УДК 004

Шамало В.В.

магистрант кафедры информатики и вычислительной техники

Кубанский государственный технологический университет

Россия, г. Краснодар

ПОЧЕМУ КЛАССИФИКАТОРЫ ПЛОХО РАБОТАЮТ НА КОРОТКИХ ТЕКСТАХ И ЧТО С ЭТИМ ДЕЛАТЬ

Аннотация: на репликах объёмом 5–15 слов трансформерные модели демонстрируют падение точности, несмотря на высокую эффективность при обработке целого абзаца. Причина кроется в механизме самовнимания (self-attention): он формирует представление слова на основе окружения, которое в коротком тексте крайне ограничено. В статье исследуется данный феномен и анализируются четыре метода его компенсации: привлечение внешнего контекста, обучение по нескольким примерам, преобразование задачи в заполнение пропусков и интеграция признаков иной природы. Приводятся рекомендации по выбору метода.

Ключевые слова: короткие тексты, классификация, самовнимание, обучение на малых выборках, промпт-методы, векторные представления.

Shamalo V.V.

Master's Student

Kuban State Technological University

Russia, Krasnodar

WHY CLASSIFIERS FAIL ON SHORT TEXTS AND WHAT IS DONE ABOUT IT

Abstract: on utterances of 5–15 words transformer models show a drop in accuracy despite performing well on a full paragraph. The cause lies in the self-attention mechanism: it builds a word representation from its surroundings, which are extremely limited in a short text. The paper examines this phenomenon and analyses four methods of compensating for it: external context,

few-shot learning, recasting the task as filling a blank, and integrating features of a different nature. Recommendations are given on choosing a method.
Keywords: short texts, classification, self-attention, few-shot learning, prompting methods, vector representations.

Введение

Основная проблема в том, что модель, уверенно обрабатывающая текст объёмом в десять предложений, показывает заметное снижение качества на реплике из шести слов. Данный разрыв воспроизводится в самых разных задачах: определение тональности, поиск оскорблений, классификация намерений.

При этом короткие тексты составляют основной массив данных в переписке, комментариях и поисковых запросах, поэтому игнорировать их невозможно, а какое-либо усиление модели проблему не решает. В данной статье объясняется, почему проседает принцип самовнимания (self-attention), и рассматриваются способы обхода.

Методы и исследования

Сначала разбирается сам механизм self-attention и причины его сбоев. Затем рассматриваются четыре группы решений, сгруппированных по источнику недостающих данных: внешние базы знаний, размеченные примеры, сама предобученная модель или признаки иной природы. В обзор вошли работы, протестированные на общедоступных датасетах.

Результаты

Откуда берётся провал. Механизм самовнимания строит представление каждого слова, взвешивая его окружение [1]. Чем длиннее текст, тем больше у модели контекстных подсказок. На реплике из шести слов опираться почти не на что: вокруг всего пять соседей, из которых три — служебные. Предобученная модель пытается компенсировать этот дефицит за счёт общих знаний [2], но лишь частично: снимать многозначность слов всё равно нечем. При этом расти в ширину

бесполезно: удвоение числа параметров добавляет эрудиции о языке вообще, а не контекста конкретной фразе.

Первый способ: привлечение внешнего контекста. Наиболее ранний подход заключается в обогащении короткого текста внешней информацией. В первоначальных работах к запросу добавлялись латентные темы, извлечённые из внешнего корпуса [3]. Впоследствии стали использовать базы знаний: в тексте идентифицируется сущность и добавляется её описание, благодаря чему классификатор восполняет контекстуальный дефицит [4]. Метод эффективен в узких предметных областях. Однако в условиях открытой переписки его продуктивность снижается: число сущностей минимально, а уровень информационного шума существенно возрастает.

Второй способ: обучение по нескольким примерам. Большие языковые модели решают задачи в режиме in-context learning, получая в составе запроса несколько аннотированных примеров и не требуя дообучения [5]. Для работы с короткими текстами данный подход экономически эффективен: разметка ста реплик требует существенно меньше ресурсов, чем десяти тысяч. Метод имеет два ключевых ограничения: примеры расходуют лимит контекстного окна, а точность классификации чувствительна к их подбору и последовательности.

Третий способ: преобразование задачи. Вместо применения классификационного слоя задача трансформируется в процедуру заполнения пропусков. Текст дополняется шаблоном вида «Этот комментарий написал ___», после чего оценивается вероятность подстановки целевых токенов [6]. Эффективность метода обусловлена его полным соответствием целевой функции предобучения модели. В последующих исследованиях процесс конструирования промптов был автоматизирован [7], а метаанализ показывает, что при ограниченном

объёме данных данный подход устойчиво превосходит традиционное дообучение [8].

Четвёртый способ: интеграция признаков иной природы. Когда семантики не хватает, на помощь приходит статистика. К нейросетевому эмбеддингу добавляют длину текста, частоту служебных слов и пунктуацию, а затем объединяют оба потока с помощью дополнительного слоя. Эти признаки легко посчитать для любого, даже самого короткого текста, и они не теряют пользы там, где сбоит смысловая часть. К слову, сравнивать короткие тексты удобнее через представления, специально обученные для сопоставления предложений [9].

Что выбрать. В условиях узкой предметной области и наличия базы знаний целесообразно применение первого метода. При ограниченном объёме аннотированных данных и доступе к крупным языковым моделям предпочтительны второй или третий подходы. При интеграции в существующую инфраструктуру оптимален четвёртый способ, не требующий структурных изменений. Общие рекомендации по тонкой настройке (fine-tuning) остаются неизменными: низкая скорость обучения и минимальное количество эпох во избежание быстрого переобучения на малых выборках [10].

Чего не решает ни один из способов. Фразы короче трёх слов недосягаемы для этих методов: контекст не подтянуть, примеры не помогают, статистика вырождается. Практичнее их не классифицировать, а выдавать статус неопределённости. Это защищает от ложных срабатываний, цена которых в модерации выше цены пропуска.

Заключение

Собрано четыре метода компенсации: внешний контекст, few-shot, заполнение пропусков и сторонние признаки. Универсального нет: выбор зависит от нехватки разметки, доменных знаний или ресурсов. Проблемой

остаются фразы короче трёх слов — для них практичнее выдавать статус неопределённости.

Использованные источники:

1. Vaswani A., Shazeer N., Parmar N. [et al.] Attention Is All You Need // Advances in NeurIPS. 2017. Vol. 30.

2. Devlin J., Chang M.W., Lee K. [et al.] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. 2019. P. 4171–4186.

3. Phan X.-H., Nguyen L.-M., Horiguchi S. Learning to Classify Short and Sparse Text and Web with Hidden Topics // Proceedings of WWW. 2008. P. 91–100.

4. Wang J., Wang Z., Zhang D. [et al.] Combining Knowledge with Deep Convolutional Networks for Short Text Classification // Proceedings of IJCAI. 2017. P. 2915–2921.

5. Brown T., Mann B., Ryder N. [et al.] Language Models Are Few-Shot Learners // Advances in NeurIPS. 2020. Vol. 33. P. 1877–1901.

6. Schick T., Schütze H. Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference // Proceedings of EACL. 2021. P. 255–269.

7. Gao T., Fisch A., Chen D. Making Pre-trained Language Models Better Few-shot Learners // Proceedings of ACL-IJCNLP. 2021. P. 3816–3830.

8. Liu P., Yuan W., Fu J. [et al.] Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in NLP // ACM Computing Surveys. 2023. Vol. 55. № 9. Art. 195.

9. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings Using Siamese Networks // Proceedings of EMNLP-IJCNLP. 2019. P. 3982–3992.

10. Sun C., Qiu X., Xu Y. [et al.] How to Fine-Tune BERT for Text Classification? // Proceedings of CCL. 2019. P. 194–206.