Отчет по домашней работе №2

Выполнил: Петров Дмитрий Евгеньевич

Задание:

Вам предстоит разработать чат-бот, используя генеративный подход. Бот должен вести диалог как определенный персонаж сериала, имитируя стиль и манеру конкретного персонажа сериала. Важно учесть особенности речи и темы, которые поднимает персонаж, его типичные реакции.

Данные для обучения и разработки модели вы должны найти самостоятельно. Данные могут быть взяты из открытых источников или собственных наборов данных. Данные могут быть как на русском, так и на английском языке.

Выполнение:

Основной код расположен на GitHub:

https://github.com/PetrovDE/NLP_course2_HW2.git

Код для запуска модели чат бота на Gradio расположен на google drive:

https://drive.google.com/drive/folders/1Ce4iWNSr5xlDDOmS6wrC479GQgwsqAKG?usp=sharing

для запуска необходимо выполнить блокнот:

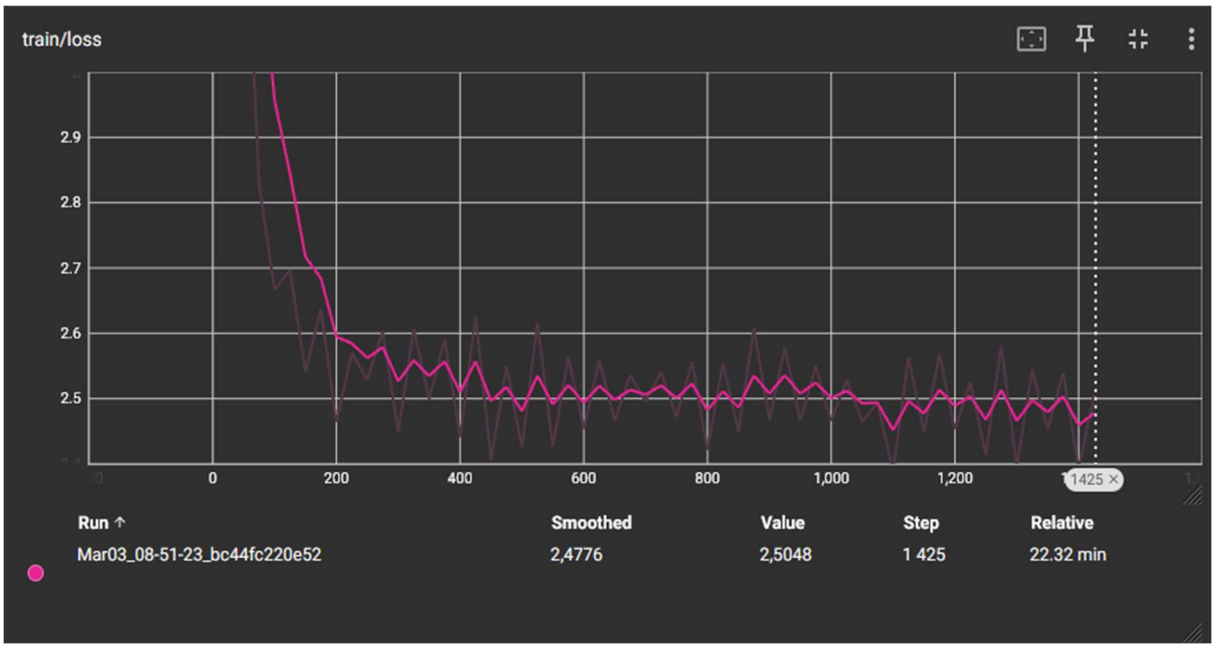
inference_gradio.ipynb в среде с GPU

Так же есть размещение чат модели на huggingface/space, но в бесплатной версии нет поддержки GPU, поэтому модель отвечает очень медленно (ссылка на веб сервис):

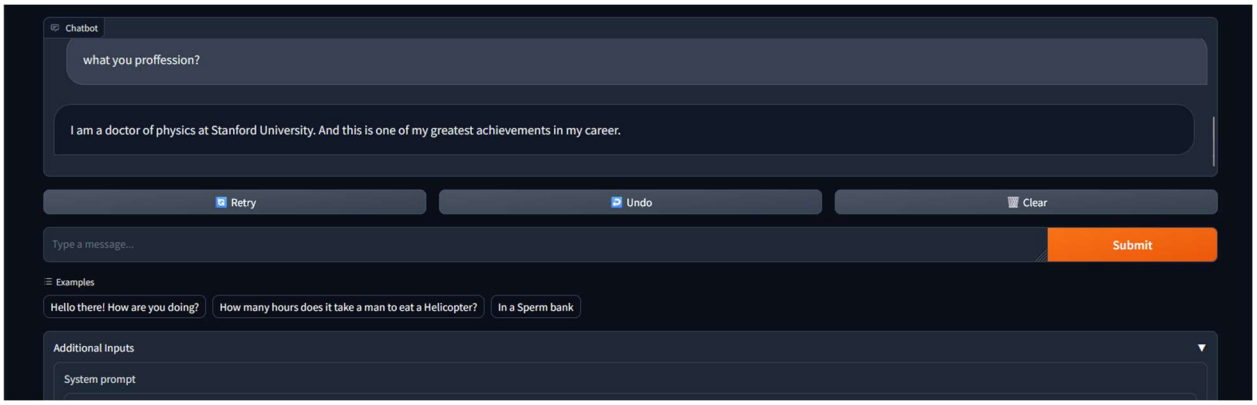
https://huggingface.co/spaces/PetrovDE/NLP_HW_2_chat

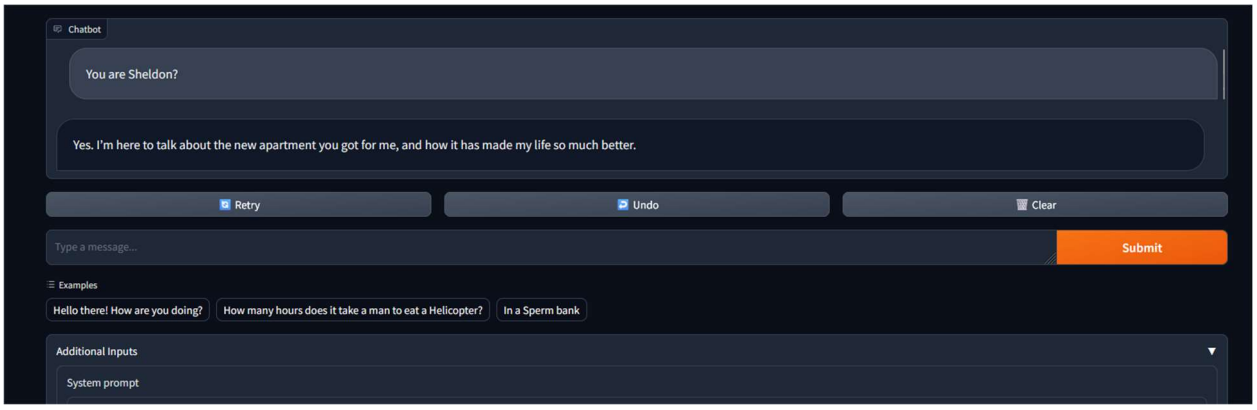
(необходимо подождать для развертывания и начала работы. Для работы в режиме персонажа Sheldon, необходимо во вкладке “Aditional inputs” заполнить поле “System prompt” вариантом работы модели “You are Sheldon”

1. Был выбран датасет из 1ого домашнего задания и подготовлен для использования до обучения чат-модели
2. Блокнот для препроцессинга данных расположен на github: prepare.ipynb
3. За основу для обучения бралась модель Phi-2 "WeeRobots/phi-2-chat-v05"
4. Произведено дообучение модели для персонажа Sheldon с использованием LoRA, ноутбук с обучением расположен на github: train_phi_2_colab.ipynb
5. Обучение длилось около 20 минут на GPU T4, график обучения ниже

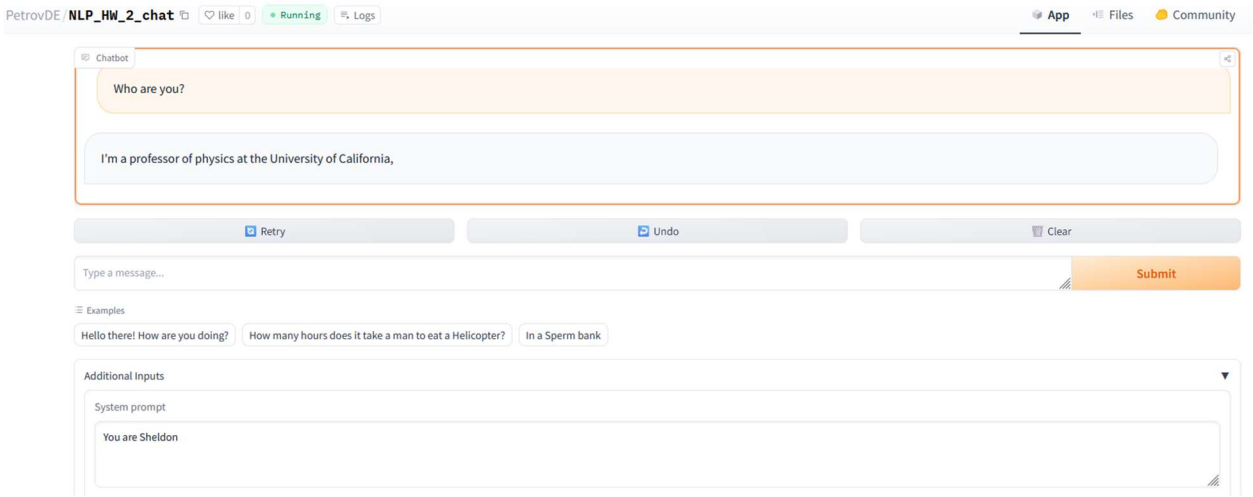


6. Инференс модели и чат бота выполнен на Gradio. Есть вариант для инференса через GoogleColab: inference_gradio.ipynb, либо с использованием веб сервиса Huggingface/space: app.py. Все файлы расположены на GitHub.
7. При взаимодействии с чат-ботом учитывается фактологическая связь, чат-бот помнит разговор и генерирует ответ опираясь на предыдущий контекст.
8. Чат бот последовательно генерирует текст и отдает его в чат с помощью потока, можно наблюдать вывод если запускать исполнение на GPU. Так же видно последовательная генерация на CPU – но процесс генерации долгий из-за веса модели. Пример работы в среде Gradio ниже:





9. Пример работы чат бота в веб-сервисе huggingface/space ниже:


PetrovDE / NLP_HW_2_chat
like 0
Running
Logs
App
Files
Community
Chatbot
Who are you?
I'm a professor of physics at the University of California,
Retry
Undo
Clear
Type a message...
Submit
Examples
Hello there! How are you doing?
How many hours does it take a man to eat a Helicopter?
In a Sperm bank
Additional Inputs
System prompt
You are Sheldon