УДК 004.89

Извлечение информации из научных статей при помощи LLM и методов mixture of experts

Крюков А. Д. (ИТМО)
Научный руководитель – кандидат химических наук, доцент Серов Н. С.
(ИТМО)

Введение. На данный момент составление больших и комплексных баз данных для в химических и биологических доменах является довольно таки трудоемкой задачей, так как в большинстве своем данные лежат в отдельно взятых статьях и для того, чтобы база данных была релевантна необходим дополнительный этап человеческой проверки для подтверждения того, что данные соответствуют действительности. В связи с этим, мы решили применить большие языковые модели для автоматизации процесса извлечения данных и методы mix-of-experts для оптимизации взаимодействия между получаемыми векторными представлениями из разных модальностей, мы провели ряд тестов применяя такие модели как Qwen, Qwen2, LLama, Mistral и различные способы смешивания представлений, получаемых из различных энкодеров.

Основная часть. Для начала мы собрали датасеты для разных модальностей, такие как данные для работы с таблицами, изображениями, схемами реакций и текстом, и предобученные энкодеры для работы с соответствующими типами данных. Далее работа разделяется на два этапа: использование энкодеров с различными способами смешивания для векторных представлений, например сложение двух векторов, сложение с использованием механизма внимание и билинейное смешивание. На этом этапе мы используем предобученные энкодеры для и ищем оптимальный вариант как смешать значения из 2 разных векторных представлений. На втором этапе мы заменяем этап соединения различных представлений на обучение адаптера, который может быть представлен полносвязной нейронной сетью, при наличии адаптера мы замораживаем веса энкодера и обучаем только адаптер. Главным преимуществом адаптера является то, что мы можем оптимально выбирать размер его входа и выхода в зависимости от того, сколько токенов на вход принимает большая языковая модель. Также использование различных архитектур адаптеров, например с использованием механизмов внимания, может нам помочь сделать векторные представления из энкодеров более репрезентативными в качестве токеном большой языковой модели.

Выводы. Мы смогли провести ряд экспериментов по применению разных энкодеров для разных модальностей, разных видов смешивания векторных представлений объектов для создания фундамента для дальнейшей разработки полноценной системы получения данных.

Список использованных источников:

1. **OmniFusion Technical Report**
   Elizaveta Goncharova, Anton Razzhigaev, Matvey Mikhalchuk, Maxim Kurkin, Irina Abdullaeva, Matvey Skripkin, Ivan Oseledets, Denis Dimitrov, Andrey Kuznetsov.
2. **Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts**

   Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, Neil Houlsby.
3. **MoE-LLaVA: Mixture of Experts for Large Vision-Language Models**

   Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, Li Yuan