Два поколения объяснительного искусственного интеллекта XAI 1.0

Аверкин Алексей Николаевич
ФИЦ «Информатика и Управления» РАН

averkin2003@inbox.ru

Национальная стратегии развития Искусственного Интеллекта на период до 2030 года

В сентябре 2019 года при принятии Стратегии ИИ президент РФ отметил, что страны с развитием искусственного интеллекта получат «преимущества, не сравнимые с ядерным оружием». И подчеркнул, что Россия имеет все шансы в этом преуспеть.

Один из основных принципов развития и использования технологий искусственного интеллекта (ИИ), приведенных в Национальной стратегии развития искусственного интеллекта на период до 2030 года является прозрачность: объяснимость работы искусственного интеллекта и процесса достижения им результатов.

Концепция объяснительного ИИ (explainable artificial intelligence – XAI) может укрепить доверие к технологии, поскольку разработчики смогут объяснять пользователям, как и почему их системы ИИ принимают те или иные решения.

Необходимость XAI

Системы искусственного интеллекта и машинного обучения (AI / ML) превзошли человеческие возможности почти во всех приложениях, где они были использованы.

ИИ начинает внедряться практически во все сферы человеческой деятельности. Эта тенденция ускоряется, и ИИ будет все больше использоваться в критически важных для безопасности системах.

К сожалению, системы искусственного интеллекта на базе нейросетей очень сложны для интерпретации и иногда допускают ошибки, и пользователи-люди не могут доверять их решениям без объяснения алгоритма принятия решения. Глубокие нейросети также являются крайне неустойчивыми системам и подвержаны атакам на алгоритмы и на данные. Для решения этих проблем и нужны системы объяснительного XAI.

5 волн ИИ

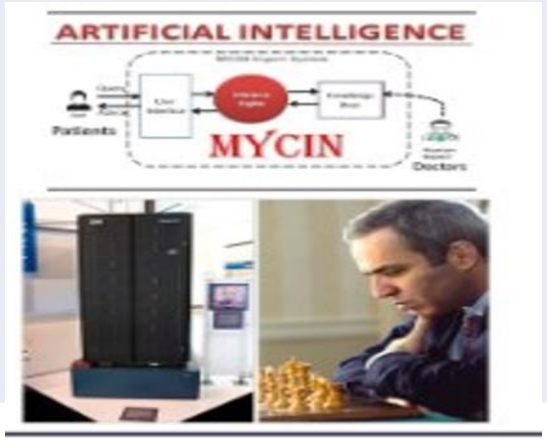
Каждое десятилетие в технологиях происходят революционные сдвиги, которые становятся новыми платформами, на которых строятся прикладные технологии.

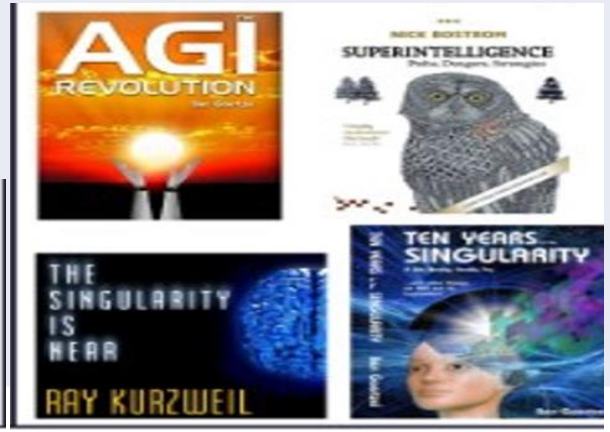
Так искусственный интеллект, перешел от экспертного обучения 1-го поколения и баз знаний, созданных вручную, к глубокому обучению 2-го поколения, использующему нейросети и большие обучающие выборки. Теперь мы вступаем в 3-е поколение ИИ, где система искусственного интеллекта может интерпретировать и объяснить алгоритм принятия решений, даже если он имеет природу черного ящика. Объяснимый искусственный интеллект являются основной частью 3-го поколения ИИ.

В 2030-х годах мы увидим ИИ 4-го поколения с машинами, которые сами будут учатся обучатся и будут динамически накапливать новые знания и навыки. К 2040-м годам ИИ 5-го поколения построит системы искусственного интеллекта с воображением, которые не будут полагаться на людей в объяснении.

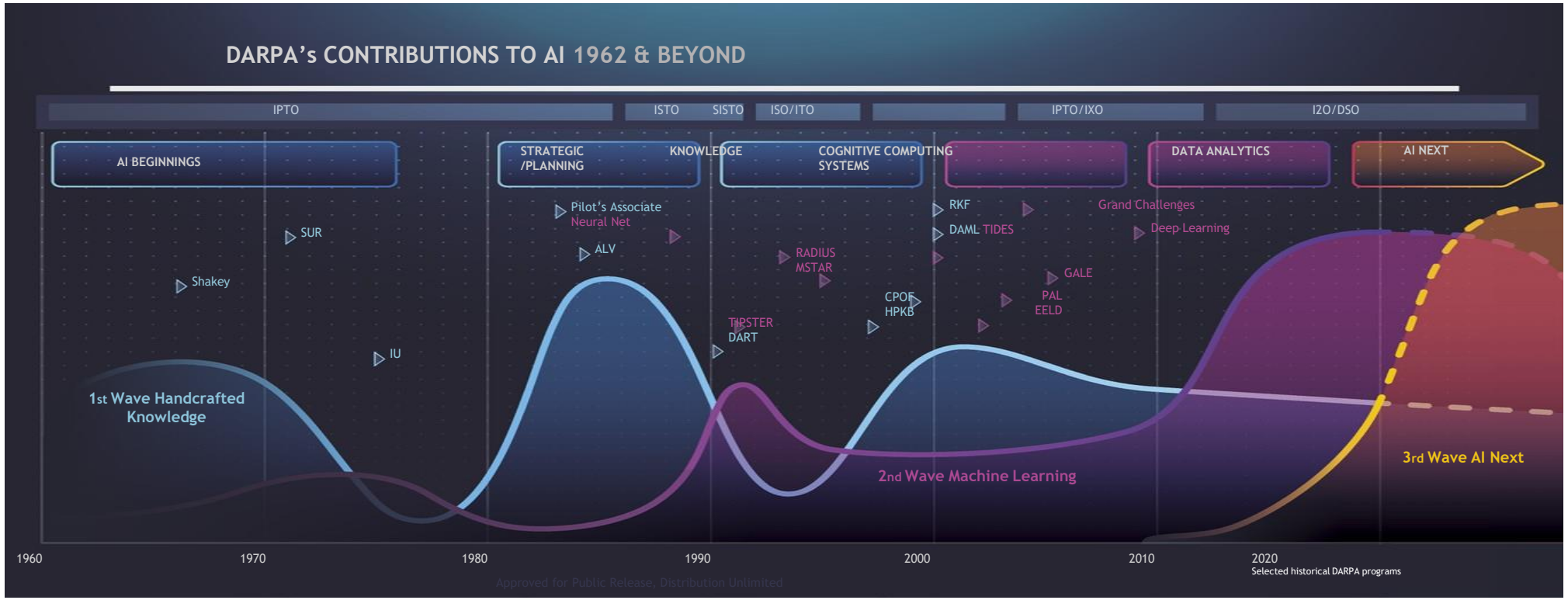
Пять волн искусственного интеллекта

Первая волна	Вторая Волна	Третья Волна	Четвертая и пятая волна
1970 – 1990 гг.	2000-2020 гг.	2020-2030 гг.	2030 ------→
Хорошо рассуждает, но есть проблемы с обучением и обобщением. Символический, эвристический, основанный на правилах	Хорошо обучается и воспринимает, но слаб в рассуждении и обобще-нии. Статистическое обучение, глубокое обучение, обработка текста и изображений	Прекрасно обучается, обобщает и рассуждает. Контекстуальная адаптация, способен объяснять решения. Способен общаться на ЕЯ. Нужны меньшие объемы данных для обучения и мини-мальный внешний контроль	Способен решать те же интеллектуальные задачи, что и человек. Сильный искусственный интеллект, ведущий к Суперинтеллекту и «технологической сингулярности»





3 волны инвестиций в ИИ агества DARPA с 1962 года



На графике показано три волны инвестиций DARPA в ИИ, высота каждой волны соответствует величине инвестиций DARPA.

Роль объяснимого искусственного интеллекта в становлении пятой промышленной революции.

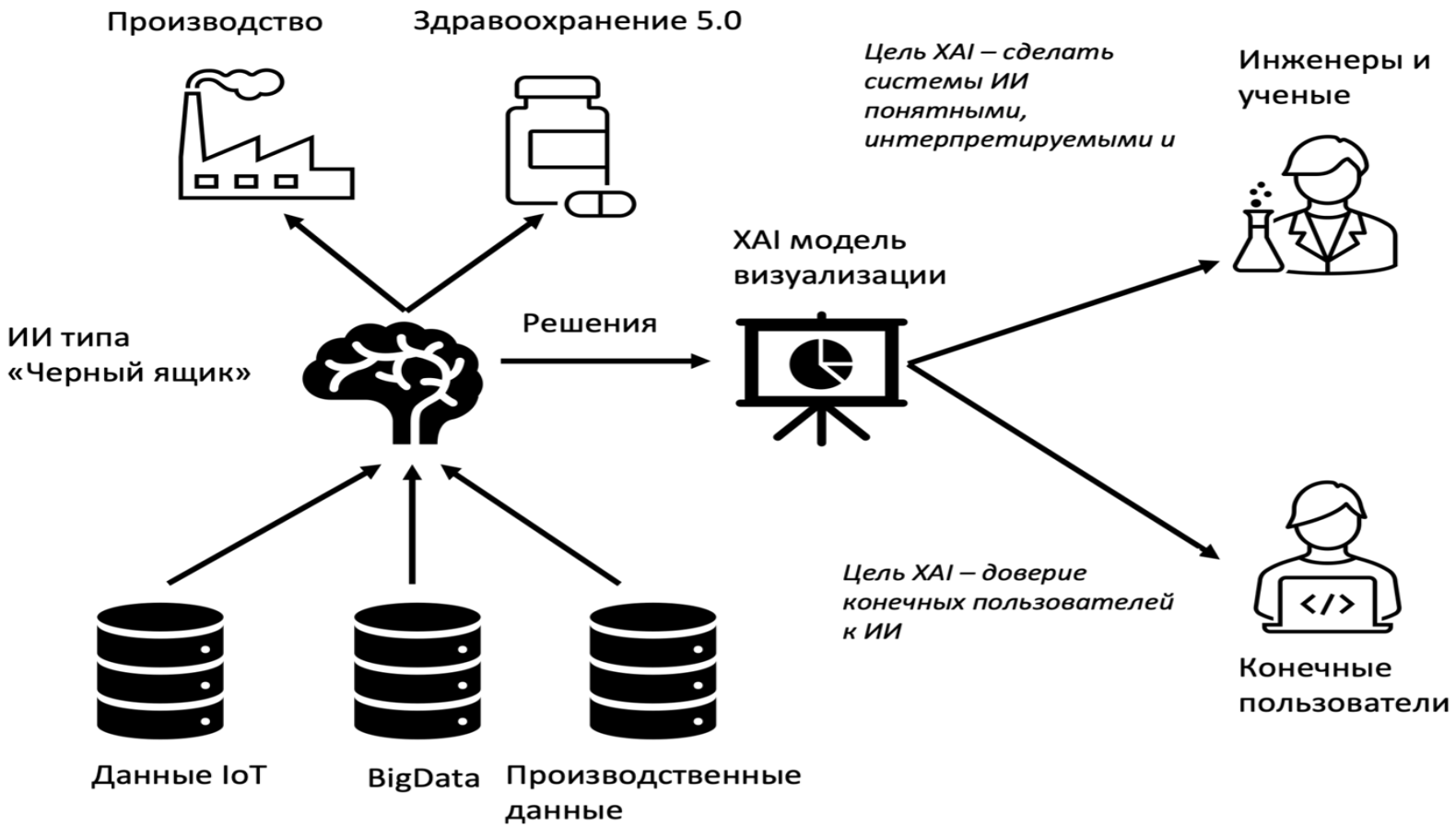
Пятая промышленная революция сосредоточится на интеллектуальном производстве, вернув человеческий интеллект на производство, позволив роботам не заменять людей, а наоборот сотрудничать и помогать им.

В производственных системах, ориентированных на человека, ключевая миссия цифровых технологий, описанная в рамках пятой промышленной революции, заключается в том, чтобы объяснить причины решений, принимаемых встроенными моделями искусственного интеллекта в системах промышленной автоматизации для того, чтобы обеспечить взаимодействие человека и технологий в производственном цикле.

Объяснимый искусственный интеллект (XAI) - один из подходов к решению этой проблемы. Цель XAI – предоставить алгоритмам искусственного интеллекта (ИИ) описательную функциональность – способность сообщить человеку об основных шагах, предпринятых для достижения решения.

Применимость XAI для Индустрии 5.0 заключается в том , что XAI помогает конечным пользователям доверять решению ИИ, в то же время XAI и позволяет инженерам или ученым полностью понять процесс работы системы ИИ.

Применимость XAI для Индустрии 5.0



Рынок объяснимого ИИ - эксклюзивный отчет 360iResearch

- Объяснительный интеллект (XAI) играет ключевую роль в развитии и применении систем искусственного интеллекта в экономике данных, способствуя их прозрачности, пониманию и доверию со стороны пользователей и заинтересованных сторон.
- XAI имеет решающее значение, особенно в приложениях, где решения, принимаемые с помощью ИИ, влияют на людей или предприятия. Растущая потребность в подотчетности и этичности ИИ, особенно в таких отраслях, как финансы, здравоохранение и юриспруденция, повышает спрос на объяснимый ИИ.
- Растущая сложность моделей ИИ и их приложений повышает спрос на объяснимость, обеспечивающую понятность и обоснованность решений. Ожидается, что постоянные инновации в области разработки новых решений в области объяснимого искусственного интеллекта поставщиками рынка создадут возможности для его роста.

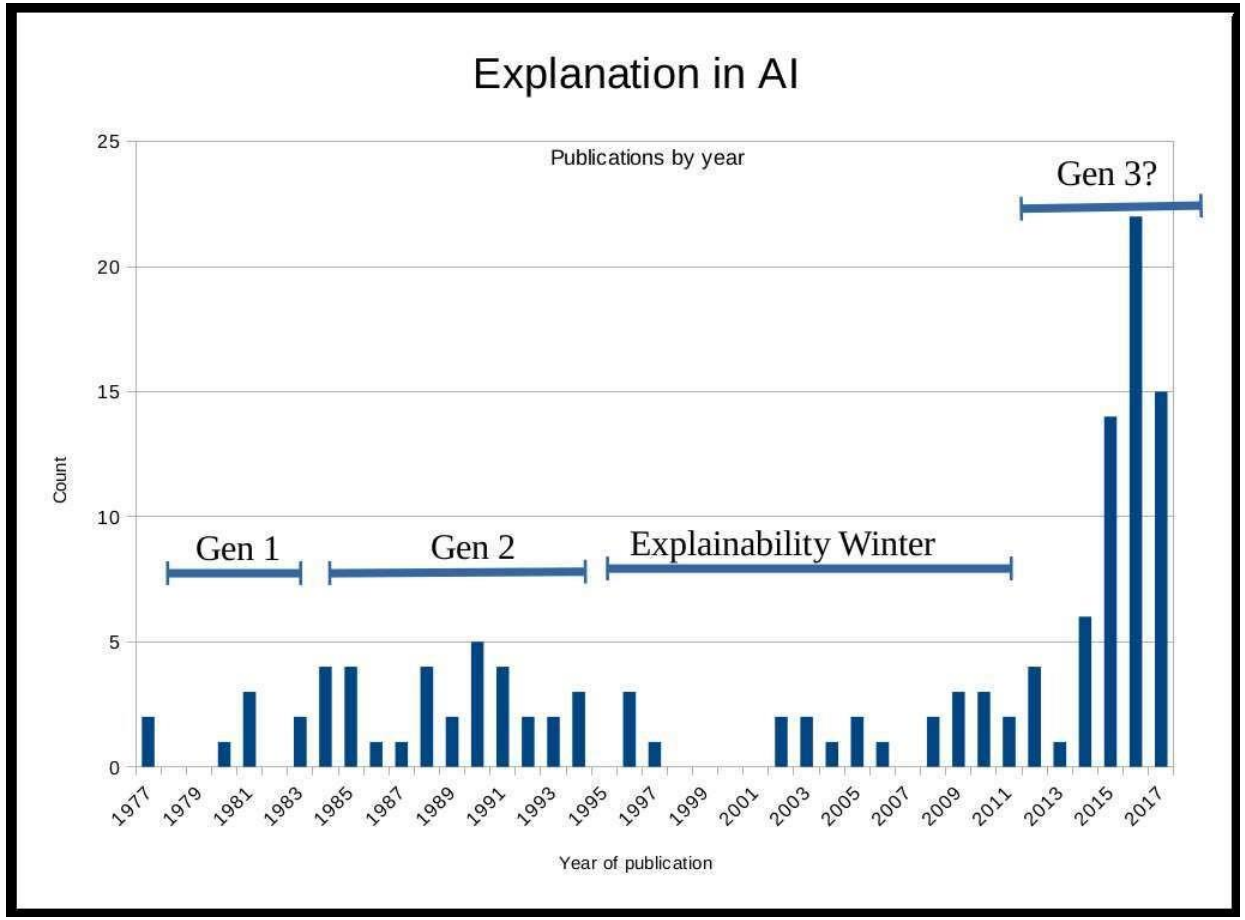
История развития объяснительного ИИ ()

Исследования в этой области можно разделить на три этапа:

- первый этап (с 1970 года) - развитие автономных экспертных систем,экспертных систем,
- второй этап (середина 1980-х годов) - переход от экспертных систем к системам, основанным на знаниях (декларативных),
- и третий (с 2010 года) - внедрение глубоких архитектур искусственных нейронных сетей, что потребовало новых глобальных исследований по построению объясняемых систем.

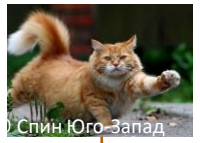
Первое и второе поколения объяснимого ИИ связаны с экспертными системами, включавшими в себя принятие решений и постановку диагнозов и содержащий представления, основанные на использовании правил и отношений. Системы имели инструментарий вопрос-ответных интерфейсов для пользователей и могли давать рекомендации и ставить диагнозы и содержали блок объяснений, дающий ответы на вопросы КАК? и ПОЧЕМУ? с помощью прохода по дереву вывода вверх и вниз.

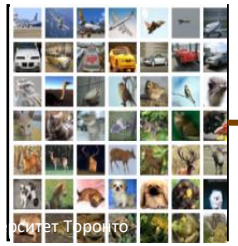
Рост числа публикаций по ИИ



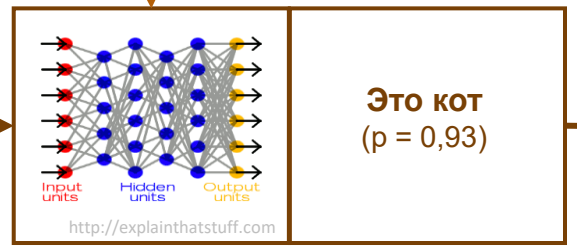
Что мы пытаемся сделать в ОИИ?

Сегодня

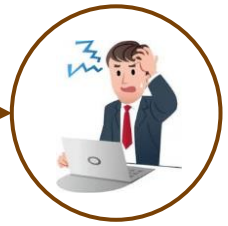




Обучение
Процесс



Это кот
(p = 0,93)



Обучение
Данные

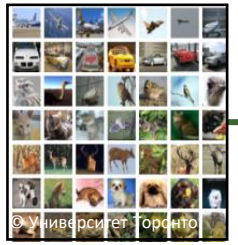
Аппроксим.
функция

Выход

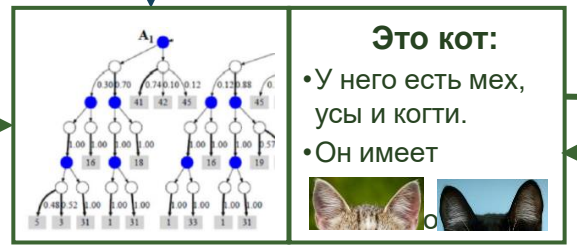
Пользователь
с
задачей

- Почему ты это сделала?
- Почему не что-нибудь другое?
- Когда тебе это удастся?
- Когда ты терпишь неудачу?
- Когда я могу тебе доверять?
- Как исправить ошибку?

Завтра



Новый
Обучение
Процесс



Это кот:
- У него есть мех, усы и когти.
- Он имеет

Обучение
Данные

Объясняемая
модель

Пояснение
Интерфейс

Пользователь с
задачей

- Я понимаю почему
- Я понимаю почему нет
- Я знаю, когда ты добьешься успеха
- Я знаю, когда ты проиграешь
- Я знаю когда тебе доверять
- Я знаю, почему ты ошиблась

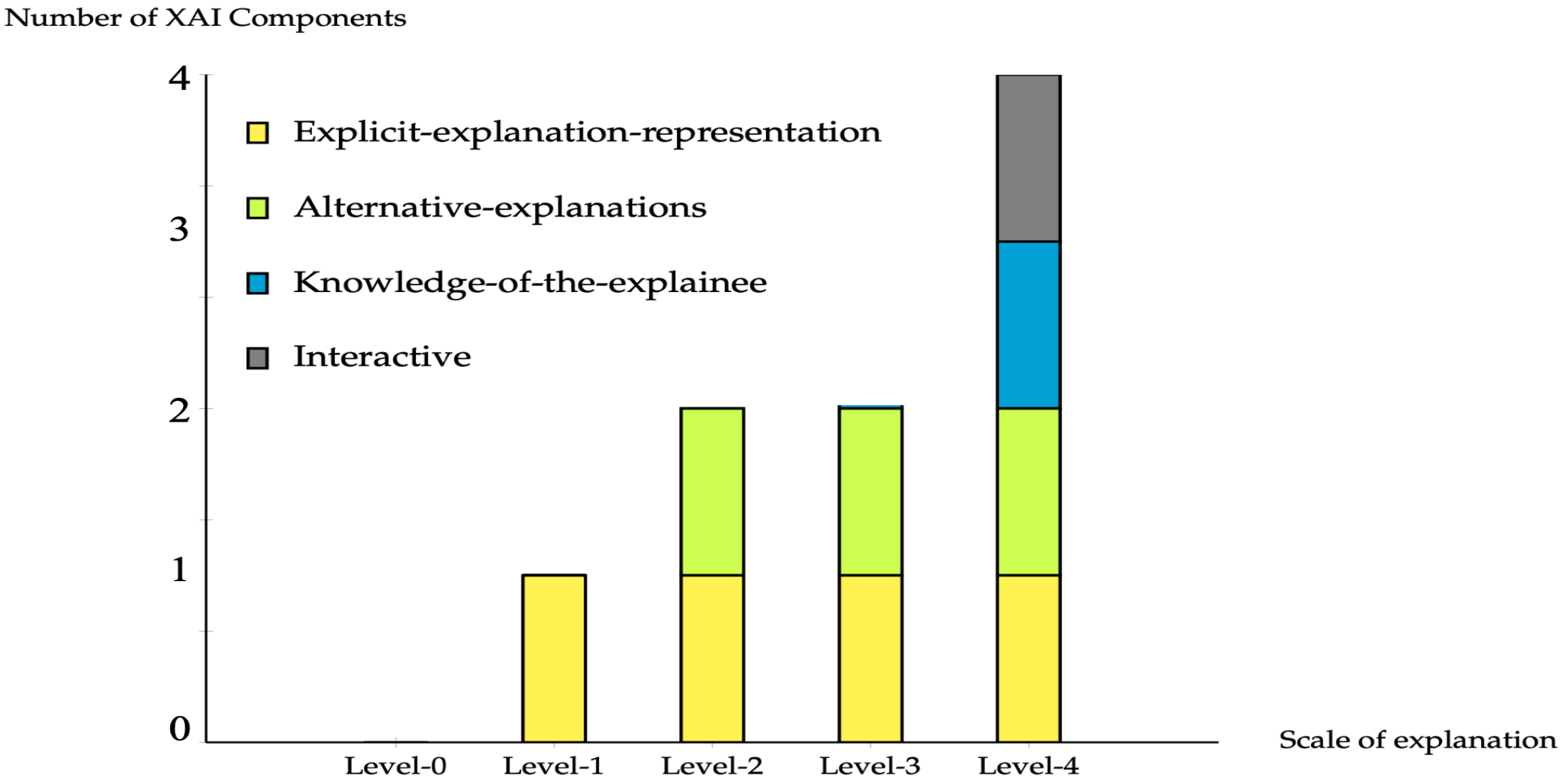
4 методических логико-математических компонента требований к ОИИ

Мы выделяем четыре основополагающих методических логико-математических компонента требований к ОИИ, составляющие:

- требования к точному представлению знаний об объяснении,
- предоставление альтернативных объяснений,
- корректировку объяснений на основе знаний лица, для которого предназначено объяснение
- использование преимуществ интерактивного объяснения.

Учитывая эти четыре компонента, составляется стратегический перечень требований к ОИИ в задачах распознавания цифровых изображений и в больших речевых моделях. Рассматривается использование ОИИ для интеграции различных атрибутов систем ИИ в целях компенсации недостатков и объединения преимуществ различных моделей, используемых для решения неоднородных задач в системах поддержки принятия решений

Основные объясняющие компоненты и их потенциальная роль в шкале объяснений



Kim, M.-Y. et al. A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 900–921. ht

Уровень 0

Приложение XAI не предоставляет никаких объяснений Модели, отнесенные к уровню 0, не предоставляют никаких объяснений вообще. По сути, это модели "черного ящика", которые не предоставляют никакой пояснительной информации пользователю. Другими словами, ожидается, что объясняющий примет или отвергнет предсказание системы без какой-либо дополнительной информации. Большинство готовых методов изучения классификаторов (например, модели глубокого обучения, методы опорных векторов или случайные леса) относятся к этому уровню. Обратите также внимание, что к этой категории относится любая прогностическая система, независимо от того, основана ли она на моделях машинного обучения или нетю

Уровень 1

Приложение XAI предоставляет подробные объяснения В моделях, относящихся к этой категории, объясняемому пользователю предоставляется один тип объяснений . Например, к этому уровню относится фреймворк, который предоставляет тепловые карты для объяснения классификации изображений . Большая часть этого подхода сосредоточена на предоставлении постфактум объяснения, которое переводит модель черного ящика, первоначально относившуюся к уровню 0, на модель уровня 1.

Уровень 2

Приложение XAI предоставляет дополнительные объяснения Уровень 2 добавляет еще одно дополнение к объяснениям, чтобы улучшить знания, передаваемые объясняемым. На этом уровне, например, система объяснения классификации изображений может предоставлять не только тепловую карту, объясняющую, что изображение классифицированного животного относится к кошкам, но и другой тип объяснения, такой как текстовое объяснение, в котором отмечается смещение прогностической модели в сторону категории (например, смещение в сторону кошек). в качестве альтернативного описания прогнозируемой классификации. Таким образом, альтернативные объяснения позволяют объясняемому получить более глубокое представление о процессе рассуждения, используемом системой для составления прогноза.

Уровень 3

Приложение с искусственным интеллектом предоставляет дополнительные объяснения, понятные для пользователя На этом уровне важную роль играет сам пользователь и его знакомство с предметной областью приложения . Этот уровень объяснения системы XAI, в дополнение к предоставлению множества объяснений, включает в себя некоторую модель знаний объясняемого в предметной области и позволяет выбрать подходящий тип объяснения в соответствии со знаниями объясняемого. Например, рассмотрим ситуацию, когда у пациента диагностировано какое-либо заболевание и система искусственного интеллекта используется для проведения потенциального лечения. В то время как терапевту требуется подробное медицинское объяснение с помощью системы искусственного интеллекта, пациент предпочел бы получить объяснение от непрофессионала относительно любых рекомендаций по альтернативному лечению.

Уровень 4

Приложение XAI в интерактивном режиме предоставляет эксплицитному пользователю дполонительные объяснения и вступает с ним в диалог.
Если предыдущие уровни (например, уровень 0, 1, 2, 3) не предусматривали возможности взаимодействия с объясняемым, за исключением, возможно, запроса альтернативных целевых объяснений (уровень 3), методы XAI, относящиеся к уровню 4, могут взаимодействовать с пользователем. Ожидается, что они будут поддерживать возможность беседы, которая позволяет объясняемому уточнить свои вопросы и опасения относительно предсказания. Другими словами, каждое взаимодействие в разговоре позволяет объясняемому получить разъяснения. Здесь система способна адаптировать свои объяснения к динамической информации, полученной в результате взаимодействия с объясняемым.. Насколько нам известно, существующие системы не имеют такой возможности взаимодействия..

Уровень 5

Как следует из прогноза развития ИИ, скоро возникнет 5-й уровень, с самообъяснением и объяснением ИИ агентами друг к другу, т.е. нейронные сети должны научиться объяснить самим себе, как они получили то или иное решение (как отмечали Ян Лекун и Ю.И. Вильзитер).

4 принципа объяснимого искусственного интеллекта-1

Национальный институт стандартов и технологий (NIST) опубликовал проект перечня принципов объяснимого искусственного интеллекта (XAI).

- **Объяснение (Explanation).** Системы ИИ должны предоставлять причины и обстоятельства, на основании которых были приняты те или иные решения. Принцип объяснения обязывает систему ИИ предоставлять объяснение в форме «свидетельства или обоснования каждого результата». Данный принцип не устанавливает никаких дополнительных требований к качеству объяснения, а лишь требует, чтобы система ИИ была способна предоставить объяснение. Стандарты таких объяснений регулируются другими принципами.

- **Значимость (Meaningful).** Системы объяснимого ИИ должны представлять объяснения, понятные отдельным пользователям.

Принцип значимости устанавливает, что получатель объяснения должен быть в состоянии понять объяснение. В документе подчеркивается, что этот принцип не предназначен для универсального применения. Пояснения должны быть адаптированы к аудитории как на групповом, так и на индивидуальном уровне. Так, например, разные типы групп пользователей могут требовать разных объяснений, а имеющиеся у них знания и опыт могут влиять на восприятие результата и его значимость.

4 принципа объяснимого искусственного интеллекта-2

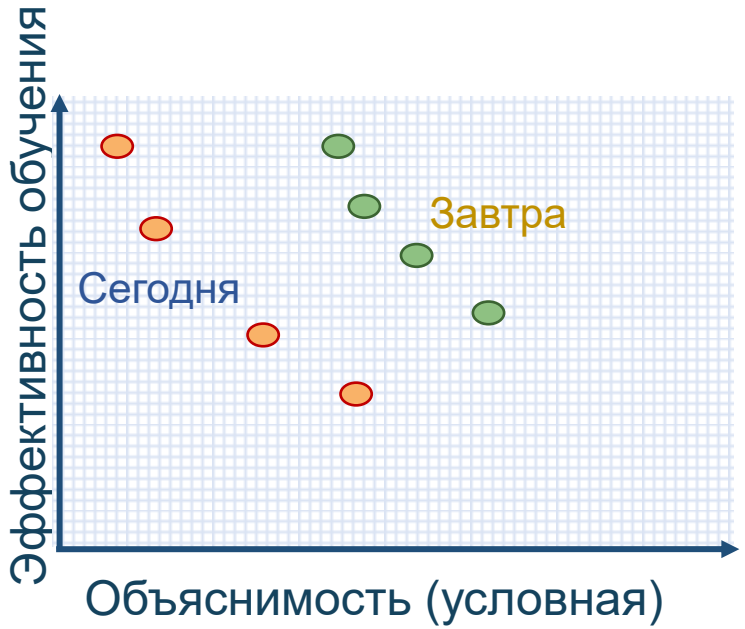
- **Точность объяснения (Explanation Accuracy).** Объяснение должно достоверно отражать суть процессов, производимых системой ИИ для генерирования результатов. Принцип точности объяснения корреспондирует с принципом значимости для регулирования качества объяснений, предусматривая точность объяснений, но не точность решений. Фактически, данный принцип является подробным разъяснением того, как система сгенерировала окончательный результат. Применение данного принципа также ставится в зависимость от контекста и конечного пользователя. Так, разные показатели точности объяснения будут представляться для разных типов групп и пользователей.

- **Пределы знаний (Knowledge Limits).** Система работает только в условиях, для которых она была разработана, или когда система достигает надлежащей достоверности в своих результатах. Принцип пределов знаний требует, чтобы система отмечала любые случаи, для которых она не была разработана. Целью этого принципа является предотвращение вводящих в заблуждение объяснений или выводов системы.

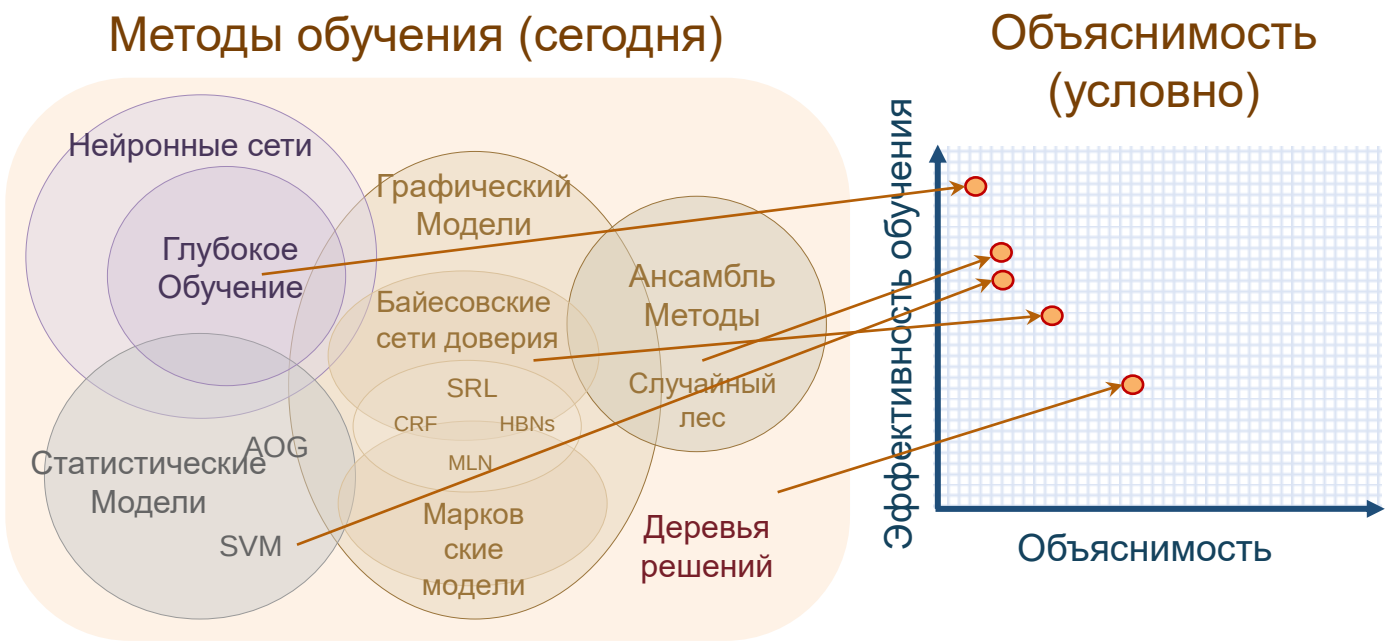
Указанные четыре принципа показывают, что решения на основе ИИ должны обладать необходимой прозрачностью, чтобы вызывать доверие к своему функционированию и уверенность в выводах системы

Цель ОИИ: производительность и объяснимость

- XAI создает набор методов машинного обучения, которые
  - создают более объяснимые модели, сохраняя при этом высокий уровень эффективности обучения (например, точность прогнозов)
  - дают возможность пользователям-людям понимать, должным образом доверять и эффективно управлять новым поколением партнеров с искусственным интеллектом



Производительность против объяснимости



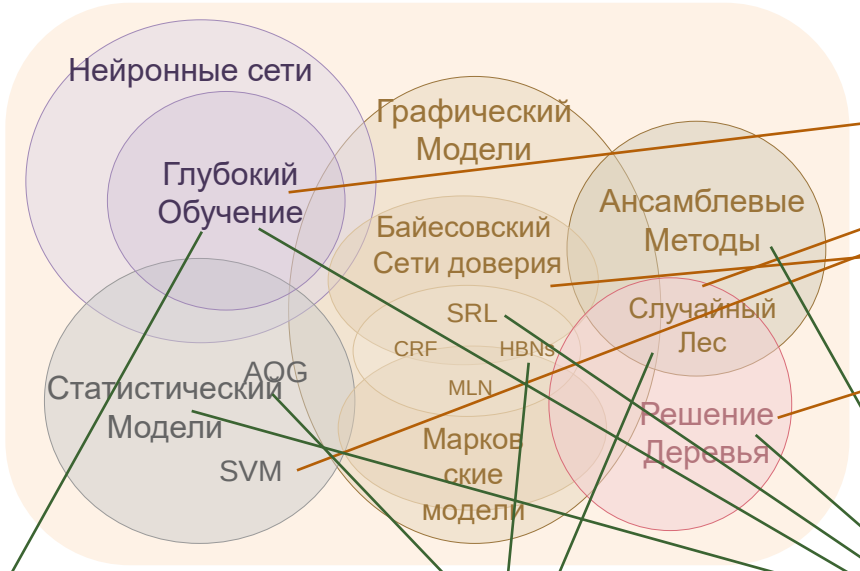
- AOG - стохастические И/ИЛИ графы
- SVM - метод опорных векторов
- MLNs - марковские логические сети
- HBNs – иерархические байесовские сети
- CRFs - условные случайные поля
- SRL - статистическое реляционное обучение

Объяснительные модели ИИ

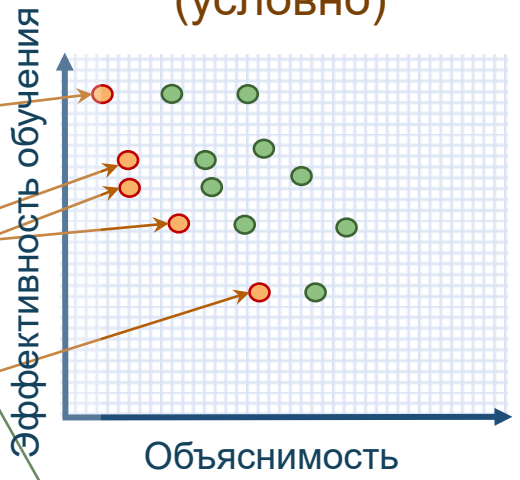
Новый Подход

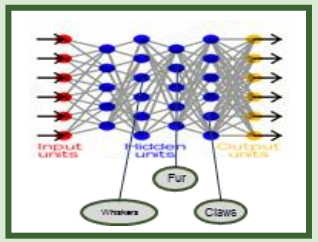
Методы машинного обучения, которые создают более объяснимые модели, сохраняя при этом высокий уровень обучаемости

Методы обучения



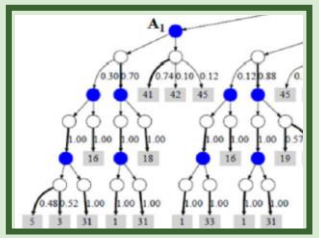
Объяснимость (условно)





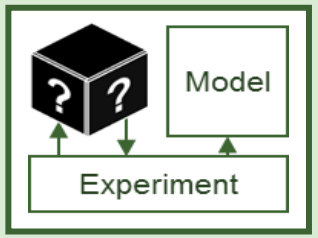
Глубокое объяснение

Модифицированные методы глубокого обучения объяснимым признакам



Интерпретируемые модели

Методы изучения более структурированных, интерпретируемых причинно-следственных моделей

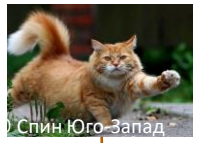


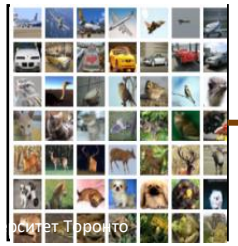
Модель индукции

Методы вывода объяснимой модели из любой модели в виде черного ящика

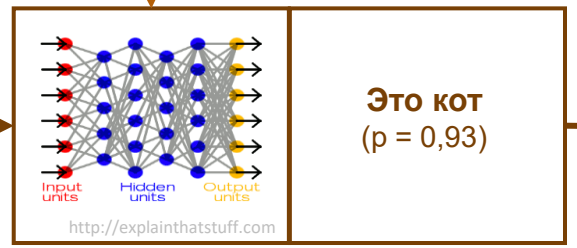
Что мы пытаемся сделать в ОИИ?

Сегодня

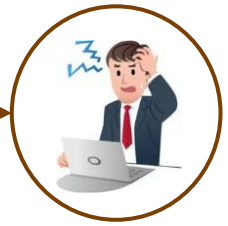




Обучение
Процесс



Это кот
(p = 0,93)



Обучение
Данные

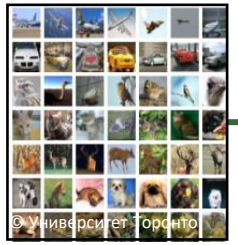
Аппроксим.
функция

Выход

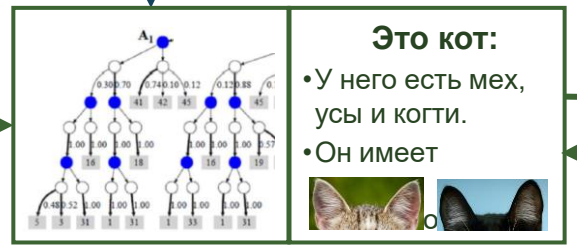
Пользователь
с
задачей

- Почему ты это сделала?
- Почему не что-нибудь другое?
- Когда тебе это удастся?
- Когда ты терпишь неудачу?
- Когда я могу тебе доверять?
- Как исправить ошибку?

Завтра



Новый
Обучение
Процесс



Это кот:
- У него есть мех, усы и когти.
- Он имеет

Обучение
Данные

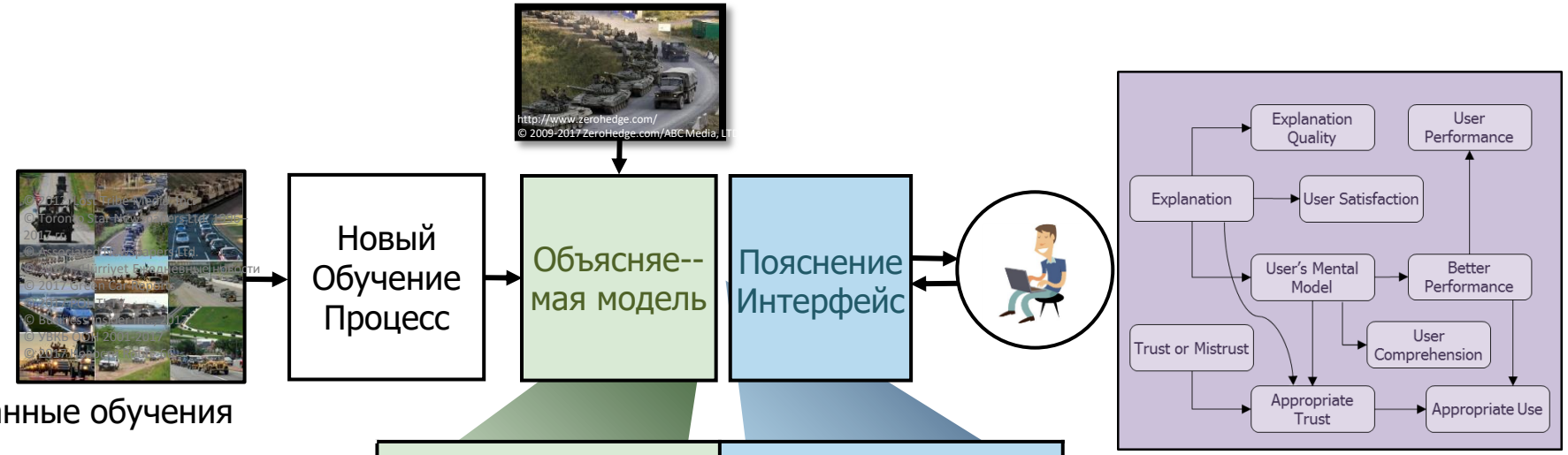
Объясняемая
модель

Пояснение
Интерфейс

Пользователь с
задачей

- Я понимаю почему
- Я понимаю почему нет
- Я знаю, когда ты добьешься успеха
- Я знаю, когда ты проиграешь
- Я знаю когда тебе доверять
- Я знаю, почему ты ошиблась

Концепция XAI в DARPA и технические подходы



IHMC

Психологическая модель объяснения

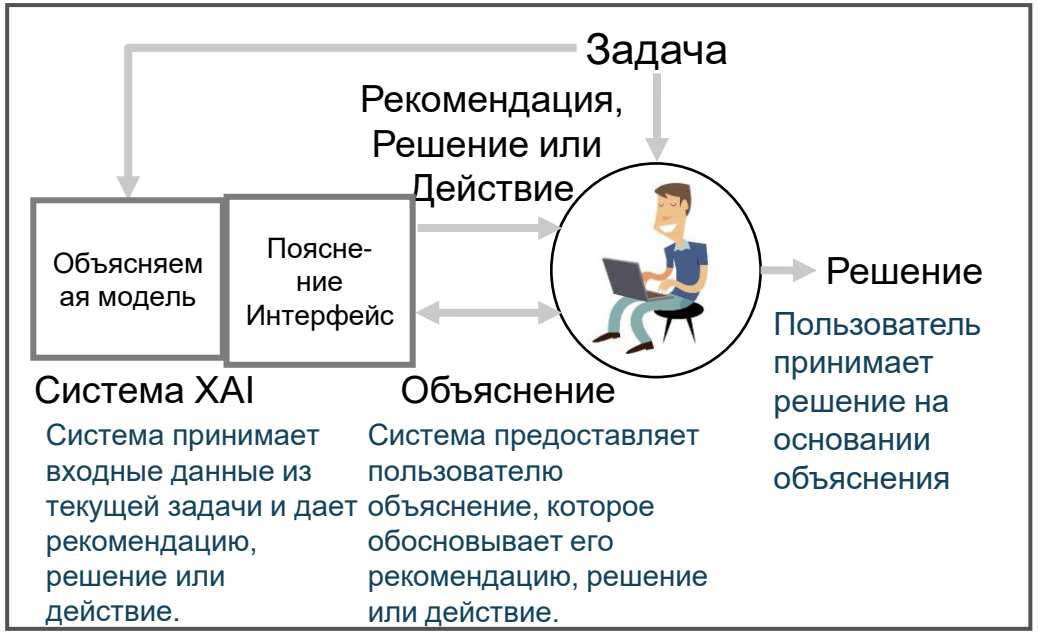
Калифорнийский университет в Беркли	Глубокое обучение	Рефлексивный и рациональный
Чарльз Ривер	Индуктивная каузальная модель	Нарративная генерация
UCLA	Теория паттернов +	Трехуровневое объяснение
OSU	Адаптивные программы	Приемочное тестирование
PARC	Когнитивное моделирование	Интерактивное обучение
CMU	Объяснимый RL (XRL)	XRL Взаимодействие
SRI	Глубокое обучение	Покажи и расскажи объяснения
Raytheon	Глубокое обучение	Аргументация и педагогика
UT Даллас	Вероятностный Логика	Диаграммы решений
Техас A&M	Мимическое обучение	Интерактивная визуализация
Rutgers	Модель индукции	Байесовское обучение

Технические подходы разработчиков XAI

Команда	Объясняемая модель	Объясняющий интерфейс	Решаемые задачи
Университет Беркли	-Объяснения постфактум путем обучения дополнительных моделей DL -Явные интроспективные объяснения (NMN) -Обучение с подкреплением (информативные развертывания, явный модульный агент)	-Рефлексивные объяснения (вытекающие из модели) -Рациональные объяснения (основанные на рассуждениях об убеждениях пользователя)	-Автономность: управление транспортным средством (BDD-X, CARLA), стратегические игры (StarCraft II) - Аналитика: визуальный контроль качества и фильтрация задач (VQA-X, ACT-X, xView, DiDeMo и т. д.)
Charles River Analytics .	Эксперимент с обученной моделью, чтобы создать объяснимую причинно-вероятностную модель программирования	Интерактивная визуализация, основанная на генерации временных, пространственных сообщений на основе причинно-следственных, вероятностных моделей	Автономность: Atari, StarCraft II Аналитика: обнаружение пешеходов (INRIA), распознавание активности (ActivityNet)
Университет штата Орегон	xDAP, комбинация адаптивных программ, глубокое обучение и объяснимость	Обеспечивает чередование визуального и ЕЯ объяснения для приемочных испытаний пилотами-испытателями на основе IFT	Автономность: одни и те же стратегии в реальном времени, основанные на специально разработанном игровом движке, поддерживающем объяснения; StarCraft II

Измерение эффективности объяснения

Структура объяснения



Мера эффективности объяснения

Удовлетворенность пользователей

- Ясность объяснения (оценка пользователей)
- Полезность объяснения (оценка пользователей)

Ментальная модель пользователя

- Понимание индивидуальных решений
- Понимание общей модели
- Оценка сильных / слабых сторон
- Прогноз "Хто он будет делать"
- Прогноз "Как мне вмешаться"

Выполнение задач

- Улучшает ли объяснение решение пользователя, выполнение задачи?
- Искусственный задачи решения введены для диагностики понимания пользователя

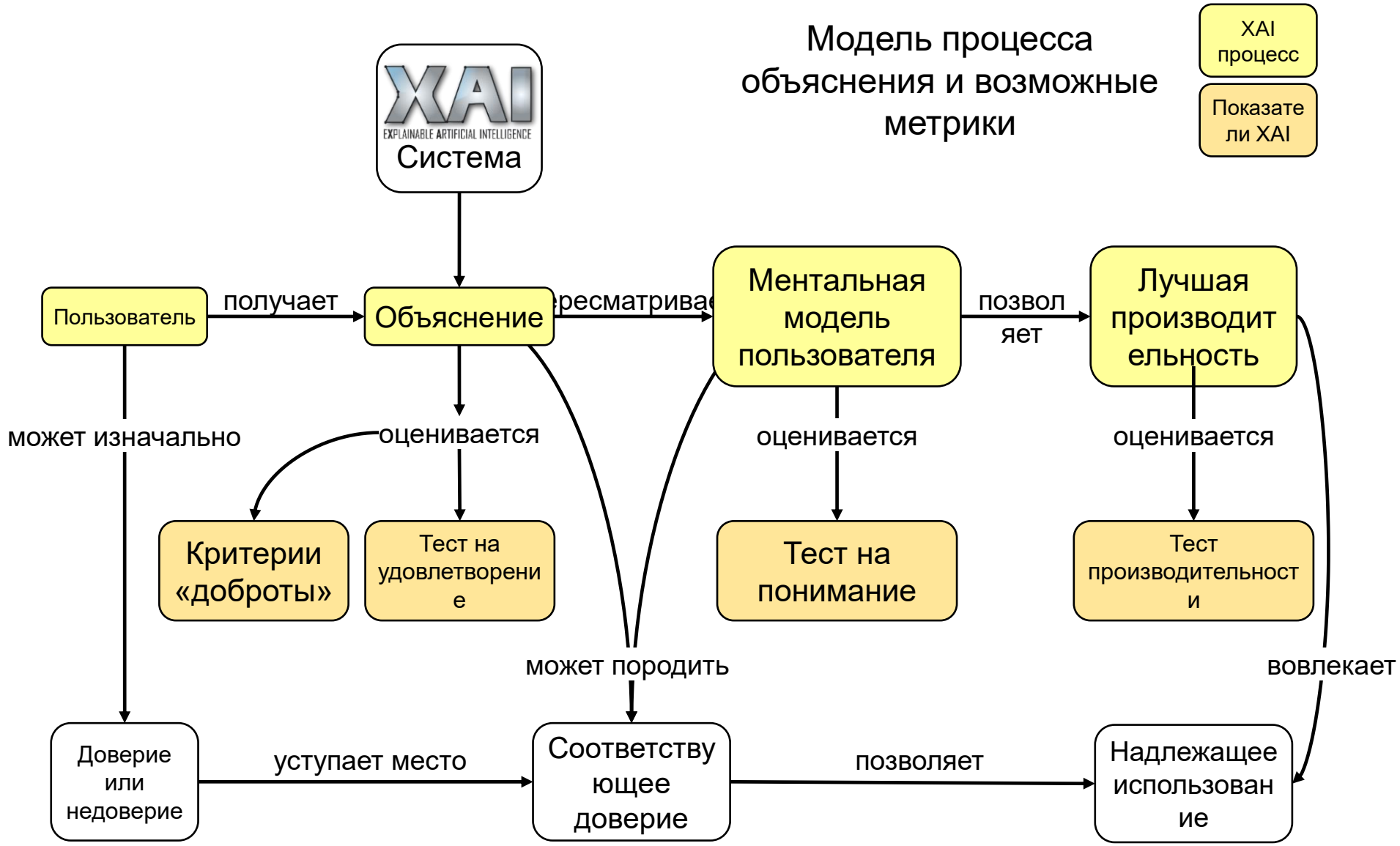
Оценка доверия

- Надлежащее использование в будущем и доверие

Исправляемость (дополнительная оценка)

- Выявление ошибок
- Исправление ошибок
- Непрерывное обучение

Концептуальная модель объяснений



Построение когнитивных метрик для XAI

Методы когнитивной науки и психологии, которые традиционно используются для вывода когнитивных процессов, лежащих в основе человеческого поведения, могут быть использованы для изучения механизмов, лежащих в основе поведения системы искусственного интеллекта (ИИ), и предоставления объяснений. Эти методы в когнитивной науке и психологии включают в себя:

- измерение и сравнение поведения системы в различных условиях;
- изучение основных факторов, объясняющих поведение, с помощью регрессионного/факторного анализа;
- создание когнитивных/прогностических моделей для обобщения или моделирования поведения

Здесь возникают измерительные процессы, связанные с совместной обработкой данных и знаний и получением на основе этой интеграции метрологически обоснованных решений, которые были названы интеллектуальными измерениями (ИнИ).

Согласно гипотезе, хорошие и удовлетворяющие пользователей объяснения позволяют им создать хорошую ментальную модель. В свою очередь, хорошая ментальная модель позволит им развить соответствующее доверие к ИИ и добиться хороших результатов при его использовании. Для оценки этой модели процесса объяснения требуется несколько типов мер.

Метрики для интеллектуальных измерений

Система XAI по существующим стандартам должна быть понятна каждому пользователю, поэтому система XAI может быть оценена с точки зрения когнитивных состояний и процессов пользователя в процессе объяснения XAI, по 7 когнитивным метрикам:

- качество объяснения,
- удовлетворенность пользователя,
- любопытство/вовлеченность пользователя,
- доверие пользователя/доверие,
- понимание пользователя,
- производительность/производительность пользователя
- управляемость/взаимодействие системы.

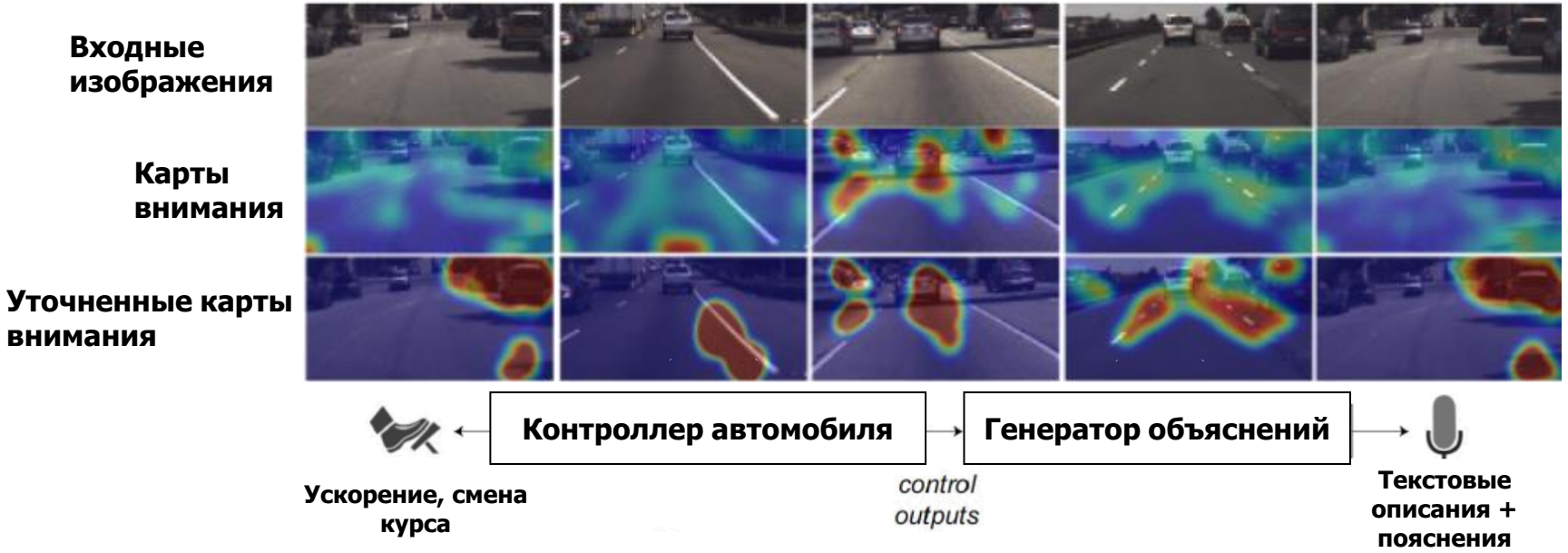
Некоторые метрики могут быть связаны друг с другом. Тип мер, подходящих для каждого показателя, может отличаться. Предлагается пять основных типов измерений, к*о*торые можно рассматривать в контексте 7 когнитивных метрик:

- субъективные измерения внешних стимулов,
- субъективные измерения внутренних состояний,
- объективные измерения когнитивных состояний/процессов,
- субъективные/объективные измерения, основанные на когнитивных моделях,
- субъективные/объективные измерения временной динамики

Объяснимый искусственный интеллект для беспилотных автомобилей

Калифорнийский университет в Беркли

Система текстового согласования, встроенная в усовершенствованные модели визуального внимания, чтобы обеспечить соответствующее объяснение поведения глубокого нейросетевого контроллера транспортного средства.



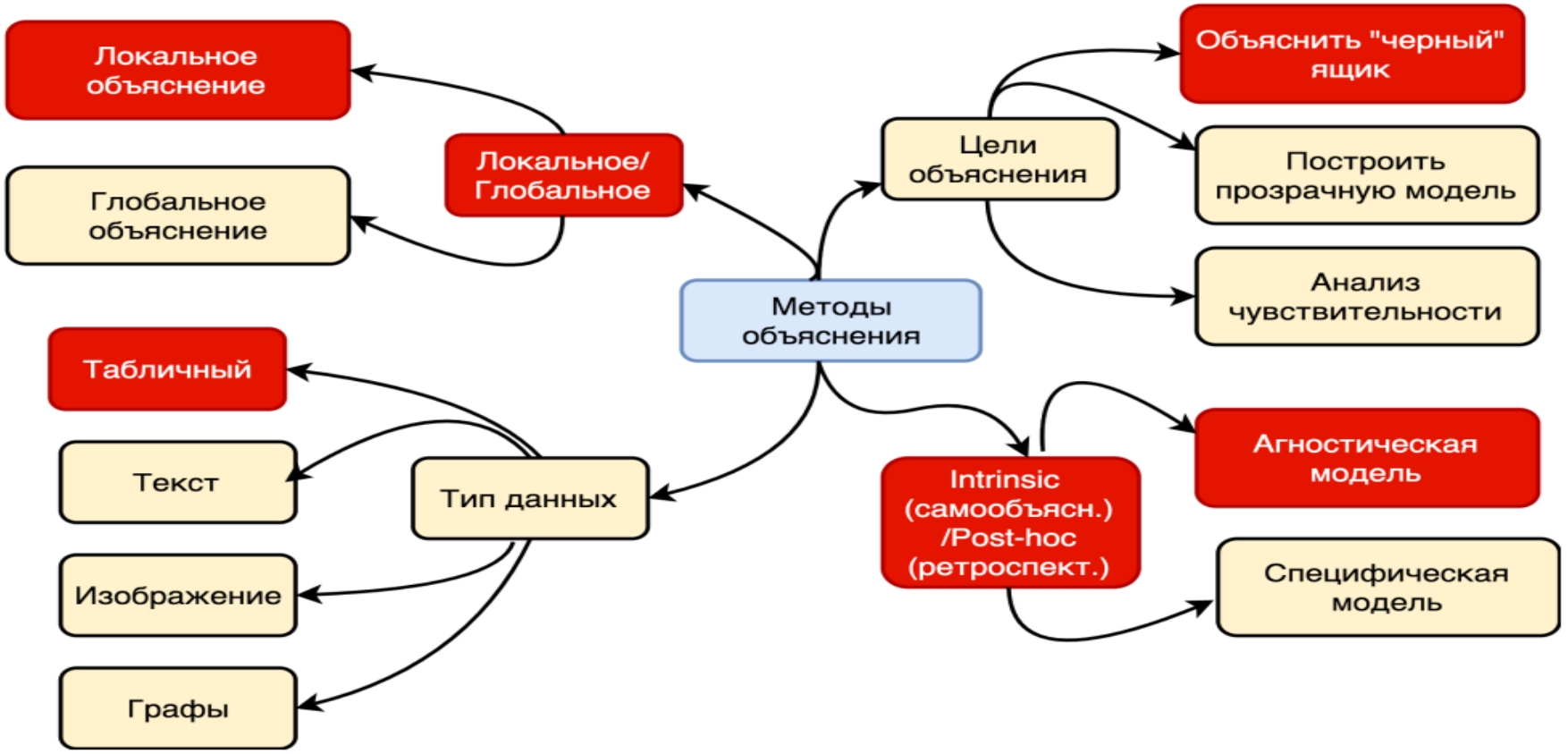
Примеры описания и обоснования действий

Описание действия	Обоснование действия
Автомобиль разгоняется	**так как** свет стал зеленым
Автомобиль медленно разгоняется	**так как** свет загорелся зеленым и движение идет
Машина едет вперед	**в качестве** движение транспорта свободно
Автомобиль переходит в левую полосу движения	**к** обойти более медленную машину перед ней

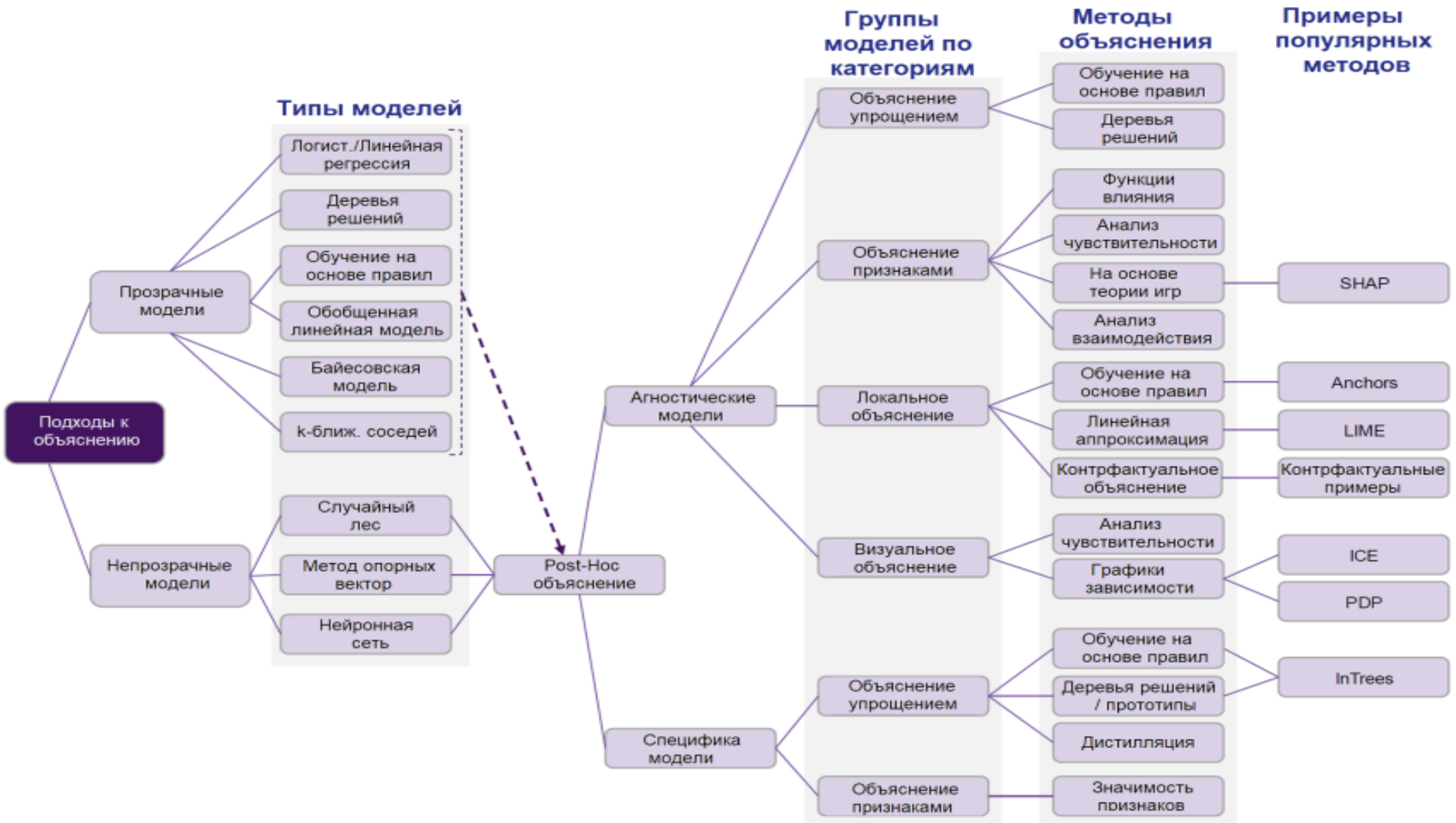
Без объяснения причин: *«Машина едет по улице»*
С объяснением: *«Машина едет по улице, потому что на ее полосе нет других машин, нет красных светофоров и знаков остановки»*

- Уточненные тепловые карты дают более сжатые визуальные объяснения и более точно отображают поведение сети
- Текстовое описание и обоснование действия обеспечивает удобную для интерпретации систему для беспилотных автомобилей.

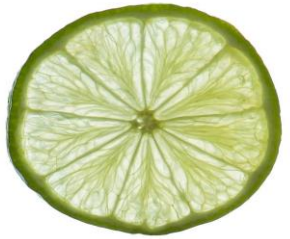
Классификация моделей для объяснений

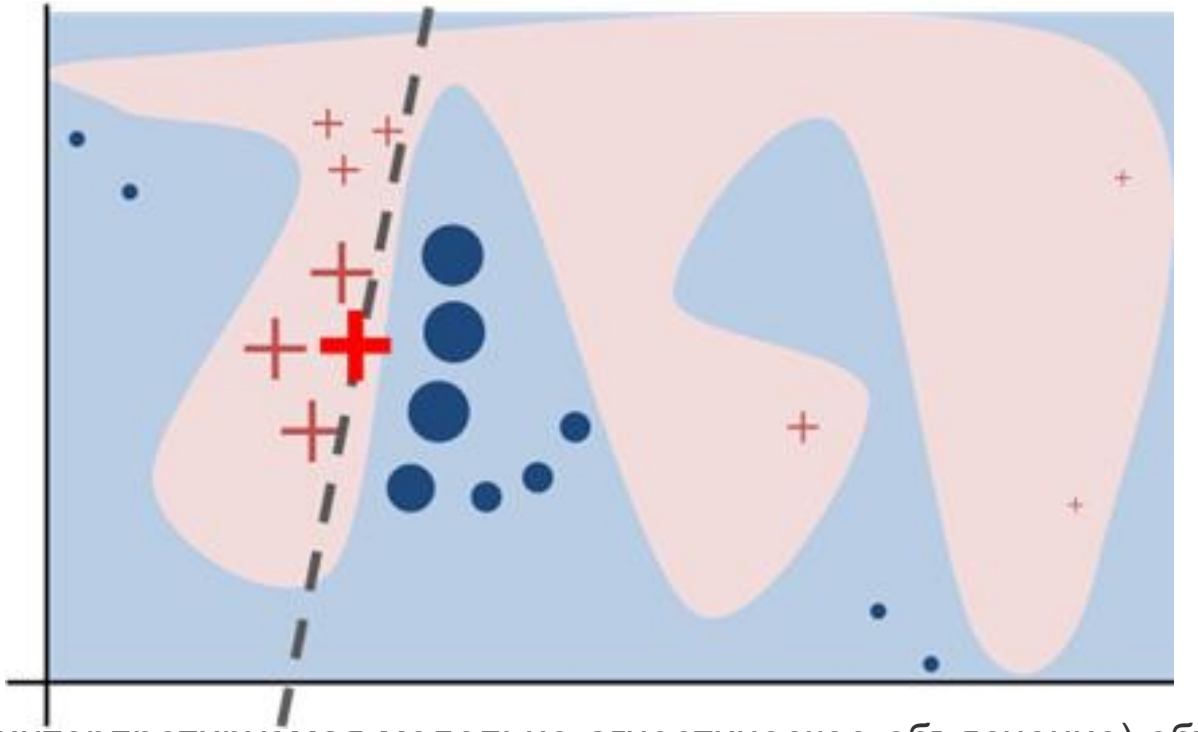


Классификация моделей XAI



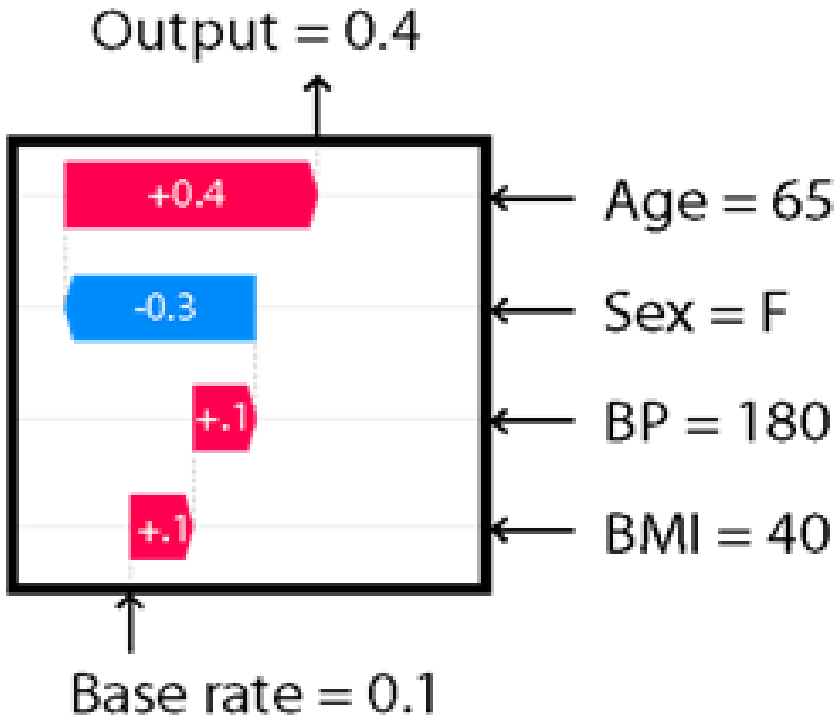
LIME локально интерпретируемое модели агностическое объяснение)





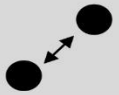
LIME (локально интерпретируемая модельно-агностическое объяснение) объясняет классификатор для конкретного единичного экземпляра и поэтому подходит для локального рассмотрения. Интуитивно понятно, что объяснение — это локальное линейное приближение поведения модели. Хотя модель может быть очень сложной в глобальном масштабе, ее легче приближать в непосредственной близости от конкретного случая.

SHAP (Аддитивное объяснение Шепли)



SHAP (Аддитивное объяснение Шепли) — еще один метод объяснения индивидуальных прогнозов. SHAP основан на игре теоретически оптимальных значений Шепли. Техническое определение значения Шепли — это “средний предельный вклад значения характеристики во все возможные коалициям...

Суррогатные методы объяснительного ИИ – это подход, основанный на построении более простой, интерпретируемой модели (линейные модели, деревья решений) для объяснения исходной сложной. Наиболее популярными методами из этой группы являются LIME и SHAP. Однако, с усложнением интерпретируемой модели вычисления на основе данных алгоритмов становятся неоправданно затратными.

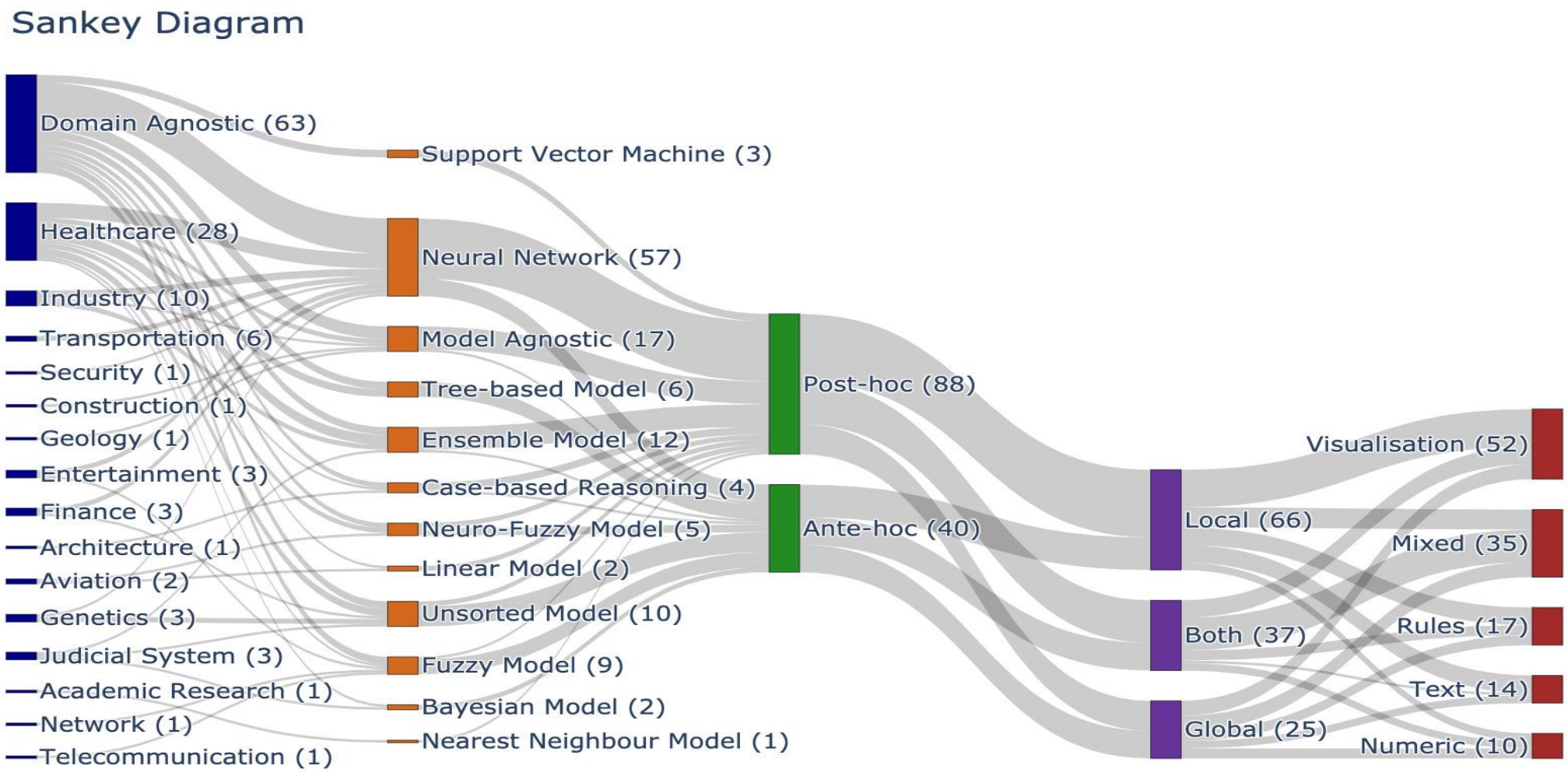
	Характеристики					
						
VG	*					
DeconvNET	*					
GBP	*					
LRP	*					
DeepLIFT	*					
CAM	*					
Grad-CAM	*					
Occlusion	*	^				
LIME	*					
SHAP	*					

Методы post hoc анализа XAI

Различные методы post hoc аназиза XAI оценивались
низкая / нет (красный)
средняя (оранжевый)
высокая / да (зеленый)

производительность на основе:
1. специфичности цели
2. пространственного разрешения
3-4. возможности локальной или глобальной воксельной зависимости
5.независимости от модели и
6.технической простоты

Распределение XAI по областям



Здравоохранение 5.0

Пятая промышленная революция сосредоточится на интеллектуальном производстве, вернув человеческий интеллект на производство, позволив роботам не заменять людей, а наоборот сотрудничать и помогать им. Здесь людям будет нужен XAI

Индустрия 5.0 также коснулась и отрасли здравоохранения. Здравоохранение 5.0 можно считать пятой промышленная революция в области здравоохранения, которая обеспечивает «массовую персонализацию». Персонализированная медицина развивается большими темпами, разрабатываются персонализированные устройства, которые могут измерять различные параметры здоровья у человека. Дистанционно также можно обрабатывать цифорвые медицинские изображения. Подобные персонализированные технологии позволяют докторам получать информацию о здоровье пациентов в режиме реального времени и онлайн.
В области здравоохранения XAI применяется в различных моделях поддержки принятия клинических решений, для анализа медицинских данных и изображений, в задачах клинической диагностики, уменьшения систематической ошибки медицинских датчиков и классификации заболеваний.

Манифест об объяснимости искусственного интеллекта в медицине

Каковы требования к XAI? Как мы можем оценить качество предоставленного объяснения? - Существуют осязаемые, реализуемые, ориентированные на пользователя требования, которые должны быть выполнены для построения системы XAI; более конкретно, существует ли необходимость измерения, интерпретации и понимания в контексте в контексте ИИ в медицине.

Если вывод системы искусственного интеллекта понятен, является ли он автоматически объяснимым? -Понимание результатов работы системы искусственного интеллекта является основой объяснимости, но это лишь одно из требований, которое должно быть объединено с удобством использования, полезностью и интерпретируемостью, чтобы получить объяснимость.

Какова роль понимания домена в достижении XAI в медицинских приложениях? - Системы на основе XAI должны начинаться с моделирования биомедицинской и клинической области, чтобы получить понимание контекста, в котором эти системы будут использоваться.

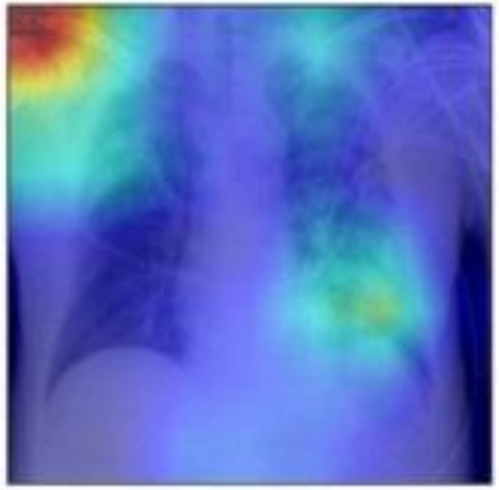
Может ли система искусственного интеллекта, не поддающаяся объяснению, заслуживать доверия? - XAI является неотъемлемым компонентом систем ИИ, заслуживающих доверия.

Всегда ли необходим XAI в медицине? - Объяснения не всегда требуются для того, чтобы модель ИИ была полезной. Функциональные спецификации, полученные в результате глубокого анализа проблемной области и пользователей, должны определять, когда требуется объяснимость и интерпретируемость

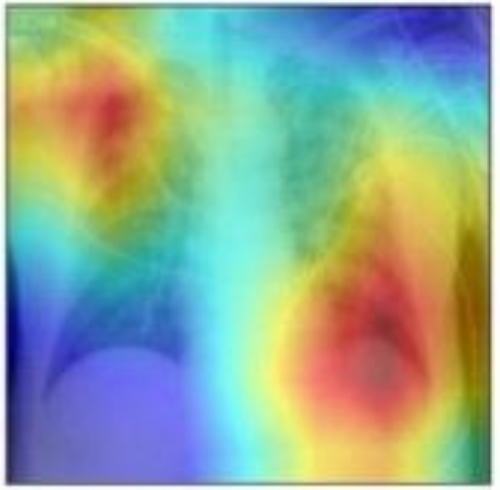
Artificial Intelligence In Medicine 133 (2022) 102423

Лучевая диагностика: COVID-19

Original Image

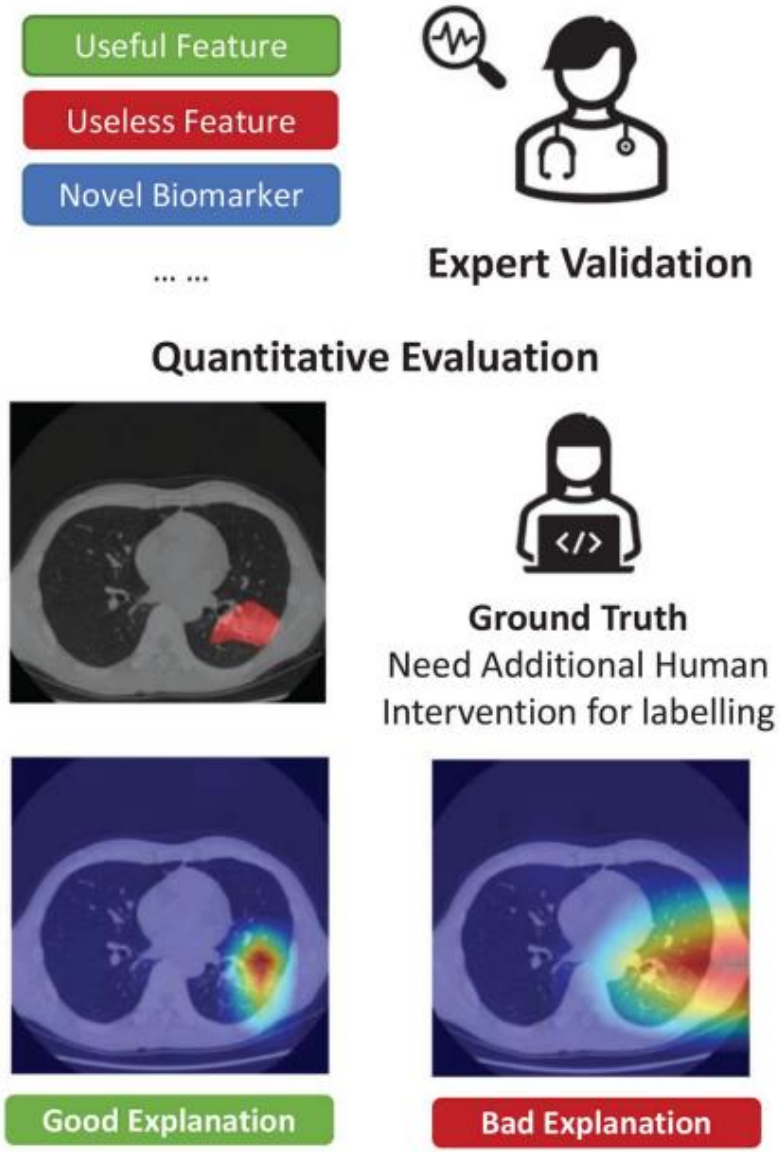


GradCAM (VGG-16)

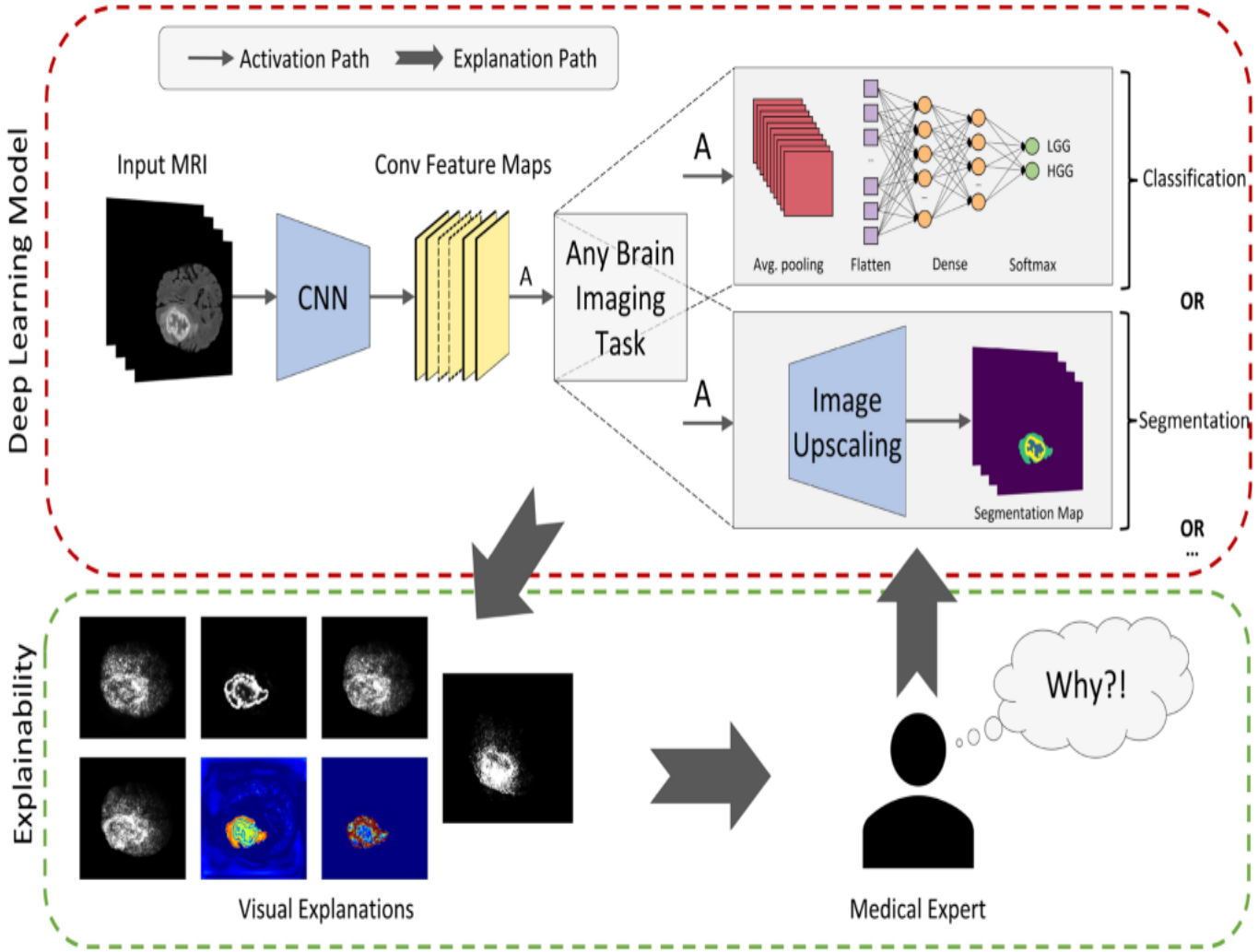


GradCAM (ResNet-19)

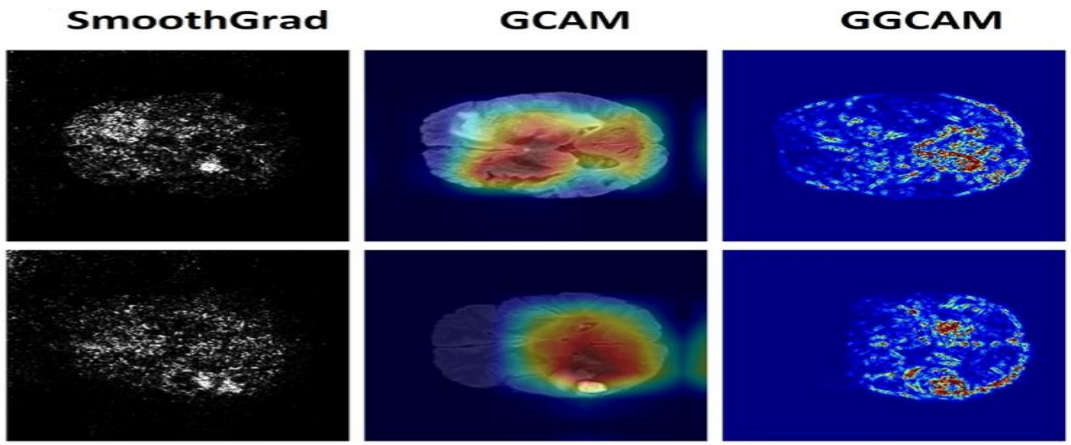
Объяснение МРТ/КТ/рентген снимков грудной клетки в диагностике ковидной пневмонии.



Лучевая диагностика: головной мозг

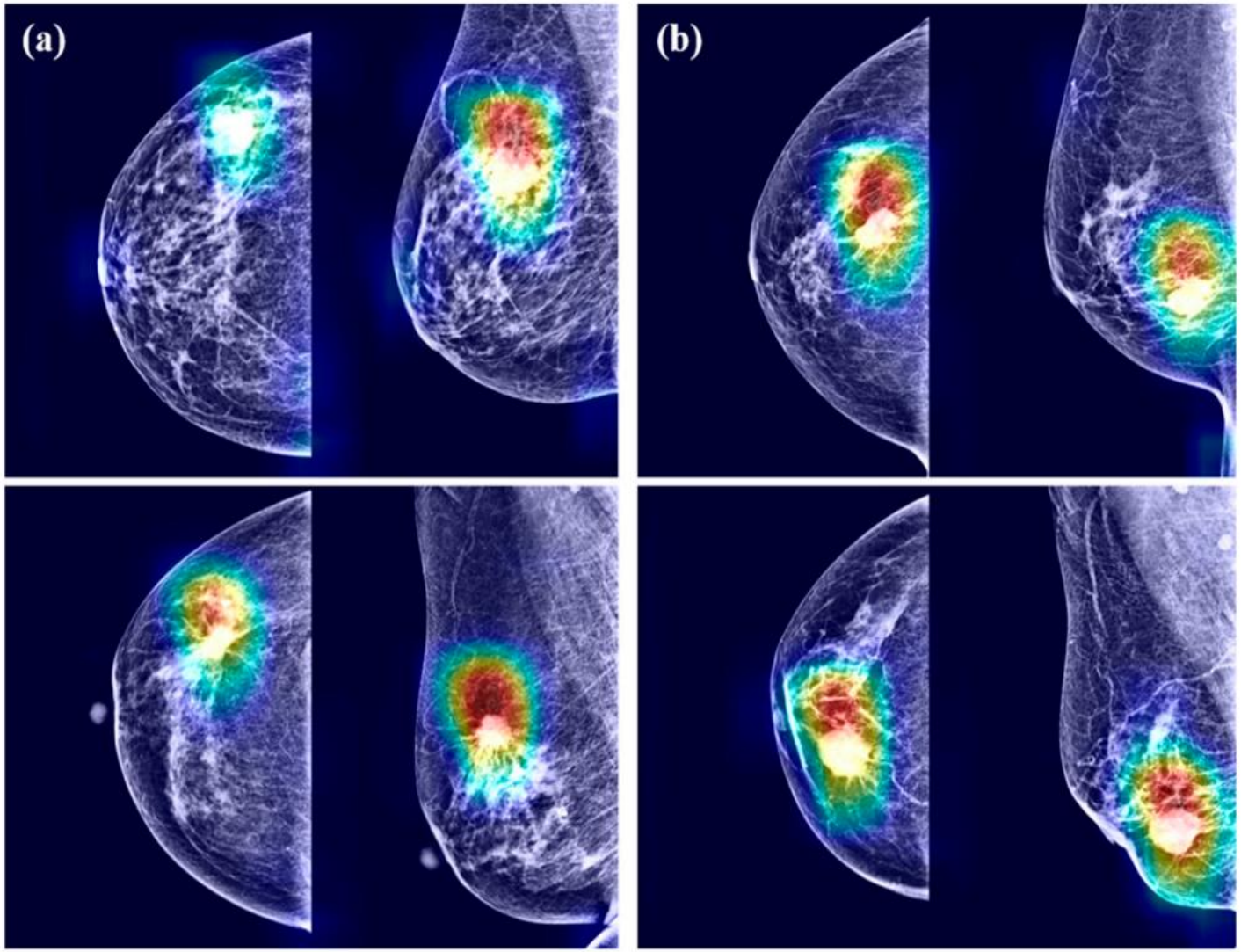


- Используются при диагностике опухолевых (глиома, глиобластома) и неопухолевых заболеваний (Болезнь Альцгеймера и т.д.) головного мозга
- Проиндексировано 42 исследования [Bas H.M. van der Velden]
- Наиболее часто используемые методы объяснения: 12-CAM, 11-Grad-CAM, 5 - LRP



Zeineldin R. A. et al. Explainability of deep neural networks for MRI analysis of brain tumors //International journal of computer assisted radiology and surgery. – 2022. – Т. 17. – №. 9. – С. 1673-1683.

Лучевая диагностика: маммография



- Применяется при визуализации новообразований молочных желёз, рака молочной железы (РМЖ)

Метод CAM для визуального объяснения результата поиска РМЖ с помощью ИНС

Algorithm Class activation mapping.

Require: Image $I^C(H,W)$; Network N
Ensure: Replace FC layer with average pooling layer in Network N
procedure CAM(I, N)
$N(I)$ ▷ Input image into network
$W_k^c \leftarrow (w_1, w_2, w_3, ..., w_k)$ ▷ Get weights from average polling layer
$F_k^c \leftarrow (f_1(x,y), f_2(x,y), f_3(x,y), ..., f_k(x,y))$ ▷ Feature map of the last convolution layer
$M_c(x,y) = \sum_k w_k^c f_k(x,y)$ ▷ Weighted linear summation
$M_c(x,y) = \frac{1}{HW} \sum_i^{H} \sum_j^{W} M_c(x,y)$ ▷ Normalize and up-sample to Network input size
$M_c(x,y) = \text{RELU}(M_c(x,y))$ ▷ Final image heat map
end procedure

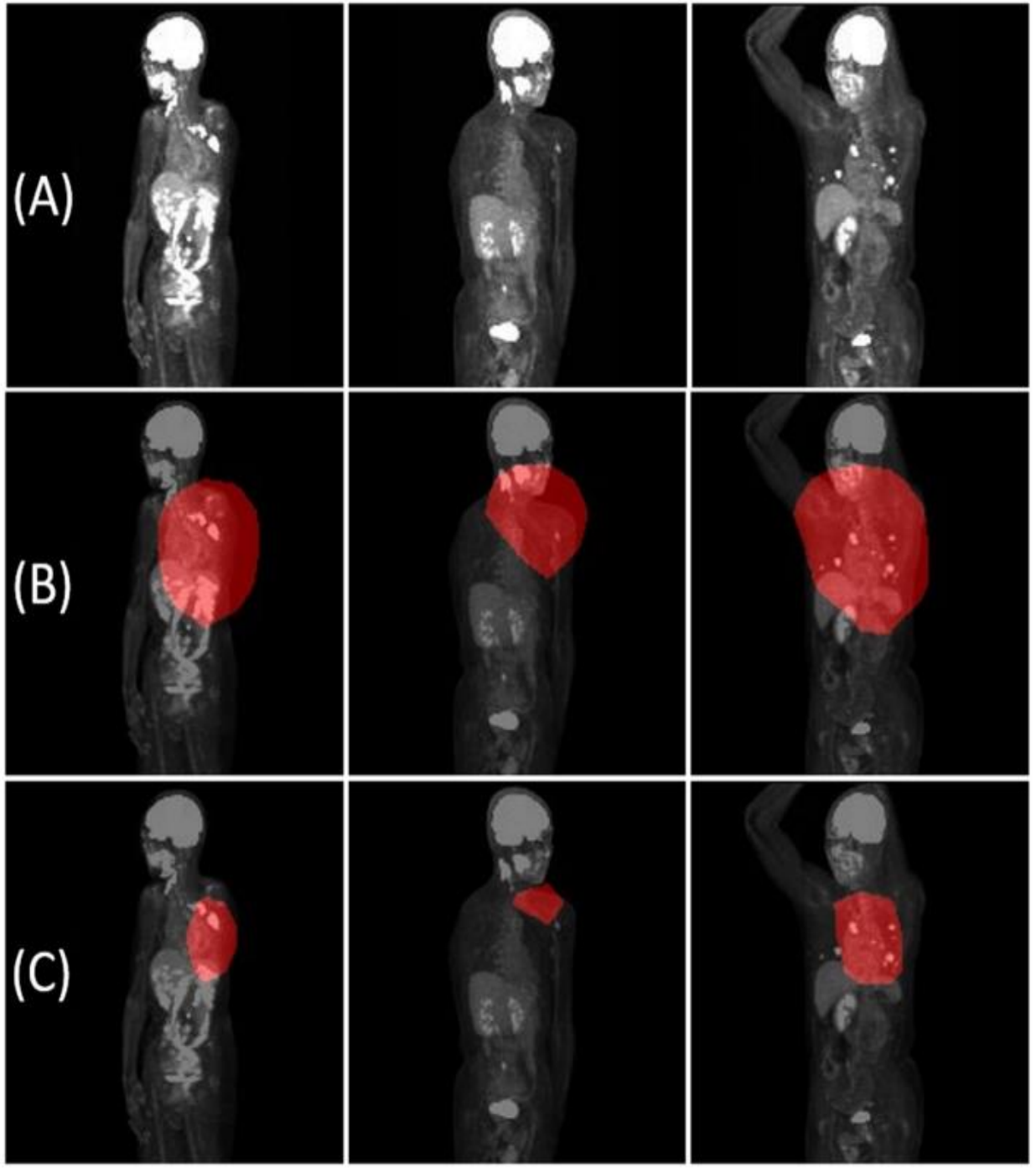
XAI & ПЭТ-КТ

Объяснение результатов работы СНС для распознавания участков накопления препарата при FDG ПЭТ-КТ [Kawauchi,2020]

A) Исходное изображение на входе СНС

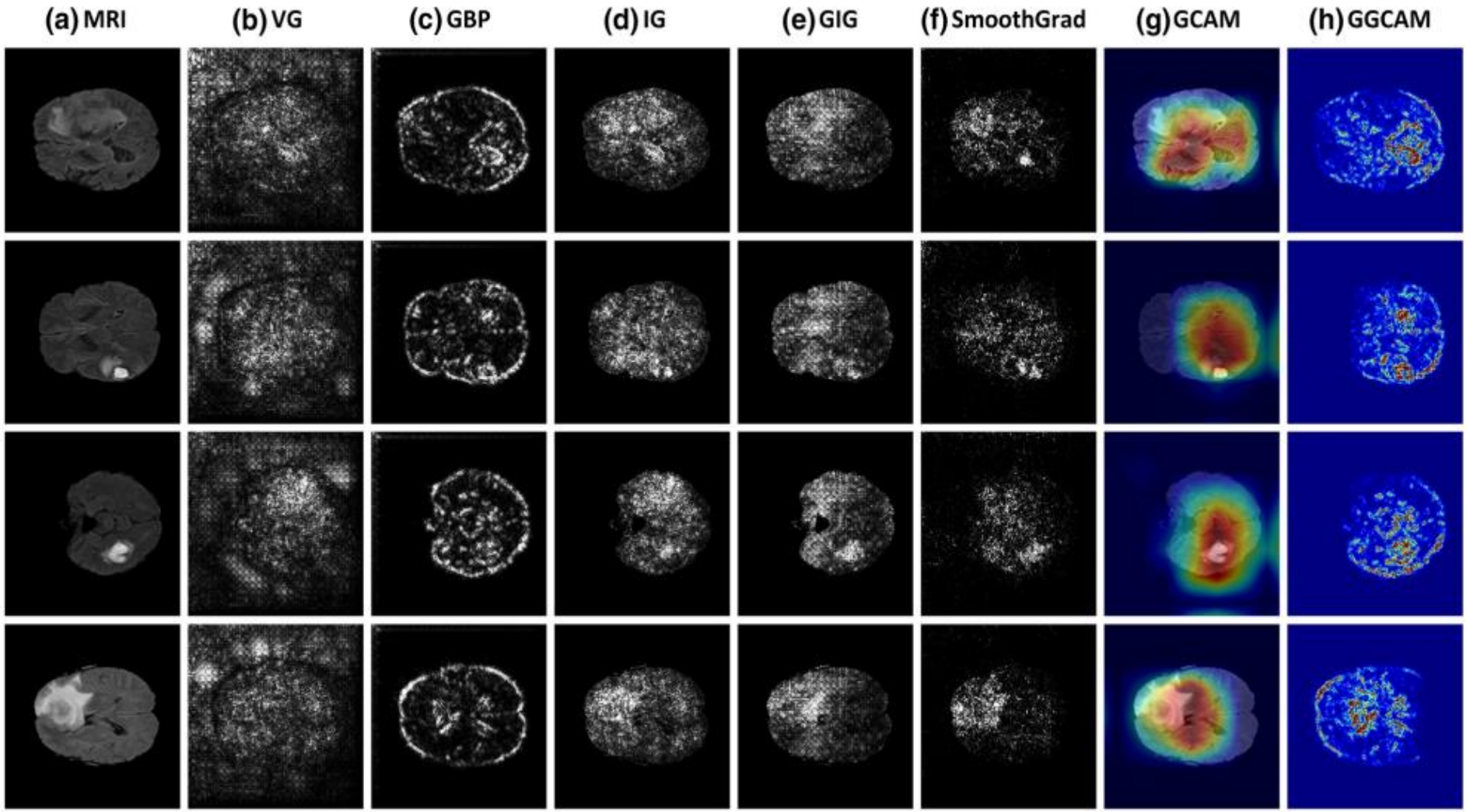
B) Объяснение с помощью метода Grad-CAM. Выделение области поглощения опухолью, порог уверенности более 70%

C) Объяснение с помощью метода Grad-CAM. Выделение области поглощения опухолью, порог уверенности более 90%



Kawauchi K. et al. A convolutional neural network-based system to classify patients using FDG PET/CT examinations //BMC cancer. – 2020. – Т. 20. – С. 1-10.

Методы XAI в объяснении предсказаний ИНС при диагностике опухолей мозга по МРТ-снимкам

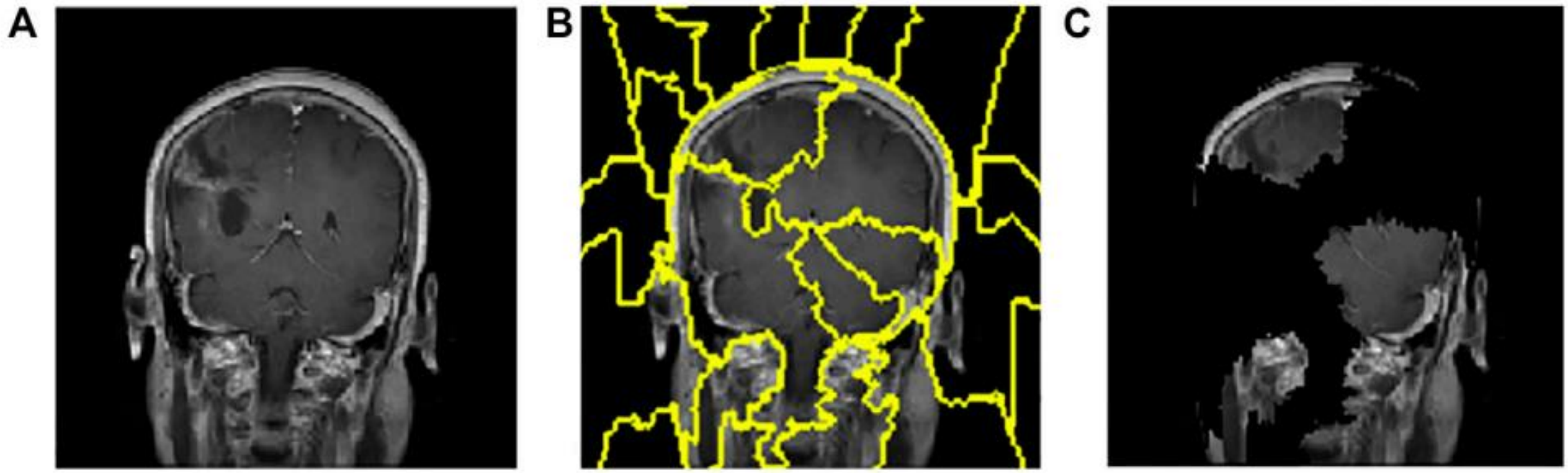


Визуальные объяснения получены с помощью методов:
b) Vanilla gradient
c) Guided backpropagation
d) Integrated gradients
e) Guided integrated gradients
f) SmoothGrad
g) Grad-CAM
h) Guided Grad-CAM

Для изображений b-f визуальные объяснения представлены saliency map, для изображений g и h - heatmap

Zeineldin R. A. et al. Explainability of deep neural networks for MRI analysis of brain tumors //International journal of computer assisted radiology and surgery. – 2022. – Vol. 17. – №. 9. – P. 1673-1683.

XAI LIME & Опухоль мозга

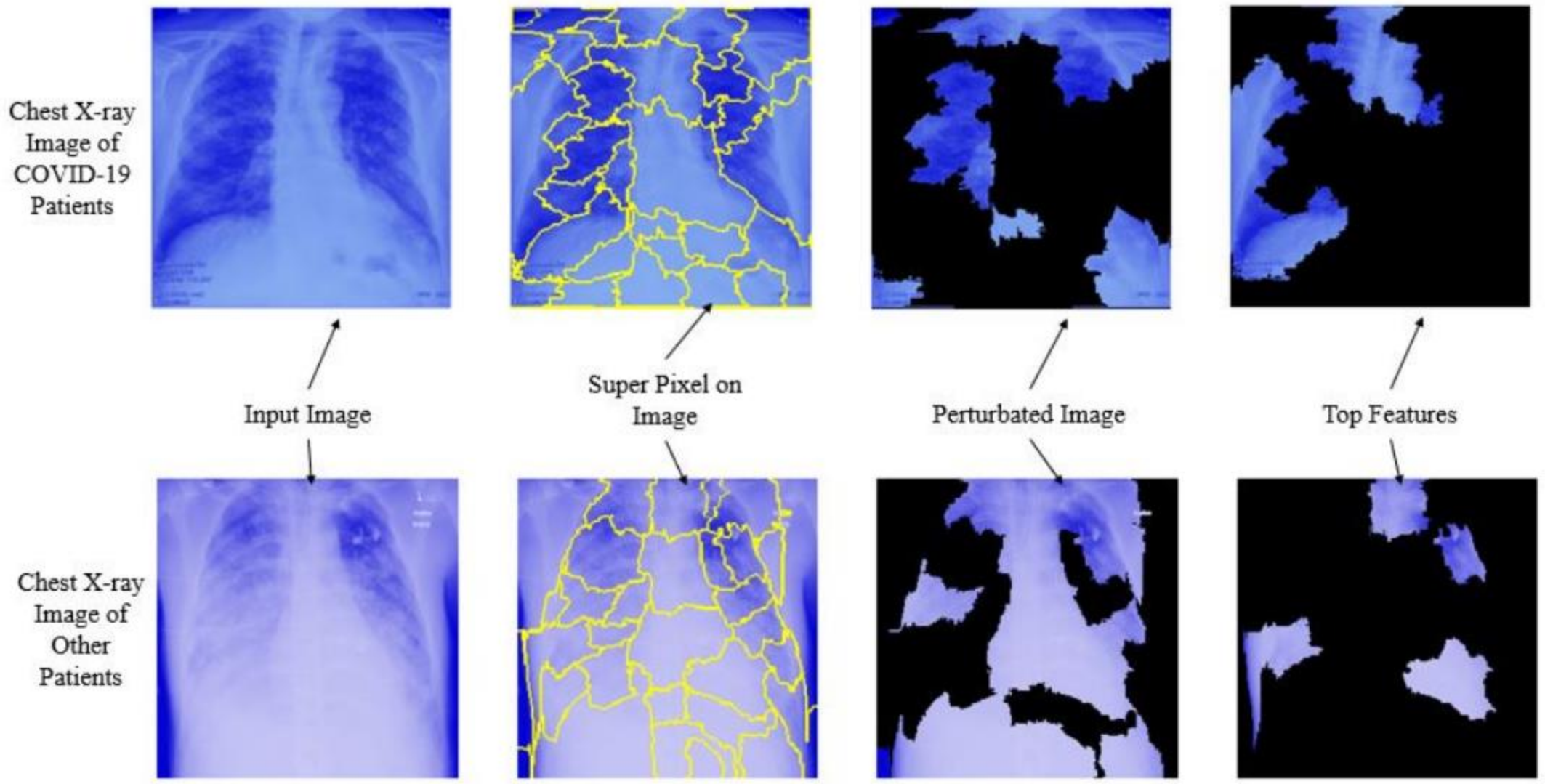


Использование метода LIME для получения объяснения работы ИНС при диагностике опухолей мозга [Gaur, 2022]

A) Входящий снимок МРТ
B) Полученные с помощью метода LIME суперпиксели
C) Объяснение с помощью метода LIME

Gaur L. et al. Explanation-driven deep learning model for prediction of brain tumor status using MRI image data //Frontiers in genetics. – 2022. – Vol. 13. – P. 448.

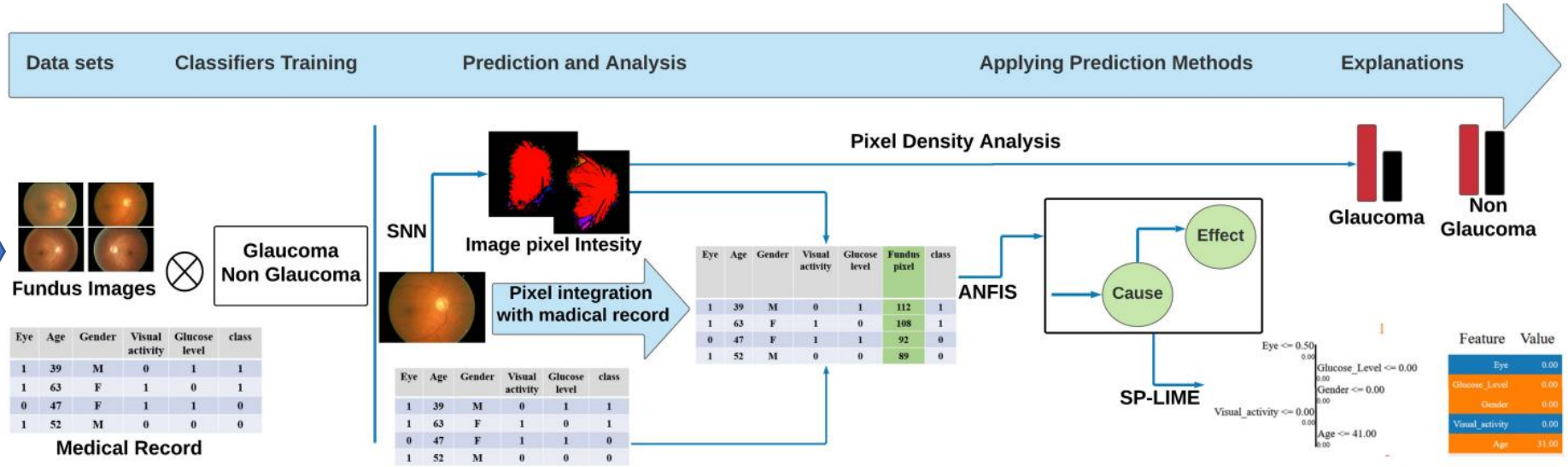
XAI LIME & Ковидная пневмония

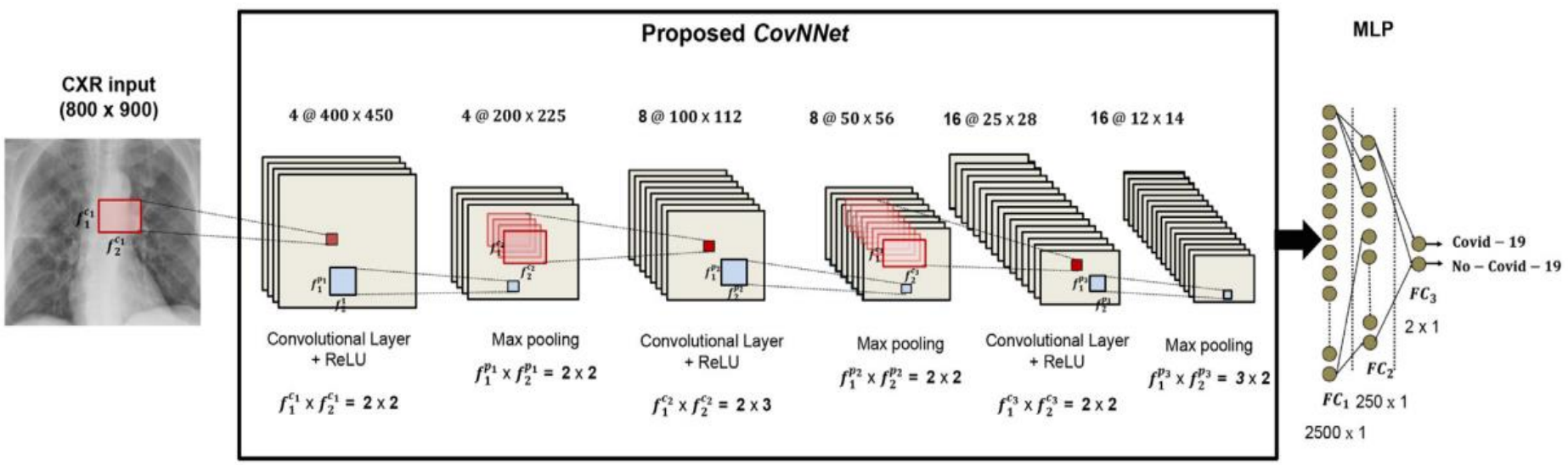


Ahsan M.M. et al. COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities // Machine Learning and Knowledge Extraction. MDPI AG, 2020. Vol. 2, № 4. P. 490–504.

Перспективы нейро-нечеткого XAI

Использование метода объяснения LIME и нейро-нечеткой системы логического вывода ANFIS для анализа изображений и записей пациентов, страдающих глаукомой





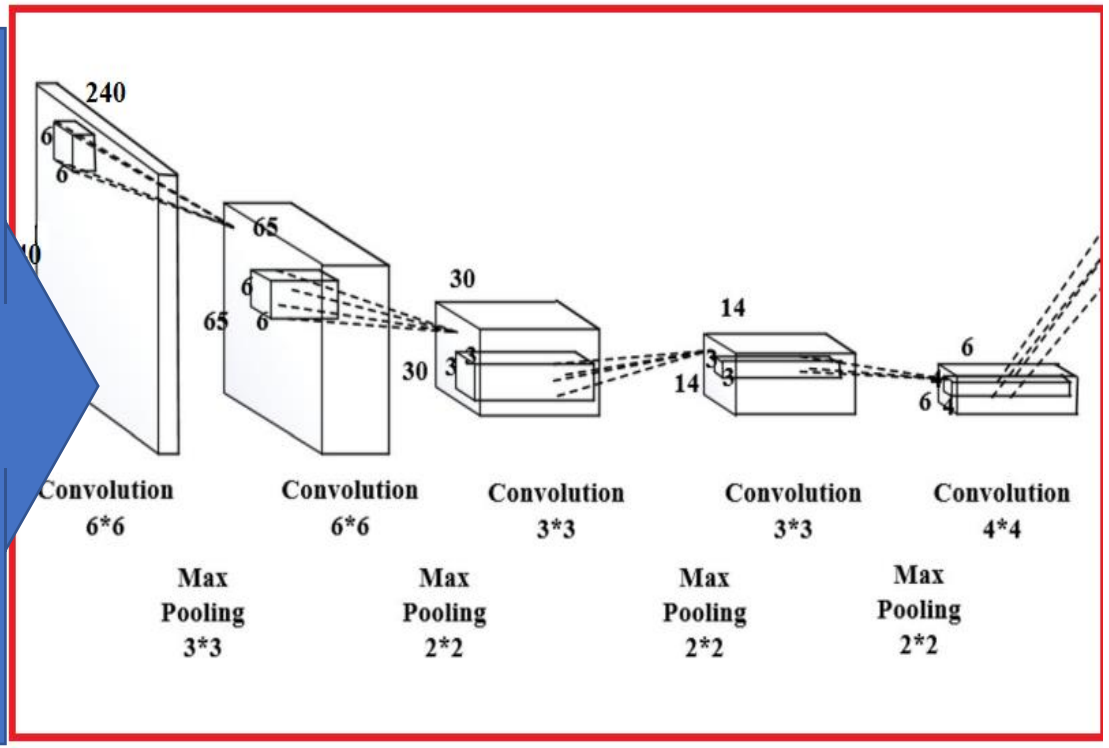
Использования нейро-нечёткой ИНС в сочетании с методом XAI CAM для анализа изображений и госпитальных записей больных COVID-19

Ieracitano C. et al. A fuzzy-enhanced deep learning approach for early detection of Covid-19 pneumonia from portable chest X-ray images // Neurocomputing. – 2022. – Т. 481.

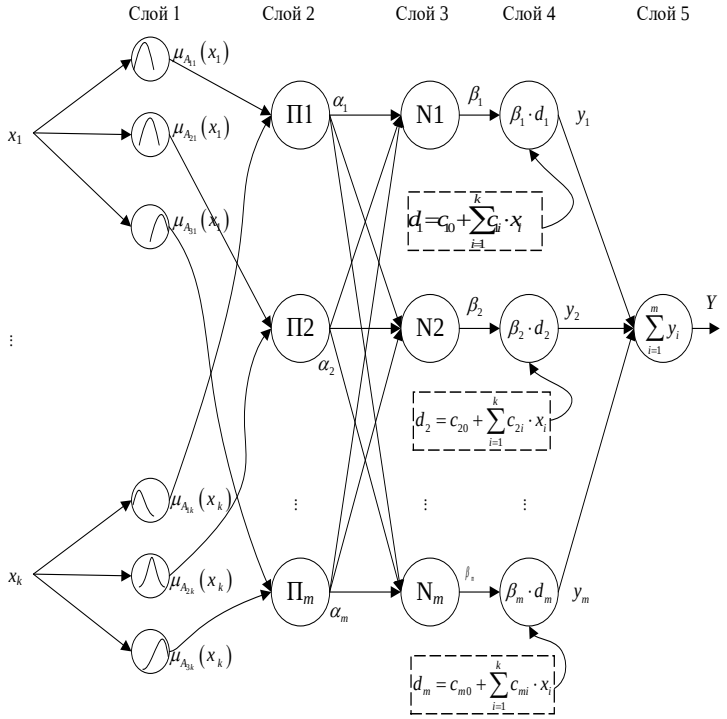
Kamal M. S. et al. Explainable AI for glaucoma prediction analysis to understand risk factors in treatment planning // IEEE Transactions on Instrumentation and Measurement. – 2022. – Т. 71.

Нейро-нечеткая система для классификации заболеваний глаукомой (уровня 2 по шкале)

Использование метода объяснения LIME и системы нейро-нечеткого логического вывода ANFIS для анализа изображений и записей пациентов с глаукомой



Свёрточные слои

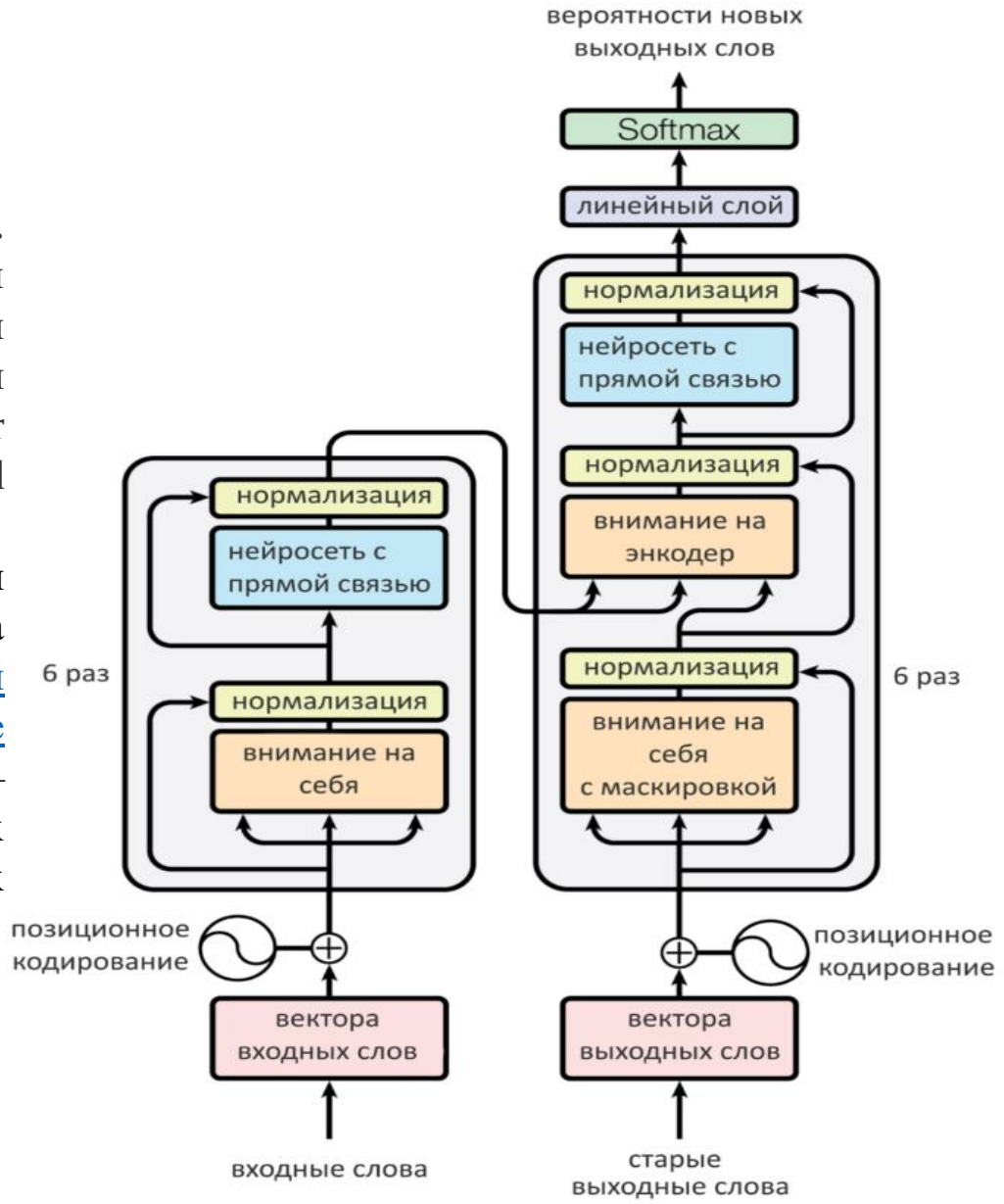


A. N. Averkin, E. N. Volkov, and S. A. Yarushev, Possibilities of Application of Neuro-Fuzzy Networks for Ophthalmologic Image Classification //PRIA JOURNAL SPECIAL ISSUE XXI NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE WITH INTERNATIONAL PARTICIPATION (CAI-2023) / SECTION 5 “FUZZY MODELS, SOFT MEASUREMENTS”, 2024

Трансформеры и большие языковые модели (LLM)

С годами языковые модели претерпели значительные изменения. Начиная с традиционных n-граммных моделей и заканчивая современными трансформерами, мы стали свидетелями экспоненциального роста их возможностей. Внедрение архитектуры Transformer, особенно таких моделей, как BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) и других, раздвинуло границы NLP до новых высот.

По аналогии с рекуррентными нейронными сетями (РНС) трансформеры предназначены для обработки последовательностей, таких как текст на естественном языке, и решения таких задач как машинный перевод и автоматическое реферирование. Генеративный предобученный трансформер (GPT) — это тип нейронных языковых моделей, впервые представленных компанией OpenAI, которые обучаются на больших наборах текстовых данных, чтобы генерировать текст, схожий с человеческим.



LLM в медицине

С момента появления в открытом доступе первой большой языковой модели (LLM) – совершенно нового класса архитектур искусственных нейронных сетей (ИНС) Chat-GPT в ноябре 2022 года, произошёл экспоненциальный рост использования LLM в различных цифровых решениях, отразившейся на всех сферах жизни современного человека.

К концу 2023 года LLM стали «мейнстримом» в области технологий искусственного интеллекта (ИИ). С ростом количества моделей LLM, появлением версий, способных работать с мультимодальными данными, всё более предпочтительным выглядит построение гибридных интеллектуальных систем с их использованием. В современной медицине LLM поддерживают значительный прогресс в разнообразных областях, включая создание интеллектуальных чат-ботов для взаимодействия с пациентами.

Применение LLM в медицинских чат-ботах позволяет значительно повысить качество первичной диагностики и консультирования, обеспечивая точность и детализацию ответов на запросы пользователей. Способность этих моделей к обработке естественного языка и генерации содержательных ответов опирается на анализ огромных массивов текстовой информации, включающей медицинские базы данных, публикации и клинические рекомендации, что позволяет им охватывать широкий спектр медицинских дисциплин

Большие языковые модели в задачах генерации автоматического описания медицинских изображений

Однако из наиболее перспективных направлений является использование LLM для автоматической генерации описаний медицинских изображений. Это включает в себя не только идентификацию аномалий на изображениях, таких как рентгеновские снимки, МРТ, или КТ, но и создание понятных для человека отчетов, которые могут быть непосредственно использованы врачами для диагностики и разработки лечебных стратегий.

На основе полученной информации и обучения на медицинских текстах LLM генерирует описание медицинских изображений. Эти описания включают информацию о состоянии анатомических структур, выявленных патологиях, возможных диагнозах и предложениях по дальнейшим обследованиям.

Ниже приведена модель автоматического создания подписей к клиническим изображениям сочетает в себе анализ радиологических сканирований со структурированной информацией о пациенте из текстовых записей. Он использует две языковые модели: Show-Attend-Tell и GPT-3 для создания полных и описательных радиологических записей.

Описание рентгенограммы грудной клетки с помощью модели на основе GPT-3

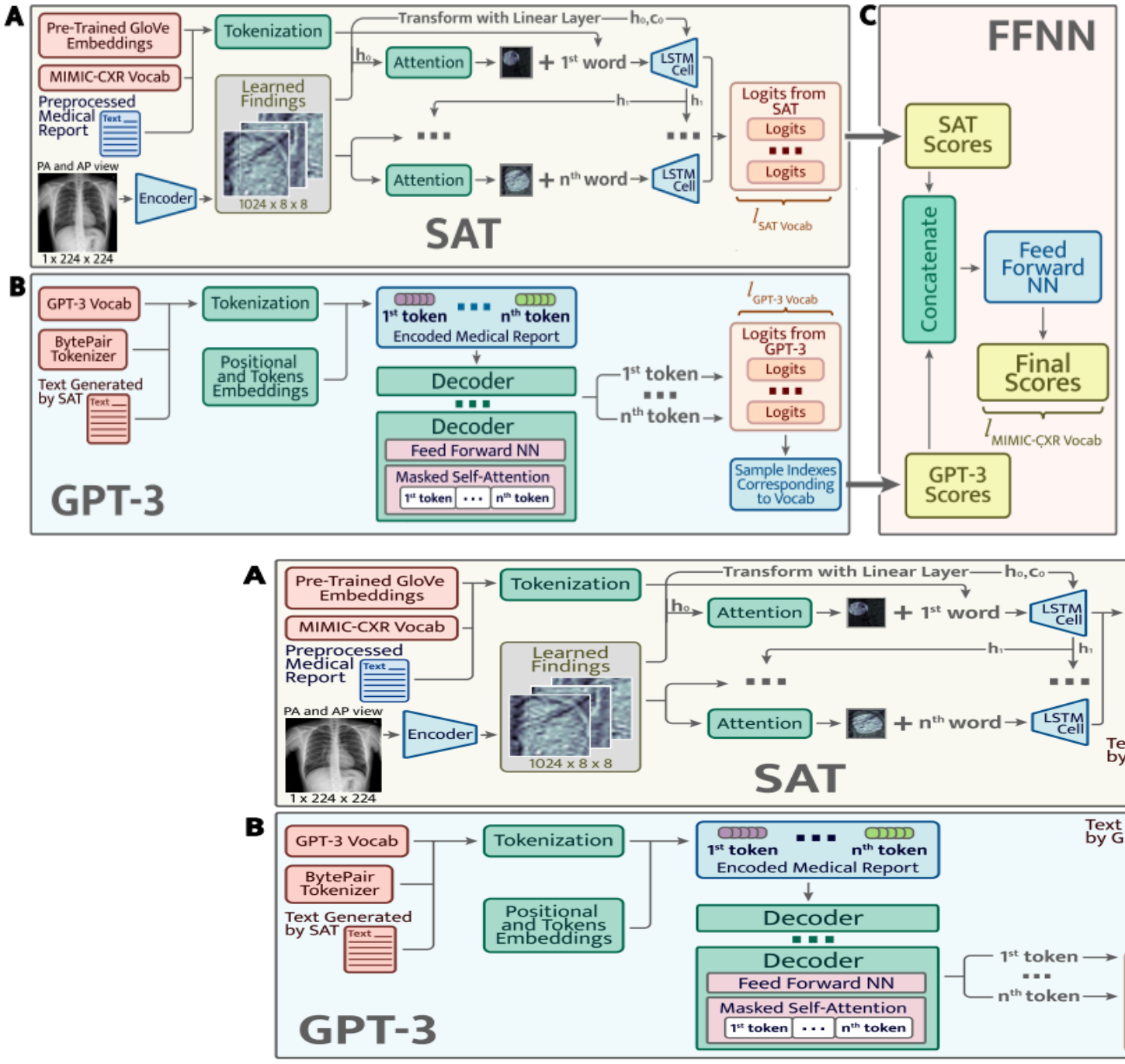
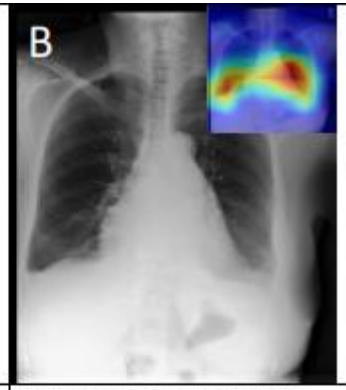


Image sample cases		
DC	No findings	Pleural effusion, Cardiomegaly, Atelectasis
Ground truth	Lungs remain well inflated without evidence of focal airspace consolidation, pleural effusions, pulmonary edema or pneumothorax.Irregularity in the right humeral neck is related to a known healing fracture secondary to recent fall. PA and lateral views of the chest at 09:55 are submitted	1. Stable bilateral small pleural effusions andatelectasis. 2. Enlarged pulmonary artery, suggesting pulmonary hypertension. Bilateral small pleural effusions and adjacent atelectasis are overall unchanged. The heart is top-normal in size, unchanged.
Approach 1	pulmonary vascularity is normal in caliber and distribution . impression : no evidence of acute pulmonary pathology with possible development of right pleural effusion .	minimal linear densities in the costophrenic angles characteristic of scarring . healed rib fractures. minimal tortuosity thoracic aorta . Multiple calcified pulmonary nodules consistent with pulmonary edema .
Approach 2	no findings. no pneumonia. no pleural effusion. no edema. there is little change and no evidence of acute cardiopulmonary disease. no pneumonia. vascular congestion. pleural effusion.of incidental note is an azygos fissure, of no clinical significance . this raises possibility of a normal variant.	pleural effusion present. lung opacity present. no edema. cardiomegaly present. atelectasis present. as compared to previous radiograph, there is an increase in extent of a pre existing small left pleural effusion with subsequent atelectasis at left lung bases. no new focal parenchymal opacities suggesting pneumonia.

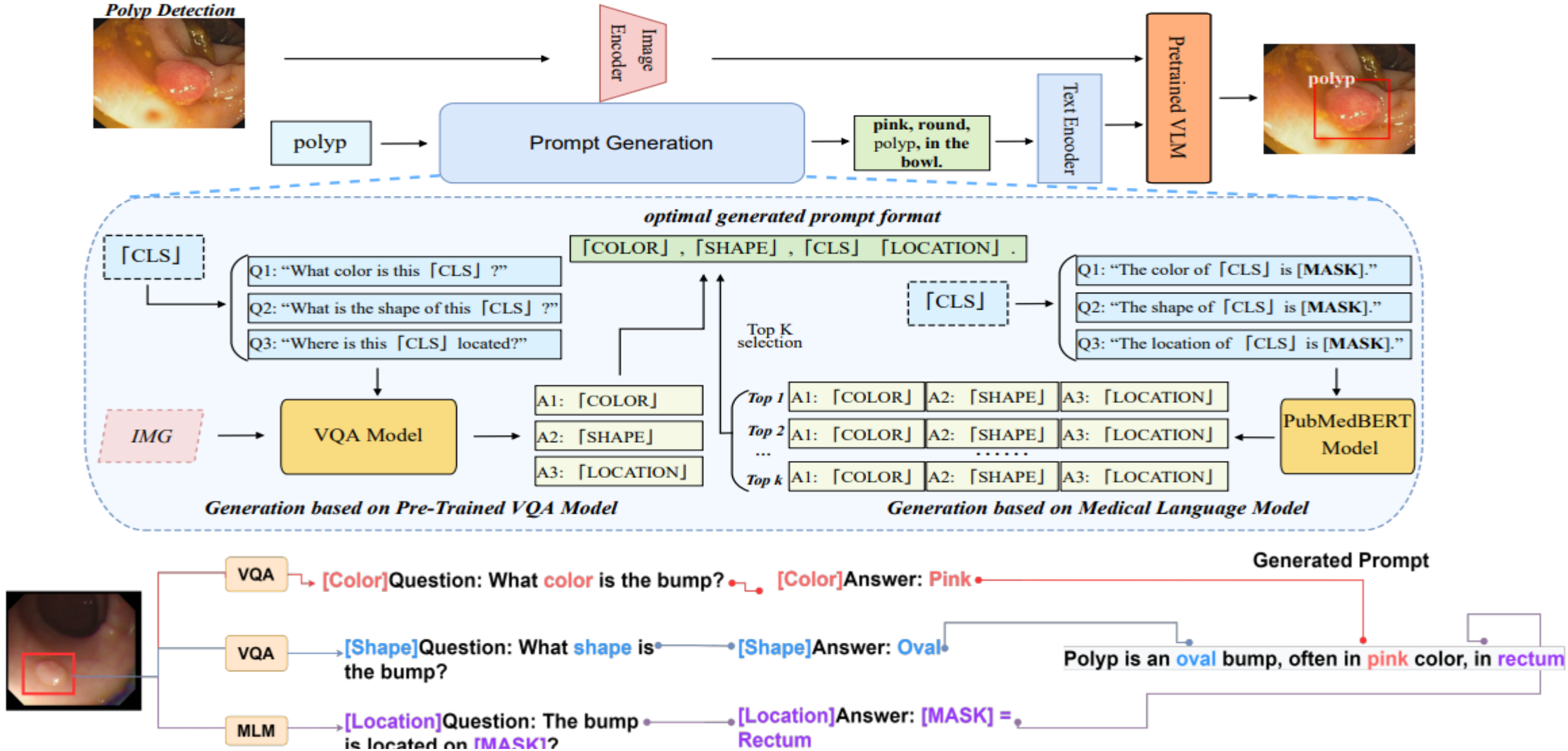
Selivanov A. et al. Medical image captioning via generative pretrained transformers //Scientific Reports. – 2023. – Vol. 13. – no. 1. – P. 4171.

Описание ренгенограммы с помощью GPT

Внутримозговое кровоизлияние (ВМК) — тяжелое нарушение мозгового кровообращения, представляющее угрозу для жизни и требующее быстрой диагностики и лечения. Хотя компьютерная томография является наиболее эффективным диагностическим инструментом для выявления кровоизлияния в мозг, ее интерпретация обычно требует опыта квалифицированных специалистов.

В работе Центра исследования цереброваскулярных заболеваний Южной Кореи, используется предварительно обученный классификатор CNN и GPT-2 для генерации текста для последовательно полученных изображений ICH CT. Первоначально CNN подвергается точной настройке путем изучения наличия ICH в общедоступных одиночных КТ-изображениях, а затем извлекает векторы признаков (т. е. матрицу) из 3D-изображений ICH CT. Эти векторы вместе с текстом вводятся в GPT-2, который обучен генерировать текст для последовательных изображений КТ. В экспериментах проведена oweнка производительности четырех моделей, чтобы определить наиболее подходящую модель субтитров к изображениям: (1) В методе на основе N-грамм ReseNet50V2 и DenseNet121 показали относительно высокие оценки. (2) В методе, основанном на внедрении, DenseNet121 показал наилучшую производительность.

Применение промптинга в задаче ответа на вопрос по изображению (Visual Question Answering)



Qin Z. et al. MEDICAL IMAGE UNDERSTANDING WITH PRETRAINED VISION LANGUAGE MODELS: A COMPREHENSIVE STUDY //The Eleventh International Conference on Learning Representations. – 2022.

Объяснимость LLM

Потребность объяснимости в LLM обусловлена растущей сложностью и внедрением систем на основе ИИ. По мере того, как LLM становятся все более изощренными, они часто действуют как «черные ящики», скрывая свои внутренние процессы принятия решений от человеческого понимания. К тому же они фактически являются интерфейсом ко всему интернету, не обладая ни моделью мира, ни семантической интерпретацией.

Статистическое предсказание не означает семантическое понимание. LLM могут не знать или не учитывать значение, цель или последствия своих слов. Это может приводить к тому, что LLM генерируют бессмысленные, нерелевантные или опасные тексты.

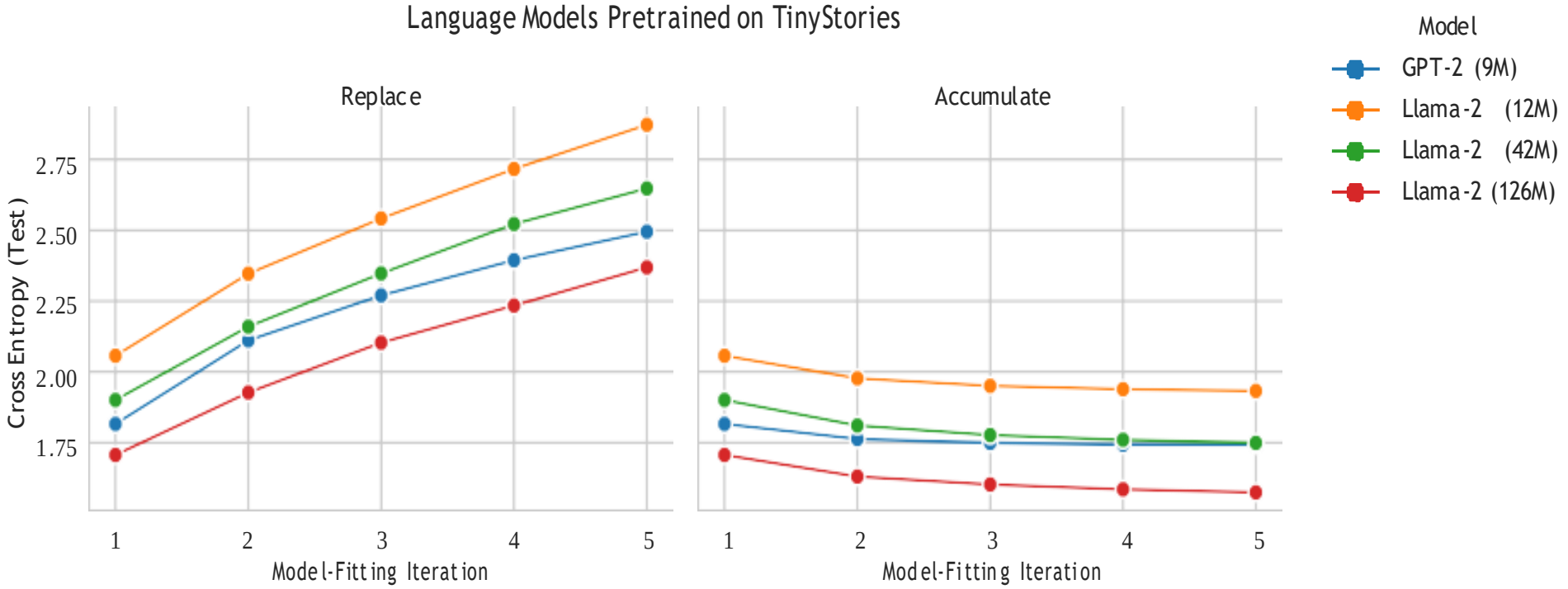
Статистическое предсказание не означает логическое рассуждение. LLM могут не следовать или не проверять правила, факты или доказательства в своих текстах. Это может приводить к тому, что LLM генерируют неправильные, противоречивые или обманчивые тексты.

Статистическое предсказание не означает творческое выражение. LLM могут не иметь или не развивать свой стиль, голос или перспективу в своих текстах. Это может приводить к тому, что LLM генерируют скучные, повторяющиеся или плагиатные тексты. трансфомеры также являются крайне неустойчивыми системам и подвержены атакам на алгоритмы и на данные.

Возникает также нейросетевой коллапс распределений при обучении на сгенерированных данных

Отсутствие прозрачности создает серьезные проблемы, особенно в таких критически важных секторах, как здравоохранение, финансы и юридические отрасли, где объяснимость имеет решающее значение для доверия и подотчетности.

Накопление данных позволяет избежать краха модели при языковом моделировании.



.Последовательности языковых моделей на основе каузальных трансформеров предварительно обучены на TinyStories. Потери при валидации по кросс-энтропии увеличиваются при замене данных (слева), но не при накоплении данных (справа). Синтетические данные сэмплировались с температурой = 1,0.

Общая структура механизмов объяснения БЯМ

Объяс-нимость БЯМ	Традиционная парадигма тонкой настройки	Локальное объяснение	Объяснение атрибуции объекта
			Объяснение, основанное на внимании
			Объяснение на основе примеров
			Объяснение на естественном языке
		Глобальное объяснение	Объяснение, основанное на исследовании
			Объяснение активации нейронов
			Объяснение, основанное на понятиях
			Механистическая интерпретируемость
		Использование объяснений	Отладочные модели
			Совершенствование моделей
	Парадигма промтов	Базовая модель	Объяснение обучения в контексте
			Объяснения в цепочке мыслей
			Инженерия представлений
		Вспомогательная модель	Объяснение роли точной настройки
			Галлюцинации и неуверенность
		Модель, использующая объяснение	Повышение точности БЯМ
			Приложения нижнего уровня
	Оценка объяснений	Парадигма тонкой настройки	
		Парадигма промтов	

Предвзятость и галлюцинации

Необходимость использования методов объяснительного ИИ и получения локальных объяснений, во-многом также исходить из двух основных проблем всех моделей LLM таких как «предвзятость» и «галлюцинации» .

Предвзятость в LLM возникает из-за систематических ошибок, заложенных в данных, используемых для их обучения. Эти данные могут отражать социальные и культурные предубеждения, дискриминационные практики, а также недостаточную представленность или искажение определенных групп населения. Как следствие, модели могут демонстрировать предвзятость в отношении расы, пола, возраста, профессии и других характеристик.

Галлюцинации в контексте LLM относятся к ситуациям, когда модель генерирует ответ, который выглядит правдоподобным, но на самом деле не соответствует действительности. Обычно такой текст включать в себя вымышленные факты, события, цитаты или ссылки, которые модель ошибочно генерирует вместо того, чтобы вывести сообщение о невозможности вывода ответа. Борьба с галлюцинациями является одним из наиболее актуальных направлений в развитии LLM , связанным с фундаментальными основами в вопросах архитектуры и принципов работы моделей данного класса.

Методы, основанные на вычислении атрибуции признаков

Наиболее часто используемыми, из всех групп методов объяснительного ИИ, являются методы, основанные на вычислении атрибуции признаков, то есть значимости каждого входного токена для прогноза модели.

В контексте подходов к объяснимости LLM суть работы данной группы методов можно формально представить следующим образом: при заданном входном запросе , являющемся некоторой упорядоченной последовательностью из токенов предварительно обученная LLM получает предсказание , при этом, методы атрибуции присваивают каждому токену оценку значимости , отражающую вклад конкретного токена в полученное предсказание модели.

Примеры визуализации объяснения LLM:

a) объяснение через MLP – блок,
b) объяснение через блок внимания