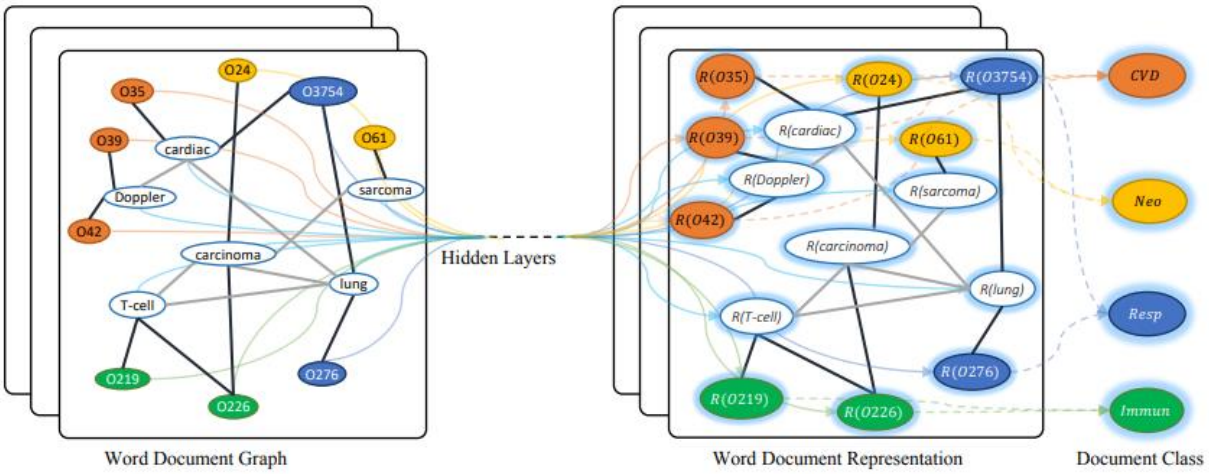

Рис. 1. Архитектура сети

На рисунке: узлы, начинающиеся с «O», — это узлы документов, остальные — узлы слов. Черные жирные ребра — это ребра документа, а тонкие серые ребра — это ребра слова. R(x) означает представление (вложение) x. Разные цвета означают разные классы документов (показаны только четыре примера классов). CVD: сердечно-сосудистые заболевания,

Neo: новообразования, Resp заболевания дыхательных путей, Immun: иммунологические заболевания.

Практическая часть

Цель данной части – сравнить нашу сеть Text GCN со следующими:

– **TF-IDF + LR**: Модель «мешка слов» с обратным взвешиванием частоты терминов по частоте документов. В качестве классификатора используется логистическая регрессия;

– **CNN**: Сверточная нейронная сеть. Мы исследовали CNN-rand, который использует случайно инициализированные встраивания слов, и CNN-non-static, который использует предварительно обученные встраивания слов;

– **LSTM**: Модель LSTM, которая использует последнее скрытое состояние в качестве представления всего текста. Мы также экспериментировали с моделью с предварительно обученными встраиваниями слов или без них.

Эксперименты будем проводить со следующими наборами:

– **20NG** (данные о новостях, распределенных по темам). Набор содержит 18 846 документов, равномерно распределенных по 20 различным категориям. Всего в обучающем наборе находится 11 314 документов, а в тестовом — 7 532 документа.

– **Ohsumed**. Набор взят из базы данных MEDLINE, которая представляет собой библиографическую базу данных важной медицинской литературы, поддерживаемую Национальной медицинской библиотекой. В этой работе мы использовали 13 929 уникальных рефератов по сердечно-сосудистым заболеваниям из первых 20 000 рефератов 1991 года. Каждый документ в наборе имеет одну или несколько связанных категорий из 23 категорий заболеваний. Поскольку мы фокусируемся на классификации текста по одной метке, документы, принадлежащие нескольким категориям, исключаются, так что остается 7400 документов, принадлежащих только одной категории. 3357

документов находятся в обучающем наборе и 4043 документа — в тестовом наборе.

– **MR**: набор данных обзоров фильмов для бинарной классификации настроений, в котором каждый обзор содержит только одно предложение. В наборе 5331 положительный и 5331 отрицательный отзыв.

Получили следующие результаты точностей:

Модель	**Наборы данных**		
	20NG	**Ohsumed**	**MR**
TF-IDF + LR	0,8319	0,5466	0,7459
CNN-rand	0,7693	0,4387	0,7498
CNN-non-static	0,8215	0,5844	**0,7775**
LSTM	0,6571	0,4113	0,7506
LSTM (pretrain)	0,7543	0,5110	0,7733
Text GCN	**0,8634**	**0,6836**	0,7674

Пример распределения данных по категориям для набора 20NG представлен на рис. 2.

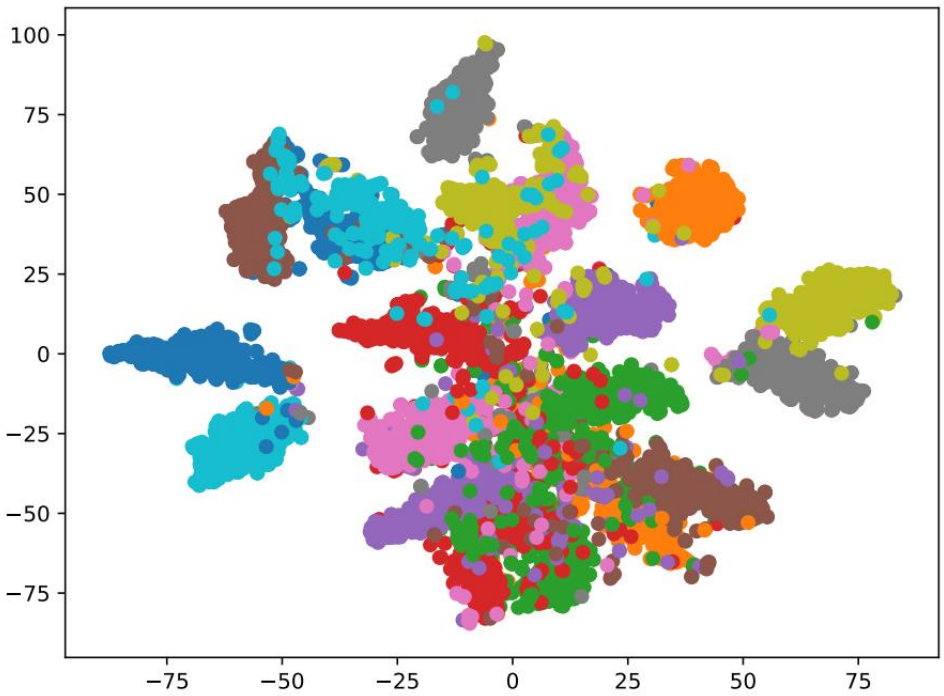

Рис. 2. Результат распределения новостей по категориям с помощью модель Text GCN

Вывод

Предложенная модель со сверточной нейронной сетью отлично подходит для решения задач текстовой классификации. Лучше всего данная модель справилась с многоклассовой классификацией, с бинарной – немного хуже.

Список использованных источников

1. Graph Models. URL: https://paperswithcode.com/methods/category/graph-models. Дата обращения – 20.05.2024.

2. Graph Convolutional Networks for Text Classification. URL: https://paperswithcode.com/paper/graph-convolutional-networks-for-text. Дата обращения – 20.05.2024.