Ассистент-3

Дмитрий Калашников, 11.04.2024

В данном задании вам предстоит подготовить и внедрить в бота инструктивную модель. В отличие от предыдущих двух версий, по итогам данного задания модель должна работать приближенно ко всем известному ChatGPT – отвечать на произвольные инструкции и соответствовать пользовательским ожиданиям.

Задание можно разбить на два подзадания:

1. Дообучение языковой модели (например, Llama 2) на некоторый набор инструкций (необязательно большой, вспомните результаты статьи LIMA, рассмотренные в лекции 6)
2. Настройка модели для быстрого и удобного инференса: приведение в оптимизированный формат, квантизация (рассматривалась на лекции 7)

Дообучение

Для дообучения языковой модели нужны либо подходящие мощности (ресурсов colab и kaggle, в теории, может не хватить), либо меньшая pretrain-модель (1.5B или 3B вместо 7B). Обратите внимание на одну из следующих моделей:

- Mistral: https://docs.mistral.ai
- Llama 2 (без VPN может не открыться): https://llama.meta.com
- Stable LM (есть 1.6B параметров): https://stability.ai/stable-lm

Можно поискать и другие модели в качестве претрейна. Советую искать сразу на hugginface, чтобы была возможность их без проблем выгрузить.

Ради практического интереса (как минимум, легче оценивать качество + полезнее) советую получить инструктивную модель именно на русском языке. Для этого помните, что чем больше llm видела текстов на русском во время претрейна, тем лучше она будет после alignment. В идеале, чтобы претрейн-модель в целом обучалась в основном на русском + английском.

Инференс

hugginface-формат моделей (директория с кучей файлов) не самый оптимальный для инференса. Чтобы модель работала быстро и удобно, написано приличное количество “инферилок” – библиотек (как правило, реализованных на C/C++ с питоновскими биндингами), приводящих модель в более оптимальный формат, помогающих с

квантизацией моделей, а также помогающих с оптимальным использованием устройств (cpu/gpu). Наиболее известные: llama.cpp (оптимальнее для cpu) и vllm (оптимальнее для gpu).

Для того, чтобы они воспользоваться, нужно собрать библиотеку. В питоновских версиях с этим обычно проблем нет, тем более на cpu, а я рекомендую вам делать инференс именно на нем – это будет работать с адекватной скоростью в случае квантизации до 4 бит, при этом качество после квантизации, как правило, критично не ухудшается. Далее – просто подставить путь к выровненной модели и наслаждаться инференсом.

llama.cpp python: https://llama-cpp-python.readthedocs.io/en/latest/
vllm: https://docs.vllm.ai/en/latest/

В чем задание

Нужно сделать так, чтобы инструктивная llm работала у вас в боте. Для этого нужно как минимум разобраться с настройкой инференса, на чем и хочется сделать акцент. Если при этом хочется дообучить модель на какой-то специфичный формат (например, писать стихи или анекдоты) – тоже welcome.

Баллы:

- Использование готовой инструктивной модели, настройка “инферилки” и внедрение в бота – 20 баллов
- Самостоятельное дообучение инструктивной модели, настройка “инферилки” и внедрение в бота – 30 баллов

Готовую инструктивную модель в нужном вам формате можно поискать на hugginface, но вот один из примеров таких моделей общего назначения на русском языке (подходит для инференса в llama.cpp):
https://huggingface.co/IlyaGusev/saiga_mistral_7b_gguf

Приятного коддинга)