Метод защиты моделей глубокого обучения цифровыми водяными знаками

Ю.Д. Выборнова
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
vybornovamail@gmail.com

Д.И. Ульянов
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
dmitryulyanovhome@gmail.com

***Аннотация*—В работе предлагается новый метод защиты авторских прав на глубокие нейронные сети. Основная идея состоит во встраивании цифрового водяного знака в нейросетевую модель путем ее дообучения на уникальном наборе изображений-триггеров. Триггерная выборка формируется путем синтеза псевдоголографических изображений и их встраивания в растровые изображения оригинального датасета, используемого при обучении исходной модели. Результаты экспериментальных исследований подтверждают работоспособность предложенного метода, при этом процесс встраивания цифровых водяных знаков не влияет на точность защищаемой модели.**

Ключевые слова— CNN, защита авторских прав, псевдоголографические изображения, цифровые водяные знаки.

1. Введение

Учитывая растущий интерес к применению искусственного интеллекта, становятся все более актуальными вопросы безопасности при хранении и передаче моделей глубокого обучения. Злоумышленники могут распространять проприетарные модели или незаконно использовать их для предоставления услуг анализа данных. Следовательно, возникает необходимость создания новых методов защиты авторских прав на данный вид интеллектуальной собственности, позволяющих доказать правообладателю факт несанкционированного копирования и распространения предобученных моделей.

Методы встраивания цифровых водяных знаков (ЦВЗ) широко использовались в последние два десятилетия как средство защиты авторских прав на мультимедийные данные (изображения, видео и аудио). Общая идея встраивания ЦВЗ во внедрении в данные защитной информации, незаметной для пользователя, но предоставляющей техническую возможность отследить и доказать факт нарушения авторских прав.

2. Предлагаемый подход

Методы встраивания ЦВЗ по принципу Black-Box [1-3] заключаются в дообучении исходной модели на так называемой триггерной выборке таким образом, чтобы модель с ЦВЗ классифицировала изображения-триггеры согласно заданным меткам и чтобы при этом результаты были отличны от прогноза исходной модели. Например, можно дообучить защищаемую модель давать намеренно неверный результат на определенных изображениях (например, классифицировать заданные изображения птиц как самолеты). Процесс проверки авторских прав на модель глубокого обучения заключается в статистической оценке результатов классификации элементов триггерной выборки.

Стоит отметить, что внесение архитектурных изменений в модель (например, добавление в выходной слой нейрона дополнительных нейронов) крайне нежелательно, поскольку в случае предоставления доступа к модели сугубо в качестве сервиса, соответствующие выходы могут блокироваться на уровне программного интерфейса.

Предлагаемый метод защиты авторских прав на глубокие нейронные сети заключается в дообучении модели на уникальном наборе изображений-триггеров, сформированном путем синтеза псевдоголографических изображений (псевдоголограмм). Псевдоголограмма представляет собой двумерный шумоподобный сигнал, который кодирует определенную битовую последовательность длины l.

Для формирования триггерной выборки сгенерированные псевдоголограммы накладываются на растровые изображения исходного набора данных с помощью аддитивной стратегии встраивания ЦВЗ. Предлагается генерировать псевдоголограммы, кодирующие последовательности $S_1, S_2, \dots, S_i, \dots, S_k$, где k – количество классов. Метка класса изображения-триггера определяется не исходной меткой избражения, на которое наложена псевдоголограмма, а индексом i последовательности, которую эта псевдоголограмма кодирует.

3. Формирование триггерной выборки

Псевдоголографический сигнал формируется на основе синтеза комплексного спектра путем размещения импульсов на двумерной плоскости в спектральной области в зависимости от битов двоичной последовательности S_i. Таким образом, если выполнить обратное дискретное преобразование Фурье, то каждый бит последовательности будет «голографически» отображаться на результирующее изображение.

Правила расположения импульсов могут быть заданы по-разному, но следует отметить, что поскольку псевдоголограмма представляет собой вещественный сигнал, полуплоскости спектра должны быть симметричны. В данной работе импульсы располагаются на двух кольцах радиусов r и $r + \Delta r$, как описано в [4]. Значения радиусов задаются по следующему правилу: $r = 0{,}36 \times l,\ \Delta r = 6$.

Изображения-триггеры формируются согласно следующему алгоритму: пусть дана псевдоголограмма W размера $N \times N$. Изображение оригинального датасета преобразуется в цветовое пространство YCbCr и приводится к размеру псевдоголограммы $N \times N$. Тогда процесс встраивания сводится к аддитивному наложению

псевдоголограммы на Y-компоненту изображения с последующей нормализацией диапазона яркости:

$$\hat{Y}_{N\times N} = \|Y_{N\times N} + W \times q\|,$$

где q-коэффициент видимости ЦВЗ.

4. Набор данных

Для экспериментов был выбран датасет “Cats and Dogs” [5], представляющий собой 24998 изображений кошек и собак. С учётом разделения набора на тренировочную, тестовую и валидационную выборки, каждая из них была разделена в пропорции 9:1 (за исключением тестовой выборки) на два датасета. Больший датасет, состоящий из 15186 тренировочных, 5061 валидационных и 2250 тестовых изображений, использовался для подготовки моделей, в которые впоследствии встраивался ЦВЗ.

Перед процедурой встраивания ЦВЗ было сгенерировано 200 псевдоголограмм на основе двух последовательностей S_1 или S_2 длины l = 50 (по 100 для каждого класса). Встраивание ЦВЗ в модель осуществлялось на изображениях меньшего датасета. При этом оригинальным изображениям без наложенных псевдоголограмм соответствовали исходные метки «кошка»/«собака», а тем же изображениям, но с нанесенной псевдоголограммой, метки были назначены в зависимости от того, какая из последовательностей, S_1 или S_2, закодирована в этой псевдоголограмме. Таким образом, размер датасета составил: тренировочная (1688x2), валидационная (563x2) и верификационная (200). Стоит отметить, что верификационная выборка представляет собой исходные псевдоголограммы (не наложенные на другие изображения) и применяется для анализа доли верно классифицированных ЦВЗ.

Примеры исходных изображений и изображений-триггеров представлены на рисунках 1 и 2 соответственно.

Рис. 1. Оригинальные изображения из датасета «Cats And Dogs»

5. Экспериментальные исследования

Для демонстрации возможности встраивания ЦВЗ в нейросети-классификаторы согласно описанному методу, были взяты 3 модели: alexnet, vgg11 и resnet18. Каждая предварительно была обучена на тренировочной части на протяжении 30 эпох.

Параметры обучения: скорость обучения: $lr = 10^{-3}$; коэффициент видимости ЦВЗ $q = 1$; функция ошибки: кросс-энтропия; метод обновления весов: стохастический градиентный спуск с $momentum = 0{,}9$; размер батча: 64; число эпох: 30. Полученные результаты представлены в таблице 1. Согласно данным таблицы исходная модель не способна различать псевдоголограммы из верификационного набора, но после процедуры встраивания ЦВЗ в модель точность распознавания псевдоголограмм близка к единице.

Рис. 2. Примеры изображений с ЦВЗ при q = 1 и l = 50

Таблица I. Точность решения исходной задачи классификации и точность извлечения цвз

Набор данных	**Alexnet**	**VGG11**	**Resnet18**
Исходная модель, test set	0,968	0,987	0,988
Исходная модель, verification set	0,500	0,545	0,495
Модель с ЦВЗ, test set	0,968	0,988	0,983
Модель с ЦВЗ, verification set	0,995	1,0	0,980

6. Заключение

Результаты экспериментальных исследований подтверждают работоспособность предложенного метода, при этом процесс встраивания цифровых водяных знаков не влияет на точность модели.

Благодарности

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

Литература

[1] Rouhani, B.D. Deepsigns: A generic watermarking framework for ip protection of deep learning models / B.D. Rouhani, H. Chen, F. Koushanfar // ArXiv e-prints, 2018.

[2] Zhang, Y.-Q. DeepTrigger: A Watermarking Scheme of Deep Learning Models Based on Chaotic Automatic Data Annotation / Y.-Q. Zhang, Y.-R. Jia, X. Wang, Q. Niu, N.-D. Chen // IEEE Access. – 2020. – Vol. 8. – P. 213296-213305.

[3] Adi, Y. Turning your weakness into a strength: Watermarking deep neural networks by backdooring / Y. Adi, C. Baum, M. Cisse, B. Pinkas, J. Keshet // Proceedings of the 27th USENIX Security Symposium (USENIX Security 18). – 2018. – P. 1615-1631.

[4] Vybornova, Y. Method for Protection of Heterogeneous Data based on Pseudo-Holographic Watermarks / Y. Vybornova // 9th International Symposium on Digital Forensics and Security (ISDFS) . – 2021. – P. 1-5.

[5] Kaggle: Dogs vs. Cats [Electronic resource]. — Access mode: https://www.kaggle.com/c/dogs-vs-cats/data (11.02.2022).