

Е. А. Басыня^{1✉}, В. Ю. Сапегин¹

Метод состязательного обучения моделей обнаружения виртуальных защищенных каналов связи на основе адаптивной генерации конфигураций Xray

¹Национальный исследовательский ядерный университет «МИФИ»,
г. Москва, Российская Федерация
e-mail: director@nii-ikt.ru

Аннотация. В работе представлен метод состязательного обучения моделей обнаружения виртуальных защищенных каналов связи, в которой генератор конфигураций на базе Xray адаптивно перебирает протокольные семейства и параметры маскировки, а бинарный детектор виртуальных защищенных каналов связи последовательно дообучается на трудных положительных примерах. Актуальность исследования определяется тем, что статические детекторы, обученные на ограниченном перечне туннелей, теряют устойчивость при появлении новых сочетаний транспорта, рукопожатия и параметров обфускации. Метод задает ограниченное, но репрезентативное пространство конфигураций Xray, правила валидации работоспособности кандидатных конфигов, порядок сбора трафика и цикл пополнения обучающей выборки трудными примерами. В статье описаны требования к экспериментальному протоколу, показателям качества и критериям остановки состязательного цикла. Практическая значимость работы состоит в переходе от разовой оценки детектора в статическом домене к воспроизводимой процедуре формирования более устойчивых моделей обнаружения виртуальных защищенных каналов связи.

Ключевые слова: виртуальные защищенные каналы связи, VPN, обнаружение VPN, Xray, состязательное обучение, анализ зашифрованного трафика

E. A. Basinya^{1✉}, V. Y. Sapegin¹

Method for adversarial training of virtual secure communication channel detection models based on adaptive Xray configuration generation

¹National Research Nuclear University MEPhI, Moscow, Russian Federation
e-mail: director@nii-ikt.ru

Abstract. The paper presents a methodology for adversarial training of virtual secure communication channel detection models. In the proposed setup, an Xray-based configuration generator adaptively enumerates protocol families and obfuscation parameters, while a binary VSCC detector is iteratively re-trained on hard positive examples. The study addresses the practical instability of static detectors trained on a narrow set of tunnel configurations. The methodology defines a constrained yet representative Xray configuration space, validity checks for generated candidates, traffic collection rules, and an iterative hard-example mining loop. The paper also specifies the experimental protocol, quality indicators, and stopping criteria for the adversarial cycle. The proposed approach is intended as a reproducible basis for building more robust virtual secure communication channel detection models.

Keywords: virtual private communication channels, VPN, VPN detection, Xray, adversarial training, encrypted traffic analysis

Введение

В условиях роста доли шифрованного сетевого обмена задача обнаружения виртуальных защищенных каналов связи приобретает для корпоративных и ведомственных сетей прикладное значение. Виртуальные защищенные каналы связи (ВЗКС) используются не только как средство легитимной защиты конфиденциальности, но и как инструмент обхода периметровой фильтрации, сокрытия командно-управляющего трафика и маскировки каналов вывода данных. По данным F.A.C.S.T. [1], использование технологий виртуальных защищенных каналов связи продолжало фиксироваться в сценариях целенаправленных атак на российские организации, а развитие нормативного регулирования мониторинга таких соединений дополнительно подчеркивается Приказом Роскомнадзора № 168 [2].

Современные исследования показывают, что даже полностью шифрованные и обфусцированные туннели сохраняют наблюдаемые статистические, транспортные и рукопожатийные особенности. В работе M. Wu и соавторов показано, что пассивная детекция *fully encrypted traffic* возможна даже для семейств, включающих VMess [3]. Хие и соавторы продемонстрировали практическую детектируемость OpenVPN и его обфусцированных конфигураций на реальном операторском трафике [4], а затем показали, что обфусцированные прокси и функционально эквивалентные ВЗКС туннели остаются уязвимыми к отпечаткам инкапсулированного TLS-рукопожатия [5]. Вместе с тем Хие и соавторы установили, что качество *encrypted traffic classification* существенно деградирует при смене сетевой среды и транспортных условий, даже если исходная модель демонстрирует высокий результат в лабораторном домене [6].

Проблема заключается в том, что большинство практических детекторов обучается на статическом перечне протоколов и фиксированных конфигураций. Такой подход создает системную асимметрию: защитная сторона оценивает модель на конечном наборе известных профилей, тогда как противоположная сторона может адаптивно варьировать транспорт, параметры рукопожатия и маскировки. В работах по *traffic analysis* уже показана уязвимость глубоких моделей к состязательным возмущениям [7], а также актуальность сертифицируемой робастности в мультимодальной постановке [8]. Обзор Guerra-Manzanares и соавторов дополнительно фиксирует слабую сопоставимость опубликованных результатов из-за зависимости от устаревших датасетов и неоднородных экспериментальных схем [9]. Дополнительно в работе Liu и соавторов продемонстрирована возможность атаки обхода системы обнаружения вредоносного трафика при доступе только к итоговой метке классификации, что подчеркивает практическую значимость оценки устойчивости детекторов в условиях ограниченной информации об атакуемой модели [10]. В отечественной литературе терминологические и прикладные основы ВЗКС, включая вопросы адаптивного управления защищенной коммуникационной инфраструктурой, последовательно рассматриваются в работах Е. А. Басыни [11–13].

Постановка задачи

Целью работы является разработка метода состязательного обучения бинарного детектора ВЗКС и легитимных сетевых потоков, в которой качество модели повышается не за счет однократного расширения датасета, а за счет управляемого цикла генерации новых конфигураций ВЗКС и последующего дообучения на наиболее трудных положительных примерах. Рассматриваются два агента. Первый агент представляет генератор конфигураций на базе Xray, второй - обычную модель бинарной детекции ВЗКС по потоковым или сеансовым признакам. На каждом шаге генератор формирует набор кандидатных конфигураций, стремясь получить работоспособные туннели, снижающие качество текущего детектора. Детектор, в свою очередь, после анализа нового трафика обновляет параметры и повышает устойчивость к найденным вариантам маскировки.

Принципиально важно, что генератор в данном методе не максимизирует абстрактное конфигурационное разнообразие. Его целевая функция должна учитывать четыре ограничения: синтаксическую корректность конфигурации, успешное установление соединения, сохранение полезной сетевой функциональности и реалистичность получаемого трафикового профиля. Иными словами, в обучающую петлю допускаются только те конфигурации, которые могут быть развернуты в лабораторном стенде без ручной доработки и генерируют наблюдаемый ВЗКС-трафик, а не формально корректный, но неработоспособный JSON. Такое требование исключает загрязнение обучающей выборки артефактами генератора и удерживает постановку задачи в прикладной плоскости.

Пространство адаптивного перебора ограничивается умеренным, но репрезентативным набором конфигураций Xray, сформированным по актуальной документации Project X [14]. В генератор включаются протокольные семейства VLESS и VMess, режимы транспортной защиты TLS и REALITY, транспортные механизмы WebSocket, gRPC, mKCP и XHTTP, а также параметры mux, padding, ALPN, SNI, path, serviceName и схемы маршрутизации. Такое ограничение позволяет, с одной стороны, не расплзтись до практически необозримого пространства вариантов, а с другой стороны, отразить типовые направления эволюции скрытых Xray-конфигураций, реально влияющие на трафиковый отпечаток.

Архитектура состязательного обучения

Генератор конфигураций целесообразно реализовывать как ограниченный поисковый или reinforcement learning-модуль, работающий поверх библиотеки допустимых трансформаций. Входом для него являются история успешных уклонений, уверенности текущего детектора, а также таблица разрешенных операций изменения конфигурации. На одном шаге генератор может сменить протокольное семейство, транспорт, режим защиты, параметры маскировки, поля заголовков или режим мультиплексирования; на следующем шаге - скорректировать маршрут и дополнительные сервисные параметры. Награда генератора должна возрастать при успешном обходе детектора, но уменьшаться при росте задержек, снижении стабильности соединения и появлении явно искусственных

артефактов трафика. Тем самым генератор учится искать не любой обход, а обход, совместимый с реальной эксплуатацией ВЗКС.

Детектор ВЗКС в методике намеренно рассматривается как архитектурно нейтральный компонент. Это может быть градиентный бустинг на статистических flow-признаках, сверточная или рекуррентная сеть на последовательностях длин и межпакетных интервалов либо трансформерная модель на кодированном представлении сессии. Ключевым требованием является не выбор конкретной нейросетевой формулы, а наличие единого признакового пространства, отражающего длины пакетов, направления, темпоральную структуру обмена, наблюдаемые handshake-атрибуты и время до вынесения решения. Дополнительно следует исключать признаки, жестко привязанные к одной версии протокола или единичной конфигурационной особенности, поскольку именно такие признаки быстрее всего разрушаются при адаптивном переборе Xray-настроек.

Состязательный цикл строится следующим образом. На начальном шаге формируется базовый корпус доброкачественного трафика и трафика ВЗКС, полученного от ограниченного числа заведомо рабочих конфигураций. После первичного обучения детектора генератор начинает выпускать новые кандидатные конфигурации, которые автоматически проверяются на разворачиваемость и успешное прохождение контрольного сценария. Далее для каждой принятой конфигурации записывается трафик, выполняется вывод модели и выделяются трудные положительные примеры, то есть такие сессии ВЗКС, которые детектор ошибочно отнес к легитимным сетевым потокам либо распознал только после недопустимо длинного окна наблюдения. Эти примеры вместе с сопоставимыми по фону доброкачественными сессиями добавляются в обучающую выборку, после чего генератор и детектор переходят к следующему витку.

Такой цикл не следует останавливать после первого же снижения ошибки. Корректнее завершать обучение при выполнении одновременно двух условий: во-первых, доля новых валидных конфигурационных семейств, которые еще обходят детектор, становится малой; во-вторых, очередное дообучение не дает заметного прироста по устойчивости на отложенном наборе ранее не встречавшихся комбинаций протокола, транспорта и параметров маскировки. Именно наличие отложенного пространства неизвестных конфигураций отличает метод от простого hard example mining на одном и том же семействе туннелей.

Экспериментальный протокол

Для практического применения метода лабораторный стенд должен рассматриваться как генератор контролируемых доменов наблюдения. В каждом эксперименте фиксируются клиентская платформа, тип сети, география серверной точки, момент времени, тип фоновой нагрузки и точка съема трафика. Минимально целесообразно варьировать настольные и мобильные клиенты, проводной и беспроводной доступ, а также несколько маршрутных сценариев до серверов в разных юрисдикциях. При этом задача остается бинарной: определяется наличие или отсутствие ВЗКС, а не конкретное приложение внутри туннеля.

Сбор данных необходимо организовывать так, чтобы исключить утечки между обучением и тестированием. Разбиение должно выполняться не только по отдельным сессиям, но и по временным срезам, платформам и целым семействам конфигураций. Если одна и та же комбинация, например VLESS-REALITY-ХНТТР с конкретным шаблоном padding, встречается и в обучении, и в тесте, то измеряется запоминание, а не переносимость. Поэтому часть конфигураций и их мутаций должна полностью удерживаться в неизвестном тестовом пуле, используемом только для итоговой оценки устойчивости.

Оценка качества детектора должна включать не только стандартные метрики классификации, такие как F1 и AUPRC, но и показатели, отражающие эксплуатационную пригодность. К ним относятся уровень ложных срабатываний на обычном TLS-трафике, время или объем наблюдения до вынесения решения, а также устойчивость к ранее не встречавшимся конфигурациям. Со стороны генератора необходимо отдельно фиксировать долю синтаксически валидных конфигов, долю реально устанавливающих соединение конфигов и долю конфигов, которые на очередной итерации обходят детектор. Такое двустороннее измерение позволяет судить не только о качестве защитной модели, но и о том, насколько быстро исчерпывается полезный потенциал адаптивного перебора.

Важным элементом предлагаемого метода является единообразие вычислительного бюджета. Все сравниваемые версии детектора должны получать одинаковое окно наблюдения, одинаковую процедуру предобработки и одинаковый набор фоновых легитимных сценариев. Аналогично, генератор на каждой итерации должен работать в фиксированном лимите числа попыток и числа допустимых модификаций одной конфигурации. Иначе выигрыш может быть обусловлен не улучшением самой модели, а банальным увеличением переборного ресурса.

Ограничения и практические риски

Главный риск состязательной схемы состоит в переобучении на узкий, пусть и динамически обновляемый, набор Xray-специфичных артефактов. Если генератор регулярно находит обходы за счет одной и той же группы признаков, детектор может стать излишне чувствительным именно к ним и ухудшить переносимость на другие семейства ВЗКС. Для предотвращения этого требуется квотирование протокольных и транспортных подпространств, контроль баланса между положительным и отрицательным классом, а также отдельная проверка на обычном HTTPS-, HTTP/2- и HTTP/3-трафике. Не менее существенен риск загрязнения датасета невалидными конфигурациями, поэтому этап автоматической валидации должен предшествовать любой записи трафика.

Второй класс ограничений связан со стоимостью эксперимента и изменчивостью внешней среды. С обновлением Xray, сменой реализаций транспорта или изменением поведения промежуточных сетевых узлов трафиковый профиль одной и той же конфигурации может заметно измениться. Следовательно, получаемая модель не должна рассматриваться как окончательно устойчивая; она представляет собой состояние детектора, достигнутое при фиксированной версии

инструментов и контролируемом наборе доменов. Рассматриваемый метод предполагает регулярное повторение цикла на новых временных срезах и обязательное версионирование всех конфигурационных шаблонов и окружений. Эксперименты при этом должны проводиться только в изолированном стенде, без вывода состязательно сгенерированных конфигураций в эксплуатационную сеть.

Заключение

Предложенный метод переводит задачу обнаружения ВЗКС из статической плоскости в постановку управляемого состязательного обучения. Научная новизна подхода состоит в совместном рассмотрении двух процессов: адаптивной генерации реалистичных Хгау-конфигураций и последовательного дообучения бинарного детектора на трудных положительных примерах. В отличие от традиционного расширения датасета, здесь новая обучающая информация формируется как результат направленного поиска конфигураций, максимально неудобных для текущей версии модели.

Практическая значимость метода заключается в формировании воспроизводимого протокола разработки более устойчивых детекторов ВЗКС. Такой протокол позволяет сравнивать модели не только по качеству в базовом домене, но и по устойчивости к ранее не встречавшимся комбинациям протокола, транспорта и маскировки. В дальнейшем разработанный метод может служить основой для полноценных экспериментальных исследований, в которых будут сопоставляться различные архитектуры детекторов и различные стратегии адаптивного генератора конфигураций.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Universal Encryptor: The number of cyberattacks in Russia increased by 140% in 2023 // F.A.C.C.T. URL: <https://www.facct.ru/media-center/press-releases/cyberattacks-2023/> (accessed: 20.04.2026).

2. Приказ Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций «О внесении изменений в Критерии оценки материалов и (или) информации, необходимых для принятия Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций решений, являющихся основаниями для включения доменных имен и (или) указателей страниц сайтов в информационно-телекоммуникационной сети «Интернет», а также сетевых адресов, позволяющих идентифицировать сайты в информационно-телекоммуникационной сети «Интернет», в единую автоматизированную информационную систему «Единый реестр доменных имен, указателей страниц сайтов в информационно-телекоммуникационной сети «Интернет» и сетевых адресов, позволяющих идентифицировать сайты в информационно-телекоммуникационной сети «Интернет», содержащие информацию, распространение которой в Российской Федерации запрещено», утвержденные приказом Федеральной службы по надзору в сфере связи, информационных технологий и массовых коммуникаций от 27 февраля 2023 г. № 25» от 08.11.2023 № 168 // Собрание законодательства Российской Федерации. 01.12.2023.

3. Wu M., Sippe J., Sivakumar D. [et al.]. How the Great Firewall of China Detects and Blocks Fully Encrypted Traffic // 32nd USENIX Security Symposium (USENIX Security 23). 2023. P. 2653-2670.

4. Xue D., Ramesh R., Jain A. [et al.]. OpenVPN Is Open to VPN Fingerprinting // 31st USENIX Security Symposium (USENIX Security 22). 2022. P. 483-500.

5. Xue D., Kallitsis M., Houmansadr A., Ensafi R. Fingerprinting Obfuscated Proxy Traffic with Encapsulated TLS Handshakes // 33rd USENIX Security Symposium (USENIX Security 24). 2024. P. 2689-2706.
6. Xie R., Cao J., Dong E. [et al.]. Rosetta: Enabling Robust TLS Encrypted Traffic Classification in Diverse Network Environments with TCP-Aware Traffic Augmentation // 32nd USENIX Security Symposium (USENIX Security 23). 2023. P. 625-642.
7. Nasr M., Bahramali A., Houmansadr A. Defeating DNN-Based Traffic Analysis Systems in Real-Time With Blind Adversarial Perturbations // 30th USENIX Security Symposium (USENIX Security 21). 2021. P. 2705-2722.
8. Yan J., Liu Z., Xie Y. [et al.]. CertTA: Certified Robustness Made Practical for Learning-Based Traffic Analysis // 34th USENIX Security Symposium (USENIX Security 25). 2025.
9. Guerra-Manzanares A., Caprolu M., Di Pietro R. A Comprehensive Review on Machine Learning-Based VPN Detection: Scenarios, Methods, and Open Challenges // Computer Science Review. 2025. Vol. 58. Article 100781. DOI: 10.1016/j.cosrev.2025.100781.
10. Liu Z., Zhao Y., Liu Z. [et al.]. A Hard-Label Black-Box Evasion Attack against ML-based Malicious Traffic Detection Systems // Network and Distributed System Security Symposium (NDSS). 2026. DOI: 10.14722/ndss.2026.230916.
11. Басыня Е. А. Система самоорганизующегося виртуального защищенного канала связи // Защита информации. Инсайд. 2018. № 5 (83). С. 10-15.
12. System of a Self-Organizing Virtual Secure Communication Channel Based on Stochastic Multi-Layer Encryption and Overlay Technologies / E. A. Basynya, Z. B. Akhayeva, D. H. Omarkhanova, G. B. Tolegenova [et al.] // Journal of Theoretical and Applied Information Technology. 2022. Vol. 100, iss. 16. P. 4918-4927.
13. Басыня Е. А. Программная реализация и исследование системы интеллектуально-адаптивного управления информационной инфраструктурой предприятия // Вестник Самарского государственного технического университета. Серия: Технические науки. 2020. Т. 28, № 1 (65). С. 6-21.
14. Xray Configuration Reference // Project X. URL: <https://xtls.github.io/en/config/> (accessed: 20.04.2026).

© Е. А. Басыня, В. Ю. Сатегин, 2026