Доклад о проектах ФГУП «ГРЧЦ» в сфере искусственного интеллекта

Титульный слайд

Уважаемый Андрей Юрьевич, уважаемые коллеги, Вашему вниманию представляется доклад о проектах ФГУП «ГРЧЦ» в сфере искусственного интеллекта.

Слайд 1. Текущий этап развития

В настоящее время во ФГУП «ГРЧЦ» уже внедрен и эксплуатируется ряд систем, в работе которых используются технологии машинного обучения и элементы ИИ. Это автоматизированные системы мониторинга средств массовых коммуникаций (АС МСМК), мониторинга и анализа социальных медиа (АС МАСМ), мониторинга аудиовизуальных ресурсов (АС МАВР), «Чистый Интернет» (АС ЧИ).

Указанные системы имеют различную архитектуру и используют в работе не связанные между собой модели данных и наборы данных. Хотя системы находятся в стадии развития, существует определенный недостаток технологических мощностей.

Предполагается реализация Единого модуля анализа - сервиса, подключаемого по стандартному программному интерфейсу к прочим автоматизированным системам направления СМК. Подключение модуля транскрибирования, то есть преобразования аудио (голоса) в текст, позволит сформировать и обрабатывать текстовые наборы данных на основе материалов, собранных в процессе мониторинга.

Слайд 2. Разработка концепций, архитектур, методик

Первым этапом в 2021 году предполагается проведение научно-исследовательских работ по двум параллельным направлениям и первый этап разработки информационной системы мониторинга интернет-ресурсов.

Первое направление НИР **(шифр ОСНОВА)** предполагает исследование возможностей ИИ для мониторинга информационного пространства в целях решения задач Роскомнадзора по выявлению запрещенной информации. Помимо непосредственно исследования предметной области, будет подготовлено нормативно-методологическое обоснование применения ИИ, проведен аудит существующих разработок ФГУП ГРЧЦ и разработана Концепция развития ИИ ФГУП ГРЧЦ.

Вторым направлением **(шифр ОКУЛУС)** является оценка возможности ИИ, использующего методы компьютерного зрения, по распознаванию запрещенной информации на изображениях и видео. Важно понимать, что в отличие от текста,

механизмы распознавания изображений компьютером принципиально отличаются от распознавания человеком, в связи с чем потребуются значительные объемы наборов данных для обучения моделей ИИ. Возможно, по результатам НИРов также потребуется корректировка планов исследований в области ИИ.

Разработка первого этапа информационной системы мониторинга интернет-ресурсов **(ИС МИР-1)** на основе технологий семантического анализа и обработки естественного языка позволит выявлять заданные сигнатуры из анализируемой информации, обнаруживать медиакампании - скоординированное продвижение (распространение) информации всех видов, размещенной в средствах массовых коммуникаций, с привязкой ко времени, выявлять их источники и прогнозировать риски их распространения.

Также важным механизмом, обеспечивающим работу системы мониторинга, является так называемый **краулер** - система автоматизированного «обхода» и сбора данных, размещаемых в средствах массовой информации и массовых коммуникаций в сети Интернет, в том числе социальных сетей, мессенджеров и веб-сайтов.

С привлечением ведущих отраслевых институтов (МФТИ, МГУ и др.) планируется разработка ряда методик по анализу информационного пространства на основании различных наборов данных. Так, по результатам анализа каналов, групп, публичных сообщений и репостов в социальных сетях, мессенджерах, публикациях и других материалах Интернет-СМИ возможно выявление манипулятивных воздействий на массовое сознание, попыток поляризации общества и т.д.

Слайд 3. Управление данными и рисками

Для внедрения технологий ИИ также необходим ряд проектов, связанных с управлением данными и рисками.

Уже начиная с текущего года, как в ходе научно-исследовательских работ, так и с использованием собственных средств ГРЧЦ, начнется формирование наборов данных для обучения нейронных сетей. Будут созданы и размечены отдельные датасеты, содержащие аудио, видеоматериалы, графические изображения, метаданные и т.д. Эта работа будет продолжена и в дальнейшем. Помимо этого, необходимы инструменты, которые позволят операторам производить разметку (например, изображений, в которых будут выделяться области с текстом, отдельные части изображений, персоны и т.д.).

Распространенной мировой практикой при работе с ИИ является управление рисками. Сбои в работе моделей машинного обучения могут вызвать большой общественный резонанс. Недостаточное тестирование или злонамеренная манипуляция входными данными может привести к нарушению работы

корпоративных сервисов, использующих облачные решения, репутационным и финансовым потерям. Атака или ошибка такого рода во время обострения общественной дискуссии могут представить ГРЧЦ (Роскомнадзор) как одну из сторон конфликта. Поэтому будет сформирована группу экспертизы проектов и оценки рисков (Red Team), которая начнет работу уже в 2021 году. К работе в группе планируется привлечь членов Экспертного совета ФГУП «ГРЧЦ» по ИИ.

Слайд 4. Управление данными и рисками (продолжение)

В случае использования нескольких моделей ИИ необходимо ведение формуляров моделей ИИ, содержащих общую информацию о технологии и границах применимости моделей, об идентифицированных рисках и о мерах, принятых разработчиками и заказчиками моделей для управления этими рисками. Предлагается разработать единый формат формуляра модели (Model Card) и использовать его для ведения единого реестра моделей машинного обучения, для коммуникации с другими ведомствами, органами власти, СМИ для коммуникации со СМИ и для поддержки позиции в публичных обсуждениях. Результатом станет повышение объяснимости принципов работы используемых моделей машинного обучения и улучшение имиджа ведомства.

Другая устоявшаяся мировая практика - повторное использование разработанных ранее данных для других моделей ИИ. Это позволяет повысить качество моделей, ускорить их разработку и внедрение, упростить поддержку и аудит. Для этого необходимо вести программно доступный реестр признаков (Feature Store). В рамках подготовки данных исходные данные преобразуются в готовые признаки для моделей и вместе с метаинформацией сохраняются в едином реестре, доступном для существующих и новых моделей.

Чем больше ИИ-моделей и источников данных, тем выше вероятность нарушения работы моделей из-за незначительных на первый взгляд модификаций процесса подготовки/обработки данных. Внедрение системы управления данными (Data Lineage) позволит отслеживать и визуализировать процессы обработки данных на пути от источника данных до модели ИИ, где они преобразовываются и очищаются. С ростом количества моделей ИИ и объема данных аналогичные технологии необходимо внедрить в работу ФГУП «ГРЧЦ».

С 2022 года планируется также реализация инструмента анализа метаданных. Метаданные – это дополнительная (в том числе скрытая) информация, которая сопровождает медиаконтент, например имя пользователя, время публикации, группа, в которой она размещена, количество лайков, просмотров, комментариев, репостов. Эти метаданные могут быть проанализированы, чтобы понять контекст контента и определить, действительно ли данный контент является угрозой.

В краткосрочной и среднесрочной перспективах специалистам в области ИИ предстоит столкнуться с состязательными атаками. Это методы, реализованные в том числе с помощью ИИ и предназначенные для «обмана» используемых моделей ИИ (к примеру, обработка фотографии человека с добавлением случайного шума делает для некоторых моделей невозможным распознавание лица). Состязательные атаки могут снизить или свести к нулю эффективность работы систем по анализу медиаконтента. Для нивелирования угрозы, детектирования и защиты от таких атак возможна дешифровка медиаконтента на основе методов машинного обучения, а в случае невозможности дешифровки - передача на модерацию эксперту. Другой подход - наращивание количества параметров в модели, однако это потребует больших вычислительных ресурсов.

Слайд 5. Исследование и анализ технологий

Научно-исследовательские и опытно-конструкторские работы по выявлению Deepfake, правовому регулированию технологий ИИ в настоящее время уже проводятся специалистами НТЦ ГРЧЦ. Дальнейшие теоретические исследования и анализ технологий ИИ планируется проводить от общих тематик к частным с привлечением ведущих отраслевых институтов в области ИИ.

В 2022 году предполагаются теоретические исследования методов повышения качества обработки текстовой информации на естественном языке (NLP) и распознавания текста в изображениях и видео.

С 2023 года мы переходим к исследованию реализуемости с использованием ИИ практических задач, относящихся к сфере деятельности предприятия:

- определение тематической направленности сайтов по результатам анализа размещенного на сайте контента;

- распознавание фейковых, взломанных аккаунтов, ботов в социальных сетях и сервисах (позволит предсказывать начало информационной атаки по паттернам поведения, определить цель атаки, даст возможность уменьшить потенциальный урон. Также позволит формировать рекомендации на модерацию и/или блокирование, добавление аккаунта в список подозрительных, выдавать рекомендаций настоящему владельцу);

- транскрибирование и разделение нескольких голосов в аудио (необходимо для корректного применения алгоритмов обработки естественного языка в видео- и аудиоконтенте для правильного распознавания диалогов и подобных материалов);

- формирование текстового описания видео и изображений (определение объектов, персон, сюжетов, прочих условий, определение связей между событиями и явлениями, сопоставление с источниками материалов в сети, осуществляющими распространение информации, позволит в автоматизированном режиме определять

место и время происходящих событий, оценивать степень их реалистичности и прогнозировать сопутствующие риски);

- распознавание сложных мультимодальных медиаматериалов (плакаты, комиксы, «мемы» и др.). Данные материалы могут содержать запрещенную информацию как прямо, так и косвенно. Автоматизированный мониторинг с использованием ИИ требует контекстного понимания интернет-культуры: недавних событий, политических взглядов, культурных убеждений, поскольку мемы часто ссылаются на другие мемы или другие онлайн-события.

Слайд 6. Применимость результатов при разработке

Результаты научно-исследовательских работ, собственных исследований ФГУП «ГРЧЦ», разработанные методики будут использованы для реализации/ рефакторинга алгоритмов автоматизированных систем ГРЧЦ, в том числе в сфере СМК:

- МАВР - автоматизированная система проверки аудиовизуальных сервисов, обеспечивающая выявление признаков нарушений в аудиовизуальных произведениях на основании анализа текстовых метаданных о них (описание произведений, отзывы и рецензии);

- МИР – информационная система, которая осуществляет поддержку принятия решений об отнесении медиа-объектов (или групп медиа-объектов) к предмету мониторинга, детектирует информационную цель, описываемую моделью угроз, оценивает угрозы существования и распространения информации, относящейся к предмету мониторинга, и прогнозирует риски возникновения и распространения такой информации;

- АРГУС (определение характера событий, происходящих на видео, выявление порнографических материалов с участием несовершеннолетних);

- ОКУЛУС – сервис, с помощью технологий компьютерного зрения, описания видео, транскрибирования аудио, распознавания текста определяющий наличие или отсутствие запрещенной информации на видео или изображениях;

- ЕМА, который позволит для получаемого на входе контента определять признаки наличия в нем нарушений по установленному набору тематик;

- добавление и развитие функционала транскрибирования аудио для использования в автоматизированных системах предприятия;

- идентификации транслируемых телерадиовещательных каналов в АСМТРВ;

- сервис валидации запрещенной и незапрещенной информации оператором-человеком;

Слайд 7. Применимость результатов при разработке (продолжение)

в сфере персональных данных технологии обработки естественного языка, data science, краулинга могут быть применены в автоматизированной системе мониторинга обработки персональных данных в сети Интернет, а также для автоматизации процесса контроля обработки ПДн в информационных системах операторов персональных данных;

для автоматизации выявления признаков нарушений в области радиоконтроля и использования радиочастотного спектра;

развития первой линии поддержки по телефону и системам мгновенного обмена сообщениями (голосовой помощник для обработки внешних обращений по телефонной связи и чат-бот, которые позволят снизить затраты, время реагирования на обращения и количество ошибок, получать более точную статистику).

Слайд 8. Целевая архитектура систем

Таким образом, от текущего состояния информационных и автоматизированных систем ФГУП «ГРЧЦ», рассмотренных на первом слайде, мы приходим к целевой обновленной архитектуре единой методологически обоснованной мультимодальной системы на основе технологий ИИ, унаследовавшей все ранее реализованные наработки и использующей в работе наборы и модели данных, реализованных на базе хранилища больших данных. Методы обработки с использованием ИИ позволят решать задачи выявления запрещенной информации, прогнозирования рисков и угроз, недостоверной информации.

Слайд 9.

Благодарю за внимание, готов ответить на ваши вопросы.