Отчёт о проделанной работе

Тема работы. Безопасное предобучение глубоких языковых моделей на синтетическом псевдоязыке

Состав коллектива:

- Горбачева Таисия Евгеньевна, t.gorbacheva@g.nsu.ru , студентка НГУ
- Бондаренко Иван Юрьевич, старший преподаватель ФиПЛ, bond005@yandex.ru, научный руководитель

Аннотация. На сегодняшний день нейросети крайне популярны и применяются во многих областях нашей жизни, в том числе для решения задач обработки естественного языка (NLP). Как известно, нейросети работают по принципу чёрного ящика, так как нельзя однозначно понять, как система принимает решение и, следовательно, гарантировать, какой итог будет после обучения. Соответственно, вопрос того, как обеспечить «безопасность» нейросети, является актуальным. Мы предположили, что одним из методов решения проблемы можно считать обучение нейросети на датасете, о котором полностью известно, какой он, то есть на том, который будет полностью создан нами. Для автоматического создания текста был выбран такой способ генерирования, как контекстно-свободная грамматика. Для проведения сравнения и оценки эффективности мы дважды предобучили трансформер RoBERTa: на сгенерированных предложениях и на выборке фраз естественного языка, которая также была подготовлена нами. Результаты тестирования показали, что модели имеют одинаковые оценки, то есть использование данных, автоматически созданных при помощи правил контекстно-свободной грамматики, даёт преимущество для «безопасности» искусственного интеллекта за счет того, что мы можем полностью контролировать состав выборки. Поскольку синтетические данные не уступают в качестве естественным, мы можем говорить о том, что на этапе предобучения модели типа RoBERTa действительно достаточно научиться распознавать только синтаксические и морфологические закономерности языка, которые могут быть успешно созданы довольно таким простым способом, как контекстно-свободная грамматика.

Научное содержание работы:

Постановка задачи. Целью данной работы является исследование того, насколько эффективно может быть безопасное предобучение нейросети на текстах синтетического псевдоязыка для решения различных вопросов обработки естественных языков. Для достижения цели необходимо предобучить две модели RoBERTa [1]: одну на сгенерированном нами датасете искусственных текстов и вторую на собранной нами выборке предложений естественного языка.

Современное состояние проблемы (на момент начала работы). На сегодняшний день нейросети крайне популярны и применяются во многих областях нашей жизни, в том числе для решения задач обработки естественного языка (NLP). Как известно, нейросети работают по принципу чёрного ящика, так как нельзя однозначно понять, как система принимает решение и, следовательно, гарантировать, какой итог будет после обучения. Соответственно, вопрос того, как обеспечить «безопасность» нейросети, является актуальным [2]. Следует пояснить, что проблема «безопасности» заключается в гарантировании надежности именно большого датасета, который используется для предварительного обучения, а не маленького, который потом применяется для донастроек, поскольку его можно проверять и вручную. Мы предполагаем, что одним из методов решения можно считать обучение нейросети на датасете, о котором полностью известно, какой он, то есть на том, который будет полностью создан нами. Для компьютерного зрения подход с использованием в предобучении синтетических данных уже применяется [3], однако, об использовании в NLP нам неизвестно.

Подробное описание работы, включая используемые алгоритмы.

Для предобучения моделей нами были подготовлены два набора данных. Первый представляет собой автоматически сгенерированные предложения по написанным нами правилам контекстно-свободной грамматики. При написании правил грамматики учитывались морфологические и синтаксические особенности русского языка, а также специфика модуля для генерации. Используя в качестве основы корпус «Тайга» [4] и тезаурус «RuWordNet» [5], мы собрали и почистили датасет из 2 477 009 уникальных слов часто с указанием их морфологической информации, эти данные в дальнейшем использовались для генерации. Таким образом, было сгенерировано два миллиона предложений, отличающихся разнообразной синтаксической и морфологической структурой. Второй набор данных представлял собой два миллиона предложений естественного языка, взятые нами из следующих источников: три новостных под-корпуса «Тайги» (Lenta.ru, Интерфакс, N+1) и тексты русской Википедии [6].

Предварительное обучение обеих моделей происходило аналогично. Для удобства введём следующие обозначения: NaturalRoBERTa — трансформер RoBERTa, предобученный на предложениях естественного языка; SyntheticRoBERTa — трансформер RoBERTa, предобученный на предложениях синтетического псевдоязыка. При написании нейросети мы оперировали библиотеками Transformers от компании Hugging Face [7], bitsandbytes [8] и PyTorch [9] для языка программирования Python. В первую очередь необходимо было обучить токенизатор. Его цель — подготовить входные данные для модели путём разбиения текста на токены и кодирования этих токенов в целые числа. Мы выбрали токенизатор байтового уровня Byte-level byte pair encoding (BBPE) из библиотеки Transformers [7]. Он разбивает слова на минимальные компоненты и объединяет наиболее часто встречающиеся. Самые распространенные слова будут представлены в словаре как один токен, в то время как менее распространенные — разбиты на два или более токенов подслов. После получения файлов токенизатора была определена конфигурация модели. Мы выбрали для предварительного обучения базовую модель RoBERTa. Эта модель, в отличие от аналогичного BERT, во время предобучения обучается только генерации пропущенного замаскированного токена (BERT также предобучался предсказанию следующего предложения). Это позволило нам в написании правил грамматики ориентироваться только на синтаксические и морфологические правила русского языка и не учитывать семантику. Кроме того, была определена функция DataCollatorForLanguageModeling(), которая должна маскировать слова в данных с вероятностью 15%. В качестве оптизиматора для экономии памяти использовался 8-разрядный оптимизатор Adam из библиотеки bitsandbytes [8]. Предобучение NaturalRoBERTa было завершено на 34 эпохе со значением функции потерь равным 8.7587. SyntheticRoBERTa завершила своё предобучение на 12 эпохе со значением функции потерь 8.6713. На рисунке 1 представлен общий график зависимости значения функции потерь на обучающей и «валидационной» выборках от эпохи обучения моделей: можно видеть, что нейросети демонстрируют практически одинаковый результат во время предварительного обучения.

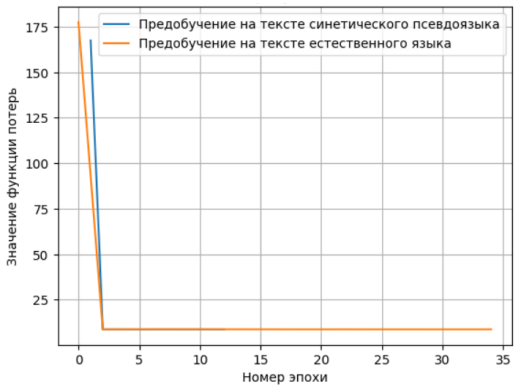

Рисунок 1 – Сравнение значений функции потерь во время предобучения трансформеров на предложениях естественного языка и на предложениях синтетического псевдоязыка

Полученные результаты. При тестировании полученных моделей использовались задания проекта RussianSuperGLUE [10] на понимание текстов. Результаты дообучения показали, что модели имеют одинаковые оценки, то есть использование выборки, сгенерированной при помощи правил контекстно-свободной грамматики, можно считать более надёжной гарантией того, что система будет более «безопасной» за счёт того, что мы абсолютно уверены в составе выборки. Также, поскольку синтетические данные не уступают в качестве естественным, мы можем говорить о том, что на этапе предобучения модели типа RoBERTa действительно достаточно научиться распознавать только синтаксические и морфологические закономерности языка. Помимо этого, использование для генерации текстов такого довольно простого способа, как контекстно-свободная грамматика является удачным, то есть для распознавания сложных закономерностей языка модели достаточно научиться их определять в простых текстах.

Эффект от использования кластера в достижении целей работы. Использование графических ускорителей (GPU) вычислительного кластера ИВЦ НГУ позволило обучить глубокие нейронные сети с механизмом внимания типа Transformer, что является очень сложной задачей, требующей больших вычислительных ресурсов.

Список литературы:

1. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Электронный ресурс] URL: https://arxiv.org/abs/1907.11692 (дата обращения: 7.04.2023).
2. Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. Concrete Problems in AI Safety [Электронный ресурс] URL: https://arxiv.org/abs/1606.06565 (дата обращения: 7.04.2023).
3. Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, Yutaka Satoh. Pre-training without Natural Images [Электронный ресурс] URL: https://openaccess.thecvf.com/content/ACCV2020/papers/Kataoka_Pre-training_without_Natural_Images_ACCV_2020_paper.pdf (дата обращения: 7.04.2023).
4. Shavrina T., Shapovalova O. To the methodology of corpus construction for machine learning: «Taiga» syntax tree corpus and parser // Proceedings of the international conference «Corpus linguistics-2017», 2017. P. 78-84.

5. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Изд-во Московского университета, 2011, 512 с.
6. Википедия — свободная энциклопедия. URL: https://ru.wikipedia.org/ (дата обращения:15.02.2023).
7. Документация библиотеки для машинного обучения Transformers для языка Python URL: https://huggingface.co/docs/transformers/index (дата обращения: 21.05.2023).
8. GitHub - TimDettmers/bitsandbytes: 8-bit CUDA functions for PyTorch. URL: https://github.com/ TimDettmers/bitsandbytes (дата обращения: 21.05.2023).
9. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas K¨opf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, 36 Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. CoRR, abs/1912.01703, 2019
10. Shavrina T., Fenogenova A., Emelyanov A., Shevelev D., Artemova E., Malykh V., Mikhailov V., Tikhonova M., Chertok A., Evlampiev A. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. P. 4717-4726.