УДК 004.8

РАЗРАБОТКА И ТЕСТИРОВАНИЕ МУЛЬТИМОДАЛЬНОЙ АУДИО-ТЕКСТОВОЙ МОДЕЛИ

Дутов Д.А. (ИТМО), **Митрофанов А.А.** (ИТМО)
Научный руководитель – Митрофанов А.А. (ИТМО)

Введение. В современном мире технологии обработки естественного языка (NLP, Natural Language Processing) играют всё более важную роль. Понимание и обработка звучащей речи (SLU, Speech Language Understanding), как ключевая часть NLP, открывает новые возможности для создания интеллектуальных систем, способных взаимодействовать с человеком на естественном языке. Разработка моделей, способных решать задачи SLU, является актуальной задачей, имеющей широкий спектр применений: от создания голосовых помощников до разработки систем автоматического перевода и анализа клиентских отзывов. Цель данной работы – создать универсальную модель для обработки аудио-текстовых запросов.

Основная часть. На начальном этапе разработки модели определён бенчмарк для тестирования — Dynamic SuperB phase 2, представляющий собой комплексную оценочную среду, включающую 180 задач различных доменов: обработки речи, анализа аудиосигналов и музыкальной информации. Бенчмарк представляет развёрнутую таксономию задач, где каждая общая задача дробится на ряд маленьких, создавая древовидный граф, на каждом конце которого представлена тестовая выборка. Такая иерархичная и подробная таксономия позволяет сформировать структурированную обучающую выборку, охватывающую широкий спектр акустических и языковых сценариев.

Разработанная модель имеет модульную архитектуру, включающую несколько ключевых компонентов:

1. Речевой энкодер WavLM [1] – глубокая нейросетевая модель для представления акустических признаков, обученная на крупных речевых корпусах.
2. Проекционный слой – адаптационный модуль, выполняющий преобразование признаков, полученных из WavLM, в формат, совместимый с языковой моделью.
3. Языковая модель Qwen 2.5 7B Instruct [2] – крупная предобученная трансформерная модель, ориентированная на обработку естественного языка и инструкционного обучения.
4. Адаптер LoRa (Low-Rank Adaptation) [3] – метод эффективного обучения больших языковых моделей с минимальным числом изменяемых параметров.

В процессе обучения веса речевого энкодера (WavLM) и языковой модели (Qwen 2.5 7B Instruct) не обновлялись, обучались только проекционный слой и адаптер LoRa. Данный подход позволил значительно снизить вычислительные затраты и упростить процесс адаптации модели к новой задаче, сохраняя при этом мощные генеративные способности предобученного LLM.

Для оценки производительности модели использовались модели-судьи — нейросетевые системы, обученные на задаче оценки качества сгенерированных ответов. Чтобы обеспечить объективную оценку, собран сбалансированный тест-кейс, включающий более 2500 аудиозаписей, аннотированных экспертами вручную. Он покрывал широкий спектр сценариев, что позволило всесторонне оценить способность модели к генерации корректных и осмысленных ответов и выбрать качественную модель-судью.

Выводы. Таким образом, разработана гибкая воспроизводимая модель, способная

адаптироваться к различным акустическим и языковым задачам, что было обеспечено предложенной архитектурой и методологией тестирования.

Важно отметить, что представленное исследование показало отличные от разработчиков бенчмаркка результаты: наиболее качественной моделью-судьёй оказалась Llama 3.3 70B Instuct с 97% точностью, в то время как создатели Dynamic SuperB phase 2 предпочитали модель GPT-4o, показавшую точность 96%.

Список использованных источников:

1. Chen S. et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing //IEEE Journal of Selected Topics in Signal Processing. – 2022. – Т. 16. – №. 6. – С. 1505-1518.
2. Yang A. et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement //arXiv preprint arXiv:2409.12122. – 2024.
3. Hu E. J. et al. Lora: Low-rank adaptation of large language models //arXiv preprint arXiv:2106.09685. – 2021.