Методология программирования систем доверенного искусственного интеллекта

С.С. Гончаров, А.В. Нечесов, Д.И.Свириденко

Москва, кластер Ломоносов, 27 мая 2024
II-Форум Технологии Доверенного Искусственного Интеллекта

Вступление

1. Развитие ИИ-технологий - одна из самых горячих тем на сегодня
2. По всей стране создаются Центры ИИ.
3. Нацеленность на построение сильного ИИ и прорывные технологии.
4. НГУ выиграл конкурс (2 волны) по тематике "Строительство и городская среда".
5. Центр ИИ НГУ - мы планируем большую конференцию по тематике Умные Города, Строительство и Энергетика.
6. Есть потребность в доверенном ИИ в Китае
7. Нужны новые прорывные технологии объяснения работы нейронных сетей.
8. Потребность в ИИ: умные города и циф. двойники

XAI - 4 принципа: Сентябрь 2021

The National Institute of Standarts and Technology (NIST)

1. Explanation: ИИ должен предоставить объяснение, которое при необходимости можно детализировать
2. Meaningful: Объяснение ИИ должно быть понятно пользователю
3. Accuracy: Объяснение должно согласовываться с алгоритмами системы
4. Limits: XAI не должен выходить за рамки своих ограничений

Пекин задал курс на XAI

Cyberspace Administration of China, CAC опубликовала 10 мая 2023 правила для ИИ-компаний и их разработок.

Суть правил:

Компании будут нести ответственность за своих чат-ботов, если:

- первоначальные источники данных нелегитимны
- за все результаты выдачи чат-ботов
- за контент, который нарушал бы основные принципы социалистической партии
- за призывы к свержению режима или действия порочащие партию

Проблемы современного ИИ

- Проблема черного ящика
  - нет прозрачности
  - нет проверяемости
  - нет объяснимости
- Проблема централизации
- Проблема аудита и верификации
- Проблема надежности
- Проблема верификации
- Проблемы безопасности
- Проблема обучения

AI-полные задачи

AI-полная задача, по аналогии с NP-полным классом задач в теории сложности, — проблема, решение которой предполагает создание «сильного AI», то есть решения главной проблемы искусственного интеллекта: сделать компьютеры такими же умными, как люди

- компьютерное зрение
- понимание естественного языка
- прохождение теста Тьюринга

AI-полные задачи

Открытые проблемы:

- построить теорию сводимости ИИ-задач
- выстроить иерархию ИИ-задач
- отсюда получить понятие суперсильного ИИ как интеллекта способного решать задачи из класса AI-hard.

Roman V. Yampolskiy, University of Louisville, USA, 2012

AI-Complete, AI-Hard, or AI-Easy – Classification of Problems in AI.

Проблемы доверия относительно уровней иерархии

Необходимо построить аналогичную иерархию и в TrAI и выявить основные характеристики доверия к системам искусственного интеллекта.

Методология программирования TrAI

База методологии:

1. Методология создания программ: программа должна всегда останавливаться и иметь полиномиальную вычислительную сложность
2. Блокчейны и защищенные БД: Исполнение и хранение
3. Обучение: логико-вероятностный модуль как дополнение к нейросетям

Гибридный ИИ как способ повышения доверия

Гибридный ИИ: комбинация подходов

- нейронных сети
- логико-вероятностные методы
- другие математические подходы

Виды взаимодействия

- последовательное (Сколтех: текст -> мат.запись -> решение)
- параллельное (с помощью логики контролируем выдачу нейросети)
- комбинированное

Доверие к системам ИИ

Можем ли мы доверять ChatGPT, GigaChat и т.д.?

ChatGPT - одна из больших языковых моделей претендующих на сильный ИИ.

Но ее сервера и данные на которых она обучалась полностью контролируются OpenAI.

Поэтому вопрос доверия (помимо структуры самой ChatGPT как LLM) это также вопрос доверия к OpenAI.

OpenAI - это компания из США, причем аффилированная с MicroSoft. Получается, что власти США, при желании, могут полностью контролировать работу ChatGPT, а тем самым производить дискриминацию пользователей в зависимости от региона, пола, вероисповедания и т.д.

Уровень централизации ИИ моделей

- автономно - да
- локальные - да
- региональные - да
- на уровне стран - да
- на уровне групп стран (Евросоюз, БРИКС) - да
- общемировые (типа ChatGPT с претензией на AGI) - децентрализация!

Открытая проблема

как внедрить децентрализацию на нижних уровнях

Атака 51 % или сколько стоит децентрализация

на 31 декабря 2023 (BTC PoW и ETH PoS

- Bitcoin PoW от 5 - 40 млрд. долл. (нужно 2.5 млн ASIC)
- Ethereum PoS от 34 млрд. долл.
- ETH PoW от 0.1 до 1 млрд. долл. (наш прогноз)
- ETC (Top 20 криптовалют) - был взломан, притом атака не превысила 1 миллиона долларов

Nuzzi, Waters, Andrade 2024

Breaking BFT: Quantifying the Cost to Attack Bitcoin and Ethereum. http://dx.doi.org/10.2139/ssrn.4727999

Языки программирования для алгоритмов ИИ

- почти все языки Тьюринг-полные Python, C++, Java и т.д.
- Проблема остановки для Тьюринг полных языков!
- Опасность зацикливания и зависания вычислимых ИИ-реализаций.
- Проблема гарантированной точности (академик С.К.Годунов, далее академик С.С.Гончаров)

PAG-теорема

Вводится понятие GNF-системы системы и определение p-вычислимости GNF-системы.

PAG-теорема Гончаров, Нечесов 2021

Пусть G –p-вычилимая GNF-система, тогда наименьшая неподвижная точка оператора Γ_G является p-вычислимой.

Goncharov, Nechesov 2021

Polynomial Analogue of Gandy's Fixed Point Theorem. https://doi.org/10.3390/math9172102

Решение проблемы P = L

Гончаров, Свириденко и Нечесов предложили ряд расширений базого языка.

$$L_0 \subseteq L_1 \subseteq L_2 \subseteq \dots \; L_n \subseteq L = P$$

Solution of the Problem P = L, 2022

Goncharov, Nechesov https://doi.org/10.3390/math10010113
Пусть $\mathbb{HW}(\mathfrak{M})$ – p-вычислимая модель сигнатуры σ, тогда:
1) сложность любой L-программы является полиномиальной.
2) для любой p-вычислимой функции существует подходящая L-программа реализующая ее.

Объектно-ориентированный p-полный язык L*

На базе языка L, построен объектно-ориентированный логический язык программирования L*, а также реализован механизм исполнения L*-программ с помощью разработанной нами виртуальной машины V.

Теорема (о консервативности расширения)
Язык L* является консервативным расширением p-полного языка L.

Замечание: Язык L* готов к применению для реализации алгоритмов искусственного интеллекта!

Goncharov, Nechesov 2022

Semantic programming for AI and Robotics
https://doi.org/10.1109/SIBIRCON56155.2022.10017077

Методология программирования 2024

Нами разработана методология программирования в высокоуровневых языках:

- по Тьюринг-полному языку L строится специальное p-полное обеднение L_M
- запрещаем в циклах FOR переприсваивание инкремента
- убираем другие операторы цикла
- оставляем рекурсивные определения только удовлетворяющие требованиям Ганди

Методология программирования

Теорема [GNS-2024] (о существовании p-полного подязыка)

Теорема 1: Пусть L — высокоуровневый Тьюринг-полный язык программирования. Пусть также все базовые функции и отношения языка L будут p-вычислимы. Тогда по языку L можно эффективно построить p-полный подязык L_M с теми же базовыми функциями и отношениями удовлетворяющий критерию сильной выразительности.

Аксиоматизация теории блокчейна

Goncharov, S.; Nechesov, A. Axiomatization of Blockchain Theory. 2023

Theorem

Модель блокчейна Bitcoin является моделью теории блокчейна T.

Theorem

Модель блокчейна Ethereum (PoW версия) является моделью теории блокчейна T.

Goncharov, Nechesov 2023

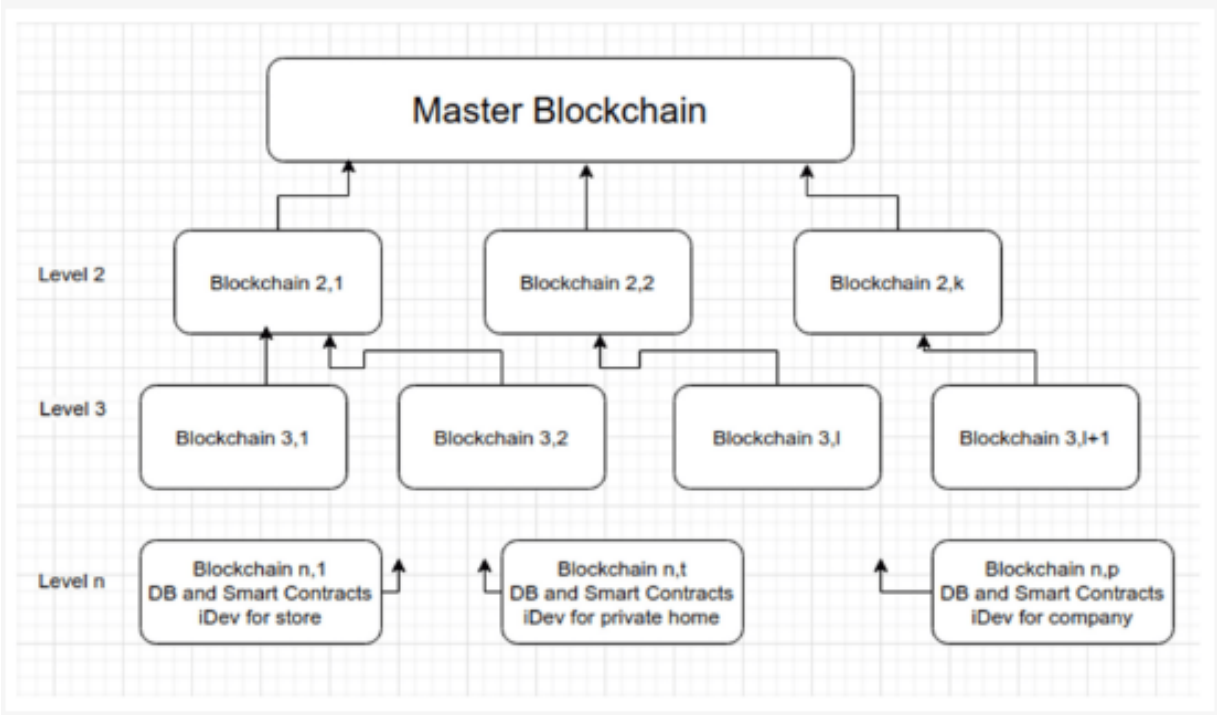
Axiomatization of Blockchain Theory. https://doi.org/10.3390/math11132966

Аксиоматизация теории блокчейна

Axioms of Blockchain Theory $\mathbb{T}$

- Blockchain axiom 1: $B(B(T)) = B(T)$
- Blockchain axiom 2: $x \in_t B(T) \to x \in_t T$
- Blockchain axiom 3: $Tree_0(B(T))$
- Blockchain axiom 4: $x \in_t BC(B(T)) \to x \in_t B(T)$
- Blockchain axiom 5: $Tree_0(BC(B(T)))$
- Equiv axiom 1: $x \equiv x$
- Equiv axiom 2: $x \equiv y \to y \equiv x$
- Equiv axiom 3: $(x \equiv y)\&(y \equiv z) \to (x \equiv z)$
- Order axiom 1: $T \leq_T T$
- Order axiom 2: $(T_1 \leq_T T_2) \vee (T_2 \leq_T T_1)$
- Order axiom 3: $(T_1 \leq_T T_2)\&(T_2 \leq_T T_3) \to T_1 \leq_T T_3$
- Order axiom 4: $\forall x \in_t T_1 \forall y\ (add(T_1, x, y) = T_2) \to T_1 \leq_T T_2$
- Zero axiom 1: $0 \in_t T$
- Zero axiom 2: $\forall x \in_t T\ \leq_b (T, 0, x)$
- Boundary axiom 1: $\forall x, y \in_t T\ \wedge (T, x, y) = \wedge(T, y, x)$
- Boundary axiom 2: $\forall x, y, z \in_t T\ \wedge (T, \wedge(T, x, y), z) = \wedge(T, x, \wedge(T, y, z))$
- Boundary axiom 3: $\forall x \in_t T\ \wedge (T, x, x) = x$
- Boundary axiom 4: $\forall x, y, z \in_t T \leq (T, \wedge(T, x, y), \wedge(T, y, z)) \to \wedge(T, x, y)$

Мульти-блокчейны



Аксиоматизация теории Мульти-блокчейнов

Нечесов, Новиков 2024

MB $\forall K \in_T T\ Blockchain(K)$

Or1 $T_1 \leq_{MB} T_1$

Or2 $(T_1 \leq_{MB} T_2)\&(T_2 \leq_{MB} T_3) \rightarrow (T_1 \leq_{MB} T_3)$

Or3 $\forall X \in_T T_1 \forall Y(add_{MB}(T_1, X, Y) = T_2) \rightarrow T_1 \leq_{MB} T_2$

Or4 $\forall X \in_T T_1 \forall b_1 \in_t X \forall b_2\ T_1 \leq_{MB} change(T_1, X, add(X, b_1, b_2))$

PC $\forall X, Y \in_T T \leq_b (T, X, Y)\&\neg(\exists Z \leq_b (T, X, Z)\& \leq_b (T, Z, Y)) \rightarrow R(X, Y)$

M1 $\forall X' \subseteq_{rootchain} X\ R(X', Y) \rightarrow R(X, Y)$

M2 $\forall Y' \subseteq_{rootchain} Y\ R(X, Y') \rightarrow R(X, Y)$

Сложность исполнения

Нечесов, Новиков 2024

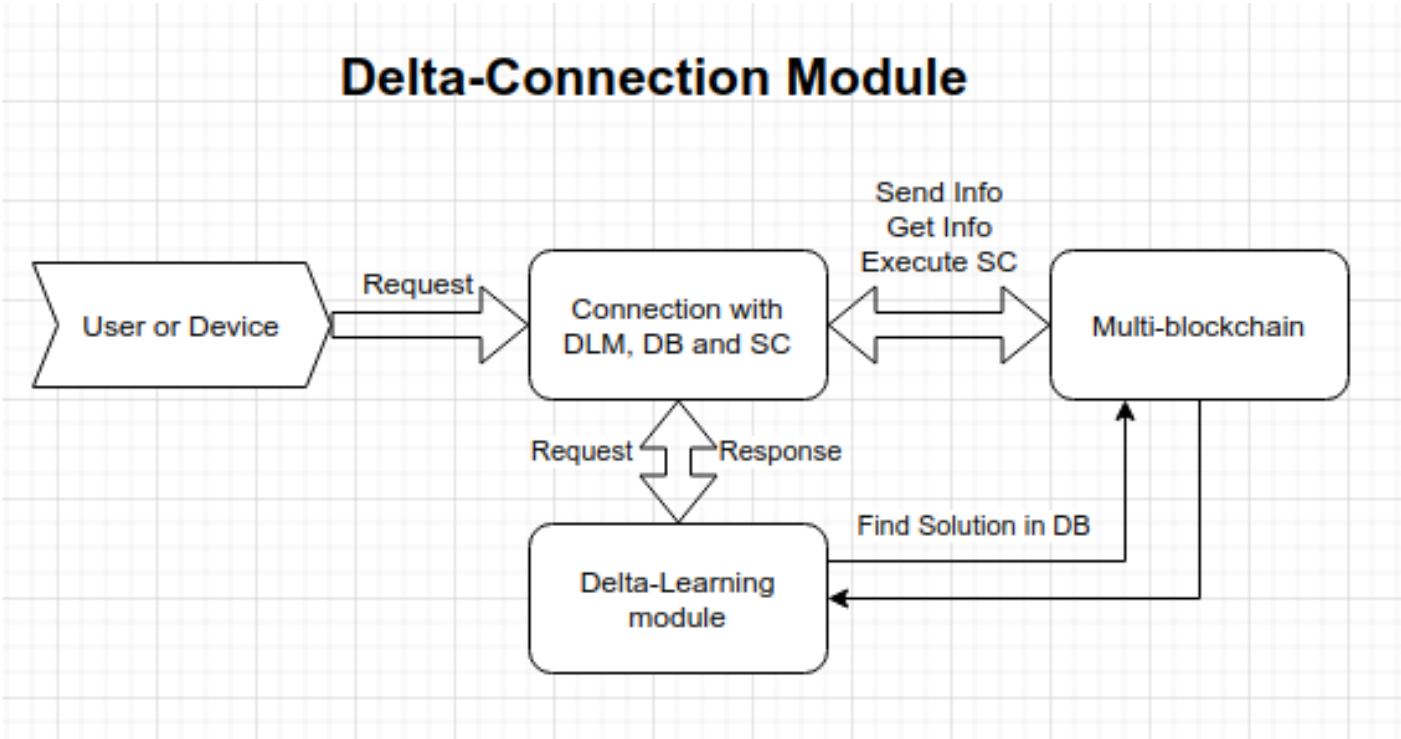
Сложность исполнения смарт-контрактов (L-программ спец. вида) на блокчейне является полиномиальной от длины входных данных и длины блокчейна.

Умные города и безопасность

Термин «умный город» используется в разных смыслах. Согласно StrategITcom, умный город определяется как «набор приложений, использующих общую безопасную инфраструктуру, центры обработки данных и хранилища данных на уровне устройств для передачи критически важных и некритических данных».

Нарушения безопасности умного города могут иметь очень серьёзные последствия — они могут быть опасными для экономики и даже для жизни, если с ними обращаться неправильно.

Интеллектуальные цифровые двойники



Goncharov; Nechesov 2023

AI-Driven Digital Twins for Smart Cities.
https://doi.org/10.3390/ecsa-10-16223

Задачный подход

ИСТОРИЯ СТАНОВЛЕНИЯ ЗАДАЧНОГО ПОДХОДА

Акад. СССР **А.Н.Колмогоров** – исчисление задач как интерпретация интуиционисткого исчисления высказываний (1932 г.), задачный подход к школьному образованию (50-60-е гг., XX век)

Г.С.Альтшуллер – автор ТРИЗ, задачный подход к инженерии (70-е гг., XX век – н.вр.)

Акад. РАН **Ю.Л.Ершов**, д.ф.н. **К.Ф.Самохвалов** – задачный подход к основаниям математики (70-80-е гг. XX века)

Акад. РАН **С.С.Гончаров**, д.ф.-м.н. **Е.Е.Витяев,** д.ф.-м.н. **Д.И.Свириденко** -- задачный подход к искусственному интеллекту (80-е гг., XX век – н.вр.)

Теория обучения и иерархия знаний

$(F(x,y), y = t(x), p)$ – это вероятностное знание

где

$F_i(x,y): \ \forall x \ \exists y \ \Phi_i(x,y) \to \Psi_i(x,y)$

Иерархия знаний

Иерархия знаний:

Мы можем говорить об иерархии знаний $(\leq_\varphi)$

$$(F_1(x,y),\ y = t_1(x),\ p_1) \leq_\varphi (F_2(x,y), y = t_2(x),\ p_2)$$

$$\Leftrightarrow$$

- $\Phi_1 \subseteq_\& \Phi_2$
- $\Psi_1 \subseteq_\& \Psi_2$
- $p_1 \leq p_2$

Теория обучения и иерархия знаний

Если мы хотим понять насколько эффективно решение $y = t(x)$, мы подставляем его вместо y:

$F_i(x, t(x)) : \ \forall x \ \Phi_i(x, t(x)) \to \Psi_i(x, t(x))$
и проверяем истинность этого утверждения на фактах, которыми мы обладаем

Оценка эффективности решения

Логическая формула (Задача) + Потенциальное решение

$\forall x\ \exists y\ \Phi(x,y) \to \Psi(x,y),\ y = t(x)$

База данных содержит все проблемы с решениями

$$F_k(x,y):\ \exists y\ \Delta_k(c_i,y) \to \Theta_k(c_i,y),\ y = t(c_i)$$

где $\Phi \subseteq_{\&} \Delta_k$

Probability value

Мы перебираем факты и считаем вероятность. Если фактов слишком много, то используем один из методов при работе с большими данными, например, RSP.

$$p(\Psi(x,t(x)) \mid \Phi(x,t(x))) = \frac{\sum_{i,k\in K} \mu(\Psi(c_i,t(c_i)))}{\sum_{i,k\in K} \mu(\Delta_k(c_i,t(c_i)))}$$

$\mu(\Phi(c_i,d_j)) = 1$, если $\Phi(c_i,d_j)$ - истина и 0 иначе.

Верификация нейронных сетей и ПО

Ануреев, Гаранина, Кондратьев из института систем информатики.

Проблема верификации

Проверка соответствия программного обеспечения его спецификации.

Верификация нейронных сетей

$$(\Phi(x), P_{Neural}(x, y), \Psi(x, y))$$

Далее с помощью Логики Хоара верифицируем нейронную сеть.

Спасибо за внимание!

Сергей Савостьянович Гончаров
Email: s.s.goncharov@math.nsc.ru
Андрей Витальевич Нечесов
Email: nechesoff@math.nsc.ru
Дмитрий Иванович Свириденко
Email: dsviridenko47@gmail.com
Центр Искусственного Интеллекта НГУ
Новосибирск, Академгородок