36. Создание мультимодальной модели, работающей с аудио-данными, на базе малых языковых моделей

Бородин Кирилл, Каримов Эльвир, Мальцева Ирина, Неминова Екатерина

Введение

В последние годы мультимодальные системы, интегрирующие различные типы данных, такие как текст и изображения, активно развиваются. Однако область, в которой используются текст и аудио, развивается медленнее, несмотря на значительный потенциал в приложениях таких систем. В данном проекте рассматривается разработка мультимодальной модели, способной обрабатывать аудио и текстовые данные.

Цели и задачи

Основная цель проекта — создать мультимодальную модель на базе малых языковых моделей для обработки текстовых и аудио данных. Модель должна выполнять инструкции в текстовом виде используя анализ аудио данных.

Для выполнения проекта поставлен ряд задач:

- Определение мультимодальной задачи, включающей аудио и текст, для создания обучающих инструкций
- Выбор подходящего датасета
- Выбор подходящей архитектуры и фреймворка для разработки
- Выбор модели кодировщика аудио данных

- Выбор LLM
- Выбор адаптера для аудио эмбеддингов
- Обучение адаптера,настройка модели и сравнение с существующими подходами (Qwen-Audio)

Методология

В данном проекте производилась адаптация мультимодальных методов, первоначально разработанных для работы с изображениями и текстом, для использования в аудио-текстовой обработке. Это включало создание и реализацию специализированных компонентов для внедрения аудио эмбеддингов в рамках фреймворка TinyLlava. На первом этапе проекта акцент был сделан на обучении адаптера с использованием новых аудио данных, чтобы упростить задачу.

Фреймворк TinyLlava, содержащий элементы для интеграции изображений и текста, был выбран в качестве основы проекта и адаптирован для нужд аудио обработки. В качестве генерирующей модели использовалась LLM Microsoft-Phi2. Ниже представлены датасеты и модели, выбранные для экспериментов в проекте.

Датасеты

1. LibriSpeech - датасет для задачи распознавания речи из аудиокниг. Длина аудио - до 32 секунд
2. Clotho - датасет для задачи audio captioning. Размер датасета - около 12 тысяч элементов, длина аудио - до 30 секунд.

Модели энкодеры аудио данных

1. XEUS - многоязычный энкодер речи, охватывает более 4000 языков и обучен на основе более чем 1 миллиона часов речи из публично доступных датасетов
2. LanguageBind - мультимодальный энкодер, обучался на 10 миллионов данных, включая видео, данные о глубине, аудио и соответствующие языковые аннотации

Результаты

- Разработаны и адаптированы компоненты для интеграции аудио модальности в фреймворк TinyLlava
- Проведено обучение MLP адаптеров (на нескольких итерациях - в виду ограничений по ресурсам)

В рамках первого эксперимента рассматривалась задача распознавания речи, был создан эксперимент по обучению адаптера для энкодера XEUS в течении 700 итераций (рисунок 1). Однако обученные эмбеддинги не показали улучшений в решении поставленной задачи.

Для оценки качества использовалась метрика Word Error Rate (WER). Результат лучшего эксперимента представлен в таблице 0.1.

Model	Word Error Rate ↓
Qwen-Audio [1]	1.8
Ours	2.7

Таблица 0.1: WER для MLP адаптера и энкодера XEUS

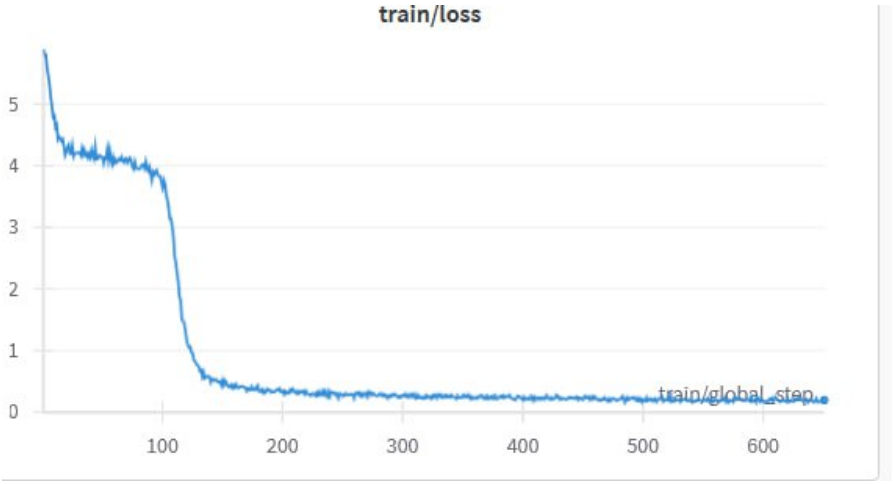

Рис. 1: Train loss для адаптера с XEUS

В рамках второго эксперимента рассматривалась задача audio captioning, был создан эксперимент по обучению адаптера для энкодера LanguageBind в течении 500 итераций (рисунок 2).

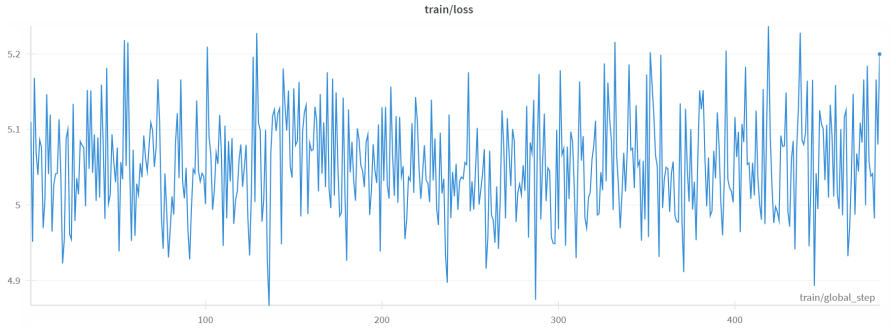

Рис. 2: Train loss для адаптера с LanguageBind

Заключение

Проект демонстрирует потенциал развития мультимодальных моделей с аудио и текстовыми модальностями, однако подчеркивает сложность задачи и необходимость дальнейших исследований и оптимизации подходов к обучению. В дальнейшем планируется продолжение экспериментов с различными конфигурациями энкодеров и улучшение качества результатов путем увеличения набора данных для обучения и вычислений.

Список литературы

[1] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. arXiv preprint arXiv:2311.07919, 2023.