ГУАП

КАФЕДРА № 44

ОТЧЕТ
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

доц., канд. техн. наук, доц.		Сергееев А. М.
должность, уч. степень, звание	подпись, дата	инициалы, фамилия

Отчет о работе на тему

Система распознавания речи

по курсу:

Основы искусственного интеллекта

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ гр. № 4142		Рябов Д.Р.
	подпись, дата	инициалы, фамилия

Санкт-Петербург 2024

1.ТЕХНИЧЕСКОЕ ЗАДАНИЕ

Итоговый продукт должен распознавать человеческую речь из аудифайла. Входные данные: аудифайл. Выходные данные: расшифровка аудифайла (текст)

2.ВЫБОР СПОСОБА РЕАЛИЗАЦИИ

Итоговый продукт будет представлен в виде нейросети, аудифайл и возвращающий текст, который содержит этот аудифайл. Разрабатываемая нейросеть будет основана на нейросети, специализированной для распознавания аудифайлов (openAI-whisper), дообученная на датасете с аудифоайлами на русском языке

3.ОПИСАНИЕ РЕАЛИЗАЦИИ

Реализация проекта происходит в несколько этапов:

1. Формирования дата сета;
2. Обучение модели;
3. Проверка работоспособности на заготовленных аудифайлах;
4. Проверка работоспособности в реальном времени

4.ДАТА СЕТ

Был использован готовый датасет от mozilla foundation, с сайта huggingface: mozilla-foundation/common_voice_11_0, содержащий в себе около 32 тысяч записей.

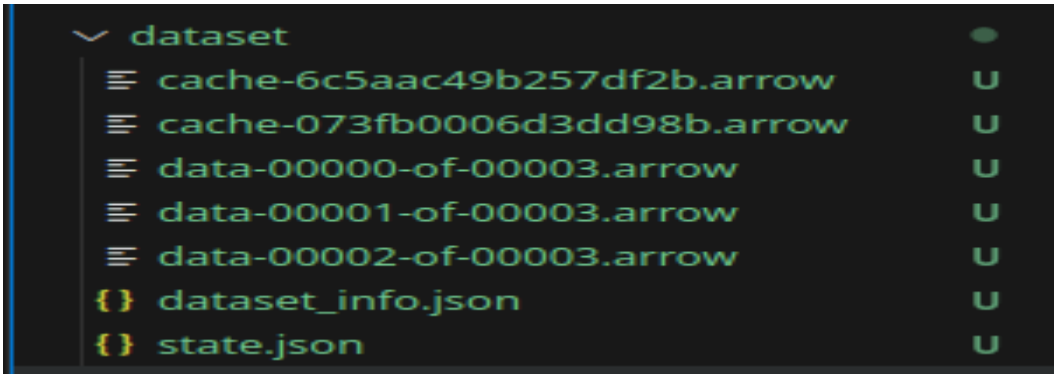

Рисунок 3 –Итоговый дата сет и файлы разметки

5.ОБУЧЕНИЕ

Поскольку датасет содержит слишком много аудиофайлов, его обучение занимало бы слишком много времени (по расчётам, около 250 часов)

Поэтому, было принято решение ограничить количество файлов до 1500, и увеличить количество поколений до 4. В итоге, обучение заняло около 6 часов (не считая ошибки, возникающие во время обучения, из-за которых приходилось начинать сначала) на домашнем сервере с использованием CUDA ядер на NVIDIA RTX 3060

6. ПРОВЕРКА НА АУДИФАЛАХ

Простые аудиофайлы нейросеть распознает довольно точно (цифру 2, например)

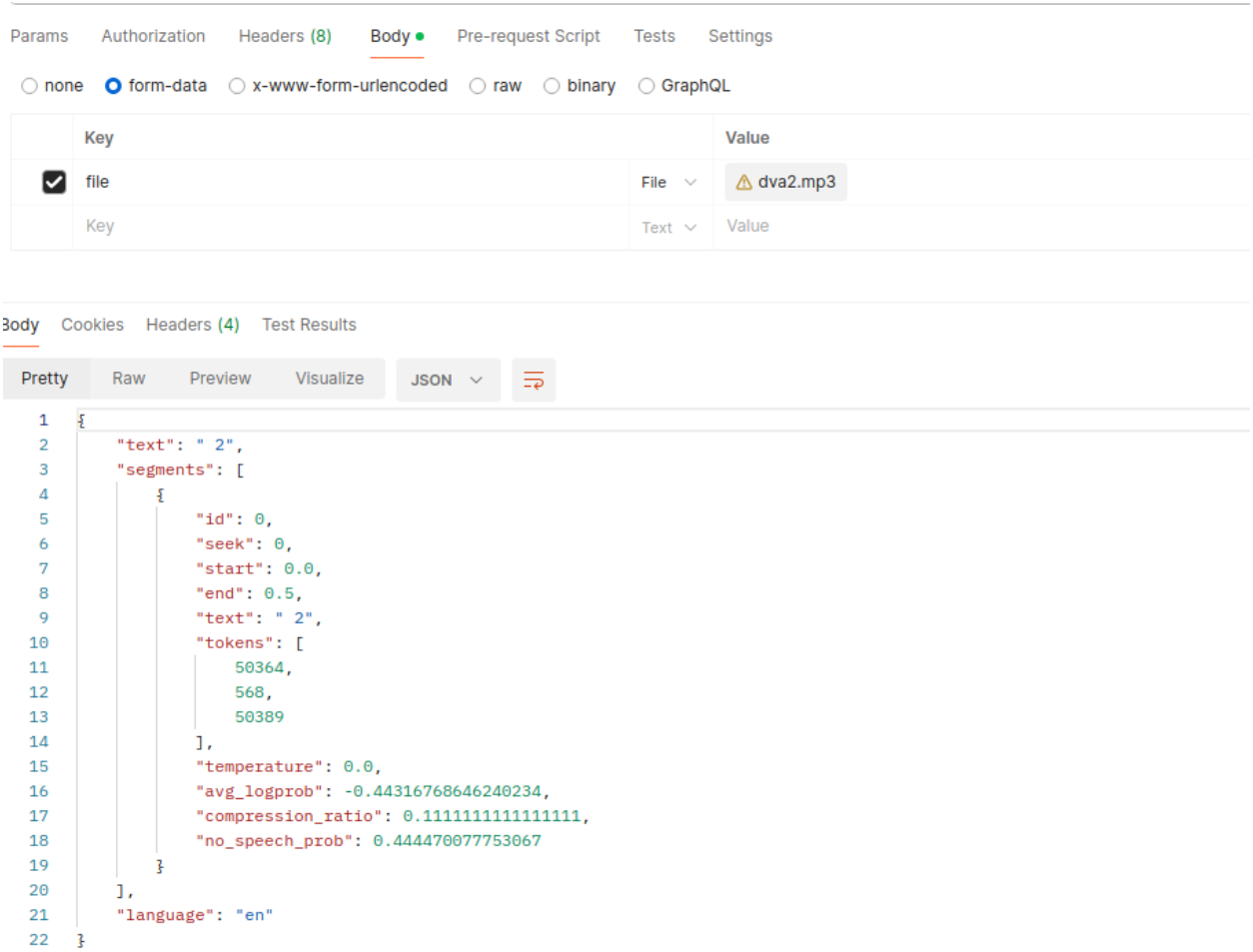

Рисунок 6 – Результат распознавания цифры 2 из аудиофайла

Но, со сложными фразами, бывают неточности:


```
{
    "text": " Доклад Международного Агентства по атомной мерке.",
    "segments": [
        {
            "id": 0,
            "seek": 0,
            "start": 0.0,
            "end": 3.5,
            "text": " Доклад Международного Агентства по атомной мерке.",
            "tokens": [
                50364,
                3401,
                2637,
                10396,
                3493,
                21207,
                1931,
                10004,
                4699,
                3450,
                1906,
                5243,
                12115,
                2801,
                2559,
                17804,
                5007,
                48231,
                8222,
                13,
                50539
            ],
            "temperature": 0.0,
            "avg_logprob": -0.41164346174760297,
            "compression_ratio": 1.15,
            "no_speech_prob": 0.10928210616111755
        }
```


Рисунок 7 — результат распознавания фразы «доклад международного агенства по атомной энергетике»

7. ЗАКЛЮЧЕНИЕ

В ходе работы над проектом “Система распознавания речи” была разработана нейронная сеть. Исходный код доступен по ссылке: https://github.com/Klutrem/ai-speech-to-text

Итоговый продукт может распознать простые аудифайлы. С более сложными возникают неточности, скорее всего, из-за небольшого датасета. Чтобы нейросеть работала более точно, необходимо её дообучить на более объемном датасете.

8. СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. Исходный датасет [Электронный ресурс]. URL:

https://huggingface.co/datasets/mozilla-foundation/common_voice_11_0/viewer/ru

2. Исходная модель openai whisper [Электронный ресурс]. URL

https://github.com/openai/whisper

3. Fine-Tune Whisper For Multilingual ASR with Transformers [Электронный ресурс]. URL https://huggingface.co/blog/fine-tune-whisper