УДК 004.89

ОБОСНОВАНИЕ ДОВЕРИЯ К ИИ ЧЕРЕЗ ФОРМАЛЬНЫЕ МЕТОДЫ ВЕРИФИКАЦИИ ПРИВАТНОСТИ

Эль-Хассан Ф., студент направления Industrial Production Engineering
Bachelor’s degree program, II курс
Миланский технический университет, г. Милан, Италия

В условиях стремительного роста использования искусственного интеллекта в критически значимых сферах – от здравоохранения и финансов до судебной экспертизы и обороны – вопрос доверия к поведению ИИ-систем становится фундаментальным. Одним из центральных элементов этого доверия является способность таких систем защищать конфиденциальную информацию, особенно когда в их обучении участвуют чувствительные или персональные данные [1].

Повсеместное распространение крупных языковых моделей, генеративных сетей и автономных решений требует новых подходов к проверке их приватности и безопасности. В этом контексте особое значение приобретают формальные методы верификации приватности, обеспечивающие строгие, математически доказуемые гарантии соответствия ИИ-моделей требованиям конфиденциальности.

Формальные методы верификации в области ИИ – это совокупность логических, алгебраических и вероятностных подходов, которые позволяют моделировать, анализировать и доказывать определённые свойства системы, не прибегая к исключительно эмпирическим или эвристическим оценкам. Для задачи приватности эти методы дают возможность подтвердить, что модель не раскрывает критически важную информацию о конкретных образцах обучающей выборки, даже если атакующий имеет значительный доступ к модели, её параметрам или предсказаниям. Такая верификация особенно важна в сценариях open-model access, характерных для современных API-сервисов и платформ открытого машинного обучения.

Одним из наиболее известных формальных подходов к обеспечению приватности является дифференциальная приватность (Differential Privacy, DP). Её идея заключается в том, что поведение модели не должно существенно меняться при добавлении или удалении одного элемента из обучающей выборки. Это формализуется с помощью параметров ε (эпсилон) и δ (дельта), которые количественно определяют допустимый уровень утечки информации. Верификация на соответствие DP-свойствам может быть проведена с использованием

математических логик, таких как логика вероятностного времени (probabilistic temporal logic), и верификационных инструментов, таких как ProVerif, EasyCrypt и CertiPriv. Эти средства позволяют формализовать процесс обучения модели как вероятностную программу и доказать, что она соответствует заданному уровню приватности.

В дополнение к дифференциальной приватности, применяются и другие формальные методы. Например, информационно-теоретический анализ, основанный на оценке взаимной информации между входными данными и выходами модели, позволяет определить, насколько сильно результат модели зависит от конкретных обучающих примеров. Методы логического вывода и абстрактной интерпретации могут быть применены для анализа поведения модели в граничных и атакующих сценариях, выявляя потенциальные утечки, обусловленные архитектурными особенностями или спецификой функций активации [2].

Особый интерес вызывает использование символьных методов анализа (symbolic analysis) и автоматизированного доказательства теорем (automated theorem proving) для изучения приватности нейронных сетей. Эти подходы позволяют прослеживать, как изменения во входных данных отражаются на выходных предсказаниях, и формально доказывать, что определённые шаблоны или группы данных не могут быть восстановлены или идентифицированы в результате предсказания. Такие методы активно исследуются в рамках программной верификации и теперь начинают проникать в область машинного обучения, где требуют дополнительных адаптаций для работы с непрерывными и стохастическими пространствами признаков.

Формальные методы также находят применение в верификации поведения моделей после их обучения, например, при дистилляции, тонкой настройке или адаптации в новых условиях. Они позволяют доказывать, что изменения параметров или архитектуры модели не приводят к нарушению ранее установленных приватностных гарантий.

Это особенно важно в условиях федеративного обучения и распределённых вычислений, где нет централизованного контроля над всеми данными, и каждый участник может потенциально влиять на модель [3].

Практическое значение формальных методов верификации приватности заключается не только в их способности выявлять и устранять уязвимости, но и в их потенциале как основы регуляторной сертификации ИИ-систем. В отличие от эмпирических тестов, результаты формальной верификации могут быть представлены как формальные доказательства соответствия нормативным требованиям – например, положениям GDPR, HIPAA, ISO/IEC 27001 и других.

Это делает их незаменимыми в сферах, где юридическая подотчетность и техническая проверяемость являются обязательными условиями для внедрения.

Тем не менее, существует ряд вызовов. Формализация сложных моделей – особенно глубоких нейросетей с миллионами параметров – требует масштабируемых, но при этом строгих и надежных методов [4].

Многие современные инструменты верификации ограничены по размеру моделей и числу состояний, которые они способны обработать.

Поэтому активные исследования сосредоточены на разработке приближённых и модульных верификационных стратегий, позволяющих анализировать крупные модели частями или с определённой степенью допущения, без потери корректности заключений.

В заключение следует отметить, что обоснование доверия к ИИ через формальные методы верификации приватности представляет собой не только технический вызов, но и стратегическое направление, обеспечивающее фундаментальную прозрачность и подотчётность ИИ-систем.

Эти методы позволяют превратить декларативные обещания приватности в формализованные, проверяемые и доказуемые свойства, обеспечивающие уверенность пользователей, заказчиков, регуляторов и общества в целом в том, что искусственный интеллект может быть как мощным, так и безопасным инструментом в цифровом мире [5].

Список литературы:

1. Shamsabadi, A. S. Label-Only Membership Inference Attacks / A. S. Shamsabadi, M. Yaghini, N. Papernot, I. Shumailov // International Conference on Machine Learning. – 2020. – Vol. 119. – P. 8774–8784.

2. Choquette-Choo, C. A. Label-Only Membership Inference Attacks / C. A. Choquette-Choo, F. Tramèr, N. Carlini, N. Papernot // International Conference on Machine Learning. – 2021. – Vol. 139. – P. 1964–1974.

3. Heikkilä, M. Differentially Private Bayesian Learning on Distributed Data / M. Heikkilä, E. Lagerspetz, S. Kaski [et al.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 3229–3238.

4. Jälkö, J. Differentially Private Bayesian Learning with Efficient MCMC / J. Jälkö, O. Dikmen, A. Honkela // Machine Learning. – 2017. – Vol. 106. – P. 919–941. – DOI 10.1007/s10994-016-5643-7.

5. Foulds, J. On the Theory and Practice of Privacy-Preserving Bayesian Data Analysis / J. Foulds, J. Geumlek, M. Welling, K. Chaudhuri // Conference on Uncertainty in Artificial Intelligence. – 2016. – P. 192–201.