ПОДТВЕРЖДЕНИЕ ПОДЛИННОСТИ ЦИФРОВЫХ ИЗОБРАЖЕНИЙ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Будко Максим Николаевич

УО «Гродненский государственный университет им. Я.Купалы», факультет математики и информатики, специальность «Компьютерная безопасность»,
Кафедра системного программирования и компьютерной безопасности.

Научный руководитель - Кадан Александр Михайлович, заведующий кафедрой системного программирования и компьютерной безопасности, кандидат технических наук, доцент

В связи с бурным развитием индустрии цифровых фотоаппаратов и повсеместным использованием камер смартфонов, цифровые изображения не только стали одним из главных форматов передачи информации, но и важной криминалистически значимой уликой. Поэтому актуальной становится задача подтверждения подлинности цифровых изображений во многих областях: информационной безопасности, страховой сфере, криминалистике и прочих.
Во введении указан объект исследования - определение подлинности цифровых фотографий.
Целью работы является исследование возможности применения методов машинного обучения в задачах классификации цифровых фотокамер (брэнд / модель) на основе сделанных ими фотографий на примере задачи подтверждения подлинности цифровых изображений в JPEG-формате.
В основной части проведен обзор существующих решений в области подтверждения цифровых изображений; рассмотрен формат хранения и передачи изображений JPEG; определена структура, основные и второстепенные маркеры формата JPEG.
Сформирована база подлинных цифровых изображений, сформирован набор признаков JPEG-файлов и проведено сравнение JPEG-форматов различных брендов цифровых фотокамер
Рассмотрен ряд методов машинного обучения, эффективность которых применительно к задаче определения подлинности цифровых фотографий исследована, а именно: метод k ближайших соседей, метод решающих деревьев, метод случайного леса, наивный байесовский классификатор.
Применение данных методов машинного обучения к выгруженным данным.
В заключении была исследована эффективность применения методов машинного обучения для определения подлинности цифровых изображений в задачах компьютерной технической экспертизы, разработан прототип системы для определения подлинности фото формата JPEG мобильных устройств на основе методов машинного обучения.
Областью возможного практического применения является проведение компьютерной технической экспертизы; использование полученных результатов в учебном процессе.

***Ключевые слова*:** программная система, методы машинного обучения, камера мобильного телефона, jpeg-формат

В современном мире информационные технологии получают широкое применение в любой сфере деятельности человека. С быстрым развитием зеркальных камер и частым использованием камер смартфонов, цифровые изображения стали одним из главных методов передачи визуальной информации. Изображения, представленные в цифровом виде, предназначены для дальнейшей обработки на компьютере или другой цифровой технике. Но в
связи с массовой доступностью инструментов для редактирования фотографий, оригинальность многих изображений ставится под сомнение. В наши дни существуют специалисты, которые могут зрительно определить, была ли фотография изменена, однако каждому человеку свойственно ошибаться, к тому же не

каждую измененную фотографию можно определить “на глаз”. Поэтому приходится применять более точные методы обнаружения редактирования изображения, например, машинное обучение, которое хорошо себя показало при решении задач классификации и предсказания в различных предметных областях. Выделяют два основных направления для оценки подлинности изображения. К первому относятся пиксельные алгоритмы, которые сравнивают рядом находящиеся пикселы изображения для поиска редактируемых областей. Второе направление основано на анализе данных, закодированных в графическом файле.

Формат JPEG

JPEG (Joint Photographic Experts Group) – популярный формат хранения и передачи с сети Интернет. Алгоритм JPEG в основном применяется для сжатия фотографий, содержащих реалистичные сцены и плавные переходы яркости и света. С другой стороны, он малоприменим для сжатия изображений с резким переходом контраста (текстовые знаки, чертежи), в таком случае используют другие форматы не использующие сжатие, такие как GIF, PNG или используют режим сжатия без потерь Lossless JPEG. Максимальный размер изображения составляет 64Kx64K пиксел. Файл JPEG обычно представлен в расширении .jpeg, .jpg, .gfif.

Большая часть маркеров начинает сегмент маркеров, состоящий из связанной группы параметров, другие же маркеры стоят в одиночку. Каждый параметр содержит либо 4 – бит, 1 байт, или 2 – байтовые коды. Такой код для параметра должен быть целым числом без знака указанной длины в битах с указанным значением параметра. Параметры длиной 4 бита должны стоять парами, а такие пары в закодированном виде образуют байт. Параметры длиной 2 байта содержат наиболее значимый байт на первом месте в упорядоченной последовательности байтов.

Формат JPEG состоит из нескольких сегментов, каждый из которых содержит информацию определенного типа, которая ограничена двумя байтами, называемыми маркерами. Каждый маркер начинается с байта-идентификатора FF 16 , сигнализирующем о начале маркера, и байта, указывающего на тип маркера. Важно отметить, что существуют маркеры, которые содержат дополнительные данные - длину информационной части маркера и самих данных.

Компьютерная криминалистика

«**Компьютерная криминалистика (форензика)** - прикладная наука о раскрытии преступлений, связанных с компьютерной информацией, об исследовании доказательств в виде компьютерной информации, методах поиска, получения и закрепления таких доказательств. Форензика является подразделом криминалистики» [10].

Отдельный раздел данной науки посвящен целостности изображений. «Для удостоверения целостности и неизменности данных используются однонаправленные хэш-функции. Эксперт, получив на исследование копию, подсчитывает с нее хэш-функцию. Если ее значение совпадает со значением, внесенным в протокол, эксперт и иные лица получают уверенность, что исследуемая копия совпадает с оригиналом с точностью до бита»[10].

Для определения целостности JPEG-изображений, можно использовать следующие признаки:

1) Exifданных изображения.

2) таблица Хаффмана, определенная производителем на основании характеристик устройства.

3) таблицы квантования, которые определены производителем.

4) размер изображения.

5) размеры эскиза, характерные для устройств определенной марки и модели. При этом эскиз имеет характерные для этой марки и модели эскизы таблицы Хаффмана и Квантования.

Затем по вышеназванным характеристикам подсчитывается хэш-функция и создается база хэшей. После чего каждая новая фотография сравнивается с тем, что есть в базе: если хэш нового изображения совпадает с хэшем в базе, это будет говорить о принадлежности данной фотографии определенному классу.

Формирование базы подлинных цифровых изображений

Данные изображений с мобильных телефонов были собраны на базе университета ГрГУ им. Янки Купалы. Благодаря чему база состоит из большого разнообразия фотографий различных производителей и моделей телефонов. Для каждого бренда мобильного телефона, был создан каталог, который включал в себя необходимый набор каталогов моделей смартфонов.

```
def getJPGpath(current, extension):
    import os
    lst_jpg = []
    for s in os.walk(current):
        print(s)
        print(s[1])
        if s[1] == []:
            #print("+++" + s[1])
            for photo_name in s[2]:
                print("---" + photo_name)
                if photo_name[-4:] in extension:
                    lst_jpg.append(s[0][2:]+'\\'+ photo_name)
    return lst_jpg
jpgPath = getJPG('.', ['.jpg', '.JPG'])
print ('\n'.join(jpgPath))
```

Листинг 1 – Формирование путей до каждого изображения

Составим набор маркеров для дальнейшего исследования изображений по этому набору

```
mark = (
    ('SOI','0xFFD8',0,'Начало изображения'),
    ('SOF0','0xFFC0',1,'Начало фрейма (базовый, ДКП)'),
    ('SOF1','0xFFC1',1,'Начало фрейма (расширенный, ДКП, код
Хаффмана)'),

    ...
    ('APPF','0xFFEF',1,'Задаётся приложением-F'),
    ('COM','0xFFFE',1,'Комментарий'),
    ('EOI','0xFFD9',0,'Конец закодированной части изображения')
    )
```

Листинг 2 – Набор маркеров JPEG-файла

Формирование набора признаков JPEG-файлов

Добавим в число признаков формата следующие характеристики:

(1-2) Название камеры, название модели;

(3) количество всех маркеров;

(4) количество маркеров начала 0xFFD8;

(5-6) длина и количество таблиц квантования 0xFFDB;

(7-8) длина и количество начала кадра, базового метода 0xFFC0;

(9-10) длина и количество таблиц кодов Хаффмана 0xFFC4;

(11-12) длина и количество начала закодированного изображения 0xFFDA.

Все используемы признаки являются основными маркерами JPEG-изображения. Сформируем указанный набор характеристик. В результате получим следующий набор признаков

```
[0, 'huawei#lyo-121', 30, 2, 0, 5, 524, 3, 290, 9, 1104, 2, 24]
[1, 'huawei#lyo-121', 30, 2, 0, 5, 524, 3, 290, 9, 1104, 2, 24]
[2, 'huawei#lyo-121', 30, 2, 0, 5, 524, 3, 290, 9, 1104, 2, 24]
...
[60, 'samsung#j5(2016)', 17, 2, 0, 2, 264, 3, 1057, 2, 836, 2, 24]
[61, 'samsung#j5(2016)', 17, 2, 0, 4, 268, 2, 34, 8, 848, 2, 24]
[62, 'samsung#j5(2016)', 17, 2, 0, 2, 264, 2, 34, 2, 836, 2, 24]
...
[124, 'xiomi#redmi note 5', 7, 2, 0, 2, 264, 2, 34, 2, 836, 2, 24]
[125, 'xiomi#redmi note 5', 7, 2, 0, 2, 264, 2, 34, 2, 836, 2, 24]
[126, 'xiomi#redmi note 5', 7, 2, 0, 2, 264, 2, 34, 2, 836, 2, 24]
```

Листинг 3 – Набор основных признаков для всех телефонов

Затем аналогично сформируем группы второстепенных характеристик.

Сравнение JPEG-форматов различных брендов цифровых фотокамер

Выберем из списка мобильных телефонов наиболее популярные модели, например: xiomi redmi 4, samsung galaxy s7, iphone 6 plus, huawei p20 lite

Сравним структуру JPEG-файлов для выбранных моделей (см рис. 2.1). В количестве маркеров для изображений данных смартфонов можно выделить как схожести, так и различия.

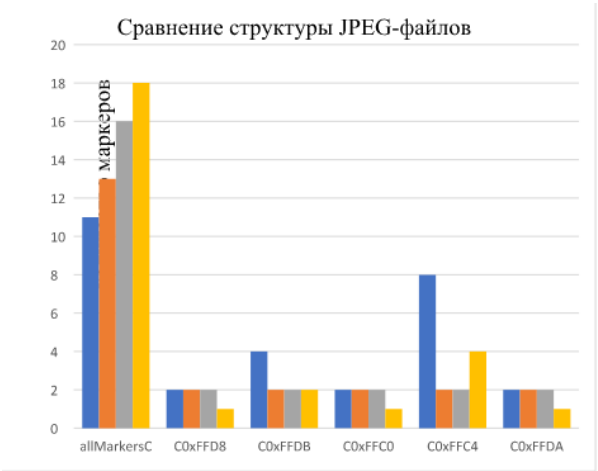

Рисунок 1 – Сравнение структуры JPEG-файлов различных брендов

В связи с полученными результатами и применяя методы машинного обучения, можно построить систему, которая будет осуществлять анализ изображений на основе базы подлинных изображений моделей смартфонов.

В результате была сформирована база, содержащая цифровые фотографии различных брендов и марок мобильных устройств. Она содержит следующие данные по каждому изображению:

1) Бренд и модель мобильного телефона;

2) Общее количество содержащихся в фотографии маркеров

3) Позиция маркеров в изображении и длина информации после них.

Для дальнейшей обработки предварительно выгрузим данные из базы. Получим названия признаков, название марки и модели телефона, признаки соответственно. Для каждого бренда и модели телефона вычислим хеш-функцию md5 и представим полученные данные в виде структуры данных словарь. После выгрузки данных из базы, нужно произвести обучение модели и сделать предсказание, для того чтобы понять, насколько эффективна данная модель для обработки новых данных. Применим процесс обучения модели и предсказания результата для следующих алгоритмов машинного обучения: дерево принятия решений, наивный байесовский классификатор, Метод k ближайших соседей и проведем сравнение полученных результатов.

Сравнение эффективности методов машинного обучения

В ходе проделанной работы, на тестовой выборке с основными и дополнительными маркерами формата JPEG, были протестированы алгоритм решающих деревьев, наивного байесовского классификатора и k-ближайших соседей. Исходя из полученных результатов, каждый из алгоритмов может использоваться для определения подлинности изображения, однако лучше всего себя показал метод решающих деревьев. Улучшения работы алгоритмов можно добиться, пополнив исходную базу изображений, сделанными на разные мобильные устройства, а так же выявив других характеристик классификации смартфонов.

Заключение

Файл JPEG содержит набор маркеров, каждый из которых начинается с байта 0xFF, который сигнализирует о начале маркера, и байта-идентификатора. Другие же маркеры состоят из дополнительных данных, состоящих из
двухбайтового поля с длинной информационной части маркера и самих данных.
У различных производителей фотокамер и смартфонов отличается тип
используемых маркеров, количество вхождений данных маркеров, а так длина
блока данных, связанного с таким маркером.

Все вышесказанное позволяет выдвинуть гипотезу, что по анализу структуры JPEG-формата возможно ответить на следующие вопросы:

- Является ли данное изображение оригинальным
- Определить бренд и модель смартфона, на который сделана фотография

На основе этих гипотез, в работе были исследованы особенности формата JPEG, связанные с особенностями программно-технической реализации мобильных устройств различных разработчиков, сформирована база цифровых изображений, а также выделены наборы характерных признаков JPEG-изображений. На полученной базе изображений были протестированы такие методы машинного обучения, как дерево принятия решений, наивный байесовский классификатор и метод k-ближайших соседей, и выбран наиболее эффективный из них. В результате было установлено, что лучший результат предсказания имеет модель, основанная на алгоритме дерева принятия решений. На основе тестовой выборки маркеров изображений, были обучены модели машинного обучения.

Литература

1. William B. JPEG:Still Image Data Compression Standard /William B., Pennebaker, Mitchell,Joan L. – Springer, 1993 – 638 с.

2. John M. Compressed Image File Formats: JPEG, PNG, GIF, XBM, BMP/John M. - Addison-Wesley Professional, 1999 – 288 с.

3. Декодирование JPEG для чайников [Электронный ресурс] / Декодирование JPEG для чайников / Хабрахабр. – Режим доступа: https://habrahabr.ru/post/102521/.–Дата доступа: 07.12.2018.

4. История создания, устройство, строение и применение графического формата JPEG[Электронный ресурс] / Историясоздания, устройство, строение и применение графического формата JPEG.– Режим доступа: http://kolpinkurs.ru/stati/jpeg.htm.–Дата доступа: 07.12.2018.

5. Метаданные в цифровой фотографии [Электронный ресурс] / Метаданные в цифровой фотографии. – Режим доступа: http://www.ixbt.com/digimage/metadxph.shtml.–Дата доступа: 08.12.2018.

6. Стандарт сжатия JPEG [Электронный ресурс] / Стандарт сжатия JPEG. | Научная библиотека. – Режим доступа: http://sernam.ru/cod_15.php – Дата доступа: 08.12.2018.

7. ISO/IEC 10918-1: 1993(E). Information technology – digital compression and coding of continuous-tone still images – requirements and guidelines [Электронныйресурс]. – Режимдоступа: https://www.w3.org/Graphics/JPEG/itu-t81.pdf. – Датадоступа: 08.12.2018.

8. Как распознать фейковые фото? [Электронный ресурс] / Как распознать фейковые фото? / BBCNews. – Режим доступа: https://www.bbc.com/ukrainian/vert-fut-russian-40486623. – Дата доступа: 09.12.2018.

9. Как определить монтаж на фото. Разоблачаем фейки, фотошоп и ретушь [Электронный ресурс] / Как определить монтаж на фото. Разоблачаем фейки, фотошоп и ретушь. – Режим доступа:https://vas3k.ru/blog/390/.– Дата доступа: 10.12.2018.

10. Федотов Н.Н. Форензика – компьютерная криминалистика – М.: Юридический Мир, 2007. – 432 с.: ил.

11. 13 онлайн-инструментов для проверки подлинности фотографий [Электронный ресурс] / 13 онлайн-инструментов для проверки подлинности фотографий. –Режим доступа: https://www.stopfake.org/13-onlajn-instrumentov-dlya-proverki-podlinnosti-fotografij/–Дата доступа: 09.12.2018.