МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра математического обеспечения вычислительных комплексов и информационных систем

Эффективное и масштабируемое трансферное обучение для обработки естественного языка

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Папшицкой Марии Алексеевны

Научный руководитель:
профессор, д. ф.-м. н. _____ Д. К. Андрейченко

подпись, дата

Зав. кафедрой МОВКиС:
д. ф.-м. н., профессор _____ Д. К. Андрейченко

подпись, дата

Саратов 2021

ВВЕДЕНИЕ

Нейронные сети лучше всего работают при обучении на больших объемах данных, но большинство помеченных наборов данных в natural language processing (NLP) имеют небольшой размер. Трансферное обучение предлагает решение: вместо того, чтобы изучать одну задачу с нуля и изолированно, модель может извлечь выгоду из большого количества текста в Интернете или других задач с обширными аннотациями. Эти дополнительные данные позволяют тренировать более крупные и выразительные сети. Однако это также резко увеличивает вычислительные затраты на обучение.

Актуальность заключается в отражении возникающих тенденций во всей области. Одна из тенденций — это отход от больших помеченных наборов данных. Поэтому было уделено внимание методу, обучающемуся только на основе помеченных наборов данных.

Внимание было уделено многозадачному обучению, полу-контролируемое обучение и самостоятельное обучение остаются для будущих разработок.

Чтобы снизить эти затраты на обучение, была поставлена цель рассмотрения метода трансферного многозадачного обучения в качестве метода, который обучается эффективнее, при этом сохраняя высокую масштабируемость.

Для достижения цели поставлены следующие задачи:

1) рассмотреть теоретические основы и математический аппарат трансферного обучения, подходов и модели Transformer;
2) рассмотреть теоретические основы дистилляции знаний;
3) рассмотреть основы работы с BERT и настройки модели;
4) изучить результаты работы и преимущества многозадачного обучения в совокупности с дистилляцией знаний.

Методологические основы эффективного и масштабируемого трансферного обучения для обработки естественного языка представлены в работах

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS;
2. Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In NeurIPS;
3. Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Improving multitask deep neural networks via knowledge distillation for natural language understanding. ArXiv, abs/1904.09482;
4. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. ArXiv, abs/1503.02531;
5. Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks;
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of deep bidirectional transformers for language understanding. In NAACL-HLT;

Теоретическая и практическая значимость бакалаврской работы. Из экспериментов видно, что Single → Multi дистилляция дает результаты неизменно лучше, чем при стандартном однозадачном или многозадачном обучении. Достижение надежного выигрыша в многозадачности при выполнении многих задач оставалось и остается трудным, поэтому метод дистилляции делает многозадачное обучение более полезным в рамках NLP. Однако, за исключением тесно связанных задач с небольшими наборами данных (например, MNLI помогает RTE), общий размер выигрыша от полученного многозадачного метода невелик по сравнению с выигрышем, обеспечиваемым transfer обучением из self-supervised задач (т.е. BERT). В целом, небольшое количество примеров из связанной задачи все еще может потерять свою значимость по сравнению с миллиардами токенов

немаркированного текста, особенно потому, что неконтролируемые задачи, выполняемые над немаркированным текстом, как правило, очень обширны (например, текст поколения) по сравнению с контролируемыми задачами (например, бинарная классификация предложений по определенному явлению, например, сантиментам). Однако на BERT можно получить большие выгоды, когда наборы данных небольшие, а задачи тесно связаны между собой.

Структура и объём работы. Бакалаврская работа состоит из введения, трех разделов, заключения, списка использованных источников и двух приложений. Общий объем работы – 92 страницы, из них 55 страниц – основное содержание, включая 40 рисунков и 5 таблиц, цифровой носитель в качестве приложения, список использованных источников информации – 42 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Общие сведения» посвящен общему справочному материалу для последующих глав. Сначала рассматриваются сети Transformer и извлечение знаний, важные методы глубокого обучения, которые будут использоваться в работе. Далее приводится обзор работы по трансферному обучению с акцентом на приложениях для обработки естественного языка. Он охватывает один из трех основных типов трансферного обучения: использование дополнительных помеченных данных (многозадачное обучение).

Трансформер – модель, которая использует механизм внимания для повышения скорости обучения. Более того, для ряда задач Трансформеры превосходят модель нейронного машинного перевода от Google. Основная задача моделирования, которую решают Transformers, — это фиксация зависимостей на большом расстоянии, что может иметь решающее значение для моделирования текста. До Transformers большинство нейронных моделей в NLP были основаны на рекуррентных нейронных сетях (RNN). RNN обрабатывают ввод по одному токену за раз, на каждом шаге обновляя представление последовательности, наблюдаемой до сих пор. Хотя они привлекательны тем, что они могут «читать» текст слева направо, как это делают люди, захват всего прочитанного контекста в одном векторе имеет ограничения. Анализ показывает, что модели «забывают» информацию издалека и, как правило, работают хуже, чем длиннее предложения. Стробирующие механизмы, такие как сети долгосрочной краткосрочной памяти, могут уменьшить этот недостаток, но только частично.

Дистилляция знаний — это метод передачи знаний из нейронной сети «учителя» в сеть «ученика». Ключевая идея состоит в том, чтобы научить ученика имитировать распределение результатов работы учителя. Дистилляция знаний может применяться для обучения эффективной небольшой модели, обучая ее на точной, но дорогостоящей в вычислительном отношении модели, такой как ансамбль или модель со

многими параметрами. В этой работе применяется дистилляция знаний для улучшения многозадачных моделей, используя однозадачные модели в качестве учителей. Также используется форма «самодистилляции» в качестве метода обучения под непосредственным руководством.

Многозадачное обучение позволяет моделям извлекать выгоду из знаний, полученных из обучающего сигнала связанных задач. По сравнению с обучением по одной задаче, это также ближе к тому виду обучения, которым пользуются люди, когда использование предыдущего опыта позволяет нам быстро осваивать новые навыки. С более практической точки зрения, разработка и развертывание одной многозадачной модели может быть проще, чем наличие множества однозадачных моделей. Многозадачные модели подвержены меньшему риску переоснащения, чем однозадачные, потому что они используют множество разнообразных входных данных.

Трансферное обучение особенно привлекательно в сочетании с нейронными сетями, потому что нейронные сети работают лучше, когда предоставляются большие наборы данных. Центральные цели трансферного обучения – создание общих систем, которые могут выполнять множество задач, и обобщаемых систем, которые работают с разнообразными данными за пределами небольшой помеченной области. В идеале многозадачное обучение приводит не только к более широко применимой системе, но и к лучшей, потому что изучение одной задачи может предоставить знания модели, которые улучшают ее способность решать связанные задачи. Это сдвигает модель, по крайней мере, в небольшой степени, к тому виду человеческого обучения, когда использование предыдущего опыта позволяет нам быстро развивать новые навыки.

Второй раздел «Задачи и наборы данных NLP» посвящен обзору задач и наборов данных.

Синтаксические задачи предполагают предсказание грамматической структуры предложения. Большинство наборов данных взято из Penn

Treebank (Marcus et al., 1993), который представляет собой большой набор статей Wall Street Journal с аннотациями синтаксических деревьев.

Маркировка части речи (POS): пометка слов с их синтаксическими категориями (например, определитель, прилагательное и т. Д.). Маркировка POS — это пример задачи маркировки тегов, где каждому слову в предложении присваивается метка.

Разделение текста на части: разделение предложения на синтаксически коррелированные части. В то время как маркировка выполняется по группам слов, фрагменты текста часто преобразуются в задачу маркировки с использованием такого метода, как кодирование BIO.

Комбинированная категориальная грамматика (Combinatory Categorial Grammar - CCG) Supertagging: помечает слова супертэгами CCG, лексическими категориями, которые кодируют информацию о структуре предиката-аргумента предложения.

Анализ зависимостей: вывод древовидной структуры, описывающей синтаксис предложения. При синтаксическом анализе зависимости слова в предложении обрабатываются как узлы в графе.

Контрольные задачи общего понимания языка. Следующий набор задач охватывает более широкий набор наборов данных, охватывающих понимание естественного языка (NLU — natural language understanding). Они получены из теста General Language Understanding (GLUE, Wang et al. (2019)), который представляет собой набор задач NLU.

CoLA: Corpus of Linguistic Acceptance (Warstadt et al., 2018). Задача состоит в том, чтобы определить, является ли данное предложение грамматическим или нет. Набор данных содержит 8,5 тыс. примеров из книг и журнальных статей по теории лингвистики.

SST: Stanford Sentiment Treebank (Socher et al., 2013). Задача состоит в том, чтобы определить, является ли предложение положительным или отрицательным по тональности. Набор данных содержит 67 тыс. примеров из обзоров фильмов.

MRPC: Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005). Задача состоит в том, чтобы предсказать, являются ли два предложения семантически эквивалентными или нет. Набор данных содержит 3,7 тыс. примеров из онлайн-источников новостей.

STS: Semantic Textual Similarity (Cer et al., 2017). Задача состоит в том, чтобы предсказать, насколько семантически похожи два предложения по шкале от 1 до 5. Набор данных содержит 5,8 тыс. примеров, взятых из новых заголовков, подписей к видео и изображениям, а также данных логического вывода на естественном языке.

QQP: Quora Question Pairs (Iyer et al., 2017). Задача состоит в том, чтобы определить, являются ли пары вопросов семантически эквивалентными. Набор данных содержит 364 тыс. примеров с сайта сообщества Quora.

MNLI: Multi-genre Natural Language Inference (Williams et al., 2018). Учитывая предложение-предпосылку и предложение-гипотезу, надо предсказать, влечет ли посылка за гипотезой, противоречит ли она гипотезе или нет. Набор данных содержит 393 тыс. примеров, взятых из десяти различных источников.

QNLI: вопрос о логическом выводе; построенный из SQuAD (Rajpurkar et al., 2016). Задача состоит в том, чтобы предсказать, содержит ли контекстное предложение ответ на вопросительное предложение. Набор данных содержит 108 тысяч примеров из Википедии.

RTE: Распознавание текстового захвата (Giampiccolo et al., 2007). Учитывая предложение предпосылки и предложение гипотезы, задача состоит в том, чтобы предсказать, влечет ли предпосылка гипотезу или нет. Набор данных содержит 2,5 тыс. примеров из серии ежегодных текстовых задач.

WNLI: схема Winograd (Levesque, 2011). Цель состоит в том, чтобы правильно выбрать антецедент местоимения из двух возможных. Схема Winograd намеренно построена так, что небольшое изменение предложения

меняет схему Winograd (например, изменение «маленького» на «большое» ниже).

Ответы на вопросы. SQuAD 1.1: Учитывая контекстный абзац (например, часть статьи в Википедии) и вопрос (например, о теме статьи), задача состоит в том, чтобы выбрать диапазон текста в абзаце, отвечающем на вопрос. Набор данных содержит 88 тысяч примеров из Википедии.

SQuAD 2.0: Эта задача добавляет в SQuAD дополнительные вопросы, ответ на которые не существует в контексте; модели должны распознавать, когда возникают эти вопросы, и не давать на них ответ. Набор данных содержит 130 тыс. примеров из Википедии (около 50 тыс. вопросов без ответа добавлено поверх SQuAD 1.1).

Другие задачи. Распознавание именованных сущностей (NER): идентификация именованных сущностей в предложении и классификация каждого как местоположения, организации, человека или разных сущностей. Используется набор данных CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) и набор данных OntoNotes (Hovy et al., 2006).

Машинный перевод: Машинный перевод — это задача автоматического преобразования текста с одного языка на другой с сохранением смысла. Используется набор данных перевода англо-вьетнамского языка из IWSLT 2015 (Cettolo et al., 2015), который содержит переводы выступлений TED.

Третий раздел «Многозадачное обучение. Описание метода» посвящен описанию метода многозадачного обучения, при котором многозадачная модель не уступает своим однозадачным аналогам. Ключевая идея – использовать дистилляцию знаний, так что однозадачные модели эффективно обучают многозадачной модели. Учитель и ученик имеют одинаковую нейронную архитектуру и размер модели, но удивительно, что ученик может превзойти точность учителя. В этой работе используются born-again networks для улучшения многозадачности. Сначала обучается однозадачная модель для каждой задачи, а затем превращается в

многозадачную. Интуитивно понятно, что многозадачная модель никогда не будет работать хуже, чем однозадачная, если она может непосредственно узнать, как однозадачная модель выполняет задачу. Модели построены на основе BERT, чтобы показать масштабируемость метода, он работает даже поверх большой и уже очень сильной модели.

Сравниваются Single → Multi (используется Single → Multi, чтобы указать на преобразование однозадачных моделей «учителей» в многозадачные модели «учеников».) born-again дистилляция с несколькими другими вариантами (Single → Single и Multi → Multi), а также исследуется выполнение нескольких раундов дистилляции (Single → Multi → Single → Multi). Выигрыш у Single → Multi больше, чем у Single → Single, что говорит о том, что дистилляция особенно хорошо работает в сочетании с многозадачным обучением. Интересно, что Single → Multi работает значительно лучше, чем Multi → Multi дистилляция. В дополнение к моделям также были обучены модели Single → Multi → Single → Multi. Однако разница с Single → Multi не была статистически значимой, что позволяет предположить, что многократная дистилляция имеет небольшую ценность.

Стандартное многозадачное обучение улучшается по сравнению с однозадачным обучением для RTE (вероятно, потому что оно тесно связано с MNLI), по другим задачам нет улучшений. В отличие от этого, дистилляция знаний Single → Multi улучшает или соответствует производительности других методов для всех задач, кроме STS, единственной задачи регрессии в GLUE. Вывод: дистилляция не подходит для задач регрессии.

В целом, ключевым преимуществом данного метода является надежность: в то время как стандартное многозадачное обучение дает смешанные результаты, Single → Multi дистилляция неизменно превосходит стандартное single-task и multi-task обучение. В некоторых испытаниях single-task обучение приводило к моделям, которые набирали довольно низкие баллы, в то время как многозадачные модели имели более надежную производительность.

ЗАКЛЮЧЕНИЕ

Область обработки естественного языка кардинально изменилась за последние несколько лет, и трансферное обучение играет центральную роль в этом изменении. Исторически сложилось так, что большая часть исследований NLP проводилась путем разработки системы, которая изучает конкретную задачу на основе данных, помеченных людьми. Напротив, теперь можно загружать предварительно обученные модели, которые можно точно настроить для обеспечения превосходной производительности при выполнении многих задач с минимальными затратами на разработку конкретных задач. Однако такой быстрый прогресс от трансферного обучения не был простым: в основе этого успеха лежит использование больших моделей, требующих обширных вычислительных ресурсов для обучения. Чтобы снизить эти затраты, целью данной работы было рассмотрение метода дистилляции обучения как метода, который был бы эффективным (позволяющим эффективно использовать вычислительные ресурсы), оставаясь при этом масштабируемым (дающими улучшения по мере роста данных и размеров моделей).

Эта цель была реализована в одной трех основных областях трансферного обучения: многозадачное обучение. Для будущих работ оставлены частично контролируемое и самостоятельное обучение.

Метод эффективен в разнообразном диапазоне задач NLP, начиная от синтаксических, таких как анализ зависимостей, до сложных задач понимания естественного языка, таких как логический вывод естественного языка и понимание прочитанного, демонстрируя универсальность метода. Он использует дистилляцию знаний для достижения эффективного многозадачного обучения с постоянным преимуществом по сравнению с однозадачными системами, даже при обучении множеству разнообразных задач. И позволяет многозадачным моделям масштабироваться для многих задач с меньшим риском ухудшения результатов из-за вмешательства задачи.

Основные источники информации:

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS;
2. Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In NeurIPS;
3. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In ICLR;
4. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. ArXiv, abs/1503.02531;
5. Tommaso Furlanello, Zachary C Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks;
6. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of deep bidirectional transformers for language understanding. In NAACL-HLT;
7. Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In EMNLP;
8. Lili Mou, Ran Jia, Yan Xu, Ge Li, Lu Zhang, and Zhi Jin. 2016. Distilling word embeddings: An encoding approach. In CIKM;
9. Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Improving multitask deep neural networks via knowledge distillation for natural language understanding. ArXiv, abs/1904.09482.
10. Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. In ICLR.
11. Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Unifying question answering and text classification via span extraction. ArXiv, abs/1904.09286.