Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

Yuntao Bai, Andy Jones, Kamal Ndousse and others.

Автор обзора-рецензии: **Иванов Данила, группа 193.**

Данная работа посвящена универсальному способу настройки предобученных языковых ассистентов с помощью обучения с подкреплением на основе оценок пользователей. У авторов получилось реализовать метод, который: 1) улучшает производительность моделей практически во всех NLP задачах, 2) полностью совместим с обучением специфичным NLP навыкам, 3) работает в онлайне, то есть модель данных постоянно адаптируется и улучшается. Все это достигается благодаря исследованиям команды в области RLHF.

Работа написана в апреле 2022 года командой авторов из лаборатории Anthropic. У статьи порядка 28 авторов, которые разделены на 3 группы по уровню вклада в статью. Три основных автора:

- **Yuntao Bai**, исследователь лаборатории Anthropic. В данной статье Yuntao Bai занимался проектированием экспериментов, а также оценкой качества работы моделей. Yuntao Bai является автором следующей схожей статьи: “A General Language Assistant as a Laboratory for Alignment”, 2021. Интересно, что, судя по публикациям, первостепенная область исследований Yuntao Bai – это физика, и только на втором месте идут NLP исследования в области ассистентов и их безопасности. Этот автор часто работает в соавторстве с **Amanda Askell**, которая тоже является автором данной статьи.
- **Andy Jones**, исследователь лаборатории Anthropic. В данной статье занимался организацией обучения с подкреплением для больших языковых моделей, а также внедрением и отладкой RLHF. Работал в паре с Kamal Ndoussse. Andy Jones является автором статей: “A General Language Assistant as a Laboratory for Alignment”, 2021; “Scaling scaling laws with board games”, 2021. Автор занимается машинным обучением в области NLP и RL, что помогло ему привнести большой вклад в рассматриваемую статью.
- **Kamal Ndoussse**, исследователь лаборатории Anthropic. В данной статье занимался организацией обучения с подкреплением для больших языковых моделей, а также внедрением и отладкой RLHF. Работал в паре с Andy Jones**.** Kamal Ndoussse является автором следующих статей: “A General Language Assistant as a Laboratory for Alignment”, 2021; “Emergent social learning via multi-agent reinforcement learning”, 2021. Автор является специалистом в области RL.

О вкладе каждого из 28 авторов более подробно можно прочитать в самой статье в разделе “Author Contributions” стр. 37.

Все авторы работают в лаборатории Anthropic, которая специализируется на безопасности, дружелюбности моделей искусственного интеллекта, и у всех авторов общая предшествующая статья, посвященная языковому ассистенту. Думаю, их следующей задачей было моделирование безопасного ассистента. Специалистов в области RL посетила идея, что в данном случае можно применить обучение на человеческих оценках. Получается, что данная статья - это не случайное открытие, а результат логического развития проектов компании Anthropic.

Основой для данной работы стала следующая статья:

- A General Language Assistant as a Laboratory for Alignment – статья, в написании которой участвовали все авторы выше. Является основой для обсуждаемой статьи, так как в ней рассматриваются языковые ассистенты в целом и исследования лаборатории Anthropic в этих моделях в частности.
- Также авторы пользуются наработками в сфере **few-shot тестирования**, **helpful и harmless формализации и измерения** и так далее. Но все статьи, в которых рассматриваются данные вопросы, являются лишь инструментами и теоретическими материалами, а не причинами появления данной статьи.

Конкурентом выступает статья ниже:

- Training language models to follow instructions with human feedback – очень близкая по смыслу и подходам статья, в которой команда из OpenAI работает над тюнингом языковых моделей с помощью RLHF подхода. Эта статья станет основой для запуска проекта ChatGPT**.**

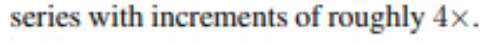
На основе данной статьи были написаны следующие работы:

- Constitutional AI: Harmlessness from AI Feedback – абсолютно логичное продолжение статьи от Yuntao Bai и команды, в котором они заменяют человеческий фитбек на AI оценки и делают модель автономной, сохранив достоинства подхода RL в виде адаптивности и универсальности.
- Scaling Laws for Reward Model Overoptimization – более глубокое и детальное изучение эффектов, возникающих во время RL обучения от лаборатории OpenAI, в том числе и тех эффектов, которые были обнаружены, но не были до конца изучены в обсуждаемой статье.

Ниже приведены сильные стороны данной статьи.

1. Главной сильной стороной является ее результат, то есть **универсальность и полезность метода,** предлагаемого в работе. Метод обучения RLHF был придуман несколько раньше этой статьи, но авторы данной работы успешно использовали его на специфичных NLP задачах. Также одним из открытий статьи является линейная связь Reward и $\sqrt{KL}$ во время обучения с подкреплением. Результаты действительно актуальные, и описанные подходы используются в наиболее современных и мощных ассистентах и чатботах, например, **ChatGPT**.
2. Одним из основных плюсов самого RL обучения является его адаптивность и самосовершенствование. В данной статье авторы смогли реализовать **online обучение**, то есть модель и датасеты обновляются каждую неделю благодаря новым взаимодействиям пользователей с ассистентом. Это несомненно преимущество данного подхода в сравнении с другими способами тюнинга.
3. Не менее важно достоинство – **большое количество иллюстраций** в статье. Такая визуальная поддержка очень помогает разобраться в новой теме. В статье используются визуализации схемы архитектуры и подхода обучения, множество графиков процесса обучения, визуализации распределений данных и так далее. Однако по графикам видно, что над статьей работало несколько человек, так как графики выполнены с разной визуальной оберткой: какие-то на белом фоне, какие-то ка синем, какие-то одной парой цветов, какие-то другой – это немного портит впечатление от статьи.

У работы есть и слабые стороны, например:

1. **Вторичность методов**. Не уверен, что это действительно можно считать недостатком, так как нужны не только статьи, в которых презентуется кардинально новый подход в машинном обучении, но и статьи, в которых применяются и исследуются уже известные подходы. Эта статья второго типа – в ней поставлена конкретная цель, в качестве решения выбран известный метод RLHF и результатом является то, что данный метод действительно прекрасно решает поставленную задачу. В статье есть новизна фактов, проверка еще непроверенных гипотез, но статья явно не революционная, хотя и на 78 страниц.
2. **Сложность воспроизведения результатов.** Стоит заметить, что с нуля такой результат получить практически невозможно, так как здесь используется онлайн человеческая экспертная разметка, а это очень сложно и дорого. Если же использовать базы данных из статьи, то обучение модели не сложное, но в статье используются довольно большие модели, поэтому из сложно запустить без необходимых ресурсов. Наконец, реализация подхода подразумевает обучение на основе RL, что является специфичным методом обучения, и далеко не каждый сможет достаточно аккуратно настроить систему, чтобы обучение достигло результатов статьи.
3. **Непонятная структуризация текста**. Статья читается тяжело, и даже не в силу своего размера, а скорее в силу неструктурированности текста и некоторых опечаток. Бывает, что новый термин используется в тексте сильно раньше, чем его определение. Блок про RLHF идет только под 4 номером, а до этого не понятно, как происходит само обучение с подкреплением и как обновляются базы данных. В некоторых местах статьи есть серьезные опечатки, например, потеряна вторая размерность в марице: 

Интересный факт.
Как бывает в реальной жизни, изобретение, которое проектировалось для использования во благо человечества, может использоваться против него, и наоборот. С подходом авторов статьи получилось также: они обнаружили, что смена знака с плюса на минус во время настройки RLHF начинает обучать модель не как безвредную, а наоборот – как вредоносную, но сохраняет естественность языка. Авторы отмечают, что предложенный ими подход может использоваться не по назначению, а именно для создания агрессивной модели, например: моделей дезинформации или мошенничества. Но самое интересное, что этот эффект был обнаружен совершенно случайно в результате ошибки в знаке во время настройки эксперимента!