# Метод защиты авторских прав на глубокие нейронные сети с помощью цифровых водяных знаков

*Ю.Д. Выборнова* [1]

[1] *Самарский национальный исследовательский университет имени академика С.П. Королёва, 443086, Россия, г. Самара, Московское шоссе, д. 34*

***Аннотация***

В статье предлагается новый метод защиты авторских прав на глубокие нейронные сети. Основная идея метода заключается во встраивании цифровых водяных знаков в защищаемую модель путем ее дообучения на уникальном наборе псевдоголографических изображений (псевдоголограмм). Псевдоголограмма – это двумерный синусоидальный сигнал, кодирующий двоичную последовательность произвольной длины. Изменяя фазу каждой синусоиды, можно формировать различные изображения-псевдоголограммы на основе одной битовой последовательности. Предлагаемая схема встраивания заключается в генерации обучающей выборки таким образом, чтобы псевдоголограммы, сформированные на основе одной последовательности, попадали в один и тот же класс. При этом каждому классу будут соответствовать различные битовые последовательности. Верификация цифровых водяных знаков осуществляется путем подачи на вход модели различных псевдоголограмм и проверки соответствия скрытой в них последовательности определенному классу. Экспериментальные исследования подтверждают работоспособность метода, а также соответствие всем критериям качества, выдвигаемым к методам встраивания цифровых водяных знаков в нейронные сети.

*Ключевые слова*: защита авторских прав, цифровой водяной знак, глубокие нейронные сети, псевдоголограмма.

*Цитирование*: **Выборнова, Ю.Д.** Метод защиты авторских прав на глубокие нейронные сети с помощью цифровых водяных знаков / Ю.Д. Выборнова // Компьютерная оптика. – 2023. – Т. 47, № 2. – С. 251-261. – DOI: 10.18287/2412-6179-CO-1193.

*Citation*: Vybornova YD. Method for copyright protection of deep neural networks using digital watermarking. Computer Optics 2023; 47(2): 251-261. DOI: 10.18287/2412-6179-CO-1193.

## *Введение*

Обучение современной глубокой нейронной сети требует проектирования архитектуры сети, сбора и предварительной обработки данных и доступа к аппаратным ресурсам, в частности, графическим процессорам (GPU), способным обучать такие модели.

Возможность обмена обученными моделями является неотъемлемой частью исследований и разработок в области искусственного интеллекта, поскольку использование уже готовых моделей для инициализации весов и дальнейшего дообучения под конкретные задачи позволяет значительно сократить вычислительные затраты.

Таким образом, обученные модели могут рассматриваться как интеллектуальная собственность, а их законные владельцы могут столкнуться с рядом проблем, вызванных отсутствием необходимых мер защиты авторских прав. Злоумышленники могут распространять проприетарные модели или нелегально использовать их для предоставления услуг анализа данных. В этой связи возникает необходимость создания методов и средств, позволяющих владельцу проследить и доказать факт несанкционированного копирования и распространения глубоких нейронных сетей.

Методы встраивания цифровых водяных знаков (ЦВЗ) широко использовались в последние два десятилетия как средство защиты авторских прав на мультимедийные данные (изображения, видео и аудио). В 2017 году [1] были предприняты первые попытки перенести основные идеи встраивания ЦВЗ для случая, когда контейнером для встраивания является глубокая нейронная сеть. Если для встраивания ЦВЗ в изображение необходимо по заданной схеме изменить значения пикселей, то в случае встраивания в нейросеть изменениям подвергаются веса обученной модели. В зависимости от доступности весов на этапе верификации разделяют два основных подхода к встраиванию ЦВЗ в глубокие модели: White-Box [1 – 9] (веса модели доступны) и Black-Box [10 – 25] (доступны только прогнозы модели).

Методы встраивания ЦВЗ по принципу Black-Box заключаются в дообучении исходной модели на так называемой триггерной выборке таким образом, чтобы модель с ЦВЗ выдавала заданную реакцию на образцы-триггеры и чтобы при этом результаты были отличны от прогноза исходной модели. Например, можно дообучить защищаемую модель давать намеренно неверный результат на определенных входных данных. Процедура проверки авторских прав на модель глубокого обучения осуществляется путем подачи на вход модели сформированного автором верификационного набора триггеров: высокая доля корректных прогнозов свидетельствует о наличии ЦВЗ.

Текущие исследования в рамках обоих подходов сосредоточены на разработке метода встраивания ЦВЗ в глубокие нейронные сети, который бы одновременно удовлетворял ряду критериев.

1) Точность. Качество решения исходных задач анализа данных не должно ухудшаться после встраивания ЦВЗ.

2) Стойкость. ЦВЗ должен быть устойчивым к дообучению модели.

3) Информационная емкость. Количество встроенной информации должно быть достаточным для того, чтобы у законного владельца имелись основания для заявления об авторских правах.

4) Безопасность. Неавторизованная сторона не должна иметь возможность обнаружить наличие ЦВЗ.

5) Достоверность. Извлечение ЦВЗ должно иметь минимальный уровень ложноотрицательных результатов, в то же время модели без ЦВЗ не должны быть ложно верифицированы.

6) Уникальность. ЦВЗ должен гарантировать однозначное соответствие между моделью и ее владельцем.

В данной работе предлагается новый метод защиты глубоких нейронных сетей, решающих задачу классификации изображений. Предлагаемый метод, основанный на новом Black-Box подходе к встраиванию ЦВЗ, удовлетворяет всем перечисленным критериям качества благодаря использованию в качестве ЦВЗ уникального набора триггеров, формируемого путем синтеза псевдоголографических изображений (псевдоголограмм).

Псевдоголографическое изображение – это двумерный сигнал, спектр которого кодирует заданную бинарную последовательность [26]. Для получения такого изображения необходимо расставить импульсы двумерного спектра в зависимости от кодируемых битов по определенному правилу (так, чтобы впоследствии последовательность нулей и единиц могла быть однозначно декодирована). Переход от спектра к псевдоголографическому изображению осуществляется с помощью обратного дискретного преобразования Фурье (ДПФ).

Использование такого сигнала в качестве ЦВЗ в рамках решаемой задачи имеет ряд преимуществ по сравнению с существующими Black-Box схемами, согласно которым триггеры формируются, как правило, на основе классических растровых изображений из наиболее распространенных датасетов.

## *1. Обзор существующих исследований*

### *Схемы встраивания White-Box*

При таком сценарии для верификации ЦВЗ требуется доступ к параметрам проверяемой модели [1 – 9]. В [1, 4] авторы встраивают бинарную последовательность ЦВЗ в конкретные параметры целевой модели путем добавления к исходной функции потерь функции регуляризации, которая вызывает статистическое смещение по этим параметрам.

В [2] авторы предлагают добавить в архитектуру сети дополнительный слой, параметры которого зависят как от весов предшествующего сверточного слоя, так и от цифровой подписи владельца. Процедура встраивания ЦВЗ в данном случае заключается в обучении модели таким образом, чтобы при попытке подделки цифровой подписи эффективность решения исходной задачи была значительно ухудшена.

В [7] авторы применяют одну из наиболее популярных стратегий встраивания в изображения и аудио, именуемую модуляцией квантования (quantization index modulation, QIM), для контроля изменений, вносимых в значения параметров модели при встраивании ЦВЗ.

Поскольку без доступа к весам проверяемой модели верификация ЦВЗ будет невозможной, методы встраивания White-Box целесообразно использовать для защиты целостности и подлинности моделей, например, в задаче защиты критических систем искусственного интеллекта от подделки [9]. Если речь идет о подтверждении авторских прав на украденную модель, то более подходящее решение может быть найдено среди методов встраивания по принципу Black-Box.

### *Схемы встраивания Black-Box*

Появление этих схем обусловлено тем фактом, что прямой доступ к весам модели глубокого обучения не всегда возможен: вместо распространения чужой нейронной сети, злоумышленники могут использовать ее в своих клиентских приложениях. В таком случае для доказательства авторских прав законный владелец может только осуществлять запросы к удаленной модели и получать в ответ вычисленные прогнозы.

Идея методов встраивания ЦВЗ в модели глубокого обучения по принципу Black-Box состоит в том, чтобы дообучить модель таким образом, чтобы на этапе прогноза при подаче на вход определенных образцов, называемых триггерами, модель с ЦВЗ выдавала ожидаемые результаты, которые при этом отличны от исходной модели без ЦВЗ [10 – 25]. ЦВЗ встраивается в целевую модель путем ее дообучения на датасете, содержащем триггеры, при этом точность решения исходной задачи не должна быть значительно снижена. Такое переобучение возможно благодаря избыточной параметризации глубоких сетей. Верификация ЦВЗ происходит путем подачи триггеров на вход удаленной модели и сравнения ее прогнозов с данными разметки. Высокая доля правильных прогнозов свидетельствует о наличии ЦВЗ.

В случае моделей классификации методы встраивания Black-Box различаются в зависимости от того, как формируются пары триггер-метка класса.

Например, существует схема встраивания, которая основана на присвоении каждому триггеру случайной

метки [10, 11]. В [12] набор триггеров состоит из абстрактных изображений. В [13] триггерные выборки образуются путем добавления некоторых искажений к исходным данным, чтобы спровоцировать сеть давать неверные предсказания.

Еще одна схема встраивания заключается в нанесении видимых меток на изображения исходного набора данных. В [14 – 17] на изображения исходного набора данных добавляется текстовый логотип, в [18, 19] – особые метки, сформированные на основе подписи владельца.

Злоумышленники могут добавить в свою систему детектор, который обнаруживает подозрительные запросы. Например, при подаче на вход сети изображения-триггера с логотипом, злоумышленник может удалить логотип. В таком случае сеть распознает исходный класс изображения, а не класс-триггер. В [20] авторы предлагают генерировать выборку триггеров, которая точно соответствует распределению исходной обучающей выборки, что исключит возможность обнаружения злоумышленником запроса от легального владельца. В [21] логотип добавляется на изображение с помощью алгоритма перекодирования таким образом, чтобы в результате логотип был визуально неразличим.

Применение методов, использующих манипуляции с изображениями исходного датасета, может сказаться на точности решения нейросетью исходной задачи классификации. Кроме того, встроенный в модель ЦВЗ может быть удален путем дообучения на наборе, в котором отсутствуют триггеры. Низкая стойкость обусловлена близостью объектов оригинального набора данных и набора триггеров: дообучение может привести к игнорированию сетью логотипа, нанесенного на триггер, и повторной классификации таких образцов в их исходные классы [27].

Еще одна проблема перечисленных методов заключается в том, что они не устанавливают прочной связи между владельцем и его ЦВЗ-набором: злоумышленник может подобрать образцы, на которых модель выдает ошибочные прогнозы, и заявить их как свой собственный ЦВЗ-набор и, соответственно, заявить о праве собственности на модель. Таким образом, возникает неопределенность при верификации ЦВЗ, которая может привести законного владельца к потере интеллектуальной собственности. Кроме того, после регистрации в специализированном удостоверяющем центре законному владельцу потребуется обеспечить надежное хранение такой выборки, размер которой, вообще говоря, может исчисляться гигабайтами данных.

Стоит отметить, что, помимо возможного снижения точности самой модели, подобные схемы не исключают возможность возникновения ложноположительных и ложноотрицательных результатов на этапе верификации ЦВЗ, а следовательно, не могут обеспечить гарантию однозначного подтверждения авторских прав владельца защищаемой нейронной сети.

Перечисленные методы направлены на обеспечение защиты глубоких нейронных сетей, решающих задачи классификации изображений. Однако существуют и другие исследования в этой области. Например, в [28 – 30] ЦВЗ встраивается в глубокие нейронные сети, предназначенные для классификации аудио. В [31 – 33] авторы предлагают методы защиты для глубоких моделей обработки изображений, которые заключаются во встраивании ЦВЗ в результат работы генеративных моделей. Такие ЦВЗ визуально неразличимы и могут быть однозначно верифицированы специально обученным декодером. Для моделей-агентов в задачах обучения с подкреплением проектируются специальные пространственно-временные последовательности действий и состояний, которые должны дать возможность правообладателю отличить их, причём необходимо учесть, что агент со встроенным ЦВЗ не должен ухудшить свои качественные показатели в исходной задаче [34].

Кроме того, в 2022 году предложены новые стратегии защиты встроенных ЦВЗ от удаления при помощи дистилляции [35]. Появились два принципиально новых подхода к решению задачи: использование специальных аппаратных ускорителей как для White-Box, так и для Black-Box методов встраивания [35] и встраивание ЦВЗ в набор данных [36], так как его сбор и обработка требуют не меньше, а зачастую и больше ресурсов, чем обучение модели.

### *2. Предлагаемый метод*

Ключевая идея предлагаемого метода заключается в использовании специального типа двумерного сигнала (псевдоголографического изображения или псевдоголограммы) в качестве цифрового водяного знака.

Псевдоголографическое изображение – это сигнал, который кодирует двоичную последовательность $S = s_1, s_2, \ldots, s_l$ длины $l$ в виде синусоидальных функций. Для получения такого изображения искусственно синтезируются компоненты комплексного спектра, которые впоследствии отображаются в двумерные синусоиды.

В зависимости от кодируемого бита последовательности $S$ спектральные импульсы последовательно размещаются на двух «кольцах» разного радиуса. Необходимо разместить импульсы таким образом, чтобы одно из колец кодировало единицы, а второе – нули, а также определить правило, по которому будет осуществляться порядок чтения / записи кодируемых битов.

Для обеспечения возможности генерации различных псевдоголограмм на основе одной и той же последовательности комплексный спектр псевдоголограммы формируется путем генерации случайных значений действительной и мнимой части для каждого спектрального импульса.

Затем путем вычисления обратного ДПФ осуществляется переход к двумерному полутоновому изображению. Подробное описание и исследование алгоритма построения псевдоголографических изображений дано в [26].

Идея синтеза таких сигналов была предложена ранее в [26] для решения задачи защиты авторских прав на векторные данные. Кроме того, дальнейшие исследования псевдоголограмм [37] показали, что они совместимы с различными популярными стратегиями встраивания в растровые изображения и последовательности видеокадров. При этом использование псевдоголограмм в качестве ЦВЗ позволяет значительно повысить стойкость исследуемых схем встраивания ЦВЗ. В этой статье предлагается исследование применимости псевдоголограмм в качестве ЦВЗ, встраиваемых в модели глубокого обучения. Для этого разработан ряд модификаций метода [26] в части генерации псевдоголограмм.

Очевидно, поскольку визуально различить псевдоголограммы, кодирующие разные двоичные последовательности, может быть затруднительно, то при верификации ЦВЗ, встроенного в нейросетевую модель, может возникнуть необходимость считывания скрытых в псевдоголограммах последовательностей для их аутентификации.

Для извлечения последовательности необходимо вычислить ДПФ псевдоголограммы и локализовать координаты спектральных составляющих с большой амплитудой (то есть импульсов), по которым двоичная последовательность может быть декодирована.

В задачах Black-Box встраивания в модели глубокого обучения необходима возможность задания правила формирования триггерных выборок, которая позволит на основе выбранного ЦВЗ-сигнала сформировать наборы изображений для обучения и валидации. Использование псевдоголографических сигналов в качестве ЦВЗ в рамках решаемой задачи имеет ряд преимуществ по сравнению с классическими схемами, перечисленными в параграфе 2, согласно которым триггеры формируются на основе растровых изображений оригинального датасета.

Одной из важнейших особенностей метода является возможность обеспечения гарантированного взаимно однозначного соответствия между моделью и ее владельцем, поскольку псевдоголограммы строятся с использованием уникального ключа-идентификатора.

Ключ-идентификатор представляет собой объединение $K$ уникальных последовательностей, на основе которых формируются наборы псевдоголограмм для каждого класса $\mathbf{S} = \|_K^1 S_i = S_1 \| \ldots \| S_K$, где $\|$ – оператор конкатенации, $K$ – число классов. Каждому набору псевдоголограмм назначается метка класса, соответствующего индексу кодируемой последовательности внутри ключа-идентификатора: последовательность с индексом $i$ получит метку $i$-го класса.

Кроме того, наличие ключа-идентификатора позволяет избежать хранения наборов псевдоголограмм, используемых для встраивания и верификации ЦВЗ: при известном алгоритме генерации значений комплексного спектра триггерная выборка может быть сгенерирована повторно на основе ключа в любой момент, когда законному владельцу будет необходимо встроить ЦВЗ или проверить право собственности на модель. Стоит также отметить, что псевдоголограммы, сгенерированные на основе последовательностей, отличных от легального ключа-идентификатора, не будут корректно распознаваться моделью, защищенной с помощью этого ключа.

Возможность генерации различных псевдоголограмм на основе одной битовой последовательности гарантированно позволяет обеспечить достаточное количество триггеров для обучения модели различать подаваемые на вход псевдоголограммы. Предлагаемый подход позволяет обеспечить формирование до $90^{3l}$ различных псевдоголограмм на основе одной битовой последовательности $S_i$.

Как будет показано далее, предлагаемый метод защиты авторских прав позволяет контролировать количество ошибок верификации с помощью порога. Кроме того, поскольку признаки псевдоголограмм сильно отличаются от изображений оригинального датасета, точность решения исходной задачи не снижается после встраивания ЦВЗ, и при этом обеспечивается достаточная стойкость к атакам, направленным на удаление встроенной информации.

Таким образом, перечисленные преимущества предлагаемого метода позволяют сделать вывод о том, что он удовлетворяет всем критериям качества, выдвигаемым к методам встраивания ЦВЗ, что будет подтверждено далее результатами экспериментов.

### *3. Формирование ЦВЗ*

В данной статье для формирования триггерной выборки предлагается новый алгоритм создания цветных псевдоголограмм, обеспечивающих достаточное количество уникальных обучающих, валидационных и тестовых триггеров для каждого класса. Цветная псевдоголограмма, кодирующая последовательность заданной длины, формируется путем объединения трех полутоновых псевдоголограмм, созданных на основе этой последовательности. Все три псевдоголограммы имеют одинаковое расположение импульсов комплексного спектра, но при этом значения данных импульсов различны и задаются на основе генератора псевдослучайных чисел (ГПСЧ).

Необходимость создания цветных псевдоголограмм обусловлена рядом факторов. Во-первых, большинство современных архитектур глубоких нейронных сетей-классификаторов спроектированы для работы с цветными изображениями. В данном случае для подачи на вход нейронной сети одноканальных полутоновых изображений потребуется их

предварительное преобразование в трехканальные. При копировании одной и той же псевдоголограммы во все три цветовых канала результирующее изображение будет представлено в оттенках серого, в то время как синтез трехканального изображения путем объединения трех различных псевдоголограмм обеспечит разнообразие цветовых оттенков. Применительно к решаемой задаче защиты нейронных сетей новый подход может рассматриваться как способ аугментации данных, обеспечивающий большее разнообразие псевдоголограмм в обучающей выборке и, как следствие, лучшую генерализацию модели при встраивании ЦВЗ.

Как уже было отмечено ранее, комплексный спектр псевдоголограммы формируется путем генерации случайных значений действительной и мнимой части для каждого спектрального импульса. Для обеспечения гарантированной уникальности каждой псевдоголограммы необходимо, чтобы на вход ГПСЧ, используемого при синтезе значений импульсов, подавались неповторяющиеся значения инициализации. Таким образом, при генерации набора из $m$ трехканальных псевдоголограмм необходимо сформировать вектор инициализации из как минимум $3m$ уникальных значений.

В данной работе в качестве ГПСЧ выбран генератор $x^2 \bmod N$ [38]. Предложенный алгоритм синтеза набора из $m$ цветных псевдоголограмм состоит из следующих шагов.

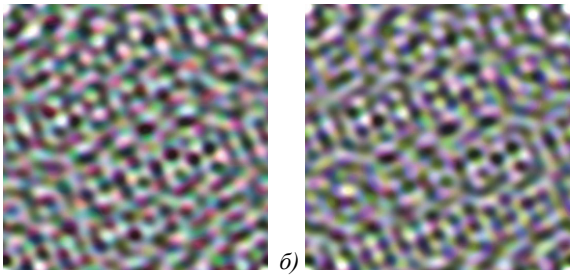
1. Сначала выбираются два простых числа $P$ и $Q$. При этом $P \equiv 3 \bmod 4$, $Q = 3 \bmod 4$. Во избежание необходимости хранения параметров такие числа могут быть сгенерированы на основе подписи владельца нейронной сети.
2. Вычисляется параметр ГПСЧ: $N = P \times Q$.
3. Формируется мультипликативная группа $Z_N^* = \{a \in Z_N \mid НОД(a, N) = 1\}$. Поскольку предполагается, что $\varphi(N) >> m$, где $\varphi(N)$ – функция Эйлера, то для генерации вектора инициализации ГПСЧ выбирается подмножество $\mathbf{A} = \{a_1, a_2, ..., a_t \mid a_k \in Z_N^*, k = \overline{1,t}\}$ из $t$ чисел взаимно простых с $N$, так чтобы $t \geq 3m$.
4. Выполняется случайная перестановка $\pi : \mathbf{A} \to \mathbf{A}$ выбранного подмножества. Первые $3m$ элементов полученной перестановки будут служить вектором инициализации $\mathbf{V}$ для генератора комплексного спектра: $\mathbf{V} = \{v_1, v_2, \ldots, v_{3m}\} = \{\pi(a_1), \pi(a_2), \ldots, \pi(a_{3m})\}$.
5. На основе полученного вектора инициализации формируются уникальные наборы значений $l$ спектральных импульсов для каждой псевдоголограммы:
   а) для $j$-й псевдоголограммы в качестве значения, инициализирующего ГПСЧ, будет взято значение $v_j : x_{j,0} = v_j$, $j = \overline{1,3m}$;
   б) далее для каждого $r$-го импульса $j$-й псевдоголограммы $f_{j,r}$, $r = \overline{1,l}$, формируется значение из 7 битов $b_{j,r}(\tau)$, $\tau = \overline{1,7}$, путем возведения предыдущего целочисленного значения генератора в квадрат по модулю $N$ и извлечения бита четности:

$$x_{j,r}(\tau) = x_{j,r}^2(\tau - 1) \bmod N;$$
$$b_{j,r}(\tau) = x_{j,r}(\tau) \bmod 2;$$

   всего на этом шаге понадобится $7l$ тактов ГПСЧ;
   в) значение угла для расчета значений действительной и мнимой части спектрального импульса формируется путем преобразования каждого 7-битного значения, полученного на предыдущем шаге, в десятичное число:

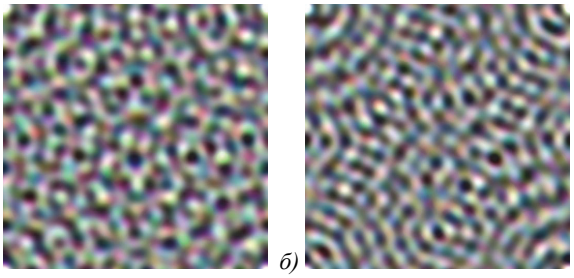
$$\alpha_{j,r} = \sum_{\tau=1}^{7} 2^{\tau-1} b_{j,r}(\tau);$$

   г) на основе полученных значений углов $\alpha_{j,r}$ рассчитываются значения действительной и мнимой части каждого импульса $f_{j,r}$: $\mathrm{Re}\, f_{j,r} = \cos \alpha_{j,r}$, $\mathrm{Im}\, f_{j,r} = \sin \alpha_{j,r}$.
6. Далее для каждого полученного спектра вычисляется обратное ДПФ, результатом которого будет полутоновая псевдоголограмма.
7. Каждые три псевдоголограммы объединяются в одно цветное изображение цветового пространства RGB. Примеры цветных псеводголограмм приведены на рис. 1 – 2.



*а)* *б)*

*Рис. 1. Пример цветных псевдоголограмм на основе одной последовательности*



*а)* *б)*

*Рис. 2. Пример цветных псевдоголограмм на основе разных последовательностей*

При формировании триггерной выборки необходимо учитывать тот факт, что псевдоголограммы, сформированные на основе одной и той же последовательности, не могут принадлежать разным классам, поскольку в этом случае псевдоголограммы представлены одними и теми же синусоидальными функциями, отличными по фазе. Таких различий недостаточно, чтобы обучить нейронную сеть

однозначно распознавать такие псевдоголограммы. Таким образом, для успешного встраивания ЦВЗ необходимо назначать псевдоголограммам, синтезированным на основе одной и той же последовательности, метку только одного класса.

### *4. Процедура встраивания ЦВЗ*

Как и другие Black-Box схемы, предлагаемый метод защиты авторских прав на нейронные сети заключается в дообучении глубокой модели на выборке, содержащей триггеры. Уникальность метода заключается в том, что триггерная выборка строится не на основе классических растровых изображений, а путем синтеза псевдоголограмм.

Для наглядности введем обозначения **W** – датасет для встраивания ЦВЗ, **T** – подмножество псевдоголограмм (триггерная выборка), **I** – подмножество оригинальных изображений. Таким образом, обучающая выборка $\mathbf{W} = \mathbf{T} \cup \mathbf{I}$.

Процедура встраивания ЦВЗ осуществляется следующим образом.

1) В первую очередь синтезируются наборы псевдоголограмм для каждого класса согласно схеме, представленной на рис. 3. Для этого:

а) Сначала формируется ключ-идентификатор **S** из $K$ последовательностей $S_i, i = \overline{1,K}$, заданной длины $l$, где $K$ – число классов.

б) Затем на основе каждой из последовательностей формируется набор из $m$ различных псевдоголограмм. Всего на этом шаге будет сформировано $m \times K = |\mathbf{T}|$ псевдоголограмм. Возможность формировать разные псевдоголограммы на основе одной последовательности обеспечивается благодаря заданию случайных значений импульсам на этапе синтеза комплексного спектра. Подробный алгоритм генерации псевдоголограмм приведен в параграфе 3.

в) Каждой псевдоголограмме, кодирующей последовательность $S_i$, назначается метка $i$-го класса.

2) Далее случайным образом выбираются $|\mathbf{I}|$ изображений оригинального датасета. Полученное подмножество **I** объединяется с подмножеством псевдоголограмм **T**.

3) Модель обучается на ЦВЗ-наборе **W** до тех пор, пока точность верификации ЦВЗ не будет достаточной для однозначного подтверждения авторских прав.

4) Процедура проверки авторских прав осуществляется путем оценки прогнозов модели на верификационном наборе псевдоголограмм, который формируется путем синтеза случайных псевдоголограмм на основе последовательностей $S_i, i = \overline{1,K}$, сгенерированных на шаге 1а). При этом псевдоголограммы должны быть отличны от элементов обучающей выборки.

Предложенная схема встраивания ЦВЗ представлена на рис. 4.

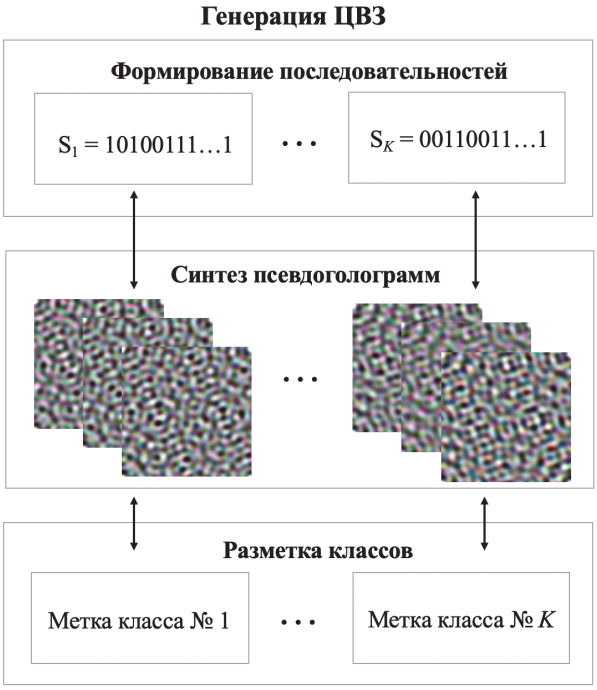



*Рис. 3. Процесс синтеза триггерной выборки*

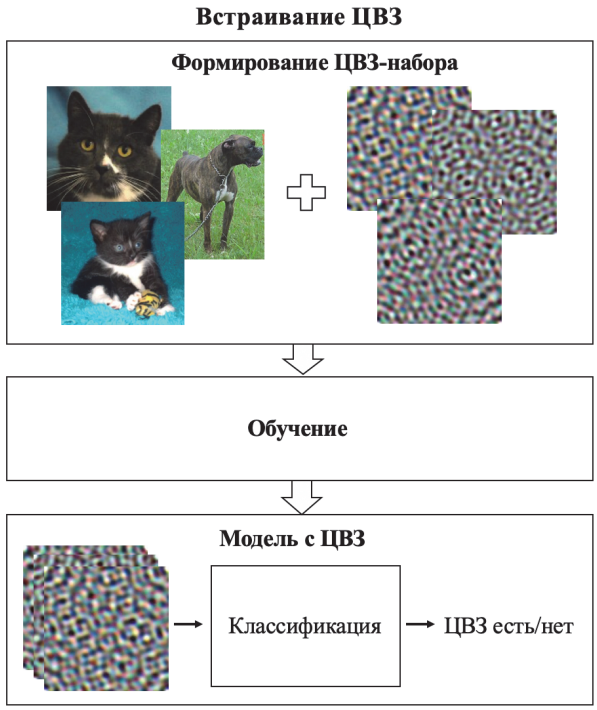



*Рис. 4. Общая схема встраивания ЦВЗ*

### *5. Экспериментальные исследования*

В качестве контейнеров для встраивания ЦВЗ были подготовлены две модели глубокого обучения. Для этого предобученная модель с архитектурой VGG11 из библиотеки Pytorch [39] была обучена на двух датасетах, CIFAR-10 и CIFAR-100 [40], с параметрами, приведенными в табл. 1. В качестве метода

оптимизации выбран стохастический градиентный спуск (Stochastic gradient descent, SGD) с фиксированным параметром шага на каждой итерации *learning_rate* $= 10^{-3}$, в качестве функции потерь – кросс-энтропия.

*Табл. 1. Параметры моделей-контейнеров*

| Набор данных | Размер батча | Число эпох | Точность на валидации |
|---|---|---|---|
| CIFAR-10 | 16 | 50 | 0,9366 |
| CIFAR-100 | 64 | 100 | 0,7791 |

*Исследование точности*

Целью данного эксперимента является оценка минимального размера триггерной выборки, достаточного как для однозначной верификации ЦВЗ, так и для сохранения исходной точности глубокой модели. Очевидно, что качество верификации может быть достигнуто путем увеличения количества псевдоголограмм в триггерном наборе, в то время как повышение точности модели на тестовом наборе может быть обеспечено путем увеличения количества элементов оригинального датасета.

Таким образом, в рамках данного эксперимента исследуется влияние размеров подмножеств $|\mathbf{T}|$ и $|\mathbf{I}|$ на точность процедуры классификации оригинальных изображений тестового датасета и процедуры верификации ЦВЗ.

Модель-контейнер обучалась на разных триггерных датасетах, сформированных для различных комбинаций параметров $|\mathbf{T}|$ и $|\mathbf{I}|$.

Для случая десяти классов (классификатор, предобученный на наборе CIFAR-10) были сгенерированы три набора псевдоголограмм с параметром $l = 20:10, 50, 100$ псевдоголограмм для каждого класса. Соответственно, размер подмножества псевдоголограмм варьировался как $|\mathbf{T}|:\{100, 500, 1000\}$.

Для случая ста классов (CIFAR-100) были сгенерированы наборы из 10, 20, 50 псевдоголограмм для каждого класса ($l = 20$). Соответственно, размер подмножества псевдоголограмм варьировался как $|\mathbf{T}|:\{1000, 2000, 5000\}$.

В качестве подмножества **I** рассмотрены случайные изображения оригинальных датасетов. Для случая десяти классов рассмотрены наборы, составляющие 10 %, 20 % и 30 % исходного датасета CIFAR-10, а для случая ста классов рассмотрены наборы, составляющие 10 %, 20 % и 30 % датасета CIFAR-100. Соответственно, в обоих случаях размер подмножества оригинальных изображений варьировался как $|\mathbf{I}|:\{5000, 10000, 15000\}$.

Помимо размеров триггерной выборки, в данном эксперименте варьировался и размер батча. Рассмотрены случаи $batch = \{16, 64, 128\}$.

Встраивание осуществлялось путем обучения моделей-контейнеров в течение 100 эпох. Для анализа эффективности встраивания с точки зрения точности решения исходной задачи классификации на каждой эпохе оценивалась точность $Acc_T$ на оригинальном тестовом наборе из датасета CIFAR-10 или CIFAR-100. Для оценки возможности модели корректно верифицировать встроенный ЦВЗ производился расчет точности модели $Acc_V$ на верификационном наборе, состоящем из 100 или 1000 псевдоголограмм (по 10 в каждом классе), сгенерированных для последовательностей $S$, используемых при формировании триггерной выборки, но при этом отличных от псевдоголограмм-триггеров.

В результате эксперимента для каждой комбинации параметров была выбрана модель с максимальным значением точности $Acc_V$. В случае, если таких моделей несколько, то среди них выбиралась модель с максимальным значением $Acc_T$.

В табл. 2 представлены результаты эксперимента для модели, обученной на CIFAR-10.

*Табл. 2. Результаты эксперимента для CIFAR-10*

| $\|\mathbf{T}\|$ | $\|\mathbf{I}\|$ | *batch* | $Acc_T$ | $Acc_V$ |
|---|---|---|---|---|
| 100 | 5 000 | 16 | 0,9193 | 0,99 |
| | | 64 | 0,9345 | 0,94 |
| | | 128 | 0,9419 | 0,82 |
| | 10 000 | 16 | 0,9213 | 0,92 |
| | | 64 | 0,9394 | 0,91 |
| | | 128 | 0,9415 | 0,64 |
| | 15 000 | 16 | 0,9265 | 1,00 |
| | | 64 | 0,9385 | 0,89 |
| | | 128 | 0,9412 | 0,75 |
| 500 | 5 000 | 16 | 0,9269 | 1,00 |
| | | 64 | 0,9392 | 1,00 |
| | | 128 | 0,9382 | 1,00 |
| | 10 000 | 16 | 0,9280 | 1,00 |
| | | 64 | 0,9431 | 0,99 |
| | | 128 | 0,9424 | 1,00 |
| | 15 000 | 16 | 0,9286 | 1,00 |
| | | 64 | 0,9401 | 0,99 |
| | | 128 | 0,9425 | 0,99 |
| 1000 | 5 000 | 16 | 0,9280 | 1,00 |
| | | 64 | 0,9389 | 1,00 |
| | | 128 | 0,9404 | 1,00 |
| | 10 000 | 16 | 0,9300 | 1,00 |
| | | 64 | 0,9412 | 1,00 |
| | | 128 | 0,9437 | 1,00 |
| | 15 000 | 16 | 0,9313 | 1,00 |
| | | 64 | 0,9424 | 1,00 |
| | | 128 | 0,9431 | 1,00 |

Результаты табл. 2 демонстрируют, что при $|\mathbf{T}| = 100$ с увеличением батча точность модели $Acc_T$ растет, но точность верификации $Acc_V$ при этом падает. При $|\mathbf{T}| = \{500, 1000\}$ точность верификации $Acc_V$

близка к единице при всех возможных комбинациях параметров.

Стоит отметить, что поскольку исходная модель-контейнер была обучена при $batch = 16$, то ее дообучение при встраивании ЦВЗ с параметрами $batch = 64$ или $batch = 128$ в результате дает более высокие показатели $Acc_T$ в сравнении с оригинальной моделью независимо от размеров $|\mathbf{T}|$ и $|\mathbf{I}|$.

При встраивании ЦВЗ с $batch = 16$ максимальная точность модели $Acc_T$ составила 0,9313, то есть исходная точность оригинальной модели не была достигнута. Однако точность модели с ЦВЗ может быть повышена путем задания порогового значения $Acc_V$, достаточного для однозначной верификации ЦВЗ. Например, при введении порога $Acc_V \geq 0{,}95$ точность модели $Acc_T$ составила 0,9343. Таким образом, при отсутствии возможности увеличения батча сохранение исходной точности модели может быть достигнуто с помощью уменьшения требуемой точности верификации $Acc_V$.

В табл. 3 представлены результаты эксперимента для случая ста классов (CIFAR-100).

*Табл. 3. Результаты эксперимента для CIFAR-100*

| $\|\mathbf{T}\|$ | $\|\mathbf{I}\|$ | batch | $Acc_T$ | $Acc_V$ |
|---|---|---|---|---|
| 1 000 | 5 000 | 16 | 0,6479 | 0,992 |
| | | 64 | 0,7370 | 0,977 |
| | | 128 | 0,7582 | 0,918 |
| | 10 000 | 16 | 0,6807 | 0,989 |
| | | 64 | 0,7458 | 0,965 |
| | | 128 | 0,7651 | 0,934 |
| | 15 000 | 16 | 0,6887 | 0,991 |
| | | 64 | 0,7442 | 0,941 |
| | | 128 | 0,7678 | 0,913 |
| 2 000 | 5 000 | 16 | 0,6480 | 1,000 |
| | | 64 | 0,7264 | 0,998 |
| | | 128 | 0,7564 | 0,993 |
| | 10 000 | 16 | 0,6742 | 0,999 |
| | | 64 | 0,7410 | 0,999 |
| | | 128 | 0,7637 | 0,987 |
| | 15 000 | 16 | 0,7059 | 0,999 |
| | | 64 | 0,7506 | 0,998 |
| | | 128 | 0,7674 | 0,994 |
| 5 000 | 5 000 | 16 | 0,6498 | 1,000 |
| | | 64 | 0,7211 | 1,000 |
| | | 128 | 0,7418 | 1,000 |
| | 10 000 | 16 | 0,6933 | 1,000 |
| | | 64 | 0,7482 | 1,000 |
| | | 128 | 0,7565 | 1,000 |
| | 15 000 | 16 | 0,6938 | 1,000 |
| | | 64 | 0,7563 | 1,000 |
| | | 128 | 0,7621 | 1,000 |

Согласно табл. 3 точность решения исходной задачи классификации $Acc_T$ заметно увеличивается с ростом $|\mathbf{I}|$. Если в одном классе 10 псевдоголограмм, то, как и для случая CIFAR-10, увеличение батча приводит к повышению точности на тестовой выборке, но при этом снижает качество верификации псевдоголографических ЦВЗ. Однако при $|\mathbf{T}| \geq 2000$ точность верификации $Acc_V$ близка к единице независимо от размера батча и размера подмножества оригинальных изображений $|\mathbf{I}|$.

Кроме того, как уже было отмечено, точность модели с ЦВЗ может быть повышена путем задания порогового значения $Acc_V$, достаточного для однозначной верификации ЦВЗ. Например, благодаря введению порогового значения $Acc_V \geq 0{,}9$ при $|\mathbf{T}| = 2000$, $|\mathbf{I}| = 15000$, $batch = 128$ модель со встроенным ЦВЗ достигает точности $Acc_T = 0{,}7772$. Стоит отметить, что при отсутствии порога (т.е. при условии $Acc_V \to 1$) максимальное значение точности верификации $Acc_V$ было достигнуто за 98 эпох для указанных параметров, а после введения порога – за 61 эпоху. Так, введение подобного ограничения позволяет не только сохранять исходную точность модели, но и значительно уменьшить время встраивания ЦВЗ.

*Информационная емкость*

Информационная емкость метода встраивания ЦВЗ может быть оценена как число встроенных бит ЦВЗ, достаточных для эффективной верификации авторских прав. Для метода, предложенного в данной статье, диапазон информационной емкости может быть оценен путем определения допустимых длин ключа-идентификатора, при которых сохраняется работоспособность алгоритмов встраивания и верификации ЦВЗ. Размер ключа зависит от числа классов защищаемой глубокой модели, а также от длины последовательностей, кодируемых в псевдоголограммах. Таким образом, $capacity = |\mathbf{S}| = l \times K$.

Целью данного эксперимента является исследование влияния длины последовательностей, кодируемых псевдоголограммами, на эффективность предложенного метода защиты авторских прав.

Для этого были сформированы наборы псевдоголограмм $\mathbf{T}$ для различных значений длины последовательности $l = \overline{10,100}$.

Аналогично предыдущему эксперименту в качестве контейнеров выбраны модели, обученные на наборах CIFAR-10 и CIFAR-100. Встраивание осуществлялось в течение 100 эпох. Оценка точности $Acc_T$ производилась на оригинальном тестовом наборе, оценка точности модели $Acc_V$ – на верификационном наборе, состоящем из уникальных псевдоголограмм, сгенерированных для последовательностей $S$ заданной длины $l$.

В результате эксперимента для каждой комбинации параметров была выбрана модель с максимальным значением точности $Acc_V$. В случае, если таких моделей несколько, то среди них выбиралась модель с максимальным значением $Acc_T$.

Параметры встраивания выбраны на основе результатов, полученных в предыдущем эксперименте:

CIFAR-10: $|\mathbf{T}| = 1000$, $|\mathbf{I}| = 15000$, $batch = 128$;

CIFAR-100: $|\mathbf{T}| = 2000$, $|\mathbf{I}| = 15000$, $batch = 128$.

Результаты эксперимента приведены на рис. 5.

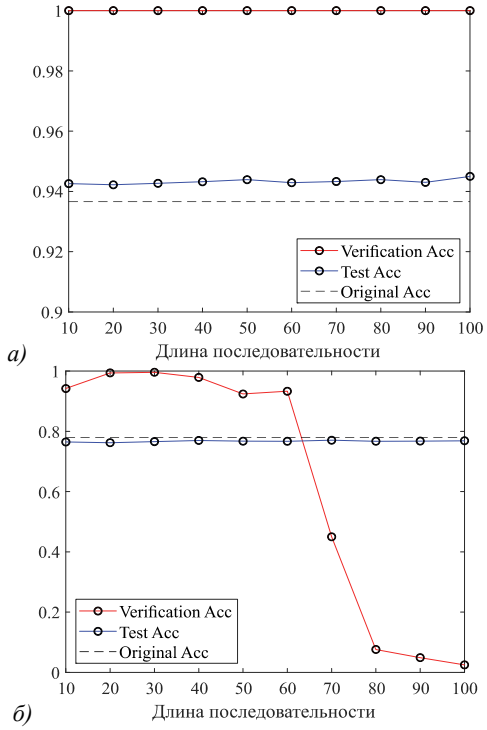



*Рис. 5. Результаты эксперимента для а) CIFAR-10; б) CIFAR-100*

Согласно полученным результатам в случае 10 классов длина последовательности практически не влияет на эффективность модели как с точки зрения точности верификации, так и с точки зрения точности модели на оригинальном тестовом наборе.

Однако для случая 100 классов при $l \geq 60$ верификация ЦВЗ оказалась невозможной, что объясняется небольшим размером набора псевдоголограмм. В отличие от случая 10 классов, где для каждого класса генерировалось 100 псевдоголограмм (суммарно $|\mathbf{T}| = 1000$), в этом случае генерировалось только 20 псевдоголограмм на класс (суммарно $|\mathbf{T}| = 2000$).

Таким образом, по результатам эксперимента может быть сделан вывод о том, что выбор параметров псевдоголограмм должен осуществляться в зависимости от количества классов защищаемой модели с учетом возможных ограничений на размер триггерной выборки.

### *Стойкость*

Целью данного эксперимента является оценка стойкости предлагаемого метода к атакам, направленным на удаление ЦВЗ путем переобучения модели.

Эксперимент построен следующим образом. Сначала в модели-контейнеры были встроены ЦВЗ с теми же параметрами, что и в предыдущем эксперименте. При этом длина кодируемых псевдоголограммами последовательностей составляет $l = 30$.

Были исследованы две атаки удаления ЦВЗ: обучение всех слоев модели (fine-tuning attack) и только последнего слоя (feature extraction attack). В обоих случаях обучение происходило в течение 100 эпох на оригинальном датасете с различными размерами. Каждый набор состоит из заданного количества случайных изображений исходного набора данных, а именно 10 %, 30 % и 50 % CIFAR-10/CIFAR-100.

На рис. 6 представлены результаты атаки fine-tuning на модель, обученную на CIFAR-10. На графике представлены значения $Acc_V$ после атаки для различных размеров обучающей выборки.

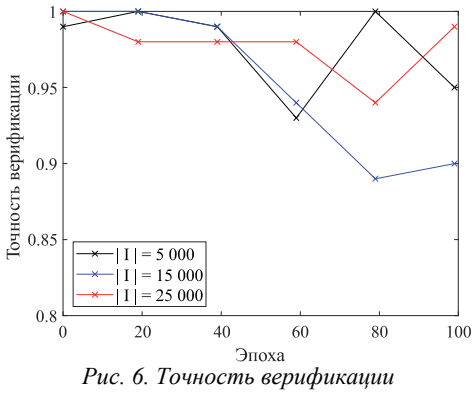



*Рис. 6. Точность верификации после атаки fine-tuning (CIFAR-10)*

Согласно полученным результатам даже в случае $|\mathbf{I}|$=25000 точность верификации $Acc_V \geq 0{,}89$, что свидетельствует о высокой устойчивости встроенного ЦВЗ к переобучению модели.

В случае атаки feature extraction точность верификации $Acc_V = 1$ независимо от размера обучающей выборки и времени обучения. Следовательно, встроенный ЦВЗ является в полной мере устойчивым к такой атаке.

На рис. 7 показаны соответствующие результаты атаки fine-tuning на модель, обученную на CIFAR-100.

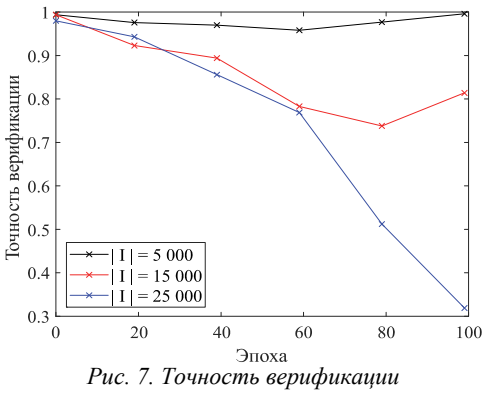



*Рис. 7. Точность верификации после атаки fine-tuning (CIFAR-100)*

Согласно результатам на рис. 6, после дообучения модели в течение 100 эпох встроенный ЦВЗ может быть корректно извлечен в случае, если $|\mathbf{I}| \leq 15000$. Однако в случае, когда набор данных злоумышленника $|\mathbf{I}|$=25000, точность верификации ЦВЗ начинает значительно снижаться после 60 эпох.

Более низкая стойкость в случае CIFAR-100 объясняется тем, что для случая CIFAR-10 каждый

класс содержит в два раза больше псевдоголограмм. Таким образом, стойкость ЦВЗ к атакам может быть повышена за счет увеличения размера набора триггеров.

Как и в предыдущем случае, встроенный ЦВЗ является в полной мере устойчивым к атаке feature extraction независимо от размера обучающей выборки и времени обучения.

По результатам проведенного эксперимента метод полностью устойчив к дообучению выходного слоя классификатора, а также обладает высокой стойкостью к дообучению всех параметров модели.

*Уникальность*

Данный эксперимент направлен на то, чтобы продемонстрировать соответствие встроенных ЦВЗ только легальному владельцу ключа-идентификатора.

Эксперимент проводился на модели, обученной на наборе CIFAR-10 с ЦВЗ, встроенным с параметрами $|\mathbf{T}| = 1000$, $|\mathbf{I}| = 15000$, $batch = 128$.

Для оценки уникальности ЦВЗ было сгенерировано 10 случайных ключей-идентификаторов $\mathbf{S}_1 \ldots \mathbf{S}^{10}$, отличных от ключа легального пользователя. Каждый ключ состоял из 10 псевдослучайных последовательностей длины $l = 30$.

На основе каждого ключа была сгенерирована верификационная выборка из 100 псевдоголограмм, которые подавались на вход модели с ЦВЗ. Эксперимент состоял в оценке точности верификации на каждой из выборок. Результаты эксперимента приведены на рис. 8.

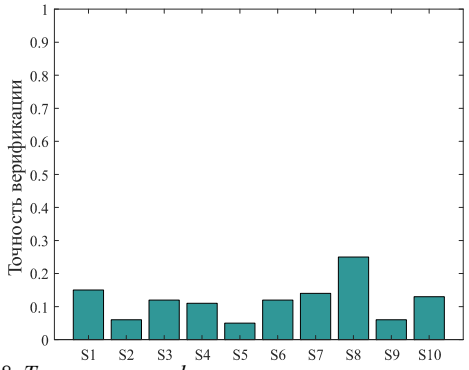



*Рис. 8. Точность верификации в случае компрометации псевдоголограмм*

Согласно результатам эксперимента, модель с ЦВЗ не распознает псевдоголограммы, которые не соответствуют последовательностям ключа-идентификатора, используемого при встраивании ЦВЗ, а следовательно, гарантируется возможность однозначной верификации авторских прав.

## *Заключение*

В статье предлагается новый метод защиты авторских прав на глубокие нейронные сети. Основная идея метода заключается во встраивании ЦВЗ в защищаемую модель путем ее дообучения на уникальном наборе псевдоголографических изображений. Предлагаемая схема встраивания заключается в генерации обучающей выборки таким образом, чтобы каждому классу соответствовали псевдоголограммы, кодирующие различные битовые последовательности.

Экспериментальные исследования подтверждают работоспособность метода. По результатам анализа результатов можно сделать вывод, что предлагаемый метод соответствует всем критериям качества, выдвигаемым к методам встраивания ЦВЗ:

1) Метод позволяет сохранить исходную точность оригинальной модели.

2) Метод демонстрирует достаточную стойкость к атакам, направленным на удаление ЦВЗ путем дообучения модели.

3) Метод позволяет формировать достаточное количество псевдоголограмм на основе последовательностей различной длины.

4) Поскольку метод не вносит искажений в модель, ЦВЗ не может быть обнаружен неавторизованной стороной.

5) Метод обеспечивает высокую точность верификации ЦВЗ, которая может контролироваться владельцем с помощью порога, в то же время модели без ЦВЗ не могут быть ложно верифицированы.

6) Благодаря заданию уникального ключа-идентификатора во время формирования триггерных наборов метод гарантирует однозначное соответствие между моделью и ее владельцем.

## *Благодарности*

Исследование выполнено за счет гранта Российского научного фонда № 21-71-00106, https://rscf.ru/project/21-71-00106/.

## *References*

[1] Uchida Y, Nagai Y, Sakazawa S, Satoh S. Embedding watermarks into deep neural networks. Proc 2017 ACM on Int Conf on Multimedia Retrieval 2017: 269-277.

[2] Fan L, Ng KW, Chan CS. Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks. Proc Advances in Neural Information Processing Systems 2019: 4714-4723.

[3] Wang T, Kerschbaum F. Robust and undetectable white-box watermarks for deep neural networks. arXiv Preprint. 2019. Source: <https://arxiv.org/abs/1910.14268>.

[4] Nagai Y, Uchida Y, Sakazawa S, Satoh S. Digital watermarking for deep neural networks. Int J Multimedia Inf Retr 2018; 7(1): 3-16.

[5] Chen H, Rohani BD, Koushanfar F. DeepMarks: A digital fingerprinting framework for deep neural networks. Proc 2019 on Int Conf on Multimedia Retrieval (ICMR '19). 2019: 105-113.

[6] Wang J, Wu H, Zhang X, Yao Y. Watermarking in deep neural networks via error back-propagation. J Electron Imaging 2020; 2020(4): 22.

[7] Kuribayashi M, Tanaka T, Suzuki S, Yasui T, Funabiki N. White-box watermarking scheme for fully-connected layers in fine-tuning model. Proc 2021 ACM Workshop on Information Hiding and Multimedia Security 2021: 165-170.

[8] Wang T, Kerschbaum F. RIGA: Covert and robust white-box watermarking of deep neural networks. Proc Web Conf 2021; 2021: 993-1004.
[9] Botta M, Cavagnino D, Esposito R. NeuNAC: A novel fragile watermarking algorithm for integrity protection of neural networks. Inf Sci 2021; 576: 228-241.
[10] Rouhani BD, Chen H, Koushanfar F. DeepSigns: A generic watermarking framework for IP protection of deep learning models. arXiv Preprint. 2018. Source: <https://arxiv.org/abs/1804.00750>.
[11] Zhang Y-Q, Jia Y-R, Niu Q, Chen N-D. DeepTrigger: A watermarking scheme of deep learning models based on chaotic automatic data annotation. IEEE Access 2020; 8: 213296-213305.
[12] Adi Y, Baum C, Cisse M, Pinkas B, Keshet J. Turning your weakness into a strength: Watermarking deep neural networks by backdooring. Proc 27th USENIX Security Symposium (USENIX Security 18) 2018: 1615-1631.
[13] Le Merrer E, Perez P, Trédan G. Adversarial frontier stitching for remote neural network watermarking. Neural Comput Appl 2020; 32: 9233-9244.
[14] Deeba F, Tefera G, She K, Memon H. Protecting the intellectual properties of digital watermark using deep neural network. 2019 4th Int Conf on Information Systems Engineering (ICISE) 2019: 91-95.
[15] Zhang J, Gu Z, Jang J, Wu H, Stoecklin MP, Huang H, Molloy I. Protecting intellectual property of deep neural networks with watermarking. Proc 2018 on Asia Conf on Computer and Communications Security 2018: 159-172.
[16] Sakazawa S, Myodo E, Tasaka K, Yanagihara H. Visual decoding of hidden watermark in trained deep neural network. 2019 IEEE Conf on Multimedia Information Processing and Retrieval (MIPR) 2019: 371-374.
[17] Wang G, Chen X, Xu C. Adversarial watermarking to attack deep neural networks. 2019 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2019: 1962-1966.
[18] Guo J, Potkonjak M. Watermarking deep neural networks for embedded systems. Proc 2018 IEEE/ACM Int Conf on Computer-Aided Design (ICCAD) 2018: 1-8.
[19] Jebreel NM, Domingo-Ferrer J, S<nchez D, Blanco-Justicia A. KeyNet: An asymmetric key-style framework for watermarking deep learning models. Appl Sci 2021; 11(3): 999. DOI: 10.3390/app11030999.
[20] Namba R, Sakuma J. Robust watermarking of neural network with exponential weighting. Proc 2019 ACM Asia Conf on Computer and Communications Security 2019: 228-240.
[21] Li Z, Hu C, Zhang Y, Guo S. How to prove your model belongs to you: A blind-watermark based framework to protect intellectual property of DNN. Proc 35th Annual Computer Security Applications Conf 2019: 126-137.
[22] Zhong Q, Zhang L, Zhang J, Gao L, Xiang Y. Protecting IP of deep neural networks with watermarking: A new label helps. Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conf 2020: 462-474.
[23] Xu X, Li Y, Yuan C. "Identity Bracelets" for deep neural networks. IEEE Access 2020; 8: 102065-102074.
[24] Zhao J, Hu Q, Liu G, Ma X, Chen F, Hassan M. AFA: Adversarial fingerprinting authentication for deep neural networks. Comput Commun 2020; 150: 488-497.
[25] Zhu R, Zhang X, Shi M, et al. Secure neural network watermarking protocol against forging attack. J Image Video Proc 2020; 2020: 37.
[26] Vybornova YD, Sergeev VV. New method for GIS vector data protection based on the use of secondary watermark. Computer Optics 2019; 43(3): 474-483. DOI: 10.18287/2412-6179-2019-43-3-474-483.
[27] Cao X, Jia J, Gong NZ. IPGuard: Protecting the intellectual property of deep neural networks via fingerprinting the classification boundary. Proc 2021 ACM Asia Conf on Computer and Communications Security (ASIA CCS '21) 2021: 14-25.
[28] Kim W, Lee K. Digital watermarking for protecting audio classification datasets. 2020 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2020: 2842-2846.
[29] Chen H, Zhang W, Liu K, Chen K, Fang H, Yu N. Speech pattern based black-box model watermarking for automatic speech recognition. 2022 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP) 2022: 3059-3063.
[30] Wang Y, Wu H. Protecting the intellectual property of speaker recognition model by black-box watermarking in the frequency domain. Symmetry 2022; 14(3): 619.
[31] Wu H, Liu G, Yao Y, Zhang X. Watermarking neural networks with watermarked images. IEEE Trans Circuits Syst Video Technol 2021; 31(7): 2591-2601.
[32] Zhang J, Chen D, Liao J, Zhang W, Feng H, Yu N. Deep model intellectual property protection via deep watermarking. IEEE Trans Pattern Anal Mach Intell 2022; 44: 4005-4020.
[33] Quan Y, Teng H, Chen Y, Ji H. Watermarking deep neural networks in image processing. IEEE Trans Neural Netw Learn Syst 2021; 32(5): 1852-1865.
[34] Chen K, Guo S, Zhang T, Li S, Liu Y. Temporal watermarks for deep reinforcement learning models. Proc 20th Int Conf on Autonomous Agents and MultiAgent Systems (AAMAS '21) 2021: 314-322.
[35] Clements J, Lao Y. DeepHardMark: Towards watermarking neural network hardware. 2022. Source: <https://www.aaai.org/AAAI22Papers/AAAI-4631.ClementsJ.pdf>.
[36] Tekgul BGA, Asokan N. On the effectiveness of dataset watermarking in adversarial settings. arXiv Preview. 2022. Source: <https://arxiv.org/abs/2202.12506>.
[37] Vybornova Y. Method for protection of heterogeneous data based on pseudo-holographic watermarks. 2021 9th Int Symposium on Digital Forensics and Security (ISDFS) 2021: 1-5.
[38] Vybornova YD. Password-based key derivation function as one of Blum-Blum-Shub pseudo-random generator applications. Procedia Eng 2017; 201: 428-435.
[39] Torchvision models subpackage. Source: <https://pytorch.org/vision/stable/models.html>.
[40] CIFAR-10 and CIFAR-100 Datasets. Source: <http://www.cs.toronto.edu/~kriz/cifar.html>.

### *Сведения об авторе*

**Выборнова Юлия Дмитриевна,** 1993 года рождения, в 2015 году окончила Самарский государственный аэрокосмический университет. В 2019 году защитила диссертацию на соискание ученой степени кандидата технических наук. Работает старшим научным сотрудником в НИЛ-55 Самарского национального исследовательского университета имени академика С. П. Королёва. Область научных интересов: защита данных, криптография, цифровые водяные знаки, обработка изображений. E-mail: *vybornovamail@gmail.com* .

*ГРНТИ: 28.23.15*
*Поступила в редакцию 15 июля 2022 г. Окончательный вариант – 23 октября 2022 г.*

# Method for copyright protection of deep neural networks using digital watermarking

*Y.D. Vybornova* [1]

[1] *Samara National Research University, 443086, Samara, Russia, Moskovskoye Shosse, 34*

***Abstract***

The article proposes a new method of copyright protection for deep neural networks. The main idea of the method is to embed digital watermarks into the protected model by retraining it on a unique set of pseudo-holographic images (pseudo-holograms). A pseudo-hologram is a two-dimensional sinusoidal signal that encodes a binary sequence of arbitrary length. By changing the phase of each sinusoid, it is possible to form various pseudo-hologram images based on a single bit sequence. The proposed approach to embedding is to generate a training sample in such a way that pseudo-holograms formed on the basis of one sequence fall into the same class. In this case, each class will correspond to different bit sequences. Verification of the digital watermark is carried out by applying various pseudo-holograms to the input of the model and checking whether the hidden sequence corresponds to a certain class. Experimental studies confirm the efficiency of the method and its compliance with all quality criteria established for the methods of neural network watermarking.

*Keywords*: copyright protection, digital watermarking, deep neural networks, pseudo-holographic image.

*Citation*: Vybornova YD. Method for copyright protection of deep neural networks using digital watermarking. Computer Optics 2023; 47(2): 251-261. DOI: 10.18287/2412-6179-CO-1193.

*Acknowledgements*: The reported study was funded by RSF (Russian Science Foundation) grant No. 21-71-00106, https://rscf.ru/en/project/21-71-00106/.

***Author's information***

**Yuliya Dmitrievna Vybornova** (b. 1993) graduated from Samara State Aerospace University in 2015, majoring in Information Security. In 2019 defended the thesis for the degree of Candidate of Technical Sciences. Currently works as a research fellow at Samara National Research University. Research interests are data protection, cryptography, steganography, and digital watermarking. E-mail: *vybornovamail@gmail.com* .

*Received July 15, 2022. The final version – October 23, 2022.*