УДК 004.942

Адаптация мультимодальных больших языковых моделей с учетом когнитивной сложности текстов

Смирнов И. З. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»),

Научный консультант - кандидат технических наук, ассистент Лаушкина А. А. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Аннотация: Визуально-текстовые мультимодальные модели (VLM) продемонстрировали высокую генеративную способность при обработке визуально-текстовой информации в разных областях знаний. Возможность создания аннотаций изображений позволяет понимать сложные данные и осуществлять навигацию по ним. Однако общего подхода для адаптации VLM к новым областям знаний не существует. Также проблемой при адаптации модели является меньшее количество аннотированных данных и различия в словарях между корпусами. Исследование предлагает возможное решение для более дешевой адаптации VLM при работе со специализированными данными за счет дополнительной регуляризации, учитывающей когнитивную сложность текстов. В частности, в работе рассмотрены методы оценки сложности текстов, основанные, учитывающие словарный состав корпуса. Такой подход к генерации текста на уровне слов позволяет использовать дополнительную информацию о распределении слов в определенной области знаний и уменьшает количество итераций во время дообучения.

Введение. Несмотря на существование VLM моделей, обученных под разные узкоспециализированные задачи, например: GeoChat, EarthGPT, LLaVA-Med, на данный момент нет общего подхода для адаптации моделей под специфические задачи [1]. При этом использование адаптированных моделей позволяет достичь большую точность при решении специализированных задач, снизить затраты, связанные с вычислениями и интеграцией.

При работе со специфическими наборами данных проблемой является относительно небольшое число аннотированных данных. Учет сложности текстов и их словарного разнообразия позволяет обучать модели более эффективно на меньшем количестве данных [2]. Существуют разные подходы для вычисления сложности текстов, в данной работе предполагается рассмотрение тех, что учитывают словарный состав целевого корпуса [3]. Таким образом в работе предлагается использовать регуляризацию, учитывающей когнитивную сложность текстов, при дообучении модели, для более эффективной адаптации генеративной модели на узкоспециализированных областях знаний.

Основная часть. В качестве основного подхода для учета информации о когнитивной сложности текстов предлагается использовать регуляризацию в виде оценки сложности текстов, которая формулируется через метрики, включающие в себя понятия незнакомых, или сложных слов, такие как индекс удобочитаемости Dale-Chall. При этом рассматривать данные слова предлагается на основании статистических данных по словам, встречающимся в целевом корпусе. Так, за знакомые слова возможно принять наиболее частотные слова корпуса. Для этого требуется учитывать частотных слов и ранжировать их, в данной работе для этого используются значения TF-IDF. Однако, проблемой данных индексов, учитывающих сложность текстов, является то, что они не учитывают то, какой сложности тексты встречаются в заданных областях, таким образом при выставлении штрафов при обучении, следует учитывать значения внутри заданного корпуса.

Такой подход позволяет учитывать специфические слова корпуса при обучении в явном виде и повышать уровень пересечения словарей при использовании слов из общих наборов данных.

При тестировании подхода рассмотрены результаты дообучения моделей на разных небольших мультимодальных наборах данных.

Выводы.

В работе предложен подход, направленный на адаптацию больших языковых моделей для работы в узкоспециализированной области знаний. Основная идея подхода заключается в дополнительной регуляризации, основанной на расчете когнитивной сложности генерируемого текста.

Источники:

1. Gao Z. et al. Mini-InternVL: A Flexible-Transfer Pocket Multimodal Model with 5% Parameters and 90% Performance. 2024.
2. Agrawal A., Singh S. Corpus Complexity Matters in Pretraining Language Models // Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2023. P. 257–263.
3. Solovyev V., Ivanov V., Solnyshkina M. Thesaurus-Based Methods for Assessment of Text Complexity in Russian. 2020. P. 152–166.

Смирнов И.З. (автор)

Подпись

Лаушкина А.А. (научный консультант)

Подпись