УДК 004.82

ОБУЧЕНИЕ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ В УСЛОВИЯХ ОГРАНИЧЕННОГО ОБУЧАЮЩЕГО НАБОРА ДАННЫХ

Чуб Вадим Сергеевич[1],
аспирант,
e-mail: vadim-chub13@mail.ru,

Цветкова Ольга Леонидовна[1],
канд. техн. наук, доцент,
e-mail: olga_cvetkova@mail.ru,
[1]*Донской государственный технический университет, г. Ростов-на-Дону, Россия*

В настоящее время для решения задач обработки естественного языка используются методы трансферного обучения на основе трансформеров, однако они могут быть требовательными к вычислительным ресурсам и памяти. Альтернативный подход, предложенный в данной статье, связан с использованием предварительно обученной языковой модели на основе рекуррентной нейронной сети LSTM (англ. Long Short-Term Memory – «долгая краткосрочная память»), позволяющей выполнить обработку естественного языка (анализ тональности) на наборе, содержащем текстовые данные. В качестве эталонного теста предложенной модели для решения задачи распознавания «языка ненависти» в тексте рассмотрены другие модели на основе трансформеров. Тест на деградацию, оценивающий устойчивость моделей к снижению производительности при уменьшении количества обучающих данных, подтверждает эффективность модели для ограниченного обучающего набора данных. Предложенная LSTM-модель на основе рекуррентной нейронной сети позволяет с нуля использовать предварительно обученную языковую модель и обеспечить более быстрое предварительное обучение в моделях на основе трансформера.

Ключевые слова: рекуррентные нейронные сети, трансферное обучение, ограниченный набор данных, обработка естественного языка

TRAINING OF RECURRENT NEURAL NETWORKS IN A LIMITED TRAINING DATASET

Tshub V.S.[1],
postgraduate student,
e-mail: vadim-chub13@mail.ru,

Tsvetkova O.L.[1],
candidate of technical sciences, associate professor,
e-mail: olga_cvetkova@mail.ru,
[1]*Don State Technical University, Rostov-on-Don, Russia*

Currently, transformer-based transfer learning methods are used to solve natural language processing problems, but they can be demanding on computing resources and memory. An alternative approach proposed in this article is related to the use of a pre-trained language model based on the recurrent neural network LSTM (Long Short-Term Memory), which allows natural language processing (tonality analysis) to be performed on a set containing text data. As a reference test of the proposed model for solving the problem of recognizing the "hate speech" in the text, other models based on transformers are considered. The degradation test, which evaluates the resilience of models to performance degradation with a decrease in the amount of training data, confirms the effectiveness of the model for a limited training dataset. The proposed LSTM model based on a recurrent neural network allows us to use a pre-trained language model from scratch and provide faster pre-learning in transformer-based models.

Keywords: recurrent neural networks, transfer learning, limited data set, natural language processing

DOI 10.21777/2500-2112-2023-4-79-85

Введение

В последние годы обработка естественного языка (англ. Natural Language Processing, NLP) быстро развивалась из-за наличия большого объема данных, вызванного распространением Интернета и доступностью более дешевых вычислительных мощностей. Был достигнут значительный прогресс в создании приложений NLP для решения таких задач, как машинный перевод, категоризация текста, классификация текста и др. Большая часть приложений NLP создана на основе популярных языков, таких как английский, испанский и другие, которые имеют доступные корпусы и аннотированные тексты. Языки, для которых доступ к большим текстовым корпусам или аннотированным фрагментам текста является ограниченным, называются малоресурсными.

Потребность в больших текстовых корпусах или наборах данных для использования преимуществ технологий NLP является препятствием, которое предотвращает или замедляет повсеместное внедрение этой технологии для языков с ограниченными ресурсами.

В последние годы методы обучения на основе трансформеров в основном преобладали, когда речь шла об эффективности технологий NLP в различных приложениях на языках с ограниченными ресурсами. К таким приложениям относятся, например, такие, которые обеспечивают обнаружение фейковых новостей на филиппинском языке [1], распознавание именованных объектов на вьетнамском языке [2] и другие. И хотя модели на основе трансформеров доказали свою эффективность в различных приложениях NLP, в настоящее время они имеют ограниченное применение, поскольку для их обучения требуется высокая вычислительная мощность и память. Модели на основе рекуррентных нейронных сетей, такие как AWD-LSTM [3], более доступны и дешевы в обучении по сравнению с моделями на основе трансформера, но уступают по ряду параметров. Доступность LSTM-модели важна для расширения распространения технологии NLP и повышения доступности ее преимуществ для решения многих задач в условиях ограниченного обучающего набора данных.

Целью статьи является создание и оценка эффективности применения языковой LSTM-модели на основе рекуррентной нейронной сети, позволяющей с нуля использовать предварительно обученную языковую модель и обеспечить более быстрое предварительное обучение в моделях на основе трансформера.

В рамках исследования можно выделить следующие задачи:

1) предложить предварительно обученную языковую LSTM-модель на филиппинском языке, которая станет базой для разработки приложений NLP для малоресурсных естественных языков;

2) сравнить LSTM-модель с другими моделями в задаче классификации «языка ненависти» в тексте, в частности, с моделями на основе трансформеров;

3) оценить производительность предложенной модели в условиях ограниченного ресурса, используя тест на деградацию, и сравнить ее с моделями на основе трансформеров.

Трансферное обучение как метод машинного обучения без опыта

Обработка естественного языка (Natural Language Processing, NLP) – это междисциплинарное научно-техническое направление лингвистики, информатики и искусственного интеллекта. В рамках данного направления исследуются методы и средства обработки и понимания естественного языка. Существующие приложения NLP реализуют следующие функции: классификацию текста (спам-фильтры электронной почты), понимание языка (умные помощники), машинный перевод (перевод текста с одного языка на другой язык), языковое моделирование (предсказание следующего слова на основе предыдущего) и др. Благодаря открытому доступу к большому объему данных и доступности вычислительных мощностей в последние годы интерес к машинному обучению существенно вырос. Машинное обучение использует большие объемы данных, которые компьютерный алгоритм использует для выявления скрытых закономерностей в этих данных. Благодаря этому машинное обучение стало эффективным подходом к решению сложных задач, поскольку нет необходимости напрямую программировать правила для решения той или иной проблемы.

Подход машинного обучения связан с обработкой больших объемов данных для получения приемлемого результата, что требует значительных вычислительных и временных ресурсов. Чтобы занять-

ся глубинным обучением нейронной сети, необходимо иметь доступ к большому очищенному набору данных и самостоятельно разработать и обучить эффективную модель. Таким образом, проекты без существенной поддержки извне невозможны по умолчанию. Один из способов решения проблемы – трансферное обучение. Трансферное обучение строится на том, что знания, накопленные в модели, подготовленной для выполнения одной задачи, могут быть перенесены на другую модель, направленную на решение другой целевой задачи. Например, знания модели распознавания человеческого лица можно использовать в качестве основы для того, чтобы узнать, является ли эмоция лица злостью, счастьем и т.п. [3]. Однако модели на основе трансформеров требовательны к вычислительным ресурсам и памяти, что вызывает необходимость совершенствования подходов и создания новых моделей обучения нейронной сети.

Исследование предложенной языковой LSTM-модели

С целью обхода высоких вычислительных требований модели на основе трансформеров проводится исследование предварительно обученной языковой LSTM-модели на конкретном языке, которая станет базой для разработки приложений NLP для малоресурсных естественных языков. Исследование проводится на примере решения задачи распознавания «языка ненависти».

Методика исследования включает следующие этапы:

1) обучить модель с применением подхода Universal Language Model Fine-tuning (ULMFiT) [5] методу на основе рекуррентной нейронной сети;

2) измерить эффективность модели в задаче классификации текста, используя набор данных о «языке ненависти» [6];

3) используя тот же набор данных, провести тест на деградацию для оценки производительности модели в условиях ограниченных ресурсов [1].

В данной статье используется инструментарий для обучения модели API fastai v2.0.13. Если в модели не упоминается конкретная конфигурация, это означает, что используется только настройка по умолчанию. Чтобы ускорить обучение, можно использовать обучение смешанной точности [7]. В обучении также используется графический процессор Tesla T4. На протяжении всего процесса обучения модели использовалась политика одного цикла графика скорости обучения [8].

ULMFiT (англ. Universal Language Model Finetuning, точная настройка универсальной языковой модели) – это эффективный метод трансферного обучения [5], который использует языковую модель, полученную из больших неразмеченных текстовых корпусов, и применяет ее в качестве основы для других задач. Этот подход оказывается эффективным, даже если целевой текстовый корпус или данные, которые будут использоваться для конкретной задачи, невелики. Этот метод состоит из трех этапов: 1) фаза предварительного обучения языковой модели или предварительное обучение языковой LSTM-модели на большом неразмеченном текстовом корпусе; 2) фаза точной настройки языковой модели или использование предварительно обученной языковой модели в качестве отправной точки для обучения модели на целевом текстовом корпусе; 3) фаза тонкой настройки классификатора текста или обучение точно настроенной модели в задаче классификации текста.

Общая схема применения подхода ULMFiT к обучению языковой модели представлена на рисунке 1.

Используемая LSTM-модель обучается в задаче моделирования языка, где на основе последовательности слов осуществляется прогноз о том, какое слово имеет наибольшую вероятность. На этапе предварительного обучения языковой модели необходимы большие неразмеченные текстовые корпусы, чтобы изучить шаблоны и структуры языка, используемые в них. Идеальный корпус текста должен быть большим, разнообразным и отражать общие свойства языка. Используемые данные для обучения – WikiText-TL-39 [1], взятые из статей тагальской Википедии. Обучающий, валидационный и тестовый наборы были объединены, и 10 % данных были случайным образом взяты в качестве набора для тестирования, а оставшиеся 90 % использовались в качестве обучающего набора. Текстовые данные прошли предварительную обработку перед использованием для обучения. Были использованы только 60 000 слов, которые чаще всего встречаются в данных. Модель обучалась в течение 20 эпох со скоростью

обучения 0,01, размером мини-выборки в 128 экземпляров и коэффициентом дропаута равным 0,5, используемым для регуляризации нейронной сети по предотвращению переобучения. Весь процесс обучения занял 26 часов.

Рисунок 1 – Общая схема применения подхода ULMFiT[1]

Используя предварительно обученную модель на первом этапе, модель дополнительно обучается на целевом текстовом корпусе, чтобы адаптироваться к языку, любым шаблонам и словарному запасу. Целевым корпусом текстов, использованным в этой статье, является набор данных о «языке ненависти» [6]. Набор данных разделен на обучающий, валидационный и тестовый. При тонкой настройке модели последний слой модели сначала обучается в течение одной эпохи со скоростью обучения 0,04. После этого все слои модели обучаются в течение 7 эпох со скоростью обучения 0,004.

Используя точно настроенную языковую модель второго этапа, для задачи классификации текста были добавлены дополнительные слои [5]. Модель была повторно обучена на целевом наборе данных «язык ненависти», но не на задаче языкового моделирования, а на задаче классификации текста. Здесь в текст включены метки (0 = нет ненависти, 1 = ненависть). Коэффициент дропаута равен 0,3, регуляризация – 0,1, а импульс – (0,8; 0,7; 0,6). Были использованы методы тонкой настройки постепенного размораживания и различительной скорости обучения [5]. В таблице 1 показан весь процесс тонкой настройки.

Таблица 1 – Набор гиперпараметров, используемых для тонкой настройки с постепенным размораживанием и избирательной скоростью обучения, равной 0,05[2]

Слои	Эпохи	Минимальная скорость обучения	Максимальная скорость обучения
последний слой	4	lr / 25	lr
последние два слоя	2	lr / ($2{,}6^4$)	lr
последние три слоя	2	lr / 2 / ($2{,}6^4$)	lr / 2
все слои	1	lr / 10 / ($2{,}6^4$)	lr / 10

Тест на деградацию [1] – это метод измерения устойчивости модели к снижению производительности при сокращении обучающих выборок. О снижении производительности сообщается в виде процентного снижения показателя задачи. Модель с медленной деградацией более эффективна в условиях ограниченных ресурсов. Используя набор данных «язык ненависти», процесс можно разделить на три этапа. В первой настройке будет использоваться весь обучающий набор или десять тысяч выборок. Вторая настройка представляет собой 50 % или пять тысяч обучающих выборок, а последняя – разделение 10 % или одна тысяча обучающих выборок. Весь процесс обучения в каждой настройке будет повторяться пять раз, и будут учитываться средние потери при тестировании и точность. В таблице 2 показаны результаты теста на деградацию.

Таблица 2 – Результаты теста на деградацию в наборе данных о «языке ненависти»[3]

Обучающих экземпляров	Величина ошибки	Доля правильно классифицированных экземпляров	Разница величины ошибки	Разница в доле правильно классифицированных экземпляров	% деградации
10 000	0,5117	76,24 %			

[1] Адаптировано из документации Fastai. Рисунок был изменен, чтобы соответствовать набору данных, используемому в этой статье.
[2] Разработано автором.
[3] Разработано автором.

5000	0,5741	72,23 %	0,0624	–4,01 %	5,26
1000	0,6291	67,64 %	0,1174	–8,6 %	11,28

После точной настройки набора данных по разжиганию ненависти LSTM-модель показала точность 76,84 % на тестовом наборе. Это на 2,08 % эффективнее лучшей базовой модели [1], но это незначительное улучшение.

Производительность предложенной LSTM-модели ниже на 4,01 % при разбиении по пять тысяч экземпляров с ухудшением на 5,26 %. В разбиении до одной тысячи точность падает на 8,6 %, в то время как деградация составляет 11,28 %.

Из этих результатов видно, что предложенная LSTM-модель более эффективна в задаче классификации разжигания ненависти по сравнению с языковыми моделями BERT (англ. Bidirectional Encoder Representations from Transformers – языковая модель, основанная на архитектуре трансформер, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач обработки естественного языка), когда используются десять тысяч обучающих экземпляров. Модель работает хуже, когда речь идет о случаях, когда данные небольшие, например, при обучающих выборках объемом 5000 и 1000. Первоначальный базовый уровень BERT [1] имеет деградацию в среднем на 3,28 % при разделении по 5000. Наихудшая модель, или самая быстрая деградация – это DistilBERT из работы [9] с деградацией 4,34 % при разделении 5000. Между тем, предложенная LSTM-модель имеет ухудшение на 5,26 % при разделении 5000. Здесь видно, что модели на основе трансформеров, такие как BERT и DistilBERT, более эффективны, когда речь идет о задачах с низким уровнем ресурсов. Это связано с тем, что модели на основе трансформеров на самом деле предназначены для извлечения более глубоких закономерностей из данных. Преимущество предложенной LSTM-модели заключается в том, что ее обучение происходит быстрее, особенно когда предварительно обученная языковая модель создается с нуля, и это можно сделать только с одним графическим процессором, в отличие от моделей на основе трансформера, в которых необходимо использовать тензорный процессор.

В таблице 3 приведены примеры случаев максимальных потерь, или прогнозы модели с высоким уровнем достоверности, которые оказываются неверными.

Таблица 3 – Результаты модели для потерь. Целевые метки и прогнозы[4]

Пример	Текст	Целевая метка	Предсказание модели
1	Mar Roxas: Kung nagpasok ka ng contra band sa airport, pa'no nagging problema ng gobyerno 'yun? Me: *facepalm* #LaglagBala	Ненависть	Нет ненависти
2	Santiago: Roxas is not on 'Daang Matuw id'. It's very dissapointing	Нет ненависти	Ненависть

Метод письма, используемый в примере 1 таблицы 3, является высококонтекстным, его значение зависит от контекста разговора, а не от буквального значения используемых слов. Другая возможная причина ошибки – использование в тексте редких слов или выражений, например, *facepalm*. Поскольку они используются редко, это означает, что они не часто встречаются в наборе данных. Соответственно модель не может полностью изучить закономерности и связи редких слов или выражений.

В примере 2 слово "dissapointing" (разочаровывает) оказывает наибольшее влияние на прогноз модели. Это редкое слово в наборе данных и оно часто используется в негативном контексте. Поскольку в обучающем наборе было всего три вхождения слова «разочарование», то все они были помечены как «ненависть». Именно по этой причине слово «разочаровывает» часто ассоциируется с меткой «ненависть», и это обстоятельство сильно влияет на предсказание модели.

[4] Разработано автором.

Заключение

В целом эффективность модели низкая, если используемые слова не являются распространенными. Для решения этой проблемы необходим более крупный и разнообразный набор данных. Одним из способов увеличения обучающего набора данных является генерация дополнительных примеров путем аугментации существующих текстов.

Чтобы получить больше шаблонов, используя модели на основе трансформеров, можно применить LSTM-модель для решения простой задачи классификации текста, чтобы повысить производительность и улучшить данные обучения. Рекомендуется использовать модели на основе трансформеров, а также несколько доступных предварительно обученных языковых моделей. Чтобы с нуля использовать предварительно обученную языковую модель и решить задачу с высокой вычислительной мощностью, также можно использовать предложенную LSTM-модель для более быстрого предварительного обучения в моделях на основе трансформера.

Список литературы

1. *Cruz J., Cheng C.* 2020. Establishing Baselines for Text Classification in Low-Resource Languages. arXiv preprint arXiv:2005.02068.
2. *Nguyen D., Nguyen A.* 2020. PhoBERT: Pre-trained language models for Vietnamese. arxiv preprint arXiv:2003.00744.
3. *Howard J., Ruder S.* Universal language model fine-tuning for text classification // In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – Melbourne, 2018. – Vol. Long Papers. – P. 328–339.
4. *Li J., Huang S., Zhang X., Fu X., Chang C.-C., Tang Z., Luo Z.* Facial Expression Recognition by Transfer Learning for Small Datasets. Part of the Advances in Intelligent Systems and Computing book series. – AISC, 2019. – Vol. 895.
5. *Micikevicius P.* et al. 2017. Mixed Precision Training. arXiv preprint arXiv:1710.03740.
6. *Cabasag N., Chan V., Lim S., Gonzales M., Cheng C.* Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing // Philippine Computing Journal. – 2019. – XIV, No. 1.
7. *Nguyen D., Nguyen A.* 2020. PhoBERT: Pre-trained language models for Vietnamese. arxiv preprint arXiv:2003.00744.
8. *Smith L., Topin N.* 2017. SuperConvergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv preprint arXiv:1708.07120.
9. *Sanh V.* et al. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

References

1. *Cruz J., Cheng C.* 2020. Establishing Baselines for Text Classification in Low-Resource Languages. arXiv preprint arXiv:2005.02068.
2. *Nguyen D., Nguyen A.* 2020. PhoBERT: Pre-trained language models for Vietnamese. arxiv preprint arXiv:2003.00744.
3. *Howard J., Ruder S.* Universal language model fine-tuning for text classification // In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. – Melbourne, 2018. – Vol. Long Papers. – P. 328–339.
4. *Li J., Huang S., Zhang X., Fu X., Chang C.-C., Tang Z., Luo Z.* Facial Expression Recognition by Transfer Learning for Small Datasets. Part of the Advances in Intelligent Systems and Computing book series. – AISC, 2019. – Vol. 895.
5. *Micikevicius P.* et al. 2017. Mixed Precision Training. arXiv preprint arXiv:1710.03740.
6. *Cabasag N., Chan V., Lim S., Gonzales M., Cheng C.* Hate speech in philippine election-related tweets: Automatic detection and classification using natural language processing // Philippine Computing Journal. – 2019. – XIV, No. 1.

7. *Nguyen D., Nguyen A.* 2020. PhoBERT: Pre-trained language models for Vietnamese. arxiv preprint arXiv:2003.00744.
8. *Smith L., Topin N.* 2017. SuperConvergence: Very Fast Training of Neural Networks Using Large Learning Rates. arXiv preprint arXiv:1708.07120.
9. *Sanh V.* et al. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.