What Prevents the Use of Formal Methods for AI Verification

Igor Anureev

The Artificial Intelligence Research Center of Novosibirsk State University
1, Pirogova str., Novosibirsk, 630090, Russia

26 march 2025

Сложность моделей ИИ

- Комбинаторный взрыв возможных входов
- Комбинаторный взрыв возможных связей входов и выходов
- Нелинейность и сложность функций активации
- Стохастичность обучения и входных данных
- Динамическое обновление моделей на новых данных
- Трудности с моделированием непрерывных пространств состояний
- Разнообразие архитектур (CNN, RNN, Transformer и др.)

Ограничения формальных методов

- Высокая вычислительная сложность проверки
- Ограниченность существующих методов в отношении масштабных моделей
- Проблемы с полнотой формальных методов
- Отсутствие единого подхода для разных типов данных (текст, изображения, звук
- Проблемы с формализацией вероятностных моделей
- Проблемы с верификацией эвристик и приближений в моделях
- Ограниченность методов для адаптивных моделей
- Невозможность предсказать поведение модели на всех возможных входах

Ресурсные ограничения

- Недостаток вычислительных ресурсов
- Высокая стоимость проведения формальной верификации
- Длительное время, необходимое для верификации больших моделей
- Отсутствие инструментов, совместимых с реальными системами ИИ
  - Несовместимость с популярными ML-фреймворками (TensorFlow, PyTorch)
- Требование к глубокому знанию формальных методов со стороны разработчиков
- Ограничения реального времени
- Проблемы с верификацией сетей, обученных на больших объемах данных

Проблемы с данными

- Шумность данных и возможность ошибок при разметке
- Невозможность охватить все возможные случаи тестовыми данными
- Этические и правовые ограничения на использование некоторых наборов данных
- Динамическое изменение входных данных в реальных приложениях
- Отсутствие формальных спецификаций для требований к данным
- Влияние смещений (bias) в данных на результаты верификации

Организационные причины

- Недостаток специалистов по формальной верификации в области ИИ
- Нежелание компаний инвестировать в сложные и дорогостоящие методы
  - Конкурентное давление
- Отсутствие стандартов формальной верификации для ИИ
- Недостаток академических исследований, направленных на интеграцию формальных методов с ИИ
- Недостаточная осведомленность разработчиков о формальных методах

Ограничения современных алгоритмов верификации

- Нехватка автоматизированных инструментов для формальной верификации ИИ
- Неэффективность SMT-решателей для больших нейросетей
- Сложность доказательства устойчивости модели к небольшим изменениям входных данных
- Отсутствие поддержки динамических изменений в моделях
- Проблемы с кодированием нейросетей в логические ограничения
- Ограниченные методы поиска контрпримеров
- Невозможность проверять модели с плавающей запятой с высокой точностью
- Проблема экспоненциального роста вычислительных требований
- Плохая адаптируемость к задачам с неполной инрформацией

Проблемы с архитектурой нейросетей

- Глубокие нейросети имеют слишком сложную зависимость между слоями
- Отсутствие верификационных методов для гибридных систем
- Разнообразие топологий нейросетей
- Формальные методы плохо адаптируются к сетям с изменяющейся структурой
- Проблемы с проверкой работы слоев нормализации и активации
- Трудности с верификацией сетей, работающих в реальном времени
- Неясность, как корректно формализовать понятие "хорошей генерации"для генеративных моделей
- Рекуррентные сети (LSTM, GRU) имеют сложные временные зависимости
- Верификация многомодальных моделей

Ограничения в математических основах

- Отсутствие общепринятой формальной логики для описания поведения ИИ
- Трудности с построением исчерпывающих спецификаций для ИИ
- Неопределенность в трактовке вероятностных моделей
- Отсутствие единой модели математического представления различных архитектур ИИ
- Формальная верификация плохо адаптирована к динамическим системам
- Сложность формальной верификации решений, основанных на эвристиках
- Формальные методы не учитывают обучение с подкреплением в реальном мире
- Верификация больших языковых моделей требует новых логических формализмов
-
-

Проблемы, связанные с интерпретируемостью ИИ

- Формальная верификация не улучшает интерпретируемость моделей
- Трудности с интерпретацией формальных доказательств корректности
- Формальные методы не могут объяснить, как модель принимает решения
- Сложность интерпретации ошибок в формальной верификации
- Проблема определения "разумных"границ допустимых отклонений
- Формальная верификация не дает гарантии, что модель "понимает"проблему
- Формальная верификация не решает проблему доверия к модели
- Формальные методы не гарантируют объяснимость решений ИИ в суде или регуляторных органах