УДК 004.89

РАЗРАБОТКА ИНСТРУМЕНТАРИЯ ДЛЯ ТЕСТИРОВАНИЯ МОДЕЛЕЙ В УСЛОВИЯХ РАСПРЕДЕЛЁННЫХ АТАК

Сорокин Н.А., Старший оператор научной роты, III курс
Военная академия связи имени Маршала Советского Союза С. М. Буденного,
г. Санкт-Петербург

С ростом масштабов и сложности систем машинного обучения, особенно в распределённых и облачных средах, вопрос устойчивости моделей к разнообразным формам атак приобретает всё более критическое значение. Если ранее основное внимание в исследовательском и прикладном сообществе уделялось точечным (локальным) атакам на модели в изолированной среде, то в современных условиях в фокусе оказываются распределённые атаки, направленные на многокомпонентные ML-системы, развёрнутые на гетерогенных инфраструктурах. Такие атаки могут происходить одновременно в нескольких точках: на этапе обучения, при передаче данных, в подсистемах предобработки или агрегации, в распределённых узлах инференса или даже в каналах взаимодействия между микросервисами. Это создаёт вызов нового уровня – необходимость разработки специализированного инструментария для тестирования моделей в условиях распределённых атак, который бы позволял симулировать, оценивать и анализировать устойчивость ИИ-систем на всех этапах их жизненного цикла и во всех точках возможного уязвимого взаимодействия [1].

Распределённая архитектура характерна для многих современных приложений машинного обучения. В системах федеративного обучения, edge-интеллекта, мультиагентных сред, корпоративных data lake-инфраструктурах или в сценариях интернета вещей (IoT) данные и вычисления распределены по множеству узлов. Это делает модель уязвимой к координированным атакам, когда злоумышленник, действуя из нескольких точек, может влиять на агрегируемые градиенты, модифицировать поданные данные, искажать коммуникационные протоколы или внедрять ложные сигналы в процесс обучения. Возникают особые формы угроз: отравление модели через отдельные узлы (data poisoning), атаки на агрегацию параметров (например, при federated averaging), перехват и подмена моделей или данных при передаче, атакующая синхронизация и скрытые коалиции внутри участников распределённой сети. Тестировать модели в таких условиях с использованием традиционных инструментов невозможно – требуется иная методология [2].

Разработка инструментария для тестирования устойчивости в условиях распределённых атак требует, прежде всего, проектирования симуляционной среды, способной воспроизводить как топологию распределённой системы, так и вероятностные сценарии атак. Такая среда должна моделировать разные типы клиентов или агентов (честные, аномальные, атакующие), различную степень связности, асинхронность взаимодействий, сетевые задержки и частичную доступность каналов. Необходимо встроить модели поведения атакующих, способных имитировать реальный вредоносный трафик, внедрение аномалий в данные, случайное или намеренное искажение весов, перехват сообщений. Разработка таких симуляторов требует междисциплинарного подхода: сочетания теории вычислительных сетей, безопасности, криптографии и моделирования поведения агентов [3].

Важным направлением является модульность атакующих стратегий. Инструментарий должен позволять конструировать сложные сценарии из элементарных атак: например, соединять backdoor-подмену на одном узле с атакой на синхронизацию в другом и с перехватом параметров в точке передачи. Это даёт возможность тестировать модели в условиях комбинированных угроз, более близких к реальности, чем упрощённые одиночные атаки. Для этого используется концепция attack graphs и сценарные DSL (domain-specific languages), позволяющие описывать атаки как композиции действий во времени и пространстве. Подобные описания могут быть затем преобразованы в автоматизированные тестовые кейсы, что облегчает систематическую проверку [4].

Кроме симуляции атак, критически важно внедрение метрик оценки устойчивости в распределённой среде. В отличие от централизованного тестирования, где достаточно оценить изменение точности или уверенности модели под воздействием атаки, в распределённой системе необходимо учитывать комплексные характеристики: устойчивость агрегатора, долю компрометированных узлов, стабильность глобальной модели при пертурбациях локальных градиентов, устойчивость к атакующим синхронизациям, сходимость модели в присутствии вредоносных клиентов и пр. Это требует разработки новых показателей – например, robustness under Byzantine ratio, stability under partial poisoning, convergence under adversarial scheduling – которые могут использоваться в качестве формальных критериев [5].

Разрабатываемый инструментарий должен обеспечивать интеграцию с существующими ML-фреймворками(например, TensorFlow Federated, PySyft, Flower, FedML), чтобы его можно было использовать в продуктивных и исследовательских пайплайнах. Он должен быть расширяемым, поддерживать plug-in архитектуру атак и защит, а также предоставлять визуализацию состояния сети, модели и качества в реальном времени. Поддержка логирования,

трассировки и воспроизводимости (например, через DVC, MLflow, Weights & Biases) является необходимым элементом, особенно в условиях командной работы и подготовки к сертификации моделей.

Отдельное внимание заслуживают встроенные средства защиты, которые могут быть активированы в ответ на смоделированные атаки: от простых механизмов исключения подозрительных клиентов до использования робастных агрегационных схем (например, Krum, Bulyan, Trimmed Mean), а также продвинутых методов доверительных вычислений – многопартийного обучения, зашифрованного инференса, использования защищённых окружений (SGX, TrustZone) [6].

Важно, чтобы инструментарий не только помогал тестировать уязвимости, но и оценивал эффективность защитных механизмов, определяя их пределы применимости и сценарии, в которых они могут быть обойдены.

Наконец, инструментарий должен быть ориентирован на автоматизацию и стандартизацию тестирования. Это предполагает наличие тестовых профилей, шаблонов угроз, интеграции с регламентами (например, EU AI Act, ISO/IEC 23894), генерации аудиторских отчётов и совместимость с нормативными процедурами оценки моделей. В перспективе такой инструментарий может стать основой для автоматизированной процедуры сертификации систем федеративного и распределённого ИИ, что критически важно для внедрения в государственные, медицинские и военные системы.

В заключение можно сказать, что разработка инструментария для тестирования моделей в условиях распределённых атак – это не просто инженерная задача, а стратегическое направление в построении доверенных и масштабируемых ИИ-систем будущего. Такой инструментарий формирует основу архитектуры доверия в распределённых вычислениях, позволяет строить более устойчивые модели, повышает прозрачность их поведения и облегчает выполнение нормативных требований. Его развитие требует междисциплинарного подхода, включающего машинное обучение, кибербезопасность, распределённые системы, симуляцию и прикладную нормативику, и имеет прямое значение для цифрового суверенитета, устойчивости цифровой экономики и защиты прав пользователей.

Список литературы:

1. Bagdasaryan, E. How To Backdoor Federated Learning / E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov // Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. – 2020. – Vol. 108. – P. 2938-2948.

2. Blanchard, P. Machine Learning with Adversaries : Byzantine Tolerant Gradient Descent / P. Blanchard, E. M. El Mhamdi, R. Guerraoui, J. Stainer // Advances in Neural Information Processing Systems. – 2017. – Vol. 30.

3. Yin, D. Byzantine-Robust Distributed Learning : Towards Optimal Statistical Rates / D. Yin, Y. Chen, R. Kannan, P. Bartlett // Proceedings of the 35th International Conference on Machine Learning. – 2018. – Vol. 80. – P. 5650-5659.

4. Xie, C. DBA : Distributed Backdoor Attacks against Federated Learning / C. Xie, K. Huang, P.-Y. Chen, B. Li // International Conference on Learning Representations. – 2020.

5. Baruch, M. A Little Is Enough : Circumventing Defenses for Distributed Learning / M. Baruch, G. Baruch, Y. Goldberg // Advances in Neural Information Processing Systems. – 2019. – Vol. 32.

6. Kairouz, P. Advances and Open Problems in Federated Learning / P. Kairouz, H. B. McMahan, B. Avent [et al.] // Foundations and Trends in Machine Learning. – 2021. – Vol. 14, № 1-2. – P. 1-210. – DOI 10.1561/2200000083.