ТЕСТОВОЕ ЗАДАНИЕ

Direct Preference Optimization: Your Language Model is Secretly a Reward Model

Теоретическая часть

Основная проблема

Современные большие языковые модели (LLMs) обучаются на больших массивах данных и благодаря этому обладают “обширными знаниями” и некоторыми способностями к рассуждению, но их поведение сложно контролировать, например, модели должны знать что из себя представляют баги в коде, но не генерировать их. Один из способов достижения управляемости состоит в сборе и использовании обратной связи от людей об ответах генерируемых моделью, а затем ее дообучения обучением с подкреплением (RLHF). Но подобные методы обладают рядом проблем, некоторые из которых решает Direct Preference Optimization (DPO).

Ключевая идея DPO

Авторы предложили отказаться от использования явной модели для вознаграждений, таким образом, основная модель начинает представлять собой как языковую модель, так и (неявное) вознаграждение. Используя reference модель и примеры хорошего и плохого ответов для запроса, DPO оптимизирует LLM через отношение вероятностей их предсказаний.

Преимущества и недостатки DPO

Преимущества заключаются в простоте в реализации и обучении; требовании меньшего объема вычислительных ресурсов; отсутствии отдельной модели для вознаграждения; большей стабильности относительно температуры. Недостатки касаются зависимости от reference модели, возможности переобучиться на тренировочных данных и “забыть” общее, чему модель училась ранее. Кроме того, неявные rewards ограничивают интерпретируемость, и вычислительные ресурсы все же требуются, особенно при увеличении размеров модели.

Преимущества и недостатки off-policy

Преимущества:

- Возможность использования открытых датасетов без необходимости самостоятельно собирать данные о предпочтениях;
- Открытые датасеты имеют большой объем и (потенциально) являются разнородными, что может позволить модели не недообучаться и сохранить “широкий кругозор”.

Недостатки:

- Модель возможно обучается на данных, фактов в которых она даже раньше не знала, так как вторая (SFT) модель могла обладать более “широким кругозором”;
- Присутствие субъективности в оценках людей, выставляющих предпочтения. Разные люди могли по-разному оценивать одни и те же ответы;
- Есть шанс потерять часть “кругозора” из-за потенциального отсутствия части данных у SFT.

Практическая часть

Ссылка на GitHub репозиторий

Проект включает несколько частей:

1. Подготовка данных
    a. Загрузка датасета Anthropic/hh-rlhf
    b. Деление на тренировочный-тестовый наборы
    c. Фильтрация элементов, содержащих один вопрос и один ответ
    d. Функции токенизации и батчинга
2. Подготовка модели
    a. Загрузка предварительно обученной модели GPT2
    b. Загрузка весов из файла
    c. PEFT через adapters без использования сторонней библиотеки - замена модуля GPT2MPL на дополнительный с линейными слоями и активационной функции
3. Обучение
    a. Функция потерь, как из статьи
    b. Функция вычисления лог вероятностей
    c. Планировщик темпов обучения (Scheduler for learning rate)
    d. Функция для одной итерации тренировки
    e. Функция для одной итерации валидации
    f. Функция проведения обучения с сохранением весов и ранней остановкой
4. Оценка результатов
    a. Сравнение предсказаний дообученной модели с “выбранными” ответами через промпт Llama 2 для выбора лучшего из них (win rate)
    b. Сравнение с GPT2 без дообучения с DPO по методу из пункта a

Результаты:

Количество обучаемых параметров уменьшилось в 2 раза за счет добавления адаптеров в модель. Однако, обучение весов с нуля требует большего времени обучения. Поэтому дообученная модель стала выводить меньше разнообразных ответов и чаще не понимает и пишет, что не понимает. Возможно следует попробовать инициализировать веса, используя веса предыдущих дообученных слоев. На ресурсах Google Collab получается запускать 40 эпох по лимитам GPU, всего получилось обучить модель на 80 эпох.

Сравнение с изначальной моделью чаще выдает ничью, поскольку оба ответа плохо попадают в контекст, и хотя изначальная модель может отвечать рациональнее, ее ответы могут быть этически не корректными.

Сравнение с выбранными (в датасете) ответами чаще склоняется к выбранным ответам.