УДК 004.891.3

**ОСОБЕННОСТИ ПРИМЕНЕНИЯ ПРЕДОБУЧЕННЫХ СВЕРТОЧНЫХ**

**НЕЙРОННЫХ СЕТЕЙ К ЗАДАЧАМ СТЕГОАНАЛИЗА ГРАФИЧЕСКИХ**

**ИЗОБРАЖЕНИЙ**

© **С. Н. Терещенко1, А. А. Перов2, А. Л. Осипов1**

*1Новосибирский государственный университет экономики и управления,*

*630099, г. Новосибирск, ул. Каменская, 56*

*E-mail:* [alosip*@*mail*.*ru](mailto:alosip@mail.ru)

*2Московский политехнический университет,*

*107023, г. Москва, ул. Б. Семеновская, 38*

*E-mail:* [perov\_artem@inbox.ru](mailto:perov_artem@inbox.ru)

Исследовано использование свёрточных нейронных сетей с целью анализа контейнера графических изображений на предмет данных, внедренных методами стеганографии. Показано, что глубокая сверточная нейронная сеть обучается классифицировать наличие скрытых данных в графических изображениях, достигая точности по метрике weighted AUC равной 0,928. Проверена гипотеза эффективности применения концепции "transfer learning” для сферы стеганографии. Эффективность предложенной технологии продемонстрирована на большом экспериментальном материале.

*Ключевые слова:* машинное обучение, сверточные нейронные сети, стеганография, стегоанализ, контейнер, классификация, графические изображения.

**Введение.** Применение стеганографических методов защиты информации позволяет реализовать один из трёх базовых критериев информационной безопасности – конфиденциальность. Принципиальным отличием стеганографии от методов шифрования данных является то, что сама информация может не подвергаться прямому преобразованию и оставаться в своём первоначальном виде, однако скрывается сам факт передачи информации, которая внедряется в контейнер. Как правило, в современной интерпретации контейнером является элемент файловой системы, предназначенный для хранения скрываемой информации.

Идея внедрять секретную информацию в сообщения и контейнеры берёт своё начало ещё до нашей эры. Тогда одним из наиболее широко применяемых способов являлось использование специальных симпатических чернил, которые проявляли себя только под воздействием какого-либо химического реактива, освещения ультрафиолетом или термического воздействия. Помимо проявляющихся чернил для внедрения секретных сообщений применялись микроточки, содержащие сообщение в максимально физически сжатом виде, трафареты, являющиеся «картой» скрытого сообщения, и акростихи, в которых контейнер был построен таким образом, что каждый символ слова являлся составляющей секретного сообщения.

Одним из первых современных методов цифровой стеганографии является LSB [1-2] метод, основанный на внедрении сообщения в наименьшие значащие биты пикселей растрового изображения, что позволяет осуществить скрытие в контейнере сообщения объемом до *H\*W\**3 бит информации, где *H* - высота изображения, а *W* - ширина. Используя палитру RGB, пиксель контейнера может спрятать по биту в каждой из компонент цвета. Заполнение младших бит каждой из компонент цвета в пикселе не изменяет визуальное восприятие цифрового изображения, однако позволяет внедрить сообщение большого объема. Такого рода методы позволяют использовать избыточность мультимедийных форматов (в качестве контейнера может применяться не только цифровое изображение, но и аудио, видео или текстовый файл). Цифровая стеганография использует глубокое обучение и нейронные сети в нескольких работах. В статье [3] посредством применения свёрточных нейронных сетей производится внедрение информации в контейнер. Использование аппарата свёрточных нейронных сетей для внедрения скрытых сообщений в контейнеры, а также определения наличия такой скрытой информации в цифровых объектах (стегоанализе) представлено ещё в ряде российских и зарубежных научных трудов [4-6], а в работе [7] авторы предложили в качестве контейнера видео файл, внедрив в него информацию.

Задачи по выявлению в контейнере внедренных сообщений решаются другим научным направлением - стегоанализом. Методы анализа контейнеров на предмет внедренной информации, основанные на сжатии данных, приведены в статьях [9-10]. В данных работах на примере графических изображений форматов BMP и JPEG, а также аудиофайлов WAV показано, что отрезки битовых последовательностей, не содержащих скрытую информацию, сжимаются лучше, чем соответствующие им отрезки со стего-внедренными данными, что с высокой вероятностью может свидетельствовать о наличии внедренной информации.

Применение математических методов стеганографии используется в [11], где авторы предлагают подход к стегоанализу, использующий предварительную фильтрацию. Результаты экспериментов, проведенные авторами в [11] показывают существенное снижение ошибки обнаружения внедренной в контейнер информации. Алгоритмы обнаружения, основанные на математических методах, предложены в работах [12-14] зарубежных учёных.

В настоящее время сфера применения технологий машинного обучения расширяется и открываются новые задачи. В частности, свёрточные нейронные сети все чаще применяются и для задач информационной безопасности. Применение свёрточных нейронных сетей в задачах информационной безопасности не является новым направлением. В статье [15] рассматривается применение технологий машинного обучения к задачам криптографии. Описывается известная в криптографии атака по побочным каналам (side-channel attack), а также упоминается несколько опубликованных атак на криптоалгоритмы, основанные на технологиях машинного обучения. Авторами предлагается атака на алгоритм AES, основанная на глубоком обучении, которая является более эффективной, чем существующие шаблонные атаки. В работе [16] также рассматривается атака по побочным каналам, основанная на технологиях машинного обучения, на цифровую подпись EdDSA, основанную на эллиптической кривой Эдвардса. В [17] автор предлагает универсальный метод статистического анализа, основанный на применении свёрточных нейронных сетей, который позволяет эффективно обнаруживать отклонения от случайности в выходных последовательностях итеративных алгоритмов шифрования на существенно меньших длинах выборок, чем применяются в аналитических и статистических методах [25].

Одной из близких работ является [18]. Авторы рассматривают возможность применения аппарата свёрточных нейронных сетей для обнаружения стеганографических вложений в цифровых изображениях, разрабатывая свою модель на основе свёрточной нейронной сети. Результаты исследования авторов демонстрируют возможность обнаружения до 85% фактов наличия стеганографических вложений, применяя при этом достаточно простую в реализации модель, которая не применяет сложных статистических алгоритмов. При построении модели нейронной сети можно использовать два подхода. Первый предполагает разработку новой архитектуры нейронной сети и ее обучение [18]. Второй — это концепция "transfer learning”, который предполагает использование готовой нейронной сети, предобученной на глобальной базе изображений ImageNet, и переобучением последних слоев на новой обучающей выборке.

Концепция "transfer learning" уже доказала свою эффективность во многих задачах. В данной работе проверяется гипотеза эффективности применения концепции "transfer learning” для сферы стеганографии. В статье рассмотрено применении архитектуры ResNet50 к задачам обнаружения внедренной информации в контейнер.

**Материалы и методы исследования.** В настоящей работе исходным материалом для исследований послужили размеченные изображения с внедренной в контейнер информацией, а также без наличия таковой, опубликованные в открытом доступе платформы Kaggle (kaggle.com). Методы исследования: теория проектирования и разработки информационных систем, программирование, методы аугментации и расширения обучающей выборки для задач компьютерного зрения, методы настройки гиперпараметров обучения моделей нейронной сети для обработки графических изображений.

**Результаты исследования.** Открытая база обучающей выборки представляет собой 300 000 изображений. Изображения размечены на четыре класса. Три класса с наличием вложенной информации методами трех алгоритмов (JMiPOD [V. Sedighi], JUNIWARD [Holub], UERD [Guo, Linjie]) и один для обычного изображения. Изображения размером 512х512 пикселей в формате JPEG. К набору изображений прилагается файл в формате "csv" с разметкой фотографий. Для обучения использовался фреймворк PyTorch, библиотека torchvision. В работе решается задача классификации. Изображения представлены в формате JPEG. Данный формат имеет свои особенности. JPEG представляет собой алгоритм сжатия файла изображения, позволяющий уменьшить его размер без потери большого количества информации. Преобразование происходит поэтапно. Сначала изображение преобразуется в YCbCr из каналов RGB. Затем DCT (аббревиатура с англ. дискретное косинусное преобразование) наносится на пиксели этих каналов, используя соответствующие коэффициенты. Изображение, закодированное с помощью алгоритма JPEG, остается в цветовом пространстве YCbCr до тех пор, пока оно не будет декодировано программным обеспечением для просмотра изображений. Имеются подходы, которые позволяют скрывать информацию в коэффициентах DCT различных каналов изображения JPEG, и поэтому полезная нагрузка (секретные данные) случайным образом распределяется между ними с учетом статистики коэффициентов DCT.

Cтепень заполненности контейнера различается. Изображения с низкой "плотностью" текстуры используются для скрытия более коротких сообщений. Для изображений с высокой "плотностью" текстуры — более длинные. Средняя длина сообщения составляет 0,4 бита на ненулевой коэффициент AC DCT.

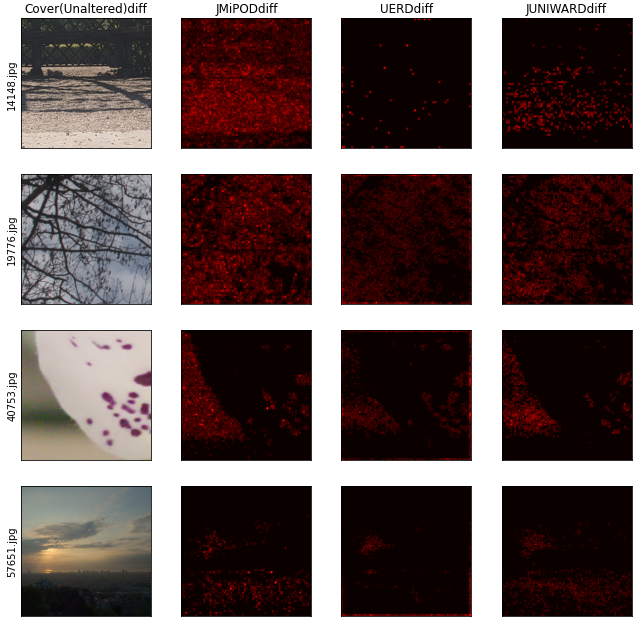
Анализ различия по пикселям в изображениях со скрытой информацией, закодированных тремя алгоритмами, от эталонного изображения по R-каналу, представлен на рис. 1.

Рисунок 1 - Отклонение пикселей от эталонного изображения по R-каналу

На рисунке 2 показано отклонение по DCT.



Рисунок 2 - Отклонение пикселей от эталонного изображения по DCT

Таким образом визуально видно, что каждый из алгоритмов оставляет свои особые отличия в кодировании пикселей. Необходимо обучить нейронную сеть определять эти признаки. Для задачи классификации в работе была использована концепция "transfer learning". Так как все современные нейронные сети используют анализ простейших графических примитивов на нижних слоях, то концепция "transfer learning" зарекомендовала себя уже на многих задачах [19-20]. Используя предобученную на большом количестве изображений нейронную сеть с переобучением последних слоев, что дает преимущество в стоимости и скорости обучения. Предобученная модель начинает процесс обучения не с полного обучения всех слоев, а с заданных паттернов обученных слоев, которые были получены при решении другой задачи, сходной по своей природе с решаемой. Предварительно обученная модель — это модель, которая была обучена на большом эталонном наборе данных (как правило несколько десятков миллионов) для решения задачи, аналогичной классификации наличия дополнительной информации в изображении. Была выбрана предобученная модель ResNet 50.

В работе предложена гипотеза, что предобученная глубинная сверточная нейронная сеть, за счет ключевой особенности по анализу простейших примитивов изображения, сможет выявить пиксельные отклонения для каждого из алгоритмов кодирования информации и выявить изображения без таковых. Ключевой особенностью многослойной сверточной нейронной сети ResNet является то, что она использует пропуск соединений или ярлыки для перехода через некоторые слои [21]. Типичные модели ResNet реализуются с двух или трехслойными пропусками, которые содержат нелинейности (ReLU) и пакетную нормализацию между ними. ResNet 50 является вариантом модели ResNet, которая имеет в общей сложности 50 слоев. В табл. 1 представлена архитектура ResNet 50. Нейронная сеть состоит из 50 слоев, сгруппированных в 5 уровней.

**Таблица 1. Архитектура ResNet 50.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Стеки слоев | Размер выхода | Слои | | |
| conv1 | 112x112 | 7×7, 64 / 2 | | |
| conv2\_x | 56×56 | 3×3 max pool / 2 | | |
| [1×1,  3×3,  1×1, | 64  64  256 ] | ×3 |
| conv3\_x | 28×28 | [ 1×1,  3×3,  1×1, | 128  128  128 ] | ×4 |
| conv4\_x | 14×14 | [ 1×1,  3×3,  1×1, | 256  256  1024 ] | ×6 |
| conv5\_x | 7×7 | [ 1×1,  3×3,  1×1, | 512  512  2048 ] | ×3 |
|  | 1×1 | average pool, softmax | | |

В работе использовался метод стохастического градиентного спуска (SGD). В качестве настройки гиперпараметра шага обучения, опытным путем было подобрано значение "lr=0.003". В качестве функции потерь была выбрана перекрестная энтропия, мультиклассовая функция оценки логарифмических потерь, описанная в работе [22].

В качестве алгоритма стохастической оптимизации был использован Адам. Общий набор изображений был разделен на три выборки: обучающая, проверочная, тестовая. Использовалась предварительно обученная (на ImageNet) модель ResNet 50 с переобучением последних слоев. В качестве метрики использовалось weighted AUC-ROC — площадь (англ. Area Under Curve) под кривой ошибок (англ. Receiver Operating Characteristic curve). Данная кривая представляет из себя линию от (0,0) до (1,1) в координатах True Positive Rate (TPR) и False Positive Rate (FPR). Пороги заданы массивом: [0.0, 0.4, 1.0]. Веса между порогами (weighted): [2, 1]. Критерий AUC-ROC устойчив к несбалансированным классам и может быть интерпретирован как вероятность того, что случайно выбранный позитивный объект будет проранжирован классификатором выше (будет иметь более высокую вероятность быть позитивным), чем случайно выбранный негативный объект [23]. На рис. 3 представлены значения метрики weighted AUC.

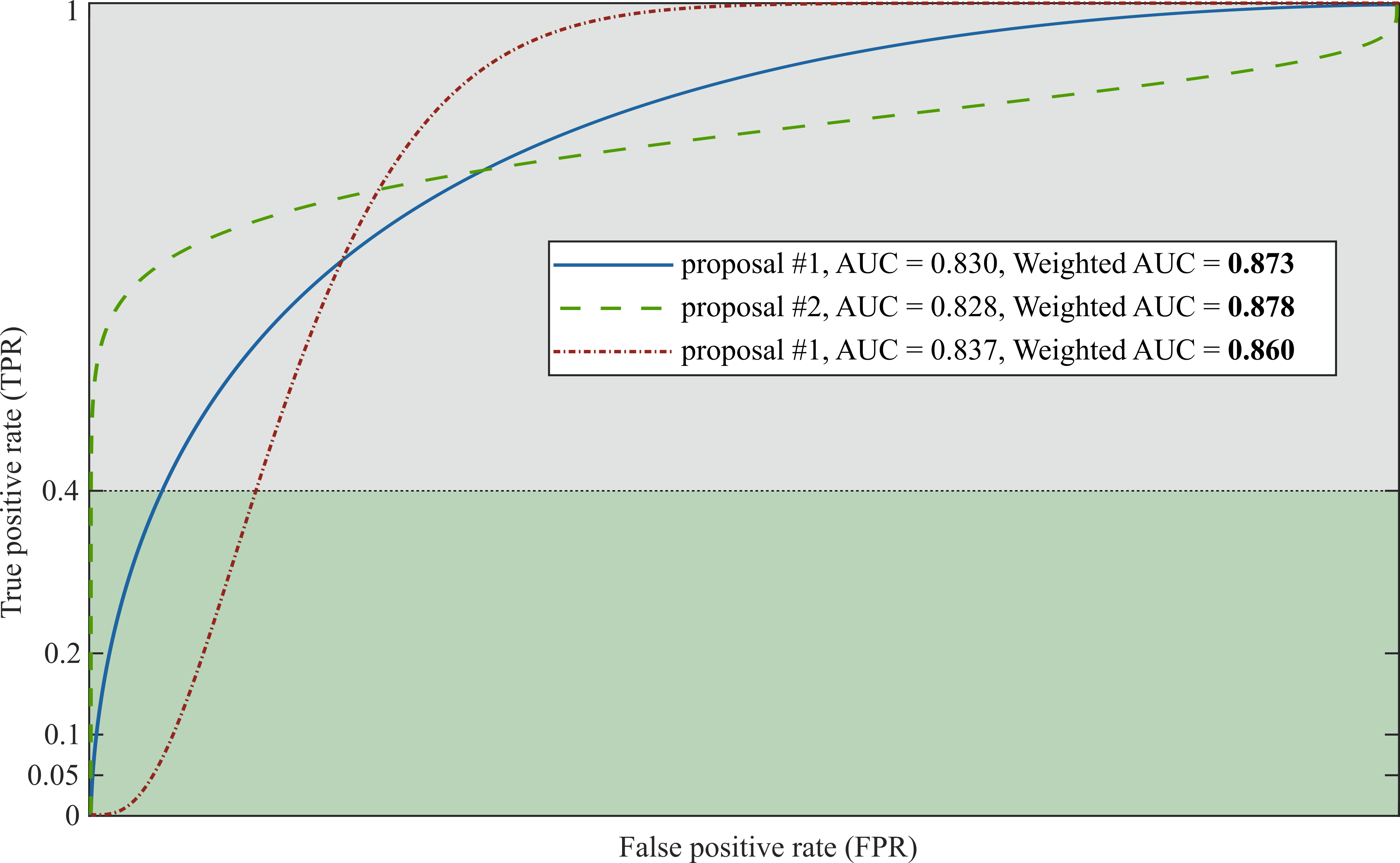


Рисунок 3 - Метрика weighted AUC [24]

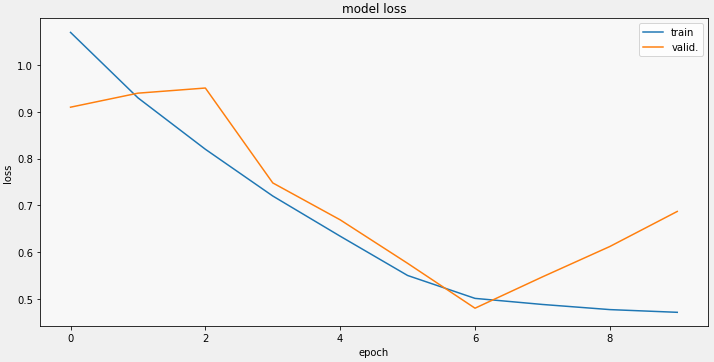
Обучение состояло из 10 эпох. Как видно на рисунке 4, на протяжении 6 эпох происходит снижение функции потерь синхронно по трейн (обучающей) и валидационной выборках. После 6 эпохи наблюдается расхождение в значениях трейн и валидации (трейн продолжает уменьшаться, а валидация начинает повышаться), что свидетельствует о начале процесса переобучения нейронной сети. 

Рисунок 4 - Значения функции потерь при обучении модели

Итоговые результаты, полученные по модели нейронной сети на тестовой выборке, показали, что точность классификации по метрике weighted AUC составила 0,928. Данные результаты позволяют сделать вывод о подтверждении гипотезы эффективности применения концепции “transfer learning” нейронных сетей для задач стеганографии.

Заключение. В данной работе демонстрируется эффективность концепции “transfer learning”, предполагающей использование предобученных (на глобальных наборах изображений) глубинных сверточных нейронных сетей, для выявления присутствия закодированной информации в изображении. Используя общедоступный набор данных из 300000 изображений эталонных и закодированных по трем наиболее известным алгоритмам стеганографии, глубокая сверточная нейронная сеть обучается классифицировать наличие скрытых данных, достигая точности по метрике weighted AUC равной 0,928. В данной работе был исследован подход использования методов глубокого обучения для автоматической классификации и обнаружения скрытой информации в изображениях. Была описана методика подготовки изображений для обучения и валидации, процедура обучения сверточной нейронной сетью и настройкой гиперпараметров обучения.

СПИСОК ЛИТЕРАТУРЫ

1. **Bhaskari L., Avadhani P., Damodaram A**. Watermartk insertion Algorithm implementation using Auxiliary carry and LSB methods / Proceedings of International Conference on Systemics, Cybernetics and Informatics, 2006. P. 666-668.

2. **Zhang, W., Zhang, X., Wang, S.** Adouble layered “plus-minus one” data embedding scheme // IEEE Signal Proc. Lett. 2007. Vol. 14, No. 11. P. 848–851.

3. **Kumar V, Laddha S., Aniker, Dogra N.** Steganography Techniques Using Convolutional Neural Networks / Review of Computer Engineering Studies, vol.7, 2020, pp.66-73.

4. **Husien, S., Badi, H.** Artificial neural network for steganography / Neural Comput & Applic 26, 2015, pp.111-116.

5. **Sharma A., Kushwaha A.** Image Steganography Scheme Using Neural Network in Wavelet Transform Domain / International Journal of Science and Research (IJSR), vol.4, Issue 12, 2015, pp. 1773-1778

6. **Полунин А.А., Яндашевская Э.А.** Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений / Труды ИСП РАН. 2020. №4.

7. **Velmurugan, K. & Subramani, Hemavathi** Video Steganography by Neural Networks Using Hash Function / Conference: 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2019, pp. 55-58

8. **Нечта И.В.** Метод стеганографического преобразования сообщения со свойством частичной неизвлекаемости // Вычислительные технологии. 2019. Т.24, № 3. С.75-87.

9. **Жилкин М.Ю.** Стегоанализ графических данных на основе методов сжатия // Сборник научных трудов «Вестник СибГУТИ». 2008. № 2. С. 62–66.

10. **Забелин М.А.** Стегоанализ аудиоданных на основе методов сжатия // Вестник СибГУТИ. 2010. № 1. С. 41-49.

11. **Монарёв В.А., Пестунов А.И.** Повышение эффективности методов стегоанализа при помощи предварительной фильтрации контейнеров // Прикладная дискретная математика. 2016. № 2(32). C. 87-99.

12. **Yang Z., Huang Y., Zhang Y.** A fast and efficient text steganalysis method // Covert Communicatin in Public Cyberspace. 2019. DOI: 10.1109/LSP.2019.2902095.

13. **Jin Z., Feng G., Ren Y., Zhang X.** Feature Extraction Optimization of JPEG Steganalysis Based on Residual Images / Signal Processing 170:107455, 2020. DOI: 10.1016/j.sigpro.2020.107455.

14. **Soto R.T., Ramos-Pollan R., Isaza G. and others.** Digital media steganalysis / Application of convolutional neural networks for Steganalysis, 2020. DOI: 10.1016/B978-0-12-819438-6.00020-7.

15. **Mohammed M. Alani.** Applications of Machine Learning in Cryptography: A Survey. 2019, 8 p. DOI: 10.1145/3309074.3309092.

16. **Weissbart L., Picek S., Batina L.** One trace is all it takes: Machine Learning-based Side-channel Attack on EdDSA / Cryptology ePrint Archive: Report 2019/358, 2019. 18 p.

17. **Перов А.А., Пестунов А.И.** О возможности применения сверточных нейронных сетей к построению универсальных атак на итеративные блочные шифры // Прикладная дискретная математика. 2020. №3 (49). С. 46-57.

18. **Полунин А.А., Яндашевская Э.А.** Использование аппарата свёрточных нейронных сетей для стегоанализа цифровых изображений // Труды ИСП РАН, т.32, вып. 4, 2020. C. 155-164.

19. Identification and Recognition of Rice Diseases and Pests Using Convolutional Neural Networks / Rahman C. R., Arko P. S., Ali M. E., Khan M. A. I., Apon S. H., Nowrin F., Wasif A. // Biosystems Engineering. June 2020. Vol. 194. P. 112-120. <https://doi.org/10.1016/j.biosystemseng.2020.03.020>.

20. Tea leaf diseases recognition using neural network ensemble / Karmokar B. C, Ullah M. S., Siddiquee Md. K., Alam K. Md. R. // International Journal of Computer Applications. March 2015. Vol. 114. No. 17. P. 27-30.

21. **He K, Zhang X, Ren Sh, Sun J.** Deep Residual Learning for Image Recognition. Available Online: <https://arxiv.org/abs/1512.03385>.

22. **Kima Y., Leea Y, Jeon M.** Imbalanced Image Classification with Complement Cross Entropy. Available Online: https://arxiv.org/pdf/2009.02189.pdf.

23. Метрики в задачах машинного обучения. Доступ: <https://habr.com/ru/company/ods/blog/328372>.

24. Alaska2 Image Steganalysis. Available Online: <https://www.kaggle.com/c/alaska2-image-steganalysis/overview/evaluation>.

25. **Osipov A.L., Bobrov L.K.** The use of statistical models of recognition in the virtual screening of chemical compounds // [Automatic Documentation and Mathematical Linguistics](https://www.elibrary.ru/contents.asp?id=36686289). 2012. Т. 46. [№ 4](https://www.elibrary.ru/contents.asp?id=36686289&selid=36686290). С. 153-158.