Заявка №: С1ИИ-217313 **Подана:** 15.05.2022

ИНФОРМАЦИЯ О ПРОЕКТЕ

Тематика проекта

Название проекта:

Развитие технологии федеративного обучения для мультимодальных данных на разных источниках

Название проекта на английском языке:

Development of Federated Learning Technology for Multimodal Data on Different Sources

Описание конечного продукта:

Когнитивные способности человека основаны на одновременном восприятии им информации от разных органов чувств: глаз (зрение), ушей (слух), языка (вкус), носа (обоняние), кожи (осязание) и вестибулярного аппарата (чувство равновесия и положения в пространстве, ускорение). Кроме того, в принятии решений человек использует накопленный ранее опыт. Таким образом, для имитации самообучения человека, системы машинного обучения должны использовать всю доступную мультимодальную (разнородную) информацию из разных источников об одних и тех же объектах и явлениях.

В интеллектуальных системах поддержки принятия решений органы чувств человека заменяют различные устройства: видеокамеры, микрофоны, сенсоры и датчики. Возможности измерений у таких систем могут значительно превосходить возможности человека как по составу, так и по точности. Однако не вся эта информация используется для анализа. Причинами могут быть: разнородность информации, конфиденциальность всей или части информации, "узкие" каналы связи до источников информации и т.п. Частично указанные проблемы решает технология федеративного обучения, предполагающая анализ данных непосредственно на источниках и обобщение результатов только на уровне моделей машинного обучения. Однако и она в настоящее время не решает проблему обучения на мультимодальных данных на разных источниках.

В результате выполнения проекта предполагается разработка и программная реализация методов мультимодальных данных, размещенных на разных источниках. Такие методы позволят существенно повысить качество машинного обучения за счет использования всей доступной информации об объектах или явлениях, что, в свою очередь, позволит получать при выполнении задач анализа и принятия решений результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.

Конечным продуктом проекта будет расширенная версия библиотеки федеративного обучения (FL4J - Federated Learning for Java) с модулями для работы с мультимодальными данными. В результате она получит следующие потребительские характеристики:

- работа с конфиденциальной информацией: персональные данные, коммерческая тайна, служебная тайна и т.п.;

- использование каналов связи с ограниченной пропускной способностью в системах, близких к реальному времени (Wi-Fi, 3G, спутниковая связь и т.п.);

- обучение на мультимодальных данных на разных источниках: измерения от датчиков, видео- и аудио-потоки, текстовая информация, структурированная информация из хранилищ данных и т.п.;

- анализ (применение обученных моделей) на мультимодальных данных на разных источниках в реальном времени: измерения сенсоров, видеоаналитика, online данные и т.п.

Примерами прикладных областей, где будут востребованы интеллектуальные системы поддержки принятия решений с таким характеристиками, являются:

- федеральные и муниципальные государственные структуры - анализ безопасности, оценка и

прогнозы социального развития, занятости населения и т.п.;

- образование – анализ обучающегося, адаптивное обучение и др.

- умные квартиры/дома/кварталы/ города - повышение безопасности, комфорта населения;

- промышленность - контроль качества, технологических процессов, аварийности и надежности оборудования т.п.

- финансовые рынки – оценка кредитоспособности клиентов, анализ и предсказание изменений котировок ценных бумаг и др.;

- медицина – рекомендации здорового образа жизни, диагностика пациентов, прогнозирование течения болезни и др.;

- страхование – оценка страхового риска, страхование автомобилей и т.п.;

- торговля – оценка потребительского спроса, персонификация предложений и услуг, прогнозирование изменения спроса и др.

- кибербезопасность - выявление инсайдеров, сетевых атак, спама и т.п.;

Требуется ли выполнение 2-го этапа (года) НИОКР?
Нет

Обоснование необходимости проведения НИОКР 2-го этапа (года)

Основное направление программы СТАРТ:
Н1. Цифровые технологии

Поднаправления:
10. Системы обработки и хранения информации. Инструменты для анализа больших данных (Big Data).

Фокусная тематика:
Инструменты для анализа больших данных (Big Data)

Приоритетные направления:
Информационно-телекоммуникационные системы

Приоритетный класс программного обеспечения:
ПО-05.06 ПО-05.06 Средства обработки Больших Данных (BigData)

Функциональные характеристики / возможности разрабатываемого ИТ-решения:
ПО-05.02.15 ПО-05.02.15 организация ввода и обработки данных из любых источников с использованием технологий ИИ

Направление в рамках Стратегии научно-технологического развития Российской Федерации:

а. Переход к передовым цифровым, интеллектуальным производственным технологиям, роботизированным системам, новым материалам и способам конструирования, создание систем обработки больших объемов данных, машинного обучения и искусственного интеллекта

Ключевые слова:

Федеративное обучение, Мультимодальные данные, Распределенные данные, Конфиденциальная информация, Персональные данные, Анализ Больших данных, Машинное обучение

Запрашиваемая сумма гранта (рублей):

4 000 000

Срок выполнения работ по проекту:

12

ИНФОРМАЦИЯ О ЗАЯВИТЕЛЕ И УЧАСТНИКАХ ПРОЕКТА

Основные сведения

Тип заявителя:

Физическое лицо

Руководитель (потенциальный) предприятия:

Филиппов Евгений Васильевич

Научный руководитель проекта:

Холод Иван Иванович

Члены проектной команды:

Сотрудник	Должность	Роль в проекте	Опыт и квалификация
Банников Алексей Александрович	Системный архитектор	Разработка архитектуры системы и отслеживание ее реализации	Более 25 лет разработки ПО
Колпащиков Максим Алексеевич	Программист (бэк-енд)	Программная реализация методов, разработка кода макета (back-end)	2+ года разработки ПО
Петухов Владимир Дмитриевич	Программист (фронт-енд)	Разработка методов и их программная реализация, разработка кода макета (front-end)	5+ лет разработки ПО
Медведев Евгений Романович	Программист (ИИ)	Разработка моделей федеративного обучения, проведение испытаний	2+ года опыт разработки в области МО
Мишанов Александр Александрович	Программист (ИИ)	Разработка моделей федеративного обучения, проведение испытаний	2+ года опыт разработки в области МО

Сагиров Сергей Владимирович	Программист (мобильные)	Разработка кода макета (mobile)	3+ года разработки мобильных приложений
Новикова Евгения Сергеевна	Инженер-исследователь	Разработка методов, разработка методик проверки, анализ реализации с точки зрения безопасности	Более 7 лет исследований в области ИИ и МО
Ефремов Михаил Александрович	Старший программист (бэк-енд)	Разработка методов и их программная реализация, разработка методик проверки, проведение испытаний	5+ лет разработки ПО
Холод Иван Иванович	Научный руководитель	Научное руководство, постановка задач, отслеживание хода работы, подготовка отчетов	Более 25 лет разработки П

Планы по привлечению новых специалистов:
Команда проекта сформирована полностью.

Для исполнителей по программе УМНИК

Подача заявки в рамках обязательств по программе «УМНИК»:
Нет

Номер контракта и тема проекта по программе «УМНИК» :

Роль исполнителя по программе «УМНИК» в заявке по программе «Старт»:

Заполняется если выбранно «Иное» в поле «Роль исполнителя по программе «УМНИК» в заявке по программе «Старт»:

Информация о заявителе

Заявитель:
Филиппов Евгений Васильевич

Дата регистрации предприятия:

Наличие в Едином реестре субъектов МСП:

Регион заявителя:
Санкт-Петербург

Выручка от реализации товаров (работ, услуг) за последний календарный год (рублей):
0

Среднесписочная численность сотрудников за последний календарный год, человек:
0

Профиль деятельности предприятия:

Заполняется если выбранно «Иное» в поле «Профиль деятельности предприятия»:

Учредители

№ п/п	Учредитель	Доля

Создано в соответствии с Федеральным законом от 2 августа 2009 г. № 217-ФЗ:
Нет

Учредитель компании по Федеральному закону от 2 августа 2009 г. № 217-ФЗ:

СОДЕРЖАНИЕ ПРОЕКТА

Аннотация проекта

Когнитивные способности человека основаны на одновременном восприятии информации от разных органов чувств. Дополнительно, человек в принятии решений использует накопленный ранее опыт. В искусственном интеллекте методы машинного обучения также применяются к данным, которые могут быть собраны из разных источников. В случае, если использование каких либо источников в обучении не возможно (например, из-за разнородности данных, конфиденциальности информации, и т.п.), то результат может быть существенно хуже.

Для решения этой проблемы может быть применена технология федеративного обучения (ФО), которая использует данные из разных источников, не передавая информацию между ними, но обмениваясь результатами обучения на них.

В рамках предлагаемого проекта планируется развитие технологии ФО для анализа мультимодальных (разнородных) данных от различных источников об одних и тех же объектах и явлениях. Полученные результаты будут реализованы в библиотеке ФО для Java (FL4J) и апробированы на реальных данных.

Принадлежность к проектам в сфере ИИ

Обоснование соответствия предмета проекта:

Когнитивные способности человека основаны на одновременном восприятии им информации от шести разных органов чувств: глаз (зрение), ушей (слух), языка (вкус), носа (обоняние), кожи (осязание) и вестибулярного аппарата (чувство равновесия и положения в пространстве, ускорение). Дополнительно, в принятии решений человек использует накопленный ранее опыт. Таким образом, для имитации самообучения человека, системы машинного обучения должны использовать всю доступную мультимодальную (разнородную) в данный момент времени информацию из разных источников.

В интеллектуальных системах поддержки принятия решений органы чувств человека заменяют различные устройства: видеокамеры, микрофоны, сенсоры и датчики. Возможности измерений у таких систем могут значительно превосходить возможности человека как по составу, так и по качеству измерений. В результате выполнения проекта предполагается разработка и реализация методов ФО мультимодальных данных, размещенных на разных источниках. Такие методы позволят существенно повысить качество машинного обучения за счет использования всей доступной информации об объектах или явлениях, что, в свою очередь, позволит получать при выполнении задач анализа и принятия решений результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека.

Технология искусственного интеллекта:

ТИИ-4 Интеллектуальная поддержка принятия решений

Обоснование выбора технологии:

В отличие от человека, который принимает решение на основе информации, поступающей от шести органов чувств и накопленного опыта, интеллектуальные системы поддержки принятия решений используют накопленную информацию и различные устройства: видеокамеры, микрофоны, сенсоры и датчики. Возможности сбора информации об окружающих объектах и происходящих явлениях у таких систем могут значительно превосходить возможности человека как по составу, так и по качеству получаемой информации. Однако, для принятия решений не всегда может быть использована вся получаемая информация, что может быть связано со следующими проблемами:

- мультимодальностью данных - для анализа изображений, текстов, измерений и другой информации используются разные методы;
- конфиденциальностью информации - не вся информация может передаваться к месту анализа для предотвращения несанкционированного доступа к ней;
- каналами с низкой пропускной способностью - время передачи по таким каналам может существенно замедлять принятие решений и обесценивать информацию.

Для решения этих проблем и повышения качества принятия решений в интеллектуальных системах может быть применена технология федеративного обучения (ФО), которая позволяет выполнять анализ данных, размещенных на разных источниках, не передавая информацию между ними, но обмениваясь только результатами обучения. В рамках данного проекта планируется развитие технологии ФО в части анализа мультимодальных данных на разных источниках. Результаты позволят существенно повысить возможности интеллектуальных систем поддержки принятия решений за счет использования всей доступной информации от разных источников об окружающих объектах или явлениях.

Технологическая задача, на решение которой направлен проект:

ЗИИ-4.03 Подготовка решений на основе открытых источников данных и неструктурированной информации, в том числе для использования в интеллектуальных системах поддержки принятия решений для решения стратегических вопросов и (или) адаптивного динамического управления сложными объектами.

ЗИИ-4.05 Управление и (или) обучение персонала и построение персонализированных карьерных или образовательных траекторий.

ЗИИ-4.07 Управление оборудованием и производственными системами на основе данных измерительных систем и исторических данных о поведении систем в различных ситуациях (включая создание систем искусственного интеллекта).

ЗИИ-4.08 Предиктивное обслуживание оборудования на основе методов математического моделирования (в том числе машинного обучения), предназначенное для снижения частоты поломок оборудования и ущерба от них, снижения затрат на диагностику и обслуживание станков и промышленного оборудования (включая создание систем искусственного интеллекта).

ЗИИ-4.09 Прогноз качества выпускаемой продукции, в частности прогноз вероятности и типов дефектов продукции, в том числе позволяющий находить и устранять причины этих дефектов (включая создание систем искусственного интеллекта).

ЗИИ-4.10 Сверхкраткосрочное прогнозирование, анализ потока данных в режиме реального времени и прогнозирование нештатных ситуаций (включая создание систем искусственного интеллекта).

ЗИИ-4.11 Поиск новых способов производства продукции или способов выпуска новой продукции путем моделирования производственного процесса для удовлетворения заданных функционально качественных параметров с помощью математических моделей, основанных на данных, в том числе моделей машинного обучения, включая исторические данные, а также данные, полученные в результате экспериментов с цифровыми двойниками производственных процессов и оборудования (включая создание систем искусственного интеллекта).

ЗИИ-4.13 Выявление аномалий производственных процессов и поиск их причин (включая создание систем искусственного интеллекта, которые должны быть основаны на алгоритмах математического моделирования, машинного обучения и исторических данных).

ЗИИ-4.14 Контроль и обеспечение производственной безопасности, основанные на анализе и моделировании поведения сотрудников (включая создание систем искусственного интеллекта, которые должны быть основаны на алгоритмах математического моделирования, машинного обучения и исторических данных).

ЗИИ-4.15 Контроль и сокращение вредных выбросов и загрязнения окружающей среды (включая создание систем искусственного интеллекта, которые должны быть основаны на алгоритмах математического моделирования, машинного обучения и исторических данных).

ЗИИ-4.17 Управление персоналом, контроль производительности, психофизического состояния и поиск возможностей оптимизации загрузки персонала (включая создание систем искусственного интеллекта, которые должны быть основаны на алгоритмах математического моделирования, машинного обучения и исторических данных).

ЗИИ-4.01 Предиктивный и прескриптивный анализ, позволяющий предсказывать

развитие ситуации на основе анализа данных и автоматизировать принятие решений в режиме реального времени (включая создание методов и моделей).

Обоснование выбора технологических зада:

Развитие технологии ФО в части анализа мультимодальных данных на разных источниках без их передачи на узел принятия решений позволит решить следующие проблемы интеллектуальных систем принятия решений:

- использование мультимодальных данных из разных источников (видеокамер, датчиков и т.п.), что позволит более комплексно оценивать текущую ситуацию и правильно принимать решения;
- использовать конфиденциальную информацию (персональные данные, служебную информацию и т.п.) без ее передачи по каналам связи 3й стороне, что позволит учитывать накопленную информацию ограниченного распространения в принятии решений;
- уменьшить время анализа за счет сокращения объема передаваемых данных и распределенного выполнения вычислений на источниках, что повысит оперативность принимаемых решений;
- использовать каналы связи с ограниченной пропускной способностью, что позволит снизить стоимость и сложность таких систем.

Полученные результаты могут использоваться при решении следующих технологических задач:

- предиктивная аналитика, позволяющая предсказывать развитие ситуации в режиме реального времени на основе анализа мультимодальных данных из разных источников, что позволит более комплексно оценивать текущую ситуацию и правильно принимать решения;
- подготовка решений, использующих данные из открытых источников и неструктурированную информацию (видео, аудио, текст и др.) для применения в интеллектуальных системах поддержки принятия решений адаптивного управления сложными объектами (транспорт, производство, и т.п.);
- предиктивный анализ работы оборудования на основании методов машинного обучения, для своевременного обслуживания с целью снижения частоты поломок оборудования и ущерба от них, а также управление оборудованием и производственными системами на основе данных от разных источников: измерительных систем (сенсоров и датчиков) и исторических данных (включая конфиденциальную информацию) о поведении систем в различных ситуациях;
- обучение и построение персонализированных карьерных и образовательных траекторий на основании анализа описаний учебных курсов, мультимедийных материалов к ним, успехов по освоению курсов и персональных данных обучающихся;
- прогноз вероятности появления и типов дефектов выпускаемой продукции, позволяющий находить и устранять причины их появления, как на основе информации о технологических процессах (от систем видеоконтроля и контроля качества), так и исторических данных о качестве продукции;
- сверх краткосрочное прогнозирование на основании анализа потока данных в режиме реального времени и прогнозирование нештатных ситуаций, за счет сокращения времени на передачу информации в аналитический центр;
- адаптивное планирование и управление производственными процессами за счет использования информации от разных источников, включая конфиденциальную информацию (персональные данные сотрудников, информацию служебного пользования и др.) в том числе при планировании производства, поставок продукции, логистики и др.
- контроль и обеспечение производственной безопасности, выявление аномалий производственных процессов и поиск их причин, основанные на анализе информации о поведении сотрудников, собираемой из разных источников, включая персональные данные;
- контроль и сокращение вредных выбросов и загрязнений окружающей среды на основе

анализа мультимодальных данных, как от измерительных станций, контролирующих параметры окружающей среды, так и информации из открытых источников и систем видеонаблюдения;

- управление персоналом, контроль производительности, психофизического состояния для повышения эффективности работы персонала на основании мультимодальных данных видеоконтроля, параметров производительности, а также персональных данных

Результат реализации проекта:

РИИ-1. Создание и (или) развитие и (или) внедрение новых технологий, программных средств или программно-аппаратных комплексов, а также их масштабирование,

Обоснование выбора результата:

Результатами проекта будет развитие и внедрении технологии ФО в части применения к мультимодальным данным на разных источниках. Результаты будут представлять собой программную реализацию в библиотеке ФО FL4J следующих методов:

- методов прямого и обратного преобразования моделей машинного обучения в унифицированный формат глобальной модели, описывающий результаты обучения в виде иерархии элементов, представляющих извлеченные из данных закономерности;
- методов комбинирования разных моделей машинного обучения, полученных на различных источниках;
- методов ФО для мультимодальных данных, полученных за единый интервал времени об одних и тех же объектах и явлениях (методы ФО для вертикально разделенных данных);
- методов применения глобальной модели к мультимодальным данным на разных источниках, позволяющих получать единый результат (оценку ситуации, прогноз и т.п.) с учетом всех доступных данных.

Программная реализация методов будет проверена на реальных мультимодальных данных о движении наземного транспортного средства, включающих в себя: данные от сенсоров (акселерометра, гироскопа, магнитометра), видеокамеры, микрофона, GPS-приемника, а также внешней информации о внешней среде (погодных условиях и др).

Научно-техническая часть проекта

Новизна предлагаемых в инновационном проекте решений:

Технология ФО была предложена компанией Google Inc в 2017. На ее основе в настоящее время ведутся разработки библиотек как с открытым исходным кодом, так и проприетарные библиотеки.

Исследование перечисленных библиотек ФО [1], показало, что ни одно из существующих решений в области ФО не позволяет работать с мультимодальными данными на разных источниках. Это существенно снижает возможности машинного обучения и подтверждает актуальность предлагаемого проекта.

Новизной предлагаемого решение является развитие технологии ФО для работы с мультимодальными данными на разных источниках и программная реализация в библиотеке FL4J. В частности:

- методов прямого и обратного преобразования моделей машинного обучения в унифицированный формат глобальной модели, будет отличаться от существующих возможностью привести к единому формату разные модели и комбинировать их в единую с учетом характеристик данных, на которых они обучены;
- методов комбинирования разных моделей машинного обучения, полученных на различных источниках, отличающихся от существующих мультимодальностью комбинируемых моделей;
- методов ФО для мультимодальных данных, полученных за единый интервал времени об одних и тех же объектах и явлениях, отличающихся от существующих мультимодальностью данных на разных источниках на которых будет осуществляться обучение;
- методов применения глобальной модели к мультимодальным данным на разных источниках, в отличие от существующих, позволяющая получать единый результат (оценку ситуации, прогноз и т.п.) с учетом всех доступных мультимодальных данных.

Реализация данных методов существенно повысит возможности применения машинного обучения, т.к. снимет ограничения на анализ данных об одних и тех же объектах и явлениях как имеющих разный формат, так и находящихся на разных источниках и не передаваемых в силу разных причин 3й стороне (например, в облако).

1. Kholod, I.; Yanaki, E.; Fomichev, D.; Shalugin, E.; Novikova, E.; Filippov, E.; Nordlund, M. Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis. Sensors 2021, 21, 167. https://doi.org/10.3390/s21010167

Способы и методы решения поставленных задач НИОКРОКР:

В научной литературе представлен ряд подходов, позволяющих обучать глобальную модель анализа на мультимодальных данных, принадлежащих разным источникам [2-5] . Однако у них есть ряд недостатков. В частности, они разработаны для обучения определенных моделей анализа, и используют достаточно ресурсоемкие криптографические преобразования для обеспечения конфиденциальности используемых для обучения данных. Использование таких криптопротоколов значительно увеличивает как время обучения, так и время логического вывода. Кроме того, поскольку такие криптопротоколы основаны на большом числе раундов одноранговой (peer-to-peer) связи между участниками обучения, практическое использование таких подходов в системах с плохой связью затруднено.

Перспективным способом обучения модели анализа на мультимодальных данных видится обучение независимых локальных моделей различного типа и их объединение путем обучения глобальной модели на основе параметров локальных моделей. Так, в [6] был предложен подход, который позволяет учитывать различные вычислительные возможности участников ФО, обучать локальные нейронные сети различной архитектуры и объединять их в одну глобальную модель. В [7] представлено развитие этого подхода, в нем для оптимизации взаимодействия между участниками ФО предлагается обмениваться предсказаниями моделей на общем неразмеченном наборе данных, а не параметрами локальных моделей. Предложенные подходы предназначены для решения проблемы обучения на данных, имеющих неоднородное распределение, т.е. Non-IID данные.

В Российской федерации проблемами ФО занимается НИУ ВШЭ, наиболее близкой к решаемой задаче является исследование, связанное с построением распределенного обучения глубоких нейронных сетей в условиях неравномерного распределения вычислительных ресурсов у клиентов ФО [8] .

В настоящем проекте будет решаться проблема объединения мультимодальных моделей, обученных на мультимодальных данных. Для этого будут решаться следующие научные и научно-технические задачи:

1. разработка способа описания мультимодальных данных на разных источниках в виде единого набора данных, синхронизированных по интервалу времени и набору объектов и/или явлений которые они описывают;
2. разработка методов прямого и обратного преобразования моделей машинного обучения в унифицированный формат глобальной модели, описывающий результаты обучения в виде иерархии элементов, представляющих извлеченные из данных закономерности;
3. развитие методов комбинирования разных моделей машинного обучения, полученных на различных источниках;
4. развитие методов ФО для мультимодальных данных, полученных за единый интервал времени об одних и тех же объектах и явлениях (методы ФО для вертикально разделенных данных);
5. развитие методов применения глобальной модели к мультимодальным данным на разных источниках, позволяющих получать единый результат (оценку ситуации, прогноз и т.п.) с учетом всех доступных данных;
6. развитие методов оценки качества глобальной модели на мультимодальных данных на разных источниках;
7. программная реализация методов в библиотеке ФО FL4J [9] для дальнейшего внедрения в интеллектуальные системы поддержки принятия решений;
8. апробация реализованных методов и алгоритмов на реальных мультимодальных данных (собранных и аннотированных вне рамок данного проекта).

Способ описания мультимодальных данных будет представлять собой набор отношений физических атрибутов реальных данных, хранящихся на источниках, и логических атрибутов набора данных, к которому будут применяться методы машинного обучения. В отличие от существующих способов описания, они будут содержать описание атрибутов для неструктурированных данных и способы их извлечения. Таким образом, отношения будут содержать следующую метаинформацию:

- источник местонахождения физических данных;
- способ преобразования физических атрибутов в логические: прямое, объединение и пивот преобразование;
- способ извлечения атрибутов из неструктурированных данных (изображений, текстов, аудио и др.);
- способ синхронизации - приведение к единому для всех данных временному отрезку;
- способ идентификации - определение тождественных объектов и/или явлений, которые данные описывают.

Способ описания мультимодальных данных будет реализован в библиотеке FL4J на базе существующей модели описания данных, что позволит применять к мультимодальным данным из разных источников все алгоритмы библиотеки.

Методы преобразования моделей машинного обучения в унифицированный формат глобальной модели [10], реализованный в библиотеке FL4J. Подобные преобразования позволят представлять разные модели машинного обучения в едином виде и в дальнейшем объединить их. Преобразования будут использовать описание мультимодальных данных на разных источниках в виде единого набора данных.

Методы комбинирования моделей машинного обучения, полученных на мультимодальных данных на разных источниках в глобальную модель. Комбинироваться будут модели преобразованные в унифицированный формат. При этом агрегация параметров моделей будет осуществляться на основании описания мультимодальных данных на разных источниках.

Методы ФО для мультимодальных данных будут основаны на методах ФО для вертикально распределенных данных и осуществлять диспетчеризацию обучения разных моделей на разных источниках на следующих этапах обучения:

1. описания мультимодальных данных на разных источниках;
2. обучение локальных моделей на каждом из источников;
3. преобразование локальных моделей в унифицированный формат глобальной модели;
4. сбор всех локальных моделей на сервере ФО;
5. комбинирование преобразованных моделей в глобальную модель;
6. рассылка комбинированной глобальной модели на клиенты ФО;
7. обратное преобразование комбинированной глобальной модели в локальные модели машинного обучения;
8. оценка качества глобальной модели ;
9. повторение шагов 2-8 до завершение обучения или пока поступают новые данные.

В отличие от существующих методов ФО новые методы будут использоваться в том числе и для неструктурированных данных разного формата: изображения, аудио, текста и др.

Методы применения глобальной модели к мультимодальным данным на разных источниках будут осуществлять локальный анализ новых мультимодальных данных на разных источниках и объединять полученные от них результаты. Методы будут выполнять следующие общие этапы:

1. отправка глобальной модели на клиенты ФО;
2. обратное преобразование глобальной модели в локальные модели машинного обучения;
3. применение моделей машинного обучения к новым данным;

4. отправка результатов применения моделей на локальных новых данных на сервер ФО;
5. формирование единого результата применения глобальной модели.

Программная реализация разработанных методов будет выполнена в библиотеке ФО FL4J, которая интегрирована с библиотекой глубокого обучения DL4J и содержит ряд “классических” алгоритмов машинного обучения. Библиотека FL4J реализует ряд уникальных решений, которые будут основой в разрабатываемых методах:

- блоковое представление алгоритмов машинного обучения, которое позволяет переносить блоки алгоритма, выполняющие вычисления над данными на источники и выполнять их параллельно как по данным, так и по задачам;
- описание информации о данных в виде набора отношений физических атрибутов реальных данных хранящихся на источниках и логических атрибутов набора данных к которому будут применяться методы машинного обучения, а также способов преобразования физических атрибутов в логические: прямое, агрегацию и пивот преобразование;
- метод трансформации алгоритмов машинного обучения как для горизонтального, так и для вертикального распределения данных, основанный на операциях из области оптимизаций компиляции, таких как перестановка и расщепление циклов (loop interchange и loop fission) [11], что позволяет практически любой алгоритм машинного обучения применить к вертикально разделяемым данным.
- унифицированную модель, описывающую результаты обучения в виде иерархии элементов, представляющих извлеченные из данных закономерности.

Апробация полученных результатов будет осуществляться на реальных данных о перемещении наземного транспорта и включающих в себя: данные от сенсоров (акселерометра, гироскопа, магнитометра), видеокамеры, микрофона, GPS-приемника, а также внешней информации о внешней среде (погодных условиях и др). Данные будут предоставлены компанией ООО “Смартилайзер Рус” в аннотированном виде. Суммарный объем записанных и аннотированных данных около 150 часов или около 500 миллионов отсчетов сенсоров. Кроме того, компания ООО “Смартилайзер Рус” предоставит программные средства для просмотра и аннотирования данных, для формирования очередей (под-наборов) для обучения и тестирования моделей и для просмотра результатов экспериментов, что позволит пополнить тестовый набор в короткие сроки.

2. R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, J. Joshi, and H. Ludwig. 2021. FedV: Privacy-Preserving Federated Learning over Vertically Partitioned Data. In Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec '21). Association for Computing Machinery, New York, NY, USA, 181–192. https://doi.org/10.1145/3474369.3486872
3. Q. Zhang, B. Gu, C. Deng, and H Huang. Secure bilevel asynchronous vertical federated learning with backward updating. arXiv preprint arXiv:2103.00958, 2021.
4. C. Wang, J. Liang, M. Huang, B. Bai, K. Bai, and H. Li. Hybrid differentially private federated learning on vertically partitioned data. arXiv preprint arXiv:2009.02763, 2020.
5. T. Chen, X. Jin, Y. Sun, and W. Yin. Vafl: a method of vertical asynchronous federated learning. arXiv preprint arXiv:2007.06081, 2020
6. D. Enmao, J. Ding and V. Tarokh. HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients.” ArXiv abs/2010.01264 (2021): n. pag.
7. Y.J.Cho, J. Wang, T. Chiruvolu, & G. Joshi, G. (2021). Personalized Federated Learning for Heterogeneous Clients with Clustered Knowledge Transfer. ArXiv, abs/2109.08119.
8. Diskin, Michael et al. “Distributed Deep Learning in Open Collaborations.” NeurIPS (2021). ArXiv, abs/2106.10207.

9. Kholod I. I. et al. FL4J—Federated Learning Framework for Java //Intelligent Distributed Computing XIV. – С. 225. DOI: 10.1007/978-3-030-96627-0_21

10. Ivan I. Kholod, Andrey V. Shorov. Unification of Mining Model for Parallel Processing. In: Proceeding of 2017 IEEE North West Russia Section Young Researchers in Electrical and Electronic Engineering Conference. (2017 ElConRusW), pp. 450–455. IEEE Xplore (2017). 10.1109@EIConRus.2017.7910588

11. Ivan Kholod, Andrey Shorov, and Sergei Gorlatch. Improving Data Mining for Differently Distributed Data in IoT Systems // The 13th International Symposium on Intelligent Distributed Computing (IDC 2019) pp 75-85 DOI: 10.1007/978-3-030-32258-8_9

Материально-техническая база, необходимая для реализации проекта (имеющаяся в наличии и/или планируемая к привлечению:

Для выполнения проекта будет использована следующая материально-техническая база:

- личные ПЭВМ членов команды для программной реализации результатов проекта;
- 2 планшета Samsung Galaxy Tab Active 2 и 1 планшет Samsung Galaxy Tab Active 3 для снятия мультимодальных данных (предоставляются безвозмездно компанией ООО “Смартилайзер Рус”);
- облачные сервисы Яндекс.Облако для отладки и апробации результатов проекта.

Задел по тематике проекта:

Библиотека ФО (Federated Learning for Java - FL4J) https://gitlab.fkti.etu.ru/fl4j/fl4j-framework - реализующая алгоритмы федеративного обучения, как в симуляционном, так и в реальном режиме, интегрированная с библиотекой Deep Learning for Java, для применения методов глубокого обучения. Библиотека FL4J реализует ряд уникальных решений, которые будут основой в разрабатываемых методах:

- блоковое представление алгоритмов машинного обучения, которое позволяет переносить блоки алгоритма, выполняющие вычисления над данными на источники и выполнять их параллельно как по данным, так и по задачам;
- описание информации о данных в виде набора отношений физических атрибутов реальных данных хранящихся на источниках и логических атрибутов набора данных к которому будут применяться методы машинного обучения, а также способов преобразования физических атрибутов в логические: прямое, агрегацию и пивот преобразование;
- метод трансформации алгоритмов машинного обучения как для горизонтального, так и для вертикального распределения данных, основанный на операциях из области оптимизаций компиляции, таких как перестановка и расщепление циклов (loop interchange и loop fission) [2], что позволяет практически любой алгоритм машинного обучения применить к вертикально разделяемым данным.
- унифицированную модель, описывающую результаты обучения в виде иерархии элементов, представляющих извлеченные из данных закономерности.

Библиотека FL4J соответствует 4-му уровню технологической готовности (УТГ) по ГОСТ Р 56861-2016: реализованы базовые части основного программного фреймворка, приняты решения в области набора программных технологий для реализации сервиса.

Ядро и алгоритмы библиотеки FL4J распространяются под лицензией Apache 2.0. Это позволяет привлекать сообщество разработчиков к расширению состава алгоритмов.

Лицензией Apache 2.0 разрешает: Коммерческое использование; Распространение; Изменение; Личное использование; Предоставление патентных прав

Лицензией Apache 2.0 требует: Упоминания авторства и лицензии в работе; Указывать изменения, внесенные в работу

От пользователей она требует только, если работа была изменена, писать об этом, и указывать исходное авторство. Лицензия отдельно оговаривает, что для производных работ нельзя использовать те же названия, если они являются торговым марками.

Компания ООО Смартилайлез Рус для выполнения проекта предоставляет:

- Стенд для изучения и сравнения фреймворков федеративного обучения, соответствующий 8-му уровню УТГ (прошел апробацию у заказчика);
- Набор аннотированных данных: около 150 часов или около 500 миллионов отсчетов сенсоров;
- Программные средства для просмотра и аннотирования данных, для формирования очередей (под-наборов) для обучения и тестирования моделей и для просмотра результатов экспериментов, соответствующие 7-му уровню УТГ (прошли апробацию в ООО Смартизайзер Рус).

Основные результаты описаны в следующих ключевых публикациях:

1. Novikova, E.; Fomichov, D.; Kholod, I.; Filippov, E. Analysis of Privacy-Enhancing Technologies in Open-Source Federated Learning Frameworks for Driver Activity Recognition. Sensors 2022, 22, 2983. https://doi.org/10.3390/s22082983
2. Efremov, M.A., Kholod, I.I., Kolpaschikov, M.A.: Java federated learning framework

architecture. In: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), pp. 306–309 (2021). DOI 10.1109/ElConRus51938.2021.9396508
3. Kholod, I.; Yanaki, E.; Fomichev, D.; Shalugin, E.; Novikova, E.; Filippov, E.; Nordlund, M. Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis. Sensors 2021, 21, 167. https://doi.org/10.3390/s21010167
4. Ivan Kholod, Andrey Shorov, and Sergei Gorlatch. Efficient Distribution and Processing of Data for Parallelizing Data Mining in Mobile Clouds. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), 11(1):2-17, Mar. 2020 DOI:10.22667/JOWUA.2020.03.31.002
5. Kholod, I., Rukavitsyn, A., Paznikov, A. et al. Parallelization of the self-organized maps algorithm for federated learning on distributed sources. J Supercomput (2020). https://doi.org/10.1007/s11227-020-03509-2
6. M. A. Efremov and I. I. Kholod, "Architecture of Swarm Robotics System Software Infrastructure," 2020 9th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 2020, pp. 1-4, doi: 10.1109/MECO49872.2020.9134247.
7. Ivan Kholod, Andrey Shorov, and Sergei Gorlatch. Improving Data Mining for Differently Distributed Data in IoT Systems // The 13th International Symposium on Intelligent Distributed Computing (IDC 2019) pp 75-85 DOI: 10.1007/978-3-030-32258-8_9
8. Kholod I. I. et al. FL4J—Federated Learning Framework for Java //Intelligent Distributed Computing XIV. – С. 225. DOI: 10.1007/978-3-030-96627-0_21

На элементы библиотеки и реализованные в ней решения и алгоритмы оформлены документы подтверждающие права интеллектуальную собственности

1. Холод И.И., Малов А.В., Родионов С.В. Способ распараллеливания интеллектуального анализа данных в вычислительной среде. // Патент на изобретение №2745018 от 18 марта 2021 г https://www1.fips.ru/registers-doc-view/fips_servlet?DB=RUPAT&DocNumber=2745018&TypeFile=html
2. Ефремов М.А., Колпащиков М.А., Табаков П.Л. Программный адаптер для управления сервером федеративного обучения. //Свид. о государств. Регистрации программы для ЭВМ № 2021669639 от 23.11.2021.
3. Ефремов М.А., Табаков П.Л. Программа для регистрации клиентов федеративного обучения. //Свид. о государств. Регистрации программы для ЭВМ № 2021669370 от 23.11.2021.
4. Ефремов М.А., Аристархов И.Е. Программа управления клиентом федеративного обучения. //Свид. о государств. Регистрации программы для ЭВМ № 2021669391 от 23.11.2021.
5. Холод И.И. Программа для распараллеливания по данным процессов интеллектуального анализа..//Свид. о государств. Регистрации программы для ЭВМ № 2016610768 от 19.01.2016.
6. Холод И.И. Программа подготовки набора функциональных блоков интеллектуального анализа данных к параллельному выполнению.//Свид. о государств. Регистрации программы для ЭВМ № 2016610772 от 19.01.2016.
7. Холод И.И. Функциональноблочная программа построения одноатрибутных классификационных правил..//Свид. о государств. Регистрации программы для ЭВМ № 2015611447 от 29.01.2015.
8. Холод И.И. Функциональноблочная программа параллельного построения одноатрибутных классификационных правил..//Свид. о государств. Регистрации программы для ЭВМ № 2015611451 от 29.01.2015.
9. Холод И.И. Функциональноблочная программа параллельного поиска ассоциативных правил..//Свид. о государств. Регистрации программы для ЭВМ № 2014660763 от 23.10.2014.

10. Холод И.И. Блоковая структура выполнения алгоритма поиска ассоциативных правил AprioriTID.//Свид. о государств. Регистрации программы для ЭВМ № 2014618030 от 12.08.2014.
11. Холод И.И. Программа блоковой декомпозиции алгоритма кластеризации kmeans.//Свид. о государств. Регистрации программы для ЭВМ № 2012610929 от 20.01.2012.
12. Холод И.И. Программа декомпозиции алгоритмов интеллектуального анализа данных.//Свид. о государств. Регистрации программы для ЭВМ № 2012610930 от 20.01.2012.
13. Холод И.И., Каршиев З.А Блоковая структура выполнения алгоритма классификации Naïve Bayes..//Свид. о государств. Регистрации программы для ЭВМ № 2012660852 от 29.11.2012.
14. Холод И.И., Накидкин А.В. Блоковая структура для параллельного выполнения алгоритмов интеллектуального анализа данных..//Свид. о государств. Регистрации программы для ЭВМ № 2012660853 от 29.11.2012.

Перспективы коммерциализации

Конкурентные преимущества создаваемого продукта, сравнение технико-экономических характеристик с основными аналогами, в том числе мировыми:

В настоящее время широко используются облачные сервисы анализа данных от IT лидеров, такие как Amazon Elastic Mapreduce, Google BigQuery, Apache Spark. Все они используют программные средства, реализующие концепцию MapReduce. Они предполагают, что данные будут передаваться в облако, где выполняется анализ. Это существенно снижает возможность их использования по следующим причинам:

- необходимо передавать данные 3й стороне;
- использовать защищенные каналы связи с хорошей пропускной способностью;
- постоянно актуализировать данные в облаке, чтобы анализ проводился на “свежих” данных.

Для преодоления данных проблем компанией Google Inc, предложена в 2017 году концепция ФО. На ее основе в настоящее время ведутся разработки библиотек с открытым исходным кодом:

•TensorFlow Federated (TFF) - от компании Google Inc (США)
https://github.com/tensorflow/federated
•Federated AI Technology Enabler (FATE) – компания Webank (Китай)
https://github.com/FederatedAI/FATE
•Paddle Federated Learning (PFL) – компания Baidu (Китай)
https://github.com/PaddlePaddle/PaddleFL
•PySyft – проект открытого сообщества OpenMined https://github.com/OpenMined/PySyft
•Flower – проект проект компании Adap GmbH (Германия) https://github.com/adap/flower
Также разработаны и проприетарные библиотеки
•Nvidia Clara Train SDK – компания Nvidia (США)
ttps://ngc.nvidia.com/catalog/containers/nvidia:clara-train-sdk
•IBM FL – компания IBM (США) https://www.ibm.com/blogs/research/2020/08/ibm-federated-learning-machine-learning-where-the-data-is/
•Swarm Learning - компания Hewlett Packard Enterprise (США)
https://github.com/HewlettPackard/swarm-learning

Был проведен сравнительные анализ представленных выше библиотек и фреймворков ФО в части поддержки различных моделей анализа, типов данных и их распределения между клиентами, наличия возможности обучения с подкреплением. Исследование [1] показало, что только библиотека FATE имеет реальную реализацию работы с вертикально распределенными данными, т.е. с данными, содержащими информацию об одних и тех же объектах и явлениях, но об их разных характеристиках на разных источниках. Однако имеется ряд ограничений, связанных с практическим применением данной библиотеки на практике. В частности, FATE реализует только две модели машинного обучения для работы с вертикальными данными: нейронные сети (алгоритм HeteroNN) и деревья решений (алгоритм HeteroGBDT). Текущая реализация алгоритма HeteroNN может анализировать данные только на двух источниках. Еще одним ограничением FATE является работа только со структурированной информацией. Таким образом, наиболее развитая в настоящее время библиотека ФО не позволяет выполнять обучение на мультимодальных данных на разных источниках. Это существенно снижает возможности машинного обучения и подтверждает актуальность предлагаемого проекта.

Новизной предлагаемого решение является развитие технологии ФО для работы с мультимодальными данными на разных источниках и программная реализация в библиотеке FL4J. Это существенно повысит возможности применения машинного обучения, т.к. снимет ограничения на анализ данных об одних и тех же объектах и явлениях, как имеющих разный

формат, так и находящихся на разных источниках и не передаваемых в силу разных причин третьей стороне (например, в облако).

1. Kholod, I.; Yanaki, E.; Fomichev, D.; Shalugin, E.; Novikova, E.; Filippov, E.; Nordlund, M. Open-Source Federated Learning Frameworks for IoT: A Comparative Review and Analysis. Sensors 2021, 21, 167. https://doi.org/10.3390/s21010167

Целевые потребительские сегменты (рынки) создаваемого продукта, их объемы, динамика и потенциал развития:

Практически каждый сегмент IT рынка или уже тесно связан с обработкой данных, или к этому неизбежно стремится.

По отчету компании IBM количество хранимых данных удваивается каждые два года.

Основными отраслями, показывающими рост использования аналитики больших данных, согласно отчету SaaS Scout, являются: медицина и фармацевтика, банковское дело, медиарынок (сервисы потокового воспроизведения, социальные сети, игровые сервисы), рынок ритейла, электроэнергетика, строительная отрасль. Анализ больших данных позволяет уменьшить издержки и сократить стоимость для конечного потребителя, получив рыночные преимущества, например в здравоохранении на 17%, согласно исследованию McKinsey.

По данным экспертов IDC, в 2021 году объем мирового рынка больших данных и бизнес-аналитики (BDA) составил $215,7 млрд, увеличившись на 10,1% относительно 2020-го. В своих расчетах аналитики учитывают коммерческие закупки оборудования, программного обеспечения и услуг, связанных с BDA. Примерно треть расходов на большие данные и бизнес-аналитику в 2021 году пришлись на три отрасли: банковский сектор, дискретное производство и профессиональные услуги. Следующие три по размеру инвестиций в BDA сегмента - непрерывное производство, телеком и правительство - добавили рынку выручку в $47 млрд по итогам 2021 года.

Ожидается, что мировой рынок машинного обучения (ML) вырастет с 21,17 млрд долларов в 2022 году до 209,91 млрд долларов к 2029 году при среднегодовом темпе роста в 38,8% в прогнозируемый период. (https://www.fortunebusinessinsights.com/machine-learning-market-102226)

В статье 2020 года в журнале Nature “The future of digital health with federated learning” рассматривается федеративная технология машинного обучения как способ избежать проблем с безопасностью чувствительных медицинских данных, а сами исследования в области федеративного обучения будут активно проводиться ближайшие 10 лет.

Согласно исследовательскому отчету “Global Federated Learning Solutions Market by Application (Drug Discovery, Industrial IoT), Vertical (Healthcare & Life Sciences, BFSI, Manufacturing, Retail & e-Commerce, Energy & Utilities), and Region - Forecast to 2028” («Рынок федеративного обучения по приложениям (...) и регионам — глобальный прогноз до 2028 года»), опубликованному MarketsandMarkets™ в начале 2022 года, в соответствии со сценарием AS-IS, размер глобального рынка федеративного обучения вырастет со 127 миллионов долларов США в 2023 году до 210 миллионов долларов США к 2028 году при совокупном годовом темпе роста (CAGR) 10,6% в течение прогнозируемого периода.
(https://www.marketsandmarkets.com/PressReleases/federated-learning-solutions.asp)

Отечественный рынок пока выглядит скромнее, но имеет потенциал к взрывному росту. Согласно данным, приведенным Ассоциацией участников рынка больших данных, объем рынка Big Data в России в 2020 году составил 10-30 млрд руб. При этом, в соответствии с усредненными прогнозами отечественных и иностранных экспертов, предполагается рост этого показателя в 10 раз – до отметки 300 млрд руб. к 2024 году.

По данным IDC Worldwide Artificial Intelligence Spending Guide, российский рынок искусственного интеллекта в 2020 году достиг 291 млн долларов США. В 2020 году наблюдался значительный рост инвестиций со стороны государственных организаций, который до 2024 года продолжится со средним ежегодным темпом 26,4%. Этот рост будет поддерживать заявленная государственная программа в области развития искусственного интеллекта

(https://www.idc.com/getdoc.jsp?containerId=prEUR247642121). Таким образом в 2023 году объем рынка может достичь 588 млн долларов США.

Основываясь на этих данных, можно спрогнозировать размер рынка ФО в Российской Федерации в 2023 году как 2,54 млн долларов США. Имея серьезный задел в виде разработанной библиотеки ФО, набора программных средств для исследований ФО и начальных наборов данных, новый стартап имеет все возможности для того, чтобы стать лидером рынка в РФ и занять долю рынка от 30% до 50% (0,76 - 1,27 млн долларов США в 2023 году)
Использование систем ФО наиболее актуально в случае:

- работа с конфиденциальными данными (персональные данные, коммерческая тайна, служебная тайна и т.п.);
- использование каналов связи с ограниченной пропускной способностью в системах близких к реальному времени (Wi-Fi, 3G, спутниковая связь и т.п.);
- возможности самообучения моделей на локальных данных (наличие обратной связи от пользователя, подтверждение прогноза и т.п.);
- необходимости выполнять прогнозирование на основе данных в реальном времени (измерения сенсоров, видеоаналитика, online данные и т.п.).

Варианты применения ФО можно разбить на три типа:

- прогнозирование (финансовые рынки, медицина, маркетинг, ритейл и др.);
- контроль сложных объектов (промышленность, безопасность, умные дома/города, беспилотный транспорт и др.);
- оценка клиентов (медицина, банковская сфера, страхование, ритейл, образование и др.).

Примерами прикладных областей где возможны подобные ситуации и могут быть востребованы полученные в проекте результаты являются:

- федеральные и муниципальные гос. структуры - анализ безопасности, социального развития, занятости населения;
- образование – анализ обучающегося, адаптивное обучение и др.
- умные квартиры/дома/кварталы/ города - повышение безопасности, комфорта населения;
- промышленность - контроль качества, технологических процессов, аварийности и надежности оборудования т.п.
- финансовые рынки – оценка кредитоспособности клиентов, анализ и предсказание изменение котировок ценных бумаг и др;
- медицина – рекомендации здорового образа жизни, диагностика пациентов, прогнозирование течения болезни и др.;
- страхование – оценка страхового риска, страхование автомобилей и т.п.;
- торговля – оценка потребительского спроса, персонификация предложений и услуг, прогнозирование изменения спроса и др.
- кибербезопасность - выявление инсайдеров, сетевых атак, спам и т.п.;

Ключевыми факторы принятия решения потребителями являются:

- осознание потенциальным заказчиком проблем, связанных с поиском закономерностей по эмпирическим данным, не поддающимся статистической обработке;
- необходимость быстрого принятия решения на основе хранимых данных;
- извлечение знаний из хранимой информации;
- тестирование собственных алгоритмов в определенной вычислительной среде;
- подбор алгоритмов решения прикладных задач машинного обучения;
- подбор вычислительной среды для реализации определенных алгоритмов;

- необходимость получения решения при невозможности раскрытия эмпирических данных и тем более сохранения их в чужом хранилище.

Описание бизнес-модели проекта и стратегии продвижения продукта на рынок:
На первом этапе результаты могут быть коммерциализированы как проекты апробации технологических решений на наборах данных заказчика. Бизнес модель реализации таких проектов - повременная оплата (Time & Materials). Предполагаемая длительность таких проектов - 1-2 месяца в случае проверки концепции, 3-9 месяцев в случае проверки концепции и интеграции технологического решения в окружение заказчика. Состав команды - 4-5 специалистов. Почасовая оплата от 2000 - 4000 руб/час. Будет предлагаться продажа лицензий на использование библиотеки алгоритмов FL4J. Модель распространения будет простой: фиксированная стоимость без ограничения времени и количества пользователей.

По мере развития продукта, бизнес модель будет меняться. После реализации функционала, обеспечивающего возможность работы с новыми видами данных, по мере накопления преднастроенных моделей, ресурсы будут предоставляться в аренду. Оплата будет повременная без детальной тарификации, примерная стоимость аренды 1-й модели на 1 месяц 1000 руб. Клиентам будут предоставляться скидки, зависящие от объема потребляемых услуг. Владельцы данных, предоставляющие свои данные для обучения моделей, будут получать процент от реализации услуг. Будет предлагаться продажа лицензий на использование фреймворка. Модель распространения будет учитывать ограничения времени и количество пользователей. Будут предлагаться услуги по созданию рабочих мест аналитика и по обеспечению доступа к наборам данных владельцев данных.

После реализации нового функционала и накопления пула клиентов будет введена повременная оплата с детальной тарификацией до 1 сек. Такая модель будет предоставляться и клиентам, и владельцам данных. Будет предлагаться продажа лицензий на использование платформы. Модель распространения будет учитывать ограничения времени и количество пользователей. Будут предлагаться услуги по созданию рабочих мест аналитика и по обеспечению доступа к наборам данных владельцев данных. Будут предлагаться услуги по подбору общедоступных данных, привлечению таких данных для обучения моделей с целью улучшения качественных и количественных показателей моделей.

Стратегия продвижения продукта на рынок.
Во-первых, планируется зарегистрировать стартап в качестве участника Сколково и через мероприятия Сколково выходить на потенциальных клиентов и инвесторов.
Во-вторых, планируется работать с профессиональными ассоциациями, такими как ассоциация участников финансового рынка ФИНТЕХ, Российский союз автостраховщиков, ассоциация потребительских кооперативов, ассоциация разработчиков программного обеспечения Руссофт и рядом других ассоциаций и объединений.
В-третьих планируется работать с городским правительством, комитетами.
Работа во всех указанных направлениях будет направлена как на сбор бизнес требований к разрабатываемому продукту, на поиск и выполнение пилотных проектов, так и на выявление потенциальных клиентов, заинтересованных как в анализе данных, так и в предоставлении своих данных для анализа.

ТЕХНИЧЕСКОЕ ЗАДАНИЕ НА ВЫПОЛНЕНИЕ НИОКР

Техническое задание на выполнение НИОКР

Цель выполнения НИОКР

Целью проекта является развитие технологии федеративного обучения (ФО) для работы с мультимодальными данными на разных источниках и ее программная реализация для внедрения в интеллектуальные системы поддержки принятия решений.

Назначение научно-технического продукта (изделия и т.п.)

Результатами проекта будут программная реализация методов ФО для мультимодальных данных, размещенных на разных источниках без их передачи 3й стороне. Таким образом, результаты проекта будут востребованы в областях использующих машинное обучение в случае:

- работы с мультимодальными данными на разных источниках: измерения от датчиков, видео- и аудио-потоки, текстовая информация, структурированная информация из хранилищ данных и т.п.;
- работы с конфиденциальными данными (персональные данные, коммерческая тайна, служебная тайна и т.п.);
- использование каналов связи с ограниченной пропускной способностью в системах близких к реальному времени (Wi-Fi, 3G, спутниковая связь и т.п.);
- необходимости выполнять прогнозирование на основе мультимодальных данных в реальном времени (измерения сенсоров, видеоаналитика, online данные и т.п.).

Примерами прикладных областей где возможны подобные ситуации и могут быть востребованы полученные в проекте результаты являются:

- федеральные и муниципальные гос. структуры - анализ безопасности, социального развития, занятости населения;
- образование – анализ обучающегося, адаптивное обучение и др.
- умные квартиры/дома/кварталы/ города - повышение безопасности, комфорта населения;
- промышленность - контроль качества, технологических процессов, аварийности и надежности оборудования т.п.
- финансовые рынки – оценка кредитоспособности клиентов, анализ и предсказание изменения котировок ценных бумаг и др;
- медицина – рекомендации здорового образа жизни, диагностика пациентов, прогнозирование течения болезни и др.;
- страхование – оценка страхового риска, страхование автомобилей и т.п.;
- торговля – оценка потребительского спроса, персонификация предложений и услуг, прогнозирование изменения спроса и др.
- кибербезопасность - выявление инсайдеров, сетевых атак, спама и т.п.

Технические требования к научно-техническому продукту (прототипу, опытному образцу), который должен быть разработан в рамках текущего этапа выполнения НИОКР

Основные технические параметры, определяющие функциональные, количественные (числовые) и качественные характеристики научно-

технического продукта, полученного в результате выполнения текущего

этап НИОКР

Функции, выполнение которых должен обеспечивать разрабатываемый научно-технический продукт

Программная реализация полученных результатов будет представлена макетами программных средств в составе:

- Сервер ФО - обеспечивающий координацию работы системы в части ФО, комбинирование моделей, обученных на мультимодальных данных на разных источниках и формирование результата применения моделей к мультимодальным данным.
- Клиент ФО - устанавливаемый на источник данных и обеспечивающий обучение локальной модели на данных, размещенных на источнике.

Сервер ФО будут включать в себя следующие модули:

- коммуницирования с клиентами ФО, выполняющий функции:
  - регистрации клиентов ФО на сервере ФО;
  - опрос доступности клиентов ФО в текущий момент времени;
  - сбор информации о данных, доступных на клиентах ФО;
  - отправку на клиенты ФО задач, которые на них должны быть выполнены;
  - получение результатов выполнения задач на клиентах ФО;
- конструктор описания метаданных, выполняющий функции:
  - отображения метаданных всех доступных на клиентах ФО мультимодальных данных;
  - формирование описание единого “виртуального” набора данных, включающие в себя мультимодальные данные на разных источниках;
- планировщик, выполняющий функции:
  - планирования процесса выполнения ФО на мультимодальных данных на разных источниках;
  - формирования задач для каждого клиента ФО, включающие в себя: адрес клиента ФО, анализируемые данные, выполняемые блоки алгоритма, локальную модель;
- диспетчер, выполняющий функции:
  - координации процесса выполнения ФО в соответствии с планом выполнения;
  - контроля выполнения процесса ФО и возникающих ошибок;
- агрегатор, выполняющий функции:
  - комбинирования моделей, полученных от клиентов ФО, в единую глобальную;
  - вычисления результата глобальной модели на основе результатов применения локальных моделей к новым мультимодальным данным на разных источниках;
  - декомпозицию глобальной модели на модели, пересылаемые на клиенты ФО;

Клиент ФО будут включать в себя следующие модули:

- коммуницирования с сервером ФО, выполняющим функции:
  - регистрации клиента ФО на сервере ФО;
  - отправки текущего статуса клиента ФО на сервер ФО;
  - отправки информации о данных, доступных на клиенте ФО;
  - приема от сервера ФО задач, которые должны быть выполнены на клиенте ФО;
  - отправка на сервер ФО результатов выполнения задач на клиенте ФО;
- описания данных, выполняющий функции:
  - описания доступных данных;
  - конфигурирования доступа к данным;
- диспетчер, выполняющий функции:
  - координации выполнения задач на клиенте ФО, полученных от сервера ФО;
  - контроля выполнения на клиенте ФО задач и возникающих ошибок;
- исполнитель, выполняющий функции:
  - выполнения задачи полученной от сервера ФО к данным, размещенным на клиенте ФО;

Количественные параметры, определяющие выполнение научно-техническим продуктом своих функций

Разрабатываемые программные средства должны обладать следующими характеристиками:

- работать с наборами данных, содержащих не менее 10 характеристик (атрибутов) распределенных по нескольким источникам;
- работать с наборами данных продолжительностью записей данных не менее 100 часов;
- поддерживать не менее 3 форматов данных;
- поддерживать не менее 8 типов источников;
- поддерживать работу на не менее 10 клиентов ФО;
- обеспечивать прирост точности при обучении на мультимодальных данных, не менее чем на 10% по отношению к данным с одного источника;
- обеспечивать снижение сетевого трафика не менее чем в 5 раз по сравнению с сетевым трафиков, возникающим при передаче данных для централизованной обработки.

Входные воздействия, необходимые для выполнения научно-техническим продуктом заданных функций

Для выполнения заданных функций должны быть обеспечены следующие информационные условия:

на источниках данных:

- размещены мультимодальные данные;
- установлены клиенты ФО;
- сконфигурирован доступ к данным;

на сервере

- установлен сервер ФО;
- введены параметры выполнения ФО:
  - метаданные набора мультимодальных данных;
  - способы разделения на обучающую и тестовую выборки;
  - перечень источников для анализа;
  - формируемая модель машинного обучения, алгоритм и параметры ее построения.

Выходные реакции, обеспечиваемые научно-техническим продуктом в результате выполнения своих функций

Выходной информацией в результате выполнения заданных функций являются:

- характеристики построенной глобальной модели машинного обучения;
- результаты применения глобальной модели машинного обучения;
- ошибки, возникающие в процессе выполнения ФО для мультимодальных данных на разных источниках.

Конструктивные требования к научно-техническому продукту, который должен быть получен в результате выполнения текущего этап НИОКР

Требования к конструкции и составным частям научно-технического продукта

Сервер ФО будет реализован с использованием следующих технологий:

- модуль коммуницирования:
  - на языке программирования Java версии 1.8 или выше;
  - протокола взаимодействия gRPC;
  - формата описания сообщений ProtoBuf.
- конструктор описания метаданных:
  - интерфейс пользователя на VueJS
- сервер на языке программирования Java версии 1.8 или выше;
- планировщик на языке программирования Java версии 1.8 или выше;
- диспетчер на языке программирования Java версии 1.8 или выше;
- агрегатор на языке программирования Java версии 1.8 или выше;

Клиент ФО будет реализован с использованием следующих технологий:

- модуль коммуницирования:
  - на языке программирования Java версии 1.8 или выше;
  - протокола взаимодействия gRPC;
  - формата описания сообщений ProtoBuf.
- модуль описания данных на языке программирования Java версии 1.8 или выше;
- диспетчер на языке программирования Java версии 1.8 или выше;
- исполнитель на языке программирования Java версии 1.8 или выше с интеграцией библиотеки глубокого обучения DeepLearning for Java

Требования к массогабаритным характеристикам научно-технического продукта

Не предъявляются

Вид исполнения, товарные формы

Программный код на языке Java, размещенный в закрытых репозиториях компании.

Требования к мощностным характеристикам научно-технического продукта – по потребляемой/производимой энергии

Требования к удельным характеристикам научно-технического продукта – на единицу производимой продукции – для машин и аппаратов

Требования к аппаратной части программных комплексов

Условия эксплуатации, использования научно-технического продукта

Иные требования к научно-техническому продукту (прототипу, опытному образцу), который должен быть разработан в рамках текущего этапа выполнения НИОКР

Требования по патентной охране

Ядро библиотеки FL4J будет выложено в доступ по принципам открытого кода. Это позволит организовать развитие базовых алгоритмов ML/FL более широким кругом разработчиков и минимизировать затраты. Оптимизированные алгоритмы работы с мультимодальными данными, программное обеспечение для организации работы владельцев данных и аналитиков как правило будет оформлено как ноу-хау компании. Программное обеспечение будет зарегистрировано в реестре Российского программного обеспечения

На полученные результаты будут оформлены права на интеллектуальную собственность:

- патенты на способы обучения на мультимодальных данных;
- свидетельства на программы ЭВМ на их программные реализации.

Перечень основных категорий комплектующих и материалов (входящих в состав разрабатываемого продукта (изделия) или используемых в процессе его разработки и изготовления)

Не планируется приобретение

Отчетность по НИОКР (перечень технической документации, разрабатываемой в процессе выполнения текущего этапа НИОКР)

- научно-технические отчеты;
- программы и методики испытаний продукции, изготовленной в соответствии с разработанной технологией;
- протоколы испытаний продукции, изготовленной в соответствии с разработанной технологией.

БЕСШОВНАЯ ПОДДЕРЖКА ПРОЕКТОВ

Платформа НТИ

Участвовал ли кто-либо из членов проектной команды в «Акселерационно-образовательные интенсивах по формированию и преакселерации команд:
Да

Участвовал ли кто-либо из членов проектной команды в программах «Диагностика и формирование компетентностного профиля человека / команды»:
Нет

Перечень членов проектной команды, участвовавших в программах Leader ID и АНО «Платформа НТИ»:

№ п/п	ФИО	LeaderId
1	Холод Иван Иванович	477257

Комментарий:
Появилась информация о возможностях продвижения научных достижений в коммерции

Фонд Сколково

Заявителю присвоен статус участника проекта «Сколково»
Нет

Предоставление заявителю грантов в рамках грантовых программ «Сколково»:
Нет

Заявитель – участник корпоративной акселерационной программы «Сколково»:
Нет

Комментарий:

РФПИ (РВК)

Заявителю предоставлены инвестиции со стороны венчурных фондов РВК:
Нет

Комментарий:

ФИОП

Заявителю предоставлена финансовая поддержка от ФИОП:
Нет

Заявителю предоставлена поддержка в рамках образовательных проектов ФИОП:
Нет

Заявителю предоставлена нормативно-техническая поддержка со стороны ФИОП:
Нет

Комментарий:

КАЛЕНДАРНЫЙ ПЛАН И СМЕТА

Календарный план

Календарный план выполнения НИОКР. 1-й годовой этап проекта:

№ этапа	Название этапа календарного плана	Длительность этапа, мес	Стоимость, руб.
1	1. Разработка и программная реализация описания мультимодальных данных на разных источниках в виде единого набора данных. 2. Разработка и программная реализация методов прямого и обратного преобразования искусственных нейронных сетей в унифицированный формат глобальной модели. 3. Разработка и программная реализация методов комбинирования разных искусственных нейронных сетей, полученных на различных источниках. 4. Разработка и программная реализация методов федеративного обучения для мультимодальных данных. 5. Разработка и программная реализация методов применения глобальной искусственной нейронной сети к мультимодальным данным на разных источниках. 6. Разработка и программная реализация методов оценки качества глобальной искусственной нейронной сети на мультимодальных данных на разных источниках.	6,00	2 000 000,00
2	1. Разработка и программная реализация методов прямого и обратного преобразования "классических" моделей машинного обучения (деревьев решений, классификационных правил, линейных функций и т.п.) в унифицированный формат глобальной модели. 2. Разработка и программная реализация методов комбинирования разных "классических" моделей машинного обучения (деревьев решений, классификационных правил, линейных функций и т.п.), полученных на различных источниках. 3. Разработка и программная реализация методов комбинирования разных моделей машинного обучения,	6,00	2 000 000,00

	включая искусственные нейронные сети, полученных на различных источниках. 4. Разработка и программная реализация методов применения глобальной модели машинного обучения к мультимодальным данным на разных источниках. 5. Разработка и программная реализация методов оценки качества глобальной модели машинного обучения на мультимодальных данных на разных источниках.		
	ИТОГО:		4 000 000

Смета

Смета затрат на реализацию проекта:

№ п/п	Наименование статей расходов:
1	Заработная плата
2	Начисление на заработную плату
3	Наборы данных
4	Материалы
5	Аренда облачных сервисов
6	Оплата работ соисполнителей и сторонних организаций
7	Прочие общехозяйственные расходы