DOI: 10.15514/ISPRAS-2026-38(4)-28

AWA: выравнивание весов на основе активаций для повышения устойчивости статических методов маркирования нейронных сетей

[1] *А.А. Акименков, ORCID: 0009-0006-7710-941X <a.akimenkov@ispras.ru>*
[1] *Д.О. Обыденков, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *А.Ю. Якушев, ORCID: 0000-0001-6089-6505 <yakushev@ispras.ru>*
[2] *Х.Н. Абуд, ORCID: 0009-0009-7131-5839 <khaled.abud@graphics.cs.msu.ru>*
[1] *А.А.Устинова, ORCID: 0009-0009-9692-2774 <alisaustinova@ispras.ru>*
[1] *Ю.В. Маркин, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25.*

[2] *Институт искусственного интеллекта МГУ имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д. 1.*

Аннотация. Рассматривается задача повышения устойчивости статических методов маркирования нейронных сетей в сценарии извлечения «белый ящик», при котором цифровой водяной знак (ЦВЗ) извлекается непосредственно из параметров модели. Основное внимание уделяется атаке согласованной перестановки весов, сохраняющей качество работы нейронной сети, но изменяющей порядок параметров, используемый при извлечении водяного знака. Для противодействия такой атаке предлагается алгоритм выравнивания весов на основе активаций AWA, который сопоставляет структурные элементы исходной и атакованной моделей по картам активаций и восстанавливает порядок весов перед извлечением ЦВЗ. В рамках работы AWA интегрируется в процедуру извлечения метода NeuralMark, поскольку данный метод обладает высокой устойчивостью к ряду атак модификации модели, но остается чувствительным к перестановке параметров. Экспериментальная оценка выполнена на нескольких наборах данных, архитектурах нейронных сетей и при разных атаках модификации модели. Результаты показывают, что применение AWA повышает устойчивость NeuralMark к атакам перестановки и сохраняет высокую устойчивость к другим рассмотренным атакам модификации модели. В частности, при атаке согласованной перестановки весов NeuralMark-AWA снижает значение метрики BER на CIFAR-10 и Caltech-101, а значение TPR@FPR увеличивается.

Ключевые слова: цифровые водяные знаки; нейронные сети; статическое маркирование; защита интеллектуальной собственности; атаки на водяные знаки; перестановка весов; выравнивание весов на основе активаций.

Для цитирования: Акименков А.А., Обыденков Д.О., Якушев А.Ю., Абуд Х.Н., Устинова А.А., Маркин Ю.В. AWA: выравнивание весов на основе активаций для повышения устойчивости статических методов маркирования нейронных сетей. Труды ИСП РАН, 2026, том 38, вып. 4, часть 2, стр. 225–244. DOI: 10.15514/ISPRAS-2026-38(4)-28.

Благодарности. Работа выполнена при финансовой поддержке государства в лице Министерства науки и высшего образования Российской Федерации (соглашение № 075-15-2024-529). Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им. В.П. Иванникова РАН – ЦКП ИСП РАН.

AWA: Activation-Based Weight Alignment for Robust Static Watermarking of Neural Networks

[1] *A.A. Akimenkov, ORCID: 0009-0006-7710-941X <a.akimenkov@ispras.ru>*
[1] *D.O. Obydenkov, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[1] *A.Yu. Yakushev, ORCID: 0000-0001-6089-6505 <yakushev@ispras.ru>*
[2] *K.N. Abud,* ORCID: *0009-0009-7131-5839 <khaled.abud@graphics.cs.msu.ru>*
[1] *A.A. Ustinova, ORCID: 0009-0009-9692-2774 <alisaustinova@ispras.ru>*
[1] *Yu.V. Markin, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

[2] *MSU Institute for AI, GSP-1, Leninskie Gory, Moscow, 119991, Russia.*

Abstract. This paper addresses the problem of improving the robustness of static neural network watermarking methods in the white-box extraction scenario, where a digital watermark is extracted directly from model parameters. The main focus is on the coordinated weight permutation attack, which preserves the predictive performance of a neural network but changes the order of parameters used during watermark extraction. To counter this attack, we propose an activation-based weight alignment algorithm, AWA (Activation-based Weight Alignment), which matches structural elements of the original and attacked models using activation maps and restores the weight order before watermark extraction. In this work, AWA is integrated into the extraction procedure of NeuralMark, since this method is robust to several model modification attacks but remains sensitive to parameter permutations. The experimental evaluation is conducted on several datasets, neural network architectures, and model modification attacks, including pruning, fine-tuning, watermark overwriting, weight shifting, and coordinated weight permutation. The results show that applying AWA improves the robustness of NeuralMark against permutation attacks while preserving high robustness against the other considered model modification attacks. In particular, under the weight permutation attack, NeuralMark-AWA reduces BER from 0.48 to 0.00 on CIFAR-10 and from 0.44 to 0.00 on Caltech-101, while TPR@FPR increases from 0.04 and 0.06 to 1.00, respectively.

Keywords: digital watermarks; neural networks; static watermarking; intellectual property protection; watermarking attacks; weight permutation; activation-based weight alignment.

For citation: Akimenkov A.A., Obydenkov D.O., Yakushev A.Yu., Abud K.N., Ustinova A.A., Markin Yu.V. AWA: Activation-Based Weight Alignment for Robust Static Watermarking of Neural Networks. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 4, part 2, pp. 225-244 (in Russian). DOI: 10.15514/ISPRAS-2026-38(4)-28.

Acknowledgements. This work was supported by Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2024-529). The results were obtained using the services of the Ivannikov Institute for System Programming (ISP RAS) Data Center.

1. Введение

Современные нейронные сети широко используются в интеллектуальных системах и требуют значительных затрат на обучение, включая большие наборы данных, вычислительные ресурсы и настройку гиперпараметров. Поэтому обученная модель представляет собой ценный объект интеллектуальной собственности. Одной из актуальных угроз является несанкционированное копирование, распространение или использование таких моделей. В отличие от традиционного программного обеспечения, авторство нейронной сети сложнее подтвердить, особенно если злоумышленник модифицирует модель с использованием различных атак.

Для защиты нейронных сетей применяются методы маркирования, предполагающие внедрение в модель скрытой информации – цифрового водяного знака (ЦВЗ), позволяющего подтвердить ее принадлежность владельцу. Такие методы обычно разделяют на динамические и статические. Динамические методы основаны на анализе реакции модели на специальные входные данные и подходят для сценария извлечения «черного ящика». Статические методы внедряют ЦВЗ непосредственно в параметры, структуру или внутренние представления сети и предполагают сценарий извлечения «белого ящика». Такие методы требуют доступа к весам модели, однако позволяют связывать водяной знак с внутренними параметрами нейронной сети и ограничивать влияние процедуры маркирования отдельными слоями.

Существующие статические методы маркирования по-разному решают задачи скрытности, устойчивости к перезаписи и защиты от подделки ЦВЗ, однако многие из них сохраняют зависимость от фиксированного порядка параметров. Наиболее проблемным преобразованием в этом случае является согласованная перестановка весов, при которой функциональность модели сохраняется, но порядок нейронов или каналов изменяется, что может нарушать процедуру извлечения ЦВЗ. Эта угроза особенно существенна для методов, извлекающих водяной знак из параметров, расположенных в фиксированном порядке. В данной работе в качестве базового метода рассматривается NeuralMark [1], поскольку он обладает высокой устойчивостью к ряду атак модификации модели и использует криптографическую хеш-функцию для защиты от подделки ЦВЗ, однако остается чувствительным к перестановке параметров. Для устранения данного ограничения предлагается алгоритм AWA, интегрируемый в процедуру извлечения NeuralMark и предназначенный для восстановления порядка структурных элементов целевого слоя перед извлечением ЦВЗ.

Основные вклады данной работы заключаются в следующем:

1. Предложен алгоритм AWA для восстановления порядка структурных элементов целевого слоя перед извлечением ЦВЗ, основанный на активационных представлениях исходной и атакованной моделей.
2. Разработана процедура выравнивания выходных фильтров и входных каналов сверточного слоя с использованием косинусной близости и решения задачи оптимального паросочетания.
3. Проведена экспериментальная оценка NeuralMark-AWA на нескольких наборах данных, архитектурах и атаках модификации модели, включая согласованную перестановку весов, а также выполнено сравнение с алгоритмом FindTheLady.

2. Обзор литературы

В данном разделе рассматриваются существующие подходы к статическому маркированию нейронных сетей. В отличие от динамических методов, основанных на реакции модели на специальные триггерные входные данные, статические методы связывают водяной знак с параметрами, структурой или скрытыми представлениями нейронной сети. Такой подход позволяет проводить проверку без обращения к поведению модели на заранее заданных входах и, в ряде случаев, ограничивать влияние маркирования отдельными слоями или подмножествами параметров. Кроме того, использование триггерных данных может влиять

на качество работы маркированной модели и создавать дополнительные ограничения при выборе данных для внедрения ЦВЗ [2].

Одним из первых методов статического маркирования нейронных сетей является подход, предложенный Uchida Y. и др. [3] (далее – Uchida). В данном методе ЦВЗ встраивается в один или несколько слоев сверточной нейронной сети путем добавления специального регуляризатора к функции потерь. Веса выбранного слоя агрегируются и преобразуются в вектор, после чего с использованием секретной матрицы-ключа владельца выполняется линейная операция для извлечения битовой последовательности. В процессе обучения модель учится не только решать основную задачу, например, классификацию изображений, но и восстанавливать заданный ЦВЗ из параметров целевого слоя. Такой подход позволяет внедрять ЦВЗ с высокой битовой емкостью без значительного влияния на качество работы нейронной сети. Однако последующие исследования показали [4], что такой способ маркирования имеет ряд ограничений. Во-первых, внедрение ЦВЗ может изменять статистическое распределение весов маркированного слоя, что позволяет обнаруживать наличие водяного знака статистически. Во-вторых, метод оказывается недостаточно устойчивым к атаке перезаписи, при которой злоумышленник пытается внедрить новый водяной знак поверх уже существующего. Кроме того, поскольку извлечение ЦВЗ зависит от фиксированного порядка параметров, подобные методы потенциально чувствительны к преобразованиям, изменяющим порядок весов.

Для повышения скрытности и устойчивости водяного знака Wang T. и Kerschbaum F. предложили метод RIGA [5]. Авторы отмечают, что регуляризаторы, используемые в существующих статических методах, изменяют распределение весов, что позволяет обнаружить наличие ЦВЗ, а линейные функции извлечения остаются уязвимыми к различным преобразованиям модели. Для устранения первой проблемы применяется состязательное обучение: дополнительный дискриминатор обучается различать маркированные и немаркированные веса, тогда как модель для внедрения ЦВЗ стремится сделать их распределения неразличимыми. Вместо использования линейной операции для извлечения ЦВЗ в RIGA предлагается использовать отдельную нейронную сеть, которая совместно обучается с целевой задачей, выполняя роль секретного ключа для извлечения. Она восстанавливает битовую последовательность из маркированных весов, а немаркированные веса с помощью специального регуляризатора отображает в случайные битовые сообщения, что предотвращает тривиальное извлечение одного и того же ЦВЗ из произвольной модели.

Другим направлением исследований является обеспечение совместимости статического маркирования с шифрованием нейронных сетей [6]. Li G. и др. [7] показали, что перестановочное шифрование параметров нарушает порядок весов и тем самым делает ЦВЗ, внедрённые классическими методами, не извлекаемыми. Для решения данной проблемы авторы предложили метод, устойчивый к перестановке параметров свёрточного слоя (далее – EncWM). Вместо использования весов в фиксированном порядке все ядра выбранного слоя усредняются с помощью операции Kernel Fusion, формируя агрегированное представление, не зависящее от их перестановки. Поскольку такая агрегация существенно уменьшает размерность пространства для внедрения, полученный вектор преобразуется сетью MappingNet в представление большей размерности, что позволяет увеличить ёмкость ЦВЗ. MappingNet и маркируемая модель обучаются совместно с использованием дополнительного регуляризатора. В обучение также включается регуляризатор, аналогичный RIGA, для предотвращения корректного извлечения ЦВЗ из немаркированной версии весов.

Fan L. и др. [8] продемонстрировали, что существующие методы статического маркирования нейронных сетей уязвимы к атаке подделывания ЦВЗ. В рамках данной атаки злоумышленник посредством оптимизации формирует поддельный ключ извлечения, который успешно проходит процедуру верификации, используя аналогичные параметры

нейронной сети. В результате для одной и той же модели могут быть подтверждены несколько различных ЦВЗ, что приводит к неоднозначности при установлении её законного владельца. Для противодействия данной угрозе авторы предложили метод на основе «цифровых паспортов», которые связывают функциональность нейронной сети с секретными данными владельца. Для этого в архитектуру модели добавляются «паспортные слои» путем замены слоев нормализации модели. Таким образом, при использовании корректного паспорта маркированная модель сохраняет исходное качество работы, тогда как применение поддельного или изменённого паспорта приводит к существенному снижению качества работы модели. Проверка прав собственности в этом случае основывается одновременно на извлечении подписи и сохранении работоспособности модели, что позволяет устранить неоднозначность, создаваемую поддельными ЦВЗ.

Дальнейшее развитие данного подхода было предложено Zhang J. и др. [9]. Авторы отметили, что метод Fan L. и др. требует модификации архитектуры модели, в частности замены исходных слоёв нормализации, что может приводить к существенному снижению точности. Для устранения данного ограничения был предложен метод, предполагающий добавление к существующим слоям нормализации отдельной «паспортно-зависимой» ветви (passport-aware normalization). Основная и паспортная ветви используют независимые статистики нормализации и параметры аффинного преобразования, причём параметры паспортной ветви вычисляются на основе цифрового паспорта и весов предшествующего слоя. Обе ветви обучаются совместно, однако при публикации модели паспортная ветвь удаляется, благодаря чему архитектура после публикации остается неизменной. При проверке авторства данная ветвь возвращается в модель: корректный паспорт обеспечивает сохранение исходного качества, тогда как использование поддельного паспорта приводит к значительному снижению качества работы маркированной модели.

Liu H. и др. [10] заметили, что существующие подходы на основе паспортных слоев остаются уязвимыми к атаке неоднозначности в случае компрометации исходного паспорта. Обладая оригинальным паспортом и параметрами модели, злоумышленник может с помощью оптимизации сформировать поддельный паспорт, отличающийся от исходного, но сохраняющий качество работы модели и позволяющий извлекать поддельный ЦВЗ. Для устранения данной уязвимости авторы предложили метод Trapdoor Normalization, основанный на необратимой проверке прав собственности. Как и в предыдущих паспортных подходах, параметры слоя нормализации связываются с паспортом и весами предшествующего слоя, благодаря чему корректность паспорта влияет на качество решения целевой задачи. Однако дополнительно в модель внедряется бинарная подпись, формируемая как хеш-функция от исходного паспорта. В результате между паспортом, внедренным ЦВЗ и параметрами модели устанавливается двунаправленная связь: поддельный паспорт может сохранить качество работы модели, но результат хэш-функции от него с высокой вероятностью не будет соответствовать внедренному ЦВЗ. Тем самым авторы обосновывают целесообразность применения криптографических хеш-функций для противодействия подделыванию ЦВЗ: лавинный эффект хеширования приводит к существенному изменению подписи даже при незначительной модификации паспорта, а необратимость функции затрудняет восстановление подходящего паспорта даже при известной подписи.

Развитием подходов, использующих хеш-функции для защиты от подделки ЦВЗ, стал метод NeuralMark, предложенный Yao Y. и др. [1]. Авторы отказались от паспортных слоёв, поскольку такие методы требуют изменения архитектуры модели, зависят от наличия слоёв нормализации и не всегда применимы к различным типам нейронных сетей. Вместо этого NeuralMark встраивает ЦВЗ непосредственно в параметры выбранного слоя, что позволяет использовать метод для произвольного типа слоев. Из секретного ключа с помощью хеш-функции SHAKE-256 формируется бинарная последовательность, одновременно выполняющая роль внедряемого ЦВЗ и фильтра для выбора параметров. Благодаря

лавинному эффекту изменение ключа приводит к непредсказуемому изменению как ЦВЗ, так и соответствующего набора весов, что препятствует градиентному восстановлению поддельного ключа. Фильтрация выполняется в несколько раундов, вследствие чего различные ЦВЗ воздействуют на слабо пересекающиеся подмножества параметров. Это снижает вероятность повреждения ЦВЗ при внедрении злоумышленником нового водяного знака и тем самым повышает устойчивость к атаке перезаписи. После фильтрации выбранные параметры объединяются посредством усреднения, что дополнительно обеспечивает устойчивость к другим типам атак. Встраивание выполняется при обучении модели с использованием дополнительного регуляризатора функции потерь.

Рассмотренные методы показывают, что развитие статического маркирования нейронных сетей направлено на устранение нескольких ключевых ограничений: статистической заметности ЦВЗ, уязвимости к перезаписи и подделке, а также зависимости процедуры извлечения от фиксированного порядка параметров. Для решения этих проблем современные методы используют состязательное обучение, паспортные механизмы, криптографические хэш-функции и специальные представления весов для однозначного подтверждения авторства. Несмотря на это, некоторые современные атаки все еще остаются крайне эффективными. Так, атака перестановки весов опирается на фундаментальное свойство симметрии нейронных сетей [11]: согласованная перестановка нейронов или каналов в соседних слоях сохраняет качество работы модели, но изменяет расположение параметров, тем самым препятствуя корректному извлечению ЦВЗ. К тому же, дополнительная опасность данной атаки определяется отсутствием традиционного компромисса между удалением ЦВЗ и сохранением качества работы модели. В отличие от таких атак, как дообучение или прореживание, перестановка не требует обучающих данных или вычислительных ресурсов. Более того, она может проводиться вслепую, без знания применённого алгоритма маркирования и даже без подтверждения наличия ЦВЗ. Yan Y. и др. [12] показали, что использование такого преобразования приводит к удалению ЦВЗ для большого количества статических методов, сохраняя при этом исходную работоспособность моделей. Li F.-Q. и др. [13] также отмечают, что такие функциональные преобразования способны без существенных затрат аннулировать доказательство владения моделью. Исходя из этого можно сказать, что устойчивость к подобным преобразованиям является необходимым условием надежного метода статического маркирования: либо метод должен извлекать ЦВЗ из перестановочно-инвариантного представления параметров, либо обеспечивать достоверное восстановление их исходного порядка перед извлечением ЦВЗ.

3. Модель угроз

В данной работе рассматривается сценарий извлечения ЦВЗ «белого ящика», в котором проверяющая сторона имеет доступ к параметрам анализируемой нейронной сети. Аналогично постановке метода NeuralMark, предполагается наличие доверенной третьей стороны, выполняющей проверку ЦВЗ: владелец модели передает ей необходимую секретную информацию, а проверяющая сторона определяет, содержится ли в спорной модели корректная метка владельца. При этом считается, что верификатор является честным и не вступает в сговор ни с владельцем, ни со злоумышленником. Злоумышленник также рассматривается в сценарии «белого ящика»: он имеет доступ к параметрам маркированной модели, знает ее архитектуру и может анализировать структуру слоев. Допускается, что атакующему известен сам факт наличия маркирования, а также слой или группа слоев, в которые может быть встроен ЦВЗ. Такая постановка является консервативной, поскольку устойчивость метода не должна основываться на сокрытии архитектуры модели или факта использования конкретной схемы маркирования.

Цель злоумышленника состоит в модификации маркированной модели таким образом, чтобы нарушить процедуру извлечения ЦВЗ или затруднить подтверждение принадлежности

модели владельцу. При этом атакованная модель должна сохранять качество на основной задаче, поскольку существенная деградация качества делает такую модель малопригодной для практического использования. Таким образом, успешной считается только такая атака, которая снижает надежность извлечения ЦВЗ при сохранении функциональности модели.

4. Методология

В данной работе предлагается алгоритм выравнивания весов AWA, предназначенный для восстановления исходного порядка структурных элементов слоя перед извлечением ЦВЗ. Его эффективность демонстрируется на основе метода NeuralMark, относящегося к классу статических методов маркирования нейронных сетей в сценарии извлечения «белый ящик».

4.1 Мотивация

NeuralMark был выбран в качестве базового метода благодаря высокой устойчивости к ряду атак и использованию криптографической функции в процессе извлечения ЦВЗ. Использование криптографической хеш-функции в NeuralMark затрудняет подделку водяного знака, однако не устраняет зависимость извлечения от порядка параметров целевого слоя. Из наиболее опасных преобразований такого типа является согласованная перестановка весов. В сверточных нейронных сетях перестановка выходных фильтров слоя может быть компенсирована соответствующей перестановкой входных каналов следующего слоя. Аналогично, порядок входных каналов целевого слоя определяется порядком выходных фильтров слоя, формирующего его вход. В результате функциональность модели и ее качество на основной задаче сохраняются, однако порядок параметров, используемый процедурой извлечения ЦВЗ, изменяется. Данная атака особенно критична для статических методов маркирования, поскольку она не требует значительных вычислительных ресурсов. В отличие от атак модификации модели, согласованная перестановка весов не предполагает процесса оптимизации, доступа к обучающей выборке или подбора сложных гиперпараметров. Формально пусть $W^l \in R^{C_l \times C_{l-1} \times k_h \times k_w}$ – весовой тензор l-ого сверточного слоя, где C_l соответствует числу выходных фильтров, а C_{l-1} – числу входных каналов. Согласованная перестановка может быть описана перестановками π_l и π_{l-1}, действующими на оси выходных фильтров и входных каналов. Тогда, преобразованный весовой тензор можно записать в форме: $\widetilde{W_{i,j,:,:}^{(l)}} = W_{\pi_l(i),\pi_{l-1}(j),:,:}^{(l)}$, где $\pi_l(i)$ задает, какой фильтр исходного слоя помещается в позицию i после перестановки, а $\pi_{l-1}(j)$ задает соответствующую перестановку входных каналов. Таким образом, она может быть реализована как простое преобразование над весами модели. Поэтому такая атака является практически реализуемой даже для злоумышленника с ограниченными ресурсами. Для методов, извлекающих ЦВЗ из фиксированного порядка весов, перестановка параметров приводит к рассинхронизации между процедурой внедрения и извлечения ЦВЗ. Даже если сами численные значения весов остаются теми же, изменение их порядка может привести к тому, что извлекаемый вектор перестанет соответствовать исходному водяному знаку. Однако, в таком случае задача восстановления порядка структурных элементов является относительно простой (например, с помощью сортировки или расчета статистик [14]), поскольку параметры модели изменены только перестановкой, а их численные значения полностью сохраняются. В более практическом сценарии злоумышленник может комбинировать перестановку весов с другими преобразованиями модели, например, прореживанием или дообучением. В таком случае параметры не только меняют порядок, но и численно отличаются от исходных, что существенно усложняет восстановление соответствия между фильтрами исходной и атакованной модели. Исходя из этого, практический метод защиты должен учитывать сценарий, в котором перестановка применяется после численной модификации параметров. Для устранения данной уязвимости в работе предлагается алгоритм AWA, интегрируемый в

процедуру извлечения NeuralMark. Основная идея состоит в том, чтобы перед извлечением ЦВЗ восстановить порядок структурных элементов целевого слоя по активационным представлениям, вычисленным на фиксированном наборе входных данных. Такой подход позволяет сопоставлять фильтры исходной и атакованной моделей по их функциональному поведению и тем самым снизить чувствительность NeuralMark к согласованной перестановке параметров.

4.2 Описание метода

Основная идея метода состоит в том, что порядок фильтров сверточного слоя может быть восстановлен не только по самим весам, но и по активациям, которые эти фильтры формируют на фиксированном наборе входных данных. В таком случае, даже если фильтры и/или каналы целевого слоя были согласованно переставлены, их можно сопоставить с фильтрами исходной модели по близости порождаемых ими активаций.

Пусть M обозначает маркированную модель, а θ – ее параметры. Для внедрения ЦВЗ выбирается один или несколько целевых сверточных слоев. Для простоты далее рассмотрим один сверточный слой с весовым тензором:

$$W \in R^{C_{out} \times C_{in} \times k_h \times k_w}, \quad (1)$$

где C_{out} – число выходных фильтров, C_{in} — число входных каналов, а k_H и k_W – размеры ядра свертки. В исходном методе веса слоя преобразуются в вектор фиксированным образом, после чего применяется фильтрация параметров на основе битовой последовательности, полученной в результате применения криптографической хэш-функции. Полная схема извлечения ЦВЗ с использованием алгоритма AWA показана на рис. 1.

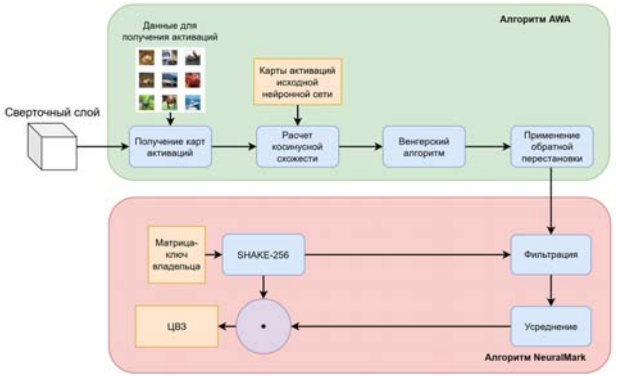

Рис. 1. Полная схема внедрения и извлечения ЦВЗ методом NeuralMark-AWA. Зеленая область обозначает алгоритм AWA, красная — исходный алгоритм NeuralMark, синие фигуры – функции, оранжевые – данные, а фиолетовая фигура – операцию матричного перемножения.
Fig. 1. Complete watermark embedding and extraction scheme of NeuralMark-AWA. The green area denotes the AWA algorithm, the red area denotes the original NeuralMark algorithm, blue shapes denote functions, orange shapes denote data, and the purple shape denotes matrix multiplication.

Пусть имеется исходная маркированная модель M и потенциально атакованная модель M'. Необходимо восстановить соответствие между фильтрами целевого слоя в M и M'. Перед публикацией модели в открытый доступ владелец выбирает фиксированный набор входных данных X_{align}, который в будущем будет использоваться для выравнивания фильтров. На основе этих данных владельцем модели вычисляются выходные активации целевого слоя (слоев):

$$Y = f_l(X_{align};\ M) \tag{2}$$

где f_l обозначает выход целевого слоя l. Если после сверточного слоя расположен слой батч-нормализации, то для формирования представлений используются активации после его применения, но до нелинейной функции активации. Затем, для каждого выходного фильтра j строится скрытое представление полученных карт активаций. Для этого соответствующая карта активаций разворачивается в вектор и нормируется по L_2-норме:

$$e_j = \frac{vec(Y_j)}{\left|\left|vec(Y_j)\right|\right|_2} \tag{3}$$

Таким образом, для исходной модели формируется матрица эталонных скрытых представлений:

$$E = [e_1, e_2, \dots, e_{C_{out}}] \tag{4}$$

Аналогично для атакованной модели M' вычисляются скрытые представления фильтров:

$$E' = [e'_1, e'_2, \dots, e'_{C_{out}}] \tag{5}$$

Для оценки соответствия между фильтрами исходной и атакованной модели используется косинусная близость:

$$S_{i,j} = e_i^T e'_j \tag{6}$$

где $S_{i,j}$ показывает близость i-го фильтра исходной модели и j-го фильтра атакованной модели. Далее задача восстановления порядка формулируется как задача оптимального паросочетания: необходимо найти такую перестановку π, которая максимизирует суммарную близость сопоставленных фильтров:

$$\pi^* = argmax_\pi \sum_{i=1}^{C_{out}} S_{i,\pi(i)} \tag{7}$$

Что эквивалентно минимизации матрицы стоимости:

$$C_{i,j} = 1 - S_{i,j} \tag{8}$$

Для решения этой задачи используется Венгерский алгоритм. В результате получается перестановка, которая задает соответствие между порядком фильтров атакованной модели и эталонным порядком исходной модели. После этого веса атакованной модели переупорядочиваются:

$$W'_{aligned} = W'_{\pi^*} \tag{9}$$

где $W'_{aligned}$ – матрица весов целевого слоя атакованной модели, приведенная к порядку исходной модели.

Для восстановления порядка входных каналов целевого слоя необходимо учитывать порядок структурных элементов, формирующих его вход. При согласованной перестановке весов перестановка фильтров предыдущего слоя приводит к соответствующей перестановке входных каналов текущего слоя. Поэтому для полного восстановления весов целевого слоя необходимо также восстановить порядок фильтров предыдущего слоя, а затем использовать найденное соответствие для переупорядочивания входных каналов целевого слоя. Исходя из этого, процедура выравнивания описанная выше должна применяться как к целевому слою, так и к слою, формирующему его вход. Для этого на фиксированном наборе входных данных сохраняются карты активаций обоих слоев: активации целевого слоя используются для восстановления порядка его выходных фильтров, а активации предыдущего слоя – для восстановления порядка входных каналов целевого слоя.

5. Эксперименты

В данном разделе описывается экспериментальная постановка, использованная для оценки качества и устойчивости рассматриваемых методов статического маркирования нейронных сетей, а также представлены результаты экспериментов. Цель экспериментов состоит в проверке трех основных свойств методов маркирования: сохранения качества модели на основной задаче, корректности извлечения ЦВЗ и устойчивости к атакам модификации модели.

5.1 Наборы данных и архитектуры

В экспериментах использовались наборы данных CIFAR-10, CIFAR-100, Caltech-101 и Caltech-256 [15, 16]. CIFAR используется как стандартный набор данных для оценки методов на задаче классификации изображений, а Caltech-101 и Caltech-256 использовались для проверки устойчивости на данных с большим числом классов и более разнообразными изображениями. Для оценки поведения методов на архитектурах с различной структурой сверточных блоков рассматривались модели нескольких семейств: AlexNet, DenseNet-121, DenseNet-161, DenseNet-169, ResNet-18, ResNet-34, ResNet-50, VGG-11, VGG-13, VGG-16 и Inception [17-21].

5.2 Методы и метрики

Модификация NeuralMark-AWA сравнивается с исходным методом NeuralMark, а также с еще несколькими подходами, в числе которых Uchida и RIGA. В тестировании также использовался метод EncWM, который по определению обладает устойчивостью к атакам согласованной перестановки весов.

Для оценки качества модели на основной задаче в экспериментах использовалась метрика точности (Accuracy), показывающая долю правильно классифицированных объектов:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(y_i = \hat{y}_i) \tag{10}$$

где N – число объектов в тестовой выборке, y_i – истинная метка класса, $\hat{y_i}$ – предсказанная моделью метка.

Для оценки корректности извлечения ЦВЗ в экспериментах использовалась метрика, определяющая долю несовпадающих битов между исходной и извлеченной битовыми последовательностями (Bit Error Rate или BER):

$$BER = \frac{1}{n}\sum_{i=1}^{n} (b_i \neq \hat{b}_i) \tag{11}$$

где n – длина битовой последовательности, b_i – i-й бит исходного ЦВЗ, а $\hat{b}_i$ – i-й бит, извлеченный из модели. Значение метрики BER, близкое к нулю, соответствует корректному извлечению ЦВЗ, тогда как значение, близкое к 0.5, соответствует случайному угадыванию. Так как разные методы могут использовать ЦВЗ различной длины, дополнительно применялась метрика TPR@x%FPR. Она показывает долю корректно обнаруженных ЦВЗ при фиксированном уровне ложноположительных срабатываний. Во всех экспериментах значение FPR было фиксировано и равно 10^{-5}, то есть TPR@10^{-5}%FPR (далее – TPR@FPR).

5.3 Гиперпараметры обучения и внедрения ЦВЗ

Во всех экспериментах в качестве основной функции потерь для задачи классификации использовалась кросс-энтропия. Обучение и внедрение ЦВЗ выполнялись в течение 200 эпох,

скорость обучения уменьшалась на 100-й и 150-й эпохах с коэффициентом $\beta = 0.1$. Для Uchida и NeuralMark использовался оптимизатор SGD с шагом градиентного спуска 0.01, а для RIGA и EncWM – AdamW со значением градиентного спуска 0.001.

Во всех экспериментах, для Uchida длина ЦВЗ составляла 256 бит, коэффициент регуляризации – 0.01. Для RIGA использовались коэффициенты регуляризации $\alpha_1 = 0.1$ и $\alpha_2 = 0.1$. Размер скрытого пространства генератора и дискриминатора 100, длина ЦВЗ – 256 бит. Для EncWM длина ЦВЗ составляла 128 бит, фактор расширения – 64, размер скрытого пространства MappingNet – 100. Для NeuralMark использовались ЦВЗ длиной 256 бит, коэффициент регуляризации $\alpha = 0.1$, количество этапов фильтрации $k = 4$, размерность ключа 512 и стандартное отклонение 1.0. Предложенный алгоритм не изменяет процесс внедрения ЦВЗ, поэтому NeuralMark с алгоритмом выравнивания (NeuralMark-AWA) использует исходные веса NeuralMark. Стоит отметить, что этап восстановления порядка весов является обязательной частью процедуры извлечения ЦВЗ и выполняется независимо от наличия атаки перестановки. Если порядок выходных фильтров или входных каналов модели не изменялся, процедура выравнивания восстанавливает тождественное или близкое к нему соответствие.

5.4 Атаки

Для проверки устойчивости методов маркирования использовались только атаки модификации модели. Атаки предобработки входных данных или атаки извлечения модели в данной работе не рассматривались, поскольку такие атаки предполагают извлечение ЦВЗ из ответов модели на специальные входные изображения.

Атака прореживания (Pruning). В экспериментах использовались как неструктурные, так и структурные атаки прореживания [22]. Неструктурное прореживание заключается в занулении отдельных весов модели. Рассматривались два варианта такой атаки: случайноепрореживание и прореживание на основе L_1-нормы. В первом случае веса выбираются случайным образом, во втором – зануляются параметры в соответствии с их величиной. Структурное прореживание направлено на удаление или зануление целых структурных элементов модели, например, фильтров или каналов сверточного слоя. Такая атака потенциально более опасна, поскольку она изменяет не отдельные элементы весового тензора, а целые группы параметров, от которых может зависеть процедура извлечения ЦВЗ. Для атак прореживания использовались следующие значения доли удаляемых параметров: 1%, 5%, 10%, 20%, 40%, 60%, 80%, 90%, 95% и 99%.

Атака дообучения (Fine-tuning). Атаки дообучения проверяют, обладает ли ЦВЗ устойчивостью к дополнительному обучению маркированной модели. Следуя сценариям, определенным в работе, предложенной Adi и др. [23], рассматривались четыре сценария дообучения: дообучение последнего слоя (Fine-tune last layer или FTLL), дообучение всех слоев (Fine-tune all layers или FTAL), переобучение последнего слоя (Re-train last layer или RTLL) и переобучение последнего слоя с дообучением всех слоев (Re-train all layers или RTAL). Эти сценарии отличаются тем, какие параметры модели обновляются и происходит ли повторная инициализация последнего слоя нейронной сети перед дообучением. Для всех атак дообучения использовался аналогичный оптимизатор, значение шага градиентного спуска составляло 0.001, а количество эпох дообучения равнялось 100.

Атака дообучения с использованием гладких меток (Label Smoothing). В этом случае модель до-обучается с регуляризацией целевых меток, при которой исходные метки классов заменяются сглаженными распределениями. Такая процедура может менять траекторию оптимизации и параметры модели, поэтому она рассматривается как отдельный вариант атаки дообучения. В экспериментах использовались значения коэффициента сглаживания 0.1, 0.2 и 0.3.

Атака перезаписи ЦВЗ (Watermark Overwrite). Атака перезаписи заключается во внедрении нового водяного знака в уже маркированную модель [3]. Цель атакующего состоит не только в добавлении собственного ЦВЗ, но и в подавлении исходного ЦВЗ владельца. Если после перезаписи исходный ЦВЗ продолжает корректно извлекаться, такая атака не считается успешной. В экспериментах перезапись выполнялась с параметрами, аналогичными параметрам исходного внедрения ЦВЗ. Количество эпох для данной атаки составляло 100.

Атака сдвига весов (Weight Shifting). Атака сдвига весов [4] направлена на изменение статистики весов целевого слоя. Поскольку статические методы маркирования часто используют агрегированные значения весов или их распределение, целенаправленный сдвиг весов может нарушить процедуру извлечения ЦВЗ. В экспериментах для атаки сдвига весов использовались коэффициенты 1.5 и 1.0. После применения сдвига выполнялось дополнительное дообучение модели в течение 100 эпох с шагом градиентного спуска 0.001.

Атака согласованной перестановки весов (Weight Permutation). Особое внимание уделялось атаке согласованной перестановки весов [12]. В экспериментах рассматривались три варианта перестановки: перестановка по оси выходных фильтров, перестановка по оси входных каналов и одновременная перестановка по обеим осям. Такой набор сценариев позволяет проверить, насколько метод зависит от исходного порядка структурных элементов сверточного слоя.

5.5 Результаты

В данном разделе описываются результаты тестирования NeuralMark-AWA, а также приводятся результаты для сравнения с существующими решениями. Для оценки методов маркирования на предмет влияния на точность работы нейронной сети в табл. 1 показаны метрики точности работы нейронной сети без ЦВЗ и с внедренным ЦВЗ.

Табл. 2 демонстрирует усредненную оценку устойчивости рассмотренных методов (за исключением атак прореживания модели). Из приведенных результатов следует, что в данной агрегированной оценке предложенная модификация демонстрирует наименьшее среднее значение BER и наибольшее значение TPR@FPR среди рассмотренных методов.

В табл. 3 показаны результаты тестирования устойчивости методов маркирования к атаке согласованной перестановки весов. Результаты демонстрируют, что атака крайне эффективна против большинства рассматриваемых методов. Устойчивость к данной атаке демонстрируют только методы EncWM и NeuralMark-AWA. На рис. 2-4 представлены усредненные результаты оценки устойчивости методов маркирования к атакам прореживания модели на наборе данных CIFAR-10. Во всех случаях метод NeuralMark-AWA демонстрирует схожую с исходным методом NeuralMark устойчивость к атаке прореживания.

Табл. 1. Усредненные результаты обучения для всех архитектур маркированных и немаркированных моделей до применения атак.
Table 1. Average training results for all architectures of watermarked and non-watermarked models before attacks.

Метод	CIFAR-10 Accuracy с ЦВЗ / без ЦВЗ	CIFAR-10 BER	CIFAR-10 TPR@FPR	Caltech-101 Accuracy с ЦВЗ / без ЦВЗ	Caltech-101 BER	Caltech-101 TPR@FPR
Uchida	0.94/0.94	0.00	1.00	0.68/0.68	0.00	1.00
RIGA	0.93/0.94	0.00	1.00	0.68/0.68	0.00	1.00
EncWM	0.93/0.94	0.00	1.00	0.67/0.68	0.00	1.00
NeuralMark	0.94/0.94	0.00	1.00	0.68/0.68	0.00	1.00
NeuralMark-AWA	0.94/0.94	0.00	1.00	0.68/0.68	0.00	1.00

Табл. 2. Усредненные результаты тестирования устойчивости статических методов маркирования к атакам модификации модели, за исключением атак прореживания.
Table 2. Average robustness results of static watermarking methods against model modification attacks, excluding pruning attacks.

Метод	Общее значение метрики BER	Общее значение метрики TPR@FPR
Uchida	0.27	0.65
RIGA	0.12	0.74
EncWM	0.11	0.76
NeuralMark	0.11	0.76
NeuralMark-AWA	0.02	1.00

Табл. 3. Результаты тестирования устойчивости методов маркирования к атаке согласованной перестановки весов.

Table 3. Robustness results of watermarking methods against the coordinated weight permutation attack.

Метод	CIFAR-10 Accuracy	CIFAR-10 BER	CIFAR-10 TPR@FPR	Caltech-101 Accuracy	Caltech-101 BER	Caltech-101 TPR@FPR
Uchida	0.94	0.34	0.33	0.68	0.33	0.33
RIGA	0.93	0.19	0.59	0.68	0.13	0.73
EncWM	0.93	0.00	1.00	0.65	0.00	1.00
NeuralMark	0.94	0.48	0.04	0.68	0.44	0.06
NeuralMark-AWA	0.94	0.00	1.00	0.68	0.00	1.00

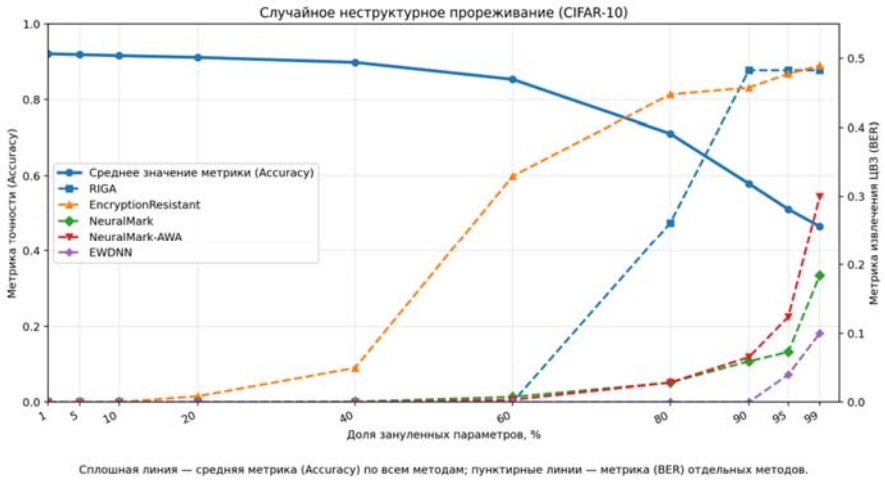

Рис. 2. Атака неструктурного случайного прореживания на методы маркирования нейронных сетей.
Fig. 2. Unstructured random pruning attack on neural network watermarking methods.

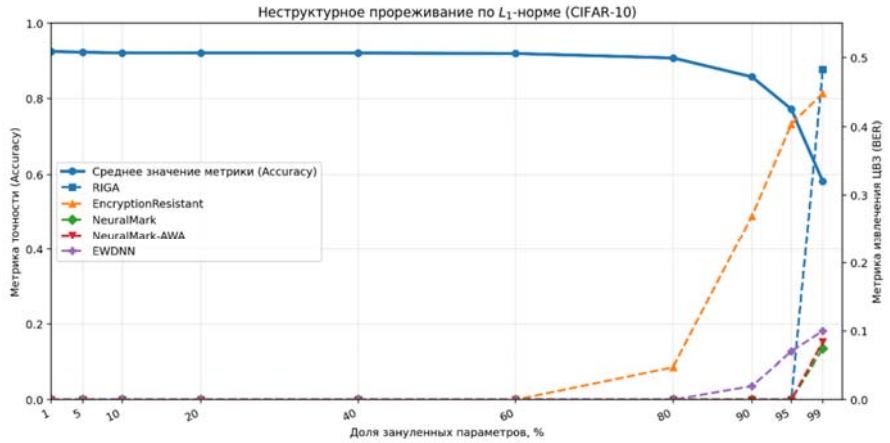

Рис. 3. Атака неструктурного прореживания по L_1-норме на методы маркирования нейронных сетей.
Fig. 3. Unstructured L1-norm pruning attack on neural network watermarking methods.

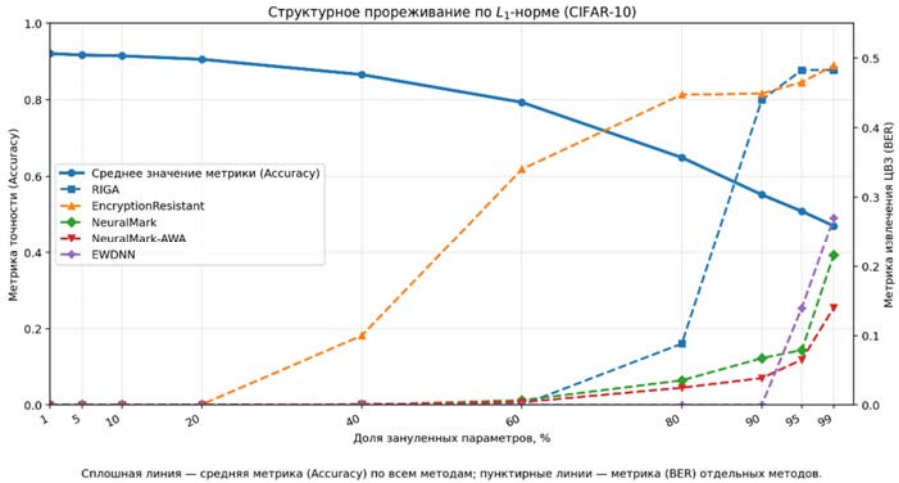

Рис. 4. Атака структурного прореживания по L_1-норме на методы маркирования нейронных сетей.
Fig. 4. Structured L1-norm pruning attack on neural network watermarking methods.

В табл. 4 показаны усредненные результаты тестирования устойчивости методов к атакам дообучения во всех описанных сценариях, включая атаку дообучения с использованием гладких меток. Из результатов видно, что атаки дообучения в различных сценариях не производят значимого влияния ни на один из рассматриваемых методов. Тестирование устойчивости к атаке перезаписи ЦВЗ показано в табл. 5. Как можно видеть, интеграция AWA не приводит к значительному снижению устойчивости метода NeuralMark к атаке перезаписи ЦВЗ. Тестирование устойчивости методов маркирования к атаке сдвига весов

показано в табл. 6. Из результатов видно, что при атаке сдвига весов метод NeuralMark-AWA сохраняет устойчивость, близкую к устойчивости исходного метода NeuralMark.

Табл. 4. Тестирование устойчивости методов маркирования к атакам дообучения, включая сценарий дообучения с гладкими метками. В таблице приведены средние значения BER и TPR@FPR по всем типам дообучения и архитектурам.

Table 4. Robustness evaluation of watermarking methods against fine-tuning and label smoothing attacks. The table reports average BER and TPR@FPR values across all fine-tuning scenarios and architectures.

Метод	CIFAR-10 BER	CIFAR-10 TPR@FPR	CIFAR-100 BER	CIFAR-100 TPR@FPR	Caltech-101 BER	Caltech-101 TPR@FPR	Caltech-256 BER	Caltech-256 TPR@FPR
Uchida	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
RIGA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
EncWM	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
NeuralMark	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
NeuralMark-AWA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

Рассмотренные эксперименты показывают, что добавление алгоритма AWA позволяет сохранить устойчивость NeuralMark к атакам прореживания, дообучения, перезаписи ЦВЗ и сдвига весов, а также существенно повышает устойчивость к согласованной перестановке параметров. Однако итоговые значения метрик BER и TPR@FPR не позволяют напрямую оценить, насколько корректно алгоритм AWA восстанавливает порядок структурных элементов слоя.

Табл. 5. Результаты тестирования устойчивости методов маркирования к атаке перезаписи ЦВЗ. В таблице приведены средние значения BER и TPR@FPR по всем архитектурам.

Table 5. Robustness results of watermarking methods against the watermark overwriting attack. The table reports average BER and TPR@FPR values across all architectures.

Метод	CIFAR-100 Accuracy	CIFAR-100 BER	CIFAR-100 TPR@FPR	Caltech-256 Accuracy	Caltech-256 BER	Caltech-256 TPR@FPR
Uchida	0.81	0.02	1.00	0.56	0.01	1.00
RIGA	0.79	0.00	1.00	0.55	0.00	1.00
EncWM	0.80	0.26	0.62	0.54	0.00	1.00
NeuralMark	0.82	0.00	1.00	0.55	0.00	1.00
NeuralMark-AWA	0.82	0.05	1.00	0.55	0.03	1.00

Табл. 6. Результаты тестирования устойчивости методов маркирования к атаке сдвига весов. В таблице приведены средние значения BER и TPR@FPR по всем архитектурам.

Table 6. Robustness results of watermarking methods against the weight shifting attack. The table reports average BER and TPR@FPR values across all architectures.

Метод	CIFAR-10 Accuracy	CIFAR-10 BER	CIFAR-10 TPR@FPR	Caltech-101 Accuracy	Caltech-101 BER	Caltech-101 TPR@FPR
Uchida	0.89	0.73	0.27	0.63	0.73	0.27
RIGA	0.77	0.32	0.33	0.62	0.35	0.27
EncWM	0.78	0.36	0.33	0.61	0.39	0.20
NeuralMark	0.90	0.00	1.00	0.64	0.00	1.00
NeuralMark-AWA	0.90	0.04	1.00	0.64	0.04	1.00

Исходя из этого, далее отдельно анализируется точность восстановления порядка весов при комбинированном сценарии, в котором после модификации модели дополнительно применяется согласованная перестановка весов, а затем оценивается точность восстановления их исходного порядка с помощью AWA.

5.6 Точность алгоритма выравнивания

Для дополнительной оценки работы предложенного алгоритма AWA был проведён отдельный анализ точности восстановления порядка весов. В отличие от предыдущих экспериментов, где основное внимание уделялось итоговой корректности извлечения ЦВЗ, в данном случае отдельно оценивалась точность сопоставления выходных фильтров и входных каналов исходной и атакованной моделей.

На рис. 5-7 представлены результаты для сценария, в котором согласованная перестановка применялась одновременно к входным каналам и выходным фильтрам сверточных слоёв после атак дообучения, перезаписи ЦВЗ и прореживания на примере архитектуры ResNet-18. В качестве базового метода выравнивания весов использовался алгоритм FindTheLady [24], основанный на оценке косинусной близости между представлениями нейронов. Полученные результаты показывают, что в рассматриваемой постановке алгоритм FindTheLady не обеспечивает надёжного восстановления исходного порядка весов, что приводит к высокой ошибке извлечения ЦВЗ. В то же время AWA сохраняет высокую точность сопоставления при атаках дообучения и перезаписи ЦВЗ, а также остаётся устойчивым при умеренном прореживании. Это подтверждает, что повышение устойчивости NeuralMark-AWA к перестановочным атакам связано с корректной работой этапа активационного выравнивания.

Таким образом, результаты дополнительного анализа показывают, что устойчивость NeuralMark-AWA к согласованной перестановке весов определяется не только сохранностью самого ЦВЗ в параметрах модели, но и корректностью восстановления порядка структурных элементов слоя перед извлечением. Предложенный алгоритм AWA обеспечивает надежное сопоставление выходных фильтров и входных каналов исходной и атакованной моделей в сценариях дообучения и перезаписи ЦВЗ, сохраняя работоспособность при умеренном прореживании.

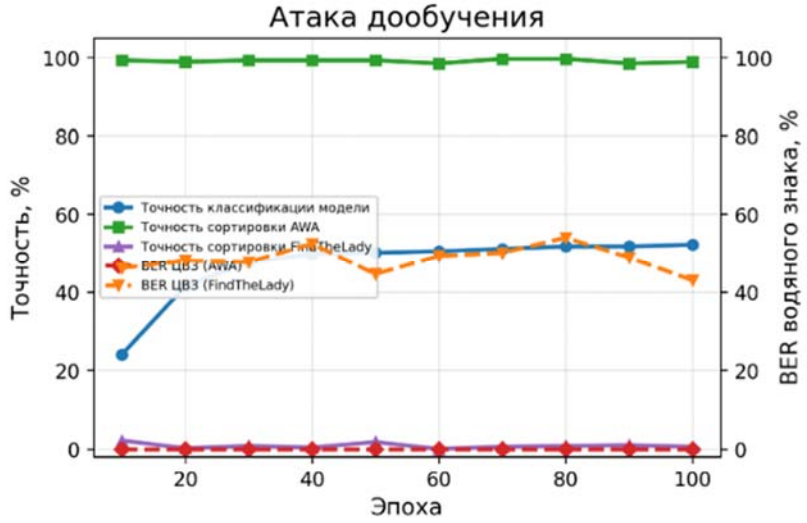

Рис. 5. Точность восстановления порядка весов после атаки дообучения и согласованной перестановки входных каналов и выходных фильтров на наборе данных Caltech-256 для архитектуры ResNet-18.

Fig. 5. Accuracy of weight order recovery after fine-tuning and coordinated permutation of input channels and output filters on the Caltech-256 dataset for the ResNet-18 architecture.

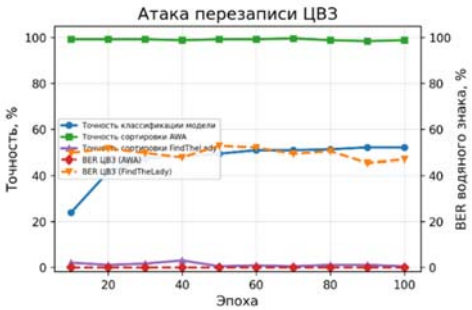

Рис. 6. Точность восстановления порядка весов после атаки перезаписи ЦВЗ и согласованной перестановки входных каналов и выходных фильтров на наборе данных Caltech-256 для архитектуры ResNet-18.
Fig. 6. Accuracy of weight order recovery after watermark overwriting and coordinated permutation of input channels and output filters on the Caltech-256 dataset for the ResNet-18 architecture.

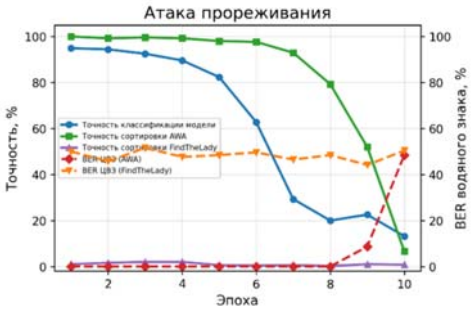

Рис. 7. Точность восстановления порядка весов после атаки прореживания и согласованной перестановки входных каналов и выходных фильтров для архитектуры ResNet-18.
Fig. 7. Accuracy of weight order recovery after pruning and coordinated permutation of input channels and output filters for the ResNet-18 architecture.

В отличие от алгоритма FindTheLady, который в рассматриваемых экспериментах не обеспечивает устойчивого восстановления порядка весов, AWA позволяет поддерживать низкое значение BER после дополнительной перестановки входных каналов и выходных фильтров. При этом снижение качества выравнивания при удалении 90–99% параметров показывает, что предложенный подход зависит от сохранности активационных представлений фильтров и может терять эффективность при существенном изменении структуры или функционального поведения модели.

6. Заключение

6.1. Ограничения

Основным ограничением предложенного алгоритма AWA является то, что он добавляет к исходному методу маркирования дополнительный этап восстановления порядка весов. В результате устойчивость всей схемы начинает зависеть не только от устойчивости самого алгоритма внедрения и извлечения водяного знака, но и от корректности работы алгоритма AWA. Если после атаки активационные представления нейронов существенно изменяются, алгоритм выравнивания может ошибочно сопоставить веса исходной и атакованной модели. В таком случае даже при сохранении части информации о водяном знаке в весах ошибка

восстановления порядка может привести к некорректному извлечению ЦВЗ. Дополнительные сложности возникают для архитектур со сложной структурой связей между слоями. В частности, в архитектурах типа DenseNet вход каждого слоя формируется не выходом одного предыдущего слоя, а объединением признаков нескольких предшествующих слоев. Поэтому восстановление порядка входных каналов требует учета нескольких источников признаков и их согласованных перестановок. Если такая структура учитывается неполно или после атаки отдельные группы признаков изменяются неравномерно, точность восстановления может значительно снизиться.

6.2. Выводы

В работе рассмотрена проблема устойчивости статических методов маркирования нейронных сетей к согласованной перестановке весов. Такая атака сохраняет функциональность модели, но изменяет порядок параметров, что нарушает извлечение ЦВЗ в методах, зависящих от фиксированного расположения весов целевого слоя. Для решения данной проблемы предложен алгоритм активационного выравнивания весов AWA, предназначенный для восстановления порядка структурных элементов слоя перед извлечением ЦВЗ.

Экспериментальная оценка показала, что интеграция AWA с NeuralMark существенно повышает устойчивость исходного метода к согласованной перестановке параметров: на CIFAR-10 значение BER снижается с 0.48 до 0.00, а TPR@FPR возрастает с 0.04 до 1.00; на Caltech-101 значение BER снижается с 0.44 до 0.00, а TPR@FPR возрастает с 0.06 до 1.00. При атаках перезаписи ЦВЗ и сдвига весов NeuralMark-AWA сохраняет TPR@FPR, равный 1.00, при небольшом увеличении BER относительно исходного NeuralMark. Это показывает, что добавление этапа активационного выравнивания существенно повышает устойчивость к перестановочной атаке и не приводит к критическому ухудшению извлечения ЦВЗ при других рассмотренных атаках. Дополнительный анализ показал, что AWA сохраняет точность сопоставления выходных фильтров и входных каналов, близкую к 100%, даже после атак дообучения и перезаписи ЦВЗ, тогда как применение алгоритма FindTheLady в этих сценариях сопровождается высокой ошибкой извлечения. Вместе с тем результаты при высоких долях удаляемых параметров в ходе прореживания (90–99%) показывают, что точность выравнивания зависит от сохранности и различимости активационных представлений фильтров. Это указывает на необходимость дальнейшего исследования предложенного подхода для более агрессивных атак и архитектур со сложной топологией связей между слоями.

Список литературы / References

[1]. Yao Y., Song J., Jin J. Hashed Watermark as a Filter: A Unified Defense Against Forging and Overwriting Attacks in Neural Network Watermarking. Proceedings of the AAAI Conference on Artificial Intelligence, 2026, vol. 40, no. № 42, pp. 35994-36002.
[2]. Kim B. et al. Margin-based neural network watermarking. International Conference on Machine Learning. PMLR, 2023, pp. 16696-16711.
[3]. Uchida Y. et al. Embedding watermarks into deep neural networks. Proceedings of the 2017 ACM on international conference on multimedia retrieval, 2017, pp. 269-277.
[4]. Lukas N. et al. Sok: How robust is image classification deep neural network watermarking? 2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022, pp. 787-804.
[5]. Wang T., Kerschbaum F. Riga: Covert and robust white-box watermarking of deep neural networks. Proceedings of the web conference 2021, 2021, pp. 993-1004.
[6]. Tian J., Zhou J., Duan J. Probabilistic selective encryption of convolutional neural networks for hierarchical services. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2205-2214.

[7]. Li G. et al. Encryption resistant deep neural network watermarking. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 3064-3068.
[8]. Fan L., Ng K. W., Chan C. S. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. Advances in neural information processing systems, 2019, vol. 32.
[9]. Zhang J. et al. Passport-aware normalization for deep model protection. Advances in Neural Information Processing Systems, 2020, vol. 33, pp. 22619-22628.
[10]. Liu H. et al. Trapdoor normalization with irreversible ownership verification. International Conference on Machine Learning. PMLR, 2023, pp. 22177-22187.
[11]. Zhou A. et al. Permutation equivariant neural functionals. Advances in neural information processing systems, 2023, vol. 36, pp. 24966-24992.
[12]. Yan Y. et al. Cracking white-box dnn watermarks via invariant neuron transforms. arXiv preprint, 2022. DOI: 10.48550/arXiv.2205.00199.
[13]. Li F. Q., Wang S. L., Zhu Y. Fostering the robustness of white-box deep neural network watermarks by neuron alignment. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 3049-3053.
[14]. Ganju K. et al. Property inference attacks on fully connected neural networks using permutation invariant representations. Proceedings of the 2018 ACM SIGSAC conference on computer and communications security, 2018, pp. 619-633.
[15]. Krizhevsky A. et al. Learning multiple layers of features from tiny images. 2009.
[16]. Fei-Fei L., Fergus R., Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. 2004 conference on computer vision and pattern recognition workshop. IEEE, 2004, pp. 178-178.
[17]. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, vol. 25.
[18]. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014. DOI: 10.48550/arXiv.1409.1556.
[19]. He K. et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.
[20]. Huang G. et al. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700-4708.
[21]. Szegedy C. et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.
[22]. Cheng H., Zhang M., Shi J. Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, vol. 46, no. 12, pp. 10558-10578.
[23]. Adi Y. et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. 27th USENIX security symposium (USENIX Security 18), 2018, pp. 1615-1631.
[24]. Trias C. D. S. et al. Find the Lady: Permutation and Re-Synchronization of Deep Neural Networks. arXiv preprint, 2023. DOI: 10.48550/arXiv.2312.14182.

Информация об авторах / Information about authors

Александр Андреевич АКИМЕНКОВ – стажер-исследователь ИСП им. В.П. Иванникова РАН, аспирант. Научные интересы: цифровые водяные знаки, обработка цифровых изображений, алгоритмы машинного обучения.

Alexander Andreevich AKIMENKOV is a research intern of the Ivannikov Institute for System Programming of the RAS and a PhD student. Scientific interests: steganography, digital image processing, machine learning algorithms.

Дмитрий Олегович ОБЫДЕНКОВ – кандидат технических наук, научный сотрудник ИСП им. В.П. Иванникова РАН. Научные интересы: методы сокрытия и защищённой передачи информации, компьютерные сети, технологии анализа сетевого трафика.

Dmitry Olegovich OBYDENKOV – Cand. Sci. (Tech.). He is a researcher of the Ivannikov Institute for System Programming of the RAS. Scientific interests: methods for information hiding and secure transmission, computer networks, technologies of network traffic analysis.

Алексей Юрьевич ЯКУШЕВ – научный сотрудник ИСП им. В.П. Иванникова РАН. Научные интересы: цифровые водяные знаки, обработка цифровых изображений, алгоритмы машинного обучения.

Aleksey Yur'evich YAKUSHEV is a researcher of the Ivannikov Institute for System Programming of the RAS. Scientific interests: watermarking, digital image processing, machine learning algorithms.

Халед Набиль АБУД – стажер исследователь Института искусственного интеллекта МГУ имени М.В. Ломоносова, аспирант. Научные интересы: состязательная устойчивость, генеративные модели, обработка цифровых изображений, глубокое обучение.

Khaled Nabil ABUD is a research intern of the MSU Institute for AI and a postgraduate student. Scientific interests: adversarial robustness, generative models, digital image processing, deep learning.

Алиса Андреевна УСТИНОВА – лаборант ИСП им. В.П. Иванникова РАН. Область научных интересов: информационная безопасность, обработка изображений, алгоритмы машинного обучения.

Alisa Andreevna USTINOVA is a laboratory assistant of the Ivannikov Institute for System Programming of the RAS. Scientific interests: information security, image processing, machine learning algorithms.

Юрий Витальевич МАРКИН – кандидат технических наук, научный сотрудник ИСП им. В.П. Иванникова РАН. Область научных интересов: информационная безопасность, анализ сетевого трафика, обработка изображений, алгоритмы машинного обучения.

Yury Vital'evich MARKIN – Cand. Sci. (Tech.). He is a researcher of the Ivannikov Institute for System Programming of the RAS. Scientific interests: information security, network traffic analysis, image processing, machine learning algorithms.