DOI https://doi.org/10.46845/2541-8254-2026-3(55)-65-65
УДК 004.855.5/004.056.53

ЗАЩИТА НЕЙРОННОЙ СЕТИ ОТ УГРОЗЫ КРАЖИ МОДЕЛИ С ПОМОЩЬЮ ИИ

Р.А. Габитов, студент,
e-mail: rostislav.gabitov@gmail.com
ФГБОУ ВО «Калининградский государственный технический университет»

В.В. Подтопельный, ст. преподаватель,
e-mail: vladislav.podtopelnyj@klgtu.ru
ФГБОУ ВО «Калининградский государственный технический университет»

В работе рассмотрена угроза кражи нейронных сетей посредством атаки Copycat CNN, при которой злоумышленник создаёт функциональную копию модели через её публичный программный интерфейс. Воспроизведена атака на модель Oracle, обученную на датасете CIFAR-10, и исследована зависимость согласованности модели-имитатора от числа запросов. Предложен детектор аномалий на основе алгоритма изолирующего леса (Isolation Forest), использующий пять признаков поведения сессии. Проведён сравнительный эксперимент по шести сценариям атаки нарастающей сложности. Определены ограничения разработанного решения и сформулированы рекомендации по его применению в реальных системах.

Ключевые слова: *нейронная сеть, кража модели, Copycat CNN, атака на программный интерфейс, обнаружение аномалий, Isolation Forest, информационная безопасность ИИ.*

ВВЕДЕНИЕ

Интенсивное развитие технологий машинного обучения сформировало новый класс угроз информационной безопасности – атаки, направленные непосредственно против нейронных сетей. Начиная с 2024 года число зафиксированных атак на ИИ-системы возросло втрое; при этом средний ущерб от инцидентов, связанных с нарушением безопасности ИИ-систем (включая кражу или утечку обученных моделей), оценивается в 14,6 млн долл. по данным аналитического отчёта DataFence Cost of a Data Breach за 2025 год [8]. Особую опасность представляет то, что 91% организаций не располагает специализированными средствами защиты программных интерфейсов своих нейронных сетей.

Атака Copycat CNN позволяет злоумышленнику создать функциональную копию модели, обратившись к её публичному API с произвольными изображениями [1]. В отличие от традиционных угроз, такие запросы неотличимы от легитимных: стандартные межсетевые экраны, системы обнаружения вторжений (IDS) и системы предотвращения утечек (DLP) не содержат правил для их идентификации.

Дополнительную значимость теме придаёт ряд публично известных прецедентов: исследователи в 2024 году продемонстрировали, что параметры коммерческой нейронной сети могут быть извлечены через стандартный API при стоимости атаки менее 2 000 рублей. В начале 2026 года компания Anthropic зафиксировала попытку промышленного шпионажа посредством дистилляции модели со стороны конкурентов – было зарегистрировано более 16 миллионов несанкционированных взаимодействий с целью кражи способностей нейросети Claude. Именно это определяет актуальность разработки специализированных методов обнаружения таких атак.

Теоретическую базу исследования составили исключительно англоязычные источники. Это обусловлено тем, что изучение специфики атак типа Copycat CNN на программные интерфейсы нейронных сетей, а также разработка методов защиты от них, на данный момент

представлены только в зарубежной научной литературе. Релевантные отечественные публикации по данной теме в настоящее время отсутствуют.

ОБЪЕКТ ИССЛЕДОВАНИЯ

Объектом исследования является модель Oracle – свёрточная нейронная сеть для классификации изображений, доступная через REST API, и атака Copycat CNN, позволяющая создать её модель-имитатор. Oracle обучен на датасете CIFAR-10: 50 000 обучающих изображений размером 32×32 пикселя (RGB) по 10 классам (самолёт, автомобиль, птица, кошка, олень, собака, лягушка, лошадь, корабль, грузовик). Архитектура Oracle включает три свёрточных блока и полносвязный классификатор с итоговой точностью ~83%.

ЦЕЛЬ И ЗАДАЧИ ИССЛЕДОВАНИЯ

Цель работы – исследование механизма атаки Copycat CNN на нейронные сети и разработка детектора аномалий для её обнаружения с оценкой эффективности в сравнении с существующими методами защиты.

Задачи исследования: 1) изучить механизм атаки Copycat CNN; 2) воспроизвести атаку и построить кривую согласованности модели-имитатора с Oracle; 3) разработать ИИ-детектор на основе Isolation Forest; 4) провести сравнительный эксперимент по шести сценариям атаки; 5) определить ограничения и область применения разработанного решения.

МЕТОДЫ ИССЛЕДОВАНИЯ

Эксперимент выполнен с использованием Python 3.14, PyTorch (обучение и инференс Oracle), scikit-learn (Isolation Forest), NumPy (генерация данных) и Flask (REST API). Атака Copycat CNN реализована согласно методологии Коррейа-Сильва [1]: на первом этапе произвольные изображения отправляются к API Oracle, ответы которого формируют обучающую выборку для модели-имитатора; на втором – имитатор обучается на накопленных парах «изображение–метка». Архитектура Oracle представлена на рисунке 1.

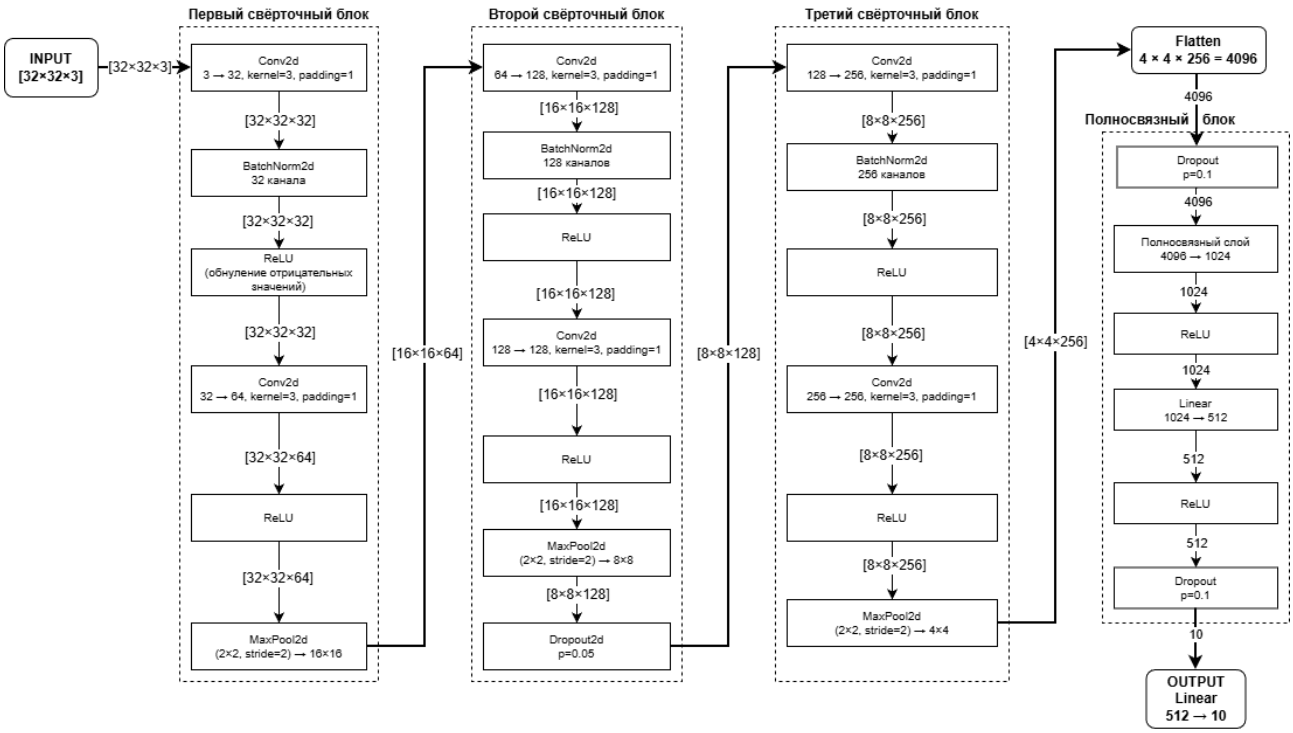

Рисунок 1 – Архитектура модели Oracle (три свёрточных блока + полносвязный)

Детектор аномалий основан на алгоритме Isolation Forest [4]. Алгоритм строит 100 деревьев изоляции и вычисляет для каждой сессии оценку аномальности. Порог срабатывания установлен на уровне 3% наиболее нетипичных сессий. На вход детектора подаётся вектор из

пяти признаков, извлекаемых из реальных ответов Oracle: число запросов в сессии, максимальное число уникальных классов в скользящем окне из четырёх запросов, энтропия распределения предсказанных классов, средняя уверенность Oracle и средний интервал между запросами. Перед подачей в модель признаки стандартизируются, чтобы численно большие признаки не подавляли малые. Выбор Isolation Forest обусловлен тремя свойствами: алгоритм обучается только на нормальном поведении (примеры атак не требуются), эффективно работает при малом числе признаков и производит предсказание менее чем за 2 мс. Обучающая выборка детектора содержала 500 примеров нормального поведения и 500 примеров поведения атакующего; допустимая доля аномалий при обучении составила 3%.

Для оценки качества детектора использованы стандартные метрики бинарной классификации:

1. Accuracy — доля всех верных решений детектора
2. Precision — насколько можно доверять блокировке (не ложная ли она)
3. Recall — какую долю реальных атак детектор поймал
4. F1-мера — баланс между Precision и Recall
5. FPR — доля легитимных пользователей, заблокированных по ошибке
6. Специфичность — доля легитимных сессий, пропущенных верно

Ниже приведен пример при 100 сессиях.

Таблица – Пример показателей работы детектора

Обозначение	Что это	Число
ИП (истинно положительные)	Атаки, которые детектор правильно заблокировал	50
ЛП (ложно положительные)	Нормальные сессии, которые детектор ошибочно заблокировал	3
ИО (истинно отрицательные)	Нормальные сессии, которые детектор правильно пропустил	47
ЛО (ложно отрицательные)	Атаки, которые детектор пропустил	0

Пример расчета метрик эффективности:

1. Accuracy (Общая точность) = 97%
   - (ИП + ИО) / (ИП + ЛП + ИО + ЛО)
   - (50 + 47) / 100 = 0,97
2. Precision (Точность обнаружения) = 94,3%
   - ИП / (ИП + ЛП)
   - 50 / (50 + 3) = 0,943
3. Recall (Полнота) = 100%
   - ИП / (ИП + ЛО)
   - 50 / (50 + 0) = 1,0
4. F1-мера = 97,1%
   - (2 × ИП) / (2 × ИП + ЛП + ЛО)
   - (2 × 50) / (2 × 50 + 3 + 0) = 0,971
5. FPR (Доля ложных блокировок) = 6,0%
   - ЛП / (ЛП + ИО)
   - 3 / (3 + 47) = 0,06
6. Специфичность = 94,0%
   - ИО / (ИО + ЛП)
   - 47 / (47 + 3) = 0,94

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Результаты атаки Copycat CNN. Модель-имитатор достигла точности ~73% при точности Oracle ~83%. Согласованность имитатора с Oracle составила ~80%, что означает

совпадение ответов в четырёх из пяти случаев. Для получения максимального результата модель-имитатор обучалась несколько раз на одних и тех же накопленных данных; из всех запусков выбирался вариант с наивысшей согласованностью, поскольку при стохастическом обучении точность между запусками может отличаться на 2–5%. Кривая согласованности в зависимости от числа запросов к Oracle представлена на рисунке 2.

Рисунок 2 – Согласованность модели-имитатора от числа запросов к Oracle

Кривая демонстрирует устойчивый рост: при 2 000 запросах – 45%, при 4000 – 52%, при 6000 – 67%, при 8000 – 74%, при 10000 – 80%. Стоимость атаки при платном API не превысила 2 000 рублей. Атака при этом невидима для стандартных средств ИБ: запросы к API неотличимы от легитимных обращений.

Пять признаков детектора и их обоснование. Каждый из пяти признаков отражает свойство, которое злоумышленник не может скрыть без ущерба для эффективности самой атаки. Первый признак – число запросов в сессии: снижение интенсивности замедляет кражу, но не устраняет аномалию. Второй – максимальное число уникальных классов в скользящем окне из четырёх запросов: для полноценной кражи необходимы все десять классов, что неизбежно проявляется в окне. Третий – энтропия распределения предсказанных классов: при равномерном покрытии всех классов энтропия возрастает до 0,9–2,0 против нормы 0,1–0,7. Четвёртый – средняя уверенность Oracle: произвольные изображения, используемые атакующим, физически хуже пользовательских, и модель отвечает на них с меньшей уверенностью (~88–91% против ~94–96%). Пятый – средний интервал между запросами: скрипт-атакующий работает с интервалами 0,005–0,05 с, тогда как живой пользователь делает паузы от 8 секунд и более.

Все пять признаков детектора были подобраны и откалиброваны специально для модели Oracle, обученной на датасете CIFAR-10 (10 сбалансированных классов, 32×32 пикселя). При переходе на датасеты с иным числом классов, другим разрешением или несбалансированной выборкой пороговые значения признаков потребуют повторной калибровки: в частности, признак «максимальное число уникальных классов в окне» напрямую зависит от общего числа классов, а признак «средняя уверенность Oracle» от сложности задачи классификации. Использование предложенного подхода на других датасетах возможно при условии перегенерации обучающей выборки детектора в условиях целевой системы.

Результаты сравнительного эксперимента. Было рассмотрено шесть сценариев атаки нарастающей сложности. Стандартное ограничение числа запросов (Rate Limit) обходится в

четырёх из шести сценариев: злоумышленник снижает интенсивность и продолжает кражу. Детектор Isolation Forest показал 100% обнаружение в сценарии «неосведомлённый атакующий», 80–90% – при знании порога, 16–42% – при знании двух признаков, 4–14% – при знании трёх, 8–16% – при знании четырёх. В сценарии адаптивной атаки с обратной связью обнаружение составило 0%. Сводные результаты приведены в таблице 1.

Таблица – Сравнение Rate Limit и Isolation Forest по сценариям атаки (А – атака обнаружена, %)

Сценарий атаки	Что знает	Rate Limit	Isolation Forest
Неосведомлённый	Ничего	Обходятся	100%
Знает 1 признак	Порог запросов	Обходятся	80–90%
Знает 2 признака	Порог + ритм	Обходятся	16–42%
Знает 2 признака	Порог + ритм + классы	Обходятся	4–14%
Знает 4 признака	Порог + ритм + классы + энтропия	Обходятся	8–16%
Адаптивный (все 5)	Все 5 признаков	Обходятся	0%

Ложные блокировки легитимных пользователей: 4–12% для «обычного» пользователя (1–2 класса, малое число запросов) и 0% для «активного» (5–7 классов, много разнообразных запросов), что свидетельствует об удовлетворительном балансе точность/отзыв при реальной эксплуатации.

Новизна предложенного подхода. В отличие от Rate Limit (один признак – счётчик запросов), Watermarking (встраивает скрытые паттерны в ответы) [9], Adaptive Misinformation [10] (искажает ответы, ухудшая пользовательский опыт) и Top1-only (скрывает вероятности, но атака продолжает работать), разработанный детектор анализирует поведенческий отпечаток сессии по пяти признакам, извлечённым из ответов самого Oracle, не искажая их для легитимных пользователей. Ключевое физическое ограничение: злоумышленник не может одновременно запрашивать все классы и сохранять типичную для легитимного пользователя энтропию – любая попытка скрыть один признак усиливает аномалию по-другому признаку, если не учитывать взаимосвязь остальных признаков.

ЗАКЛЮЧЕНИЕ

Атака Copycat CNN представляет реальную угрозу для коммерческих систем на основе нейронных сетей: при затратах менее 2 000 рублей злоумышленник получает модель с ~80% согласованностью с оригиналом. Стандартное ограничение числа запросов неэффективно против умеренно осведомлённого атакующего: четыре из шести исследованных сценариев его тривиально обходят.

Разработанный детектор на базе Isolation Forest надёжно обнаруживает неадаптивные атаки и сохраняет работоспособность при частичной осведомлённости злоумышленника (сценарии 1–5), однако уязвим к полностью адаптивным атакам с обратной связью (сценарий 6, обнаружение 0%). Устранение этой уязвимости требует динамической адаптации признаков или привлечения дополнительных источников данных (например, телеметрии сетевого уровня).

Детектор следует рассматривать как один из слоёв эшелонированной защиты, а не как самостоятельное решение. Практическая рекомендация: совмещать Isolation Forest с ограничением числа запросов (Rate Limit) и механизмами встраивания водяных знаков в ответы модели.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Correia-Silva J.R., Berriel R.F., Badue C., de Souza A.F., Oliveira-Santos T. Copycat CNN: Stealing Knowledge by Persuading Confession with Random Non-Labeled Data// arXiv, 2018. – URL: https://arxiv.org/abs/1806.05476 (дата обращения: 18.03.2026)

2. Correia-Silva J. R., Berriel R. F., Badue C., De Souza A. F., Oliveira-Santos T. Copycat CNN: Are Random Non-Labeled Data Enough to Steal Knowledge from Black-box Models? // arXiv, 2021. – URL: https://arxiv.org/abs/2101.08717 (дата обращения: 18.03.2026)

3. Correia-Silva J. R., Berriel R. F., Badue C., De Souza A. F., Oliveira-Santos T. Copycat CNN: Are Random Non-Labeled Data Enough to Steal Knowledge from Black-box Models? // arXiv, 2021. – URL: https://arxiv.org/abs/2101.08717 (дата обращения: 18.03.2026)

4. Liu F.T., Ting K.M., Zhou Z.-H. Isolation Forest // Data Mining, 2008. – URL: https://www.researchgate.net/publication/224384174_Isolation_Forest (дата обращения: 18.03.2026)

5. Zhao K., et al. A Systematic Survey of Model Extraction Attacks and Defenses: State-of-the-Art and Perspectives // arXiv, 2025. – URL: https://arxiv.org/abs/2508.15031 (дата обращения: 18.03.2026)

6. Xu A., Lam S. System Design Interview: An Insider's Guide. – Byte Code LLC, 2020. – Vol. 2. – 269 p.

7. Correa-Silva J.R. Stealing_DL_Models / Example_of_use // GitHub, 2021. – URL: https://github.com/jeiks/Stealing_DL_Models/tree/master/Example_of_use (дата обращения: 18.03.2026)

8. Cost of a Data Breach: AI Data Breaches Cost $14.6M in 2025 // DataFence, 2025. – URL: https://www.datafence.ai/blog/cost-of-ai-data-breaches-2025 (дата обращения: 18.03.2026).

9. Chen H., Rouhani B. D., Fan X., Kilinc O. C., Koushanfar F. Performance Comparison of Contemporary DNN Watermarking Techniques // arXiv, 2019. – URL: https://arxiv.org/abs/1811.03713 (дата обращения: 18.03.2026)

10. Kariyappa S., Qureshi M. K. Defending Against Model Stealing Attacks with Adaptive Misinformation // arXiv, 2020. – URL: https://arxiv.org/abs/1911.07100 (дата обращения: 18.03.2026)

PROTECTION OF NEURAL NETWORKS AGAINST MODEL THEFT USING AI

R.A. Gabitov, student,
e-mail: rostislav.gabitov@gmail.com
Kaliningrad State Technical University

V.V. Podtopelny, Senior Lecturer,
e-mail: vladislav.podtopelnyj@klgtu.ru
Kaliningrad State Technical University

The paper examines the threat of neural network theft via the Copycat CNN attack, in which an adversary creates a functional copy of a target model by querying its public API with arbitrary images. The attack was reproduced against the Oracle model trained on the CIFAR-10 dataset, and the agreement between the surrogate and Oracle was studied as a function of query count. An anomaly detector based on the Isolation Forest algorithm was proposed, utilizing five behavioral session features extracted from real Oracle responses. A comparative experiment across six attack scenarios of increasing sophistication was conducted. The limitations of the proposed solution are identified and practical recommendations for its deployment in real-world systems are formulated.

Key words: *neural network, model theft, Copycat CNN, API attack, anomaly detection, Isolation Forest, AI security.*