Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет Санкт-Петербургская школа физико-математических и компьютерных наук
Департамент информатики

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему

«Методы оценки сложности текстовых данных для ускорения обучения языковых моделей с помощью обучения по плану»

Направление подготовки 01.03.02 Прикладная математика и информатика

Выполнил студент группы БПМ171С, 4 курса,
образовательной программы «Прикладная математика и информатика»
Сурков Максим Константинович

Научный руководитель:
? Ямщиков Иван Павлович

Санкт-Петербург
2021

Оглавление

Аннотация

Современные системы обработки естественного языка активно используют глубокие нейронные сети (BERT, GPT-3), которые требуют значительных ресурсов для их обучения. За последние годы было разработано множество подходов для решения данной проблемы. Одним из них является обучение по плану. Данный метод состоит из двух составляющих: оценки сложности тренировочных данных и алгоритма их семплирования. Основной целью данной работы является исследование метрик оценки сложности текстовых данных в контексте обучения по плану и влияния данного метода ускорения обучения на скорость сходимости языковых моделей на задачах предобучения и классификации. В процессе исследований было предложено и адаптировано несколько подходов из разных областей математики. Также были реализованы производительные алгоритмы вычисления найденных метрик на больших объемах данных в несколько десятков миллинов тренировочных примеров. Объемный набор экспериментов на задачах предобучения и классификации выявил наиболее эффективные метрики для использования в обучении по плану. В то же время было установлено, что обучение по плану негативно влияет на скорость сходимости на задаче предобучения, однако не уступает базовому подходу (обучению без плана) на задаче классификации. Также был рассмотрен важный частный случай обучения языковой модели на шумном наборе тренировочных данных. Сравнительный анализ показал ускорение обучения до двух раз на первых 10% обучения при применении обучения по плану с наиболее эффективной метрикой.

Ключевые слова: обработка естественного языка, обучение по плану, теория информации, оценка сложности текстовых данных

Modern state-of-the-art natural language processing systems use deep neural networks (BERT, GPT-3) that require many resources for training. Several techniques have been developed for the last ten years. One of them is curriculum learning, which consists of two parts, namely data complexity evaluation and sampling. The main purpose of this work is to research metrics of text complexity in the context of curriculum learning and explore the influence of curriculum learning on training time on pre-training and classification tasks. Several approaches from different mathematics fields were suggested and adapted during the research. Moreover, efficient algorithms for calculating given metrics on large datasets of several tens of millions of samples were implemented. Extensive experiments highlighted the most efficient metrics for use in curriculum learning. At the same time, it was established that curriculum learning negatively affects convergence time on pre-training task, but not inferior to the basic solution (learning without curriculum) on the classification task. Also, training on a noisy training dataset was considered. Comparative analysis showed a double reduction in training time on the first 10% of training using curriculum learning with the most effective metric.

Keywords: natural language processing, curriculum learning, information theory, text complexity estimation

Введение

На сегодняшний день существует множество сфер, где активно применяется обработка естественного языка. Например, в разработке голосовых помощников, алгоритмов фильтрации текста и машинного перевода. Возникающие задачи необходимо решать эффективно с точки зрения качества модели и скорости работы системы. За основу многих подходов взят механизм внимания [1]. На его базе были разработаны модели, такие как BERT [2], GPT-3 [10] и многие другие. Данные сети имеют высокое качество, однако, за это приходится платить существенным временем обучения. В рамках данной работы исследуется влияние обучения по плану на примере тренировки модели BERT, так как она является одной из самых популярных моделей, имеет высокую точность и сравнительно небольшой размер для удобства постановки экспериментов. Также стоит отметить, что для обучения модели используют объемные корпуса данных, состоящие из нескольких десятков миллионов примеров, для которых нужны производительные алгоритмы их обработки.

Таблица 1: Количество примеров в тренировочных корпусах данных

Корпус данных	Размер
Wikipedia	3-600M
BookCorpus	74M
Hyperpartisan News Detection	600k-2M
sentiment140	1.6M
IWSLT	200-230k
QQP	364k
MNLI	393k

Процесс тренировки модели состоит из двух основных частей. Первая заключается в предобучении сети на задаче Masked Language Modelling [2], которая состоит в восстановлении предложения после замены 15% случайных слов на пробелы. Предобучение занимает несколько недель исполнения кода на дорогостоящих графических процессорах. Второй этап представляет из себя задачу дообучения языковой модели, например на задачу классификации, и требует несколько дней даже на элементарных задачах, таких как определение спама или грубой речи [11]. Одним из методов ускорения обучения моделей является обучение по плану [4]. При его применении данные сортируются по сложности, а затем семплируются с помощью заранее определенного алгоритма, который учитывает порядок данных. Данный подход хорошо себя показал во многих областях машинного обучения [6, 7, 12], однако в обработке естественного языка существует лишь ограниченное число успешных работ [3, 5]. Более того, на данный момент нет исследований влияния обучения по плану на скорость сходимости модели на задачах предобучения и классификации. Также в существующих работах

авторы уделяют большое внимание алгоритмам семплирования данных, а метрики берут из достаточного узкого множества, которое можно значительно расширить, применив различные сферы компьютерных наук. Это позволяет обозначить широкое поле для исследований подходов к оценке сложности текстов и предположить, что существует метрика, которая позволит значительно ускорить обучение модели на вышеуказанных задачах. Однако, в последнее время стали появляться работы с отрицательными результатами применения обучения по плану на задачах компьютерного зрения [13], что показывает спорную репутацию данного подхода к ускорению. В то же время, в тех же статьях авторы находят частные случаи применения обучения по плану на практике. Таким образом, можно выделить обширную сферу исследований нетривиального вопроса применимости обучения по плану к ускорению тренировочного процесса языковых моделей на задачах предобучения и классификации.

Цель и задачи

Целью данной работы является ускорение обучения языковой модели BERT с помощью обучения по плану за счет применения улучшенной метрики сложности текстовых данных на задачах классификации и предобучения

- Предложить метрики оценки сложности текста
- Реализовать производительные алгоритмы вычисления предложенных метрик на больших корпусах данных
- Сравнить найденные метрики
- Исследовать влияние найденных метрик на скорость обучения языковой модели BERT на чистых и шумных тренировочных данных

Достигнутые результаты

Структура работы

1. Обзор литературы

1.1. Возникновение обучения по плану

Точная дата возникновения обучения по плану не известна, но можно выделить логическое начало в работе Bengio 2009 года [4], в которой было показано, что обучение по плану может привести к улучшению качества моделей машинного обучения. Авторы поставили несколько экспериментов, одним из которых является опыт по обучению классификатора геометрических фигур. Было обнаружено, что если сначала предъявить модели более простые примеры (квадраты, круги, равнобедренные треугольники) перед стандартным алгоритмом обучения, то итоговое качество возрастет. Этот простой пример подчеркивает большой потенциал обучения по плану к улучшению существующих алгоритмов в машинном обучении.

1.2. Применение обучения по плану в смежных сферах машинного обучения

Обучение по плану активно применялось в разных областях машинного обучения в течение последних нескольких лет. Например, Hacohen и Weinshall в 2019 году [7] применили данный метод к задачам компьютерного зрения. Они предложили модельную метрику оценки сложности картинок, которая считается следующим образом. Рассматривается независимая модель, предобученная на датасете ImageNet. Далее сложность примера определяется как уверенность модели в своем предсказании. Наконец, ученые использовали лестничный алгоритм семплирования в паре с предложенной метрикой. В результате был получен прирост в скорости обучения и в качестве итоговой модели.

Обуение по плану применяется в и классическом глубоком обучении. Mermer и др. [12] предложили способ автоматической оценки сложности векторных данных для решения задачи классификации. Для этого авторы для каждого примера строят распределение вероятностей классов двумя способами.

1. на базе k ближайших соседей. Рассмотрим мультимножество меток соседей, тогда вероятность i-го класса равна доле соседей с данной меткой

2. на базе ансамбля. Обучим k независимых классификаторов стандартным алгоритмом. Далее для каждой метки определим среднюю предсказанную классификаторами вероятность того, что данный пример имеет рассматриваемую метку

В итоге сложность примера вычисляется как энтропия построенного распределения. Авторы рассмотрели 36 датасетов, на многих из которых обучение по плану выиграло у стандартного алгоритма обучения. Заметим, что подход, основанный на метках соседей, невозможно применить к текстам в явном виде, так как примеры из естественного языка не имеют векторной структуры. Однако можно рассмотреть пространство эмбеддингов. Но данный способ будет зависеть от метода получения векторного представления текстов. Более того, размеры датасетов обучения языковых моделей состоят из большого числа примеров, и поиск k ближайших соседей может требовать большого количества времени, что недопустимо при решении задачи ускорения обучения. Те же самые проблемы возникают и при применения подхода, основанного на ансамблировании.

Важно отметить, что при применении обучения по плану можно получить и отрицательный результат. Так Wu и др. [13] выявили негативное влияние обучения по плану на скорость обучения широкого множества глубоких нейронный сетей на задаче классификации картинок. Авторы применили данный подход к более чем сотне архитектур, среди которых ResNet и VGG-19. Авторы рассмотрели несколько метрик сложности данных, которые сильно коррелировали с величиной $s(x_i, y_i)$ последней эпохи t , после которой модель w правильно отвечает на пример (x_i, y_i) вплоть до последней эпохи T (предсказание модели $\hat{y}_w(t)$ совпадает с реальной меткой y_i примера), для которого считается сложность (формула (1)).

$$s(x_i, y_i) = \min_{t^*}\{\hat{y}_w(t)_i = y_i, \forall t^* \leq t \leq T\} \tag{1}$$

Было использовано семейство "префиксных" семплеров (рис. 1), а именно возрастающих функций, которые определяют долю легких примеров $g(t)$ в конкретный момент обучения t. Таким образом, данные семлперы в момент времени t строят батч данных, равноверятно выбираея примеры из первых $g(t)$ семплов отсортированного по метрике сложности датасета. В результате, ученые показали, что при использовании данной конфигурации обучения по плану не приводит ни к улучшению качества итоговой модели, ни к ускорению обучения. Также было рассмотрено два важных частных случая, а именно обучение на шумных данных и обучение с ограниченным числом тренировочных шагов. При добавлении 20% шума в тренировочный корпус, обучение по плану позволяет улучшить точность модели на 10%. Влияние же подхода на скорость обучения исследовано не было.

1.3. Применение обучения по плану в обработке естественного языка

В обработке естественного языка существует ограниченное число существенных

Название	Формула $g_{(a,b)}(t)$
log	$Nb + N(1-b)\left(1 + .1\log\left(\frac{t}{aT} + e^{-10}\right)\right)$
exp	$Nb + \frac{N(1-b)}{e^{10}-1}\left(\exp\left(\frac{10t}{aT}\right) - 1\right)$
step	$Nb + N\left\lceil\frac{x}{aT}\right\rceil$
polynomial	$Nb + N\frac{(1-b)}{(aT)^p}t^p \quad - \; p = 1/2, 1, 2$

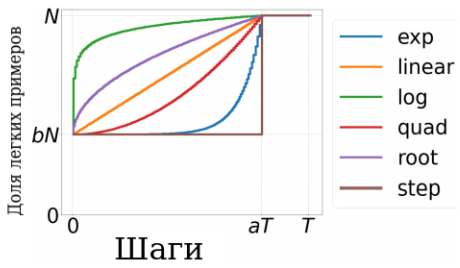

Рис. 1: Семейство функций для семплирования данных

результатов в контексте применения обучения по плану. Вероятно, это связано с тем, что естественный язык состоит из слов и предложений, не имеющих строгую структуру, которую сложно формально описать. Более того, на данный момент наука не понимает всех процессов, происходящих внутри современных языковых моделей.

Platanios и др. [3] исследовали влияние обучения по плану на задаче машинного перевода на скосрость сходимости нейронных сетей. Авторы рассмотрели две метрики сложности текстов: длину и вероятность правдоподобия, которая вычисляется по формуле (2), где s_i – текст или набор токенов, w_k^i – слово или токен, $p(x)$ – доля токенов x во всем датасете.

$$d(s_i) = -\sum_{k=1}^{N} \log p(w_k^i) \tag{2}$$

Они показали, что правдоподобие не имеет никаких преимуществ по сравнению с длиной с точки зрения скорости обучения моделей. В качестве семплеров был взят префиксный семплер с функцией $c(t)$ (формула (3)), вычисляющий долю простых примеров, доступных для построения батча.

$$c(t) = \min\left(1, \sqrt{t\frac{1-c_0^2}{T} + c_0^2}\right) \tag{3}$$

T – общее число тренировочных шагов, c_0 – доля простых примеров, доступных в самом начале обучения (авторы используют $c_0 = 0.01$)

В итоге, ученые добились улучшения качества модели на 2.2 BLEU и ускорения обучения на 70%.

Xu и др. [5] предложили альтернативный способ применения обучения по плану в обработке естественного языка на задаче Natural Language Understanding. Их метод оперирует понятием модельной оценки сложности текстов. Сложность примеров меняется в процессе обучения в зависимости от качества модели на момент применения метрики к примеру. Авторы предлагают алгоритм, который в цикле повторяет следующую процедуру. Тренировочный корпус данных разделяется на несколько частей. Затем, для каждого блока независимо обучается новая модель, которая инициализируется весами текущей глобальной модели. После этого оценивается сложность всех

примеров как сумма уверенностей всех моделей по всем блокам кроме блока, в котором находится данный пример (рис. 2).

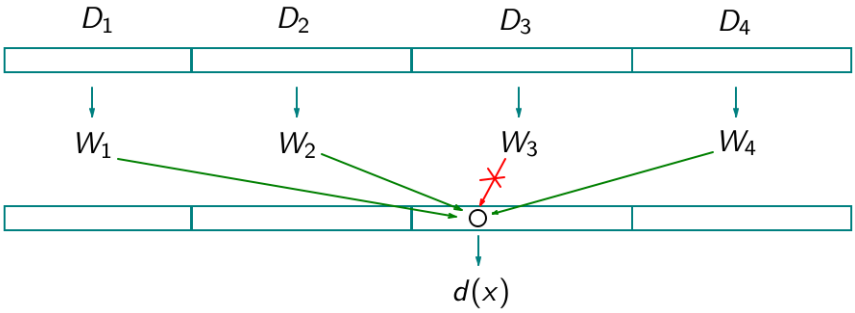

Рис. 2: Алгоритм вычисления модельной метрики сложности текста. На данной схеме датасет делится на 4 части, на каждой из которых учится независимая модель BERT с начальными весами текущей глобальной модели. Сложность примеров в блоке под номером 3 вычисляется как сумма уверенностей моделей W_1, W_2, W_4.

Наконец, весь датасет сортируется в соответствии с найденными сложностями примеров, и текущая глобальная модель обучается на новой эпохе, обрабатывая примеры в порядке возрастания сложности. Данный подход позволяет улучшить точность итоговой модели на 1.5%, но требует существенно больше времени.

В 2017 году Kocmi и др. [8] провели сравнтиельный анализ двух подходов к ускорению обучения: minibatch bucketing и обучение по плану. Первый подход отличается от стандартного алгоритма обучения нейронных сетей только построением батчей, каждый из которых содержит данные, слабо отличающиеся по заранее определенному показателю (например, батч из предложений с не более чем пятью глаголами). Для обучения по плану авторы используют три метрики.

1. длина предложения

2. максимальный частотный ранг слова – для каждого слова вычисляется количество его вхождений в тренировочный корпус данных. Затем все слова сортируются в порядке убываения частоты. Рангом слова называется его позиция в данном отсортированном массиве. Сложность предложения определяется как максимальный ранг по всем словами в данном предложении

3. количество конъюнкций (например, эту метрику можно определить как количество союзов)

Исследование показало, что обучение по плану позволяет увеличить качество итоговой модели на 1 BLEU, но без уменьшения времени обучения. Более того, модель, обучаемая с помощью обучения по плану достигает 85%-го качества модели, обученной базовым алгоритмом, в два раза медленнее.

Важный вклад в исследование обучения по плану внесли Zhang и др. [14] в 2018 году. В качестве метрик оценки сложности текста авторы рассматривают модельную метрику, длину, максимальный частотный ранг и средний частотный ранг. В данной работе ранг слова определяется аналогично работе Kocmi и др. [8]. Модельная метрика определяется ошибкой вспомогательной модели, заранее обученной на задачу машинного перевода стандартным алгоритмом. Принципиальным отличием данной работы является выбор методов построения батчей. Ученые не рассматривают префиксные семплеры, а используют алгоритмы, с течением времени изменяющие распределение вероятности взять пример в текущий батч. Авторы показали, что обучение по плану очень чувствительно к выбору гиперпараметров. Среди нескольких десятков конфигураций лишь некоторые позволяют получить ускорение обучения модели до 30% без потери точности. Также было установлено, что длина не является удачной метрикой для обучения по плану, а именно она приводит к замедлению обучения модели до двух раз и уменьшению точности модели до 4.2 BLEU.

1.4. Существующие метрики оценки сложности текстов

Таким образом, был рассмотрен ряд метрик, активно используемых в обучении по плану: длина [3, 8, 14], вероятность правдоподобия [3], модельная [5], максимальный [8, 14] и средний [14] частотный ранг.

На первый взгляд кажется, что можно придумать еще несчетное количество подходов. Действительно, существует большое количество метрик оценки сложности текстов. Это показал Kurdi в своей статье 2020 года [9], в которой решал задачу определения уровня английского языка, необходимого для прочтения текста. Для этого он строил множество признаков входного текста для их передачи на вход классификатору. Ученый рассмотрел несколько десятков методов определения сложности текста понятных для человека: фонологические, морфологические, лексические, синтаксические признаки и многие другие. Это позволило решить задачу с высокой точностью. Несмотря на хорошее качество полученной модели, вопрос применимости данных метрик к обучению по плану остается открытым. Однако на него можно найти ответ в работе Frans van der Sluis и Egon L. van den Broek [16]. Они рассмотрели два набора данных, Wikipedia и Simple English Wikipedia (упрощенная версия Wikipedia, состоит из статей меньшей длины, написанных более простым языком), и показали что лингвистические метрики плохо коррелируют с реальной сложной текстов. Таким образом, классические методы оценки сложности текстовых данных имеют меньший приоритет для рассмотрения в сранении с метриками, основанными на статистических подходах.

Оценка сложности текстов тесно связана с количеством информации в них. Этот

объкт изучает теория информации. В 2006 году Ау и др. [15] предложили четыре метода оценки сложности конечных дискретных систем. К сожалению, в чистом виде данные подходы невозможно применить к текстам, так как они (подходы) получают на вход некоторую совместно распределенную случайную величину. Однако эти метрики можно адаптировать, о чем будет рассказано в последующих главах.

1.5. Выводы

- Нет существующих исследований влияния обучения по плану на скорость сходимости на не менее важных задачах предобучения языквоых моделей и классификаии текстов
- Рассмотрено узкое множество метрик оценки сложности текстов, которое можно расширить, используя методы из смежных областей математики и информатики, таких как теория информации и информационный поиск
- Большинство работ применяют обучение по плану для улучшения качества модели, но не скорости ее обучения
- Можно заметить, что во многих работах обучение по плану приводит к уменьшению тренировочного времени и увеличению точности моделей только в определенных конфигурациях, которые сильно зависят от гиперпараметров, задачи, модели и корпуса данных. Важной деталью является тот факт, что все эксперименты были проведены на чистых наборах данных. Такая ситуация редко встречается при решении прикладных задач и реализации реальных проектов. Например, крупные компании тратят большие деньги для очистки тренировочных данных. Таким образом, важность исследования обучения языковых моделей на шумных корпусах данных очевидна, однако работ, освещяющих данный вопрос найдено не было

2. Глава1

3. Глава2

4. Глава3

5. Глава4

6. Заключение

Список литературы

[1] Attention is all you need / Ashish Vaswani, Noam Shazeer, Niki Parmar et al. // arXiv preprint arXiv:1706.03762. — 2017.

[2] Bert: Pre-training of deep bidirectional transformers for language understanding / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. — 2018.

[3] Competence-based curriculum learning for neural machine translation / Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig et al. // arXiv preprint arXiv:1903.09848. — 2019.

[4] Curriculum learning / Yoshua Bengio, Jérôme Louradour, Ronan Collobert, Jason Weston // Proceedings of the 26th annual international conference on machine learning. — 2009. — P. 41–48.

[5] Curriculum learning for natural language understanding / Benfeng Xu, Licheng Zhang, Zhendong Mao et al. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. — 2020. — P. 6095–6104.

[6] Curriculum learning for reinforcement learning domains: A framework and survey / Sanmit Narvekar, Bei Peng, Matteo Leonetti et al. // Journal of Machine Learning Research. — 2020. — Vol. 21, no. 181. — P. 1–50.

[7] Hacohen Guy, Weinshall Daphna. On the power of curriculum learning in training deep networks // International Conference on Machine Learning / PMLR. — 2019. — P. 2535–2544.

[8] Kocmi Tom, Bojar Ondrej. Curriculum learning and minibatch bucketing in neural machine translation // arXiv preprint arXiv:1707.09533. — 2017.

[9] Kurdi M Zakaria. Text Complexity Classification Based on Linguistic Information: Application to Intelligent Tutoring of ESL // arXiv preprint arXiv:2001.01863. — 2020.

[10] Language models are few-shot learners / Tom B Brown, Benjamin Mann, Nick Ryder et al. // arXiv preprint arXiv:2005.14165. — 2020.

[11] MITRE at SemEval-2019 task 5: Transfer learning for multilingual hate speech detection / Abigail S Gertner, John Henderson, Elizabeth Merkhofer et al. // Proceedings of the 13th International Workshop on Semantic Evaluation. — 2019. — P. 453–459.

[12] Mermer Melike Nur, Amasyali Mehmet Fatih. Scalable Curriculum Learning for Artificial Neural Networks // IPSI BGD TRANSACTIONS ON INTERNET RESEARCH. — 2017. — Vol. 13, no. 2.

[13] Wu Xiaoxia, Dyer Ethan, Neyshabur Behnam. When Do Curricula Work? // arXiv preprint arXiv:2012.03107. — 2020.

[14] An empirical exploration of curriculum learning for neural machine translation / Xuan Zhang, Gaurav Kumar, Huda Khayrallah et al. // arXiv preprint arXiv:1811.00739. — 2018.

[15] A unifying framework for complexity measures of finite systems / Nihat Ay, Eckehard Olbrich, Nils Bertschinger, Jürgen Jost // Proceedings of ECCS / Citeseer. — Vol. 6. — 2006.

[16] van der Sluis Frans, van den Broek Egon L. Using complexity measures in information retrieval // Proceedings of the third symposium on information interaction in context. — 2010. — P. 383–388.