**Тема ВКР**: “Интеллектуальная система робота-ассистента на основе анализа мультимодальных данных”

**Актуальность ВКР**

На данный момент происходит очень стремительное развитие технологий искусственного интеллекта и робототехники. Именно поэтому интеллектуальные системы, способные анализировать и обрабатывать информацию из различных источников, включая звук, изображения и текст становятся все более востребованными. Роботы-ассистенты, способные взаимодействовать с людьми на основе анализа таких мультимодальных данных, могут эффективно помогать в широком спектре задач и сфер, включая медицину, образование, бытовые услуги и производство.

Применение мультимодального анализа в робототехнике позволяет создавать более гибкие и адаптивные системы, способные лучше понимать контекст взаимодействия с человеком и эффективнее реагировать на его потребности и запросы. Например, такая система может анализировать речь, тем самым определяя нужны пользователя и предоставляя подходящую поддержку или информацию.

Интеграция мультимодального анализа в систему робота-ассистента открывает широкие перспективы для развития персонализированных и контекстно-ориентированных решений в области робототехники и искусственного интеллекта.

Таким образом, можно сказать, что ВКР имеет высокую актуальность и может внести значительный вклад в развитие современных технологий робототехники и искусственного интеллекта, а также в повышение уровня комфорта и эффективности взаимодействия между человеком и роботом в различных сферах жизни.

**Цели и задачи ВКР**

Целью ВКР является повышение эффективности обработки мультимодальных данных при помощи нейросетевых технологий.

Задачи ВКР:

* Изучение подходов обработки отдельных модальностей данных (изображение, текст, речь)
* Спроектировать архитектуру интеллектуальной системы
* Разработка компонентов обработки данных в интеллектуальной системе (сервис распознавания лиц, сервис распознавания речи, языковая модель генерации ответа, сервис синтеза речи)
* Интеграция разработанных компонентов обработки данных в интеллектуальной системе между собой

**Объект и предмет исследования ВКР**

Объектом исследования являются мультимодальные данные, представляющие собой информацию, собранную из различных источников, таких как текст, и изображения.

Предметом исследования является анализ мультимодальных данных.

Анализ мультимодальных данных позволяет получить более полное представление о контексте и содержании информации, что в свою очередь способствует более точным и глубоким выводам и решениям в различных областях, включая искусственный интеллект, робототехнику, обработку естественного языка, компьютерное зрение и другие.

**Изучение подходов обработки изображений**

В рамках ВКР мною были исследованы следующие подходы для обработки изображений:

Метод главных компонент (Principal Component Analysis, PCA)

PCA используется для снижения размерности признаков лица и выделения наиболее информативных компонентов.

Локальные бинарные шаблоны (Local Binary Patterns, LBP)

Метод позволяет описывать локальные текстурные особенности на изображениях с использованием простых и вычислительно эффективных операций.

Методы глубокого обучения (Deep Learning)

Представляют собой класс методов машинного обучения, основанных на искусственных нейронных сетях с большим количеством слоев

Алгоритмы локализации и выравнивания лица

Это методы компьютерного зрения, которые применяются для обнаружения и точного позиционирования лица на изображении, а также для его выравнивания по стандартным критериям.

Лицевые дескрипторы

Это числовые представления лица, которые содержат информацию о его характеристиках и особенностях. Эти дескрипторы являются ключевыми для идентификации лиц, анализа их особенностей и сравнения между собой.

Методы геометрического моделирования лиц

Это подходы к описанию и анализу лиц, основанные на их геометрических характеристиках, таких как расположение ключевых точек, форма, пропорции и другие геометрические параметры.

Совместные статистические модели (Joint Statistical Models)

В контексте распознавания лиц совместные статистические модели используются для моделирования совместного распределения различных лицевых характеристик или признаков лиц.

**Изучение подходов обработки речи**

* Рекуррентные нейронные сети
* Сверточные сети
* End-2-end модели

**Изучение подходов генерации текста**

В рамках ВКР мною были исследованы следующие подходы для генерации текста:

Марковские модели

Основаны на идее, что вероятность появления следующего символа или слова в тексте зависит только от некоторого конечного числа предыдущих символов или слов.

Рекуррентные нейронные сети (RNN)

Это класс нейронных сетей, способных обрабатывать последовательные данные, сохраняя состояние (память) и используя его для анализа последующих элементов последовательности.

LSTM (Long Short-Term Memory)

Длинная краткосрочная память является эффективным инструментом для генерации текста, так как он способен эффективно запоминать долгосрочные зависимости и использовать их при генерации последующих слов.

Преобразовательное представление текста (Transformer)

Transformer стал одной из наиболее важных и влиятельных разработок в области генерации текста и обработки естественного языка (Natural Language Processing, NLP) благодаря своей способности эффективно моделировать долгосрочные зависимости и взаимодействия между словами в тексте

**Изучение подходов синтеза речи**

* Конкатенативный способ
* Нейросетевые подходы
* End-2-end модели

**Проектирование архитектуры интеллектуальной системы**

Функциональные требования:

1. Ассистент должен уметь распознавать и интерпретировать запросы пользователя, выраженные в естественном языке.
2. Ассистент должен уметь распознавать человека, задающего вопрос
3. Система должна поддерживать интеграцию со сторонними сервисами для расширения функциональности
4. Система должна иметь способность отвечать на вопросы пользователя с помощью аудио

Нефункциональные требования:

1. Робот должен обрабатывать запросы пользователя в реальном времени с минимальной задержкой
2. Система должна быть способна масштабироваться для обработки увеличивающегося количества задач без потери производительности
3. Взаимодействие должно проходить по зашифрованному каналу связи
4. Интерфейс робота должен быть интуитивно понятным и легким в использовании для людей всех возрастов и уровней технической грамотности

**Разработка сервиса распознавания лиц**

Перед тем как перейти к описанию функциональности реализованного мною сервиса необходимо определиться с тем, что такое распознавание лиц и какие библиотеки можно использовать для данной прикладной задачи.

Распознавание лиц — это технология, позволяющая идентифицировать личность человека на основе его лица. Прогресс в этой области прошел долгий путь от ранних методов до современных алгоритмов глубокого обучения. На данный момент стандартный набор процессов (пайплайн, pipeline) решающий проблему распознавания лиц будет следовать следующей схеме (рисунок).

Широкие возможности для использования предобученных сверточных моделей и удобные абстрактные обертки для их использования предоставляет библиотека DeepFace для языка Python, именно она использовалась при разработке вышеупомянутого сервиса.

Архитектура БД с хранением изображений:

id - уникальный идентификатор записи в БД, name - имя человека, uid - название папки, в которой лежат фотографии лица человека

Функциональность:

* получение стрима изображения
* попытка найти и распознать лицо на полученном изображении (в случае успеха имя распознанного человека пушится в redis)
* функция add\_face, вызовом которой можно взять текущее фото, найти на нем лицо и сохранить в базу

**Разработка сервиса распознавания речи**

При выборе модели для распознавания речи, в первую очередь, внимание было приковано к быстродействию модели, по этой причине в качестве модели была выбрана Conformer-CTC модель. Данная архитектура модели была выбрана по следующим причинам:

* Данная модель совмещает в себе возможность трансформер улавливать взаимосвязи в длительных последовательностях и умение сверточных сетей хорошо понимать локальные зависимости
* модель использует CTC и не является авторегрессионной, что сильно увеличивает скорость распознавания, так как его можно проводить параллельно

**Разработка языковой модели генерации ответа**

Перед тем как перейти к рассмотрению функций реализованного мною сервиса рассмотрим небольшую теоретическую справку и существующие решения в языковых моделях.

Большие лингвистические модели, обученные на больших массивах текстов позволяют поддерживать диалог на высоком уровне с собеседником в роли ассистента. Но в связи с большим размером моделей усложняется их тренировка. Поэтому для прикладных задач удобнее использовать предобученные модели. Они обладают рядом особенностей, которые делают их важными инструментами в обработке естественного. Эти модели обучаются на огромных объемах текстовых данных и это позволяет им развивать глубокое понимание языка и его нюансов, что ведет к высокому качеству генерации и понимания текста. Кроме того, благодаря обучению на разнообразных данных, большие модели способны эффективно обрабатывать и генерировать тексты различных жанров и стилей, например в виде диалога с пользователем, что важно для робота ассистента.

Для разработки сервиса с языковой моделью генерации ответа использовалась большая языковая модель LLaMA 3, выпущенная Meta\* (запрещена и признана в РФ экстремистской) AI в феврале 2023 года и фреймворк с открытым исходным кодом для создания приложений на основе больших языковых моделей LangChain. LangChain предоставляет инструменты и абстракции для расширения возможностей настройки, повышения точности и актуальности генерируемой информации.

Функциональность:

* прием уведомлений от сервисов распознавания лиц и речи через redis
* в случае распознания лица и речи происходит генерация ответа на заданный пользователем вопрос и передача ответа в сервис синтеза речи
* в случае распознания только лица происходит генерация приветствия и передача текста в сервис синтеза речи
* когда распознана речь, но не распознано лицо, генерируется ответ, чтобы узнать имя человека. Если человек сказал и подтвердил имя, его лицо сохраняется путем вызова функции add\_face в сервисе распознавания лиц

**Разработка сервиса синтеза речи**

Основной моделью синтеза речи была выбрана модель - xTTS. xTTS - основана на другой архитектуре Tortoise, данная модель является объединением принципов работы других моделей - декодеров (GPT) и диффузионных моделей. Авторегрессионный декодер трансформирует входные текстовые токены в речевые токены, которые представляют собой промежуточное представление между текстом и аудио, после чего применяется конфронтационная модель, позволяющая отранжировать ответы сгенерированные декодером, путем сравнивания различных вариантов выхода и выбирая из них лучший. Затем диффузионная модель переводит отобранные токены обратно в логмел спектрограммы, которые затем передаются далее на вокодер, который переводит спектрограммы в звуковую волну.

**Интеграция разработанных компонентов между собой**

Все сервисы используемые для работы системы были разработаны с помощью языка Python. Для межсерверного взаимодействия использовался фреймворк gRPC. Данный фреймворк был выбран из-за высокой скорости работы, это обеспечивается при помощи бинарной сериализации, которая работает в несколько раз быстрее, чем использование JSON. Так же, фреймворк построен на протоколе HTTP/2, т.е имеет возможность поддерживать постоянное соединение, что было использовано в нашей системе. В качестве балансировщика нагрузки и reverse proxy был использован Nginx. В качестве сессионного хранилища выбран Redis - NoSQL хранилище вида ключ/значение, что хорошо подходит для хранения информации о сессии.

**Результаты и выводы ВКР**

Полученные результаты приведены на слайде 14