УДК 004.056.57

ОБНАРУЖЕНИЕ СПУФИНГ-АТАК НА ОСНОВЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ И ЦВЕТОВЫХ ПРОСТРАНСТВ

Михалева Ирина Анатольевна, Салита Даниил Сергеевич

Алтайский государственный университет, Барнаул
i.mixalyova@yandex.ru, d.s.salita@gmail.com

DETECTION OF SPOOFING ATTACKS BASED ON CONVOLUTIONAL NEURAL NETWORKS AND COLOR SPACES

Mikhaleva IrinaA., Salita Daniil S.

Altai State University, Barnaul
i.mixalyova@yandex.ru , d.s.salita@gmail.com

Аннотация. В статье описана разработка метода обнаружения спуфинг-атак на основе сверточных нейронных сетей и цветовых пространств.

Были разработаны две архитектуры сверточных нейронных сетей (CNN), с применением различных цветовых пространств. Первая архитектура имеет один входной слой для одного цветового пространства. Вторая архитектура имеет два параллельных входных слоя для двух цветовых пространств. Были обучены следующие модели: RGB, CMYK, YCbCr, HSV, RGB+YCbCr, RGB+HSV, RGB+CMYK, YCbCr+CMYK, YCbCr+HSV, HSV+CMYK. Модели обучались на двух датасетах: OpenForensics Dataset и Dataset Nvidia & StyleGAN. Было выявлено, что все разработанные модели показывают достаточно высокие результаты в обнаружении сгенерированных изображений. Результаты работы могут быть применены для обнаружения сгенерированных изображений с помощью сверточных нейронных сетей и цветовых пространств.

Ключевые слова: атака на представление, CNN, обнаружение сгенерированных изображений, цветовые пространства, генерация изображений

Abstract. The article describes the development of a method for detecting spoofing attacks based on convolutional neural networks and color spaces.

Two convolutional neural network (CNN) architectures were developed, using different color spaces. The first architecture has one input layer for one color space. The second architecture has two parallel input layers for two color spaces. The following models were trained: RGB, CMYK, YCbCr, HSV, RGB+YCbCr, RGB+HSV, RGB+CMYK, YCbCr+CMYK, YCbCr+HSV, HSV+CMYK. The models were trained on two datasets: OpenForensics Dataset and Dataset Nvidia & StyleGAN. It was found that all developed models show fairly high results in detecting generated images. The results of the work can be applied to detect generated images using convolutional neural networks and color spaces.

Keywords: representation attack, CNN, detection of generated images, color spaces, image generation

Для цитирования: *Михалева И.А., Салита Д.С.* Обнаружение спуфинг-атак на основе сверточных нейронных сетей и цветовых пространств // Проблемы правовой и технической защиты информации. 2025. No13. С. 42–49.

For citation: *Mikhaleva I.A., Salita D.S.* Detection of spoofing attacks based on convolutional neural networks and color spaces. *Legal and Technical Problems ofInformation Security*. 2025. No. 13. P. 42–49.

С развитием различных технологий искусственного интеллекта нейронные сети значительно продвинулись вперед, позволяя создавать реалистичные сгенерированные изображения. Эти технологии находят применение в различных областях, включая искусство, развлечения и медиа. Однако с развитием этих методов возникает новая проблема — обнаружение сгенерированных изображений, что имеет критическое значение для обеспечения информационной безопасности. Сгенерированные изображения используют как метод кибератаки для дальнейшего проникновения в системы идентификации или аутентификации пользователей с целью получения несанкционированного доступа к охраняемой информации. Одним из методов обнаружения сгенерированных изображений являются сверточные нейронные сети (CNN).

Цветовые пространства играют ключевую роль в детектировании сгенерированного изображения. С помощью преобразования изображения в различные цветовые пространства можно выявить характерные признаки, указывающие на то, является изображение реальным или же сгенерированным [1]. Цветовое пространство RGB представляет собой аддитивную цветовую модель, в которой основные цвета красного (R), зеленого (G) и синего (B) цвета складываются вместе различными способами для воспроизведения широкого спектра цветов.

Сверточные нейронные сети, обученные на RGB, могут выявлять локальные несоответствия в цветовых каналах (например, шум, артефакты сжатия). RGB отлично подходит для обнаружения клонирования (copy-move) и шумовых аномалий.

Для определения пикселей цвета кожи на изображении в цветовом пространстве RGB можно использовать нормализованную цветовую гистограмму, которая настраивается с учетом изменений яркости для обеспечения четкости. Для сегментации кожи в пространстве RGB используют следующие значения:

$$R > 95, G > 40, B > 20, \# \quad (1)$$

$$(\max(R, G, B) - \min(R, G, B)) < 15, \# \quad (2)$$

$$|R - G| > 15, R > G, R > B, \# \quad (3)$$

где R, G, B — значения красного, зеленого и синего каналов соответственно.

Пороговые значения подобраны так, чтобы выделять оттенки, характерные для человеческой кожи, и отсеивать большинство фонов [2].

Объяснение выбранных значений:

– с помощью значений в формуле (1) отсекаются слишком темные и не характерные для кожи цвета;

– с помощью формулы (2) производится исключение серых пикселей, то есть между компонентами должен быть достаточный разброс;

– в формуле (3) используются следующие неравенства, так как для большинства оттенков кожи красный цвет является доминирующим.

Также для детектирования кожи на изображении используют следующие формулы:

$$R > 220, G > 210, B > 70, \# \quad (4)$$

$$|R - G| \leq 15, B < R, B < G. \# \quad (5)$$

Эти формулы предназначены для выделения на фотографии только очень светлой кожи.

Цветовое пространство CMYK — это субтрактивная цветовая модель, используемая в технологии цветной печати. Координаты цвета в этой системе можно описать следующим образом: голубой (Cian), пурпурный (Magenta), желтый (Yellow) и черный (Key color, blacK). В этой модели интенсивность каждого цвета задается в процентах от 0 до 100, где 0 — это отсутствие цвета, 100 — это максимальное значение его интенсивности.

Цветовое пространство CMYK может быть использовано для идентификации кожи на изображении. Пурпурный и желтый цвета могут быть использованы, для описания цвета кожи. Однако, для успешной идентификации реального изображения лица человека, использование только цветового пространства CMYK может быть недостаточным. Цветовое пространство CMYK может быть полезно для анализа печатных подделок, так как отражает особенности полиграфии, может выявлять артефакты, связанные с преобразованием из RGB в CMYK (например, неестественное распределение чернил). Компоненты значений цветовой модели CMYK, которые следует использовать для обнаружения кожи [3]:

$$K < 205, \# \quad (6)$$

$$0 \le C \le 0{,}05, \# \quad (7)$$

$$0{,}089 < Y < 1, \# \quad (8)$$

$$0 \le \frac{C}{Y} < 1, \# \quad (9)$$

$$0 \le \frac{Y}{M} < 4{,}8, \# \quad (10)$$

Обоснование выбора значений:

– черная компонента не должна быть слишком высокой (кожа не бывает слишком темной по этому параметру);

– голубой компонент очень низкий — кожа практически не содержит голубого оттенка;

– желтая компонента должна быть выражена, но не минимальна и не максимальна;

– отношение голубого к желтому должно быть меньше 1 (голубого всегда меньше, чем желтого);

– отношение желтого к пурпурному находится в определенных границах, что отражает баланс между этими цветами в оттенках кожи.

Цветовое пространство YCbCr используется для кодирования и передачи изображений в цифровом виде. Цветовая модель YCbCr делит изображения на яркостные (Y-каналы) и цветные (Cb и Cr-каналы), где Cb — это компонент цветоразностности синего цвета, а Cr — компонент цветоразностности красного цвета. YCbCr представляет собой способ преобразования и кодирования изображений из RGB. Использование YCbCr применяется в JPEG-сжатии, поэтому помогает выявлять артефакты повторного сжатия.

Применение цветового пространства YCbCr может быть полезным в обнаружении сгенерированных изображений [4]. В сгенерированных изображениях часто бывает, что распределение яркости может быть не естественным. Например, сгенерированные изображения могут иметь некоторые аномалии в тенях и светах изображения. Все это отражается в текстуре канала Y. Связано это с тем, что GAN сети иногда создают слишком ровные, или наоборот, слишком резкие переходы яркости в изображении. На таких изображениях отсутствуют естественные световые блики и тени. Помимо этого, в цветовых каналах Cb и Cr могут наблюдаться слишком однородные, гладкие, без шумов области или наоборот, в некоторых областях изображения показатель шума будет иметь запредельное значение.

Кроме того, в реальных изображениях яркость и цветность тесно связаны между собой определенными закономерностями,

отражающими признаки реального изображения. В сгенерированных же изображениях наблюдается не типичное распределение значение между каналами Y, Cb и Cr. Иными словами, Разделение яркости (Y) и цветности (Cb, Cr) упрощает обнаружение размытия, клонирования и вставок в сгенерированных изображениях.

Для определения принадлежности пикселя к коже используются пороговые значения компонент Cb и Cr. Пиксель считается относящимся к коже, если значения Cb и Cr попадают в определенный диапазон [2]:

$$77 \leq Cb \leq 127, \# \quad (11)$$
$$133 \leq Cr \leq 173. \# \quad (12)$$

Данные границы позволяют выделить цвет кожи с высокой точностью, при этом исключая фон и объекты с цветами, нехарактерными для кожи.

Цветовое пространство HSV — это модель, которая описывает изображения с помощью трех параметров: тон (hue), насыщенность (saturation) и яркость (value). В сгенерированных изображениях будет наблюдаться дисбаланс между насыщенностью и яркостью [5]. Также с помощью HSV можно обнаружить кожу на фотографии. Для этого средний пиксель в области кожи должен удовлетворять следующим условиям [2]:

$$0° \leq H \leq 25°, \# \quad (13)$$
$$0{,}2 \leq S \leq 0{,}6, \# \quad (14)$$
$$V \geq 40, \# \quad (15)$$

где H — тон, S — насыщенность. V — яркость. Если пиксель на фотографии попадает в указанные диапазоны, то он считается кожей.

Объяснение выбранных значений:

– в диапазоне от 0 до 25 градусов представлены оттенки от красного до желтого, что является характерными оттенками для кожи;

– с помощью выбранного диапазона насыщенности от 0,2 до 0,6 исключаются слишком бледные и слишком насыщенные цвета, не характерные для цвета человеческой кожи;

– с помощью установленной яркости в формуле (15) исключаются слишком темные области, которые могут являться фоном или одеждой на изображении.

Преобразование между цветовыми моделями HSV и RGB является нелинейным. Значение оттенка (H) может быть неточным для распознавания, если интенсивность в цветовой модели HSV низкая.

Для обнаружения сгенерированных изображений была разработана архитектура сверточной нейронной сети (CNN). На основе разработанной архитектуры были созданы 4 модели сверточных нейронных сетей, использующих 4 цветовых пространства RGB, CMYK, YCbCr и HSV соответственно. Все эти модели имеют схожую архитектуру.

Архитектура модели состоит из следующих компонентов:

1) входной слой принимает изображения размером 256x256 пикселей, и количеством каналов, в зависимости от цветовой модели (4 канала для цветового пространства CMYK, 3 канала для цветовых пространств RGB, YCbCr и HSV);

2) сверточные слои. Первый сверточный блок состоит из двух последовательных сверточных слоев с 64 фильтрами размером (3x3), с использованием padding='same' для сохранения размерности карты признаков. Второй сверточный блок имеет два последовательных сверточных слоя с 128 фильтрами размером (3x3). Третий сверточный блок имеет два последовательных сверточных слоя с 256 фильтрами размером (3x3). Четвертый сверточный блок имеет два последовательных сверточных слоя с 512 фильтрами размером (3x3). Во всех сверточных слоях используется функция активации ReLU [6,7] и BatchNormalization;

3) пулинговый слой [8]. В каждом сверточном слое используется MaxPooling2D с окном 2x2;

4) полносвязаные слои. Используется два полносвязаных слоя с функцией активации ReLu, Dropout с значением 0.3 и L2-регуляризацией с коэффициентом 0.001;

5) выходной слой. Используется один полносвязный слой с функцией активации sigmoid.

Для оптимизации модели используется функция потерь и оптимизатор Adam с шагом обучения 0,0001. Визуализация архитектуры представлена на рисунке 1.

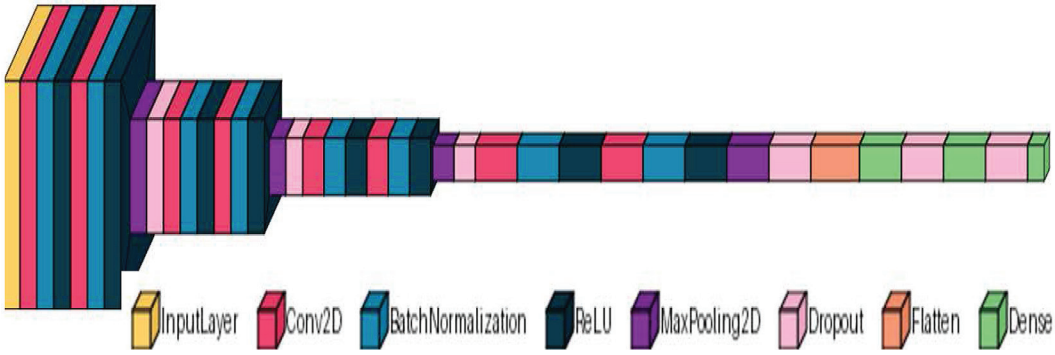

Рисунок 1. Визуализация архитектуры сверточной нейронной сети с одним цветовым пространством

Обучение каждой модели производилось в течение 10 эпох, процесс обучения одной модели занял около 5 часов. Для того чтобы улучшить способность обнаружения сгенерированных изображений, была разработана архитектура сверточной нейронной сети с использованием различных комбинаций двух цветовых моделей с использованием общего полноразмерного слоя.

Архитектура нейронной сети представляет собой модель, которая состоит из двух параллельных ветвей [9], где каждая ветвь обрабатывает изображения в определенном цветовом пространстве. Затем происходит объединение этих ветвей с использованием общего полноразмерного слоя. Разработанная архитектура имеет два параллельных входных слоя. Ветви принимают изображения размерами 128x128 пикселей и с 3 или 4 каналами, в зависимости от используемой цветовой модели. Архитектуры двух ветвей аналогичны друг другу и состоят из следующих компонентов:

– сверточные слои. Имеются четыре сверточных былока: с 64 фильтрами, 128 фильтрами, 256 фильтрами и 512 фильтрами размерами 3x3. Все сверточные слои используют функцию активации ReLU и padding='same' для сохранения размерности карты признаков;

– нормализация слоя. После каждого сверточного слоя применяется Batch Normalization для стабилизации и ускорения обучения;

– пулинговый слой. Затем после Batch Normalization используется MaxPooling2D с окном 2x2 для уменьшения размерности изображения;

– полносвязные слои. Используется один полносвязный слой с 256 нейронами с функцией активации ReLu и Dropout с значением 0.5.

Далее с помощью слоя concatenate происходит объединение с выходов двух ветвей. Объединенные признаки проходят через два полноразмерных слоя с 512 нейронами и функцией активации ReLU, с последующим Dropout с коэффициентом 0,3 и L2-регуляризацией с коэффициентом 0,001. После используется выходной слой с функцией активации sigmoid. Для оптимизации используется функция потерь binary crossentropy, оптимизатор Adam с шагом обучения 0,0001.

Визуализация архитектуры представлена на рисунке 2. Из рисунка видно, что модель имеет два входа для двух цветовых пространств (слои InputLayer). Также на рисунке изображен момент объединения двух ветвей (слой Concatente), после которого следует общий полноразмерный слой (Dense).

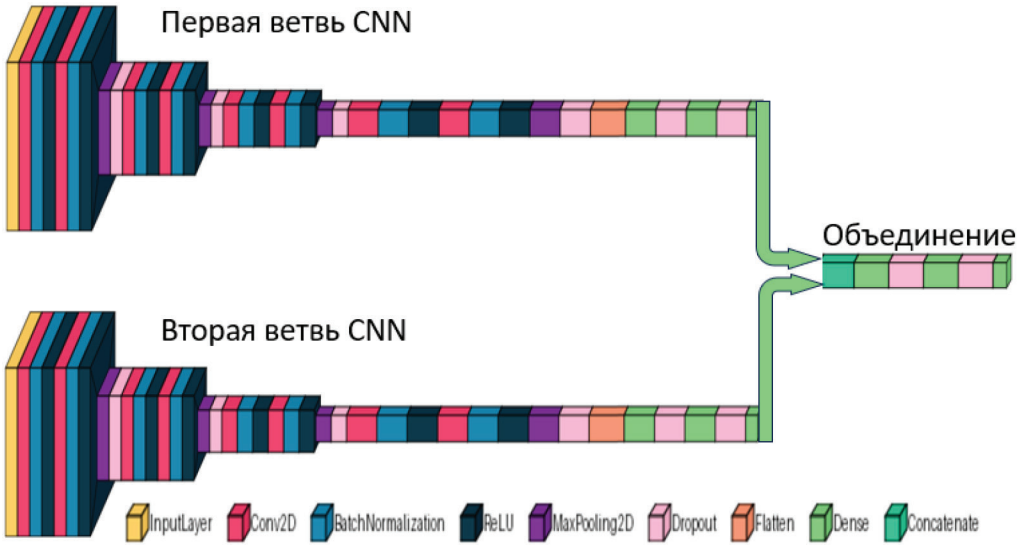

Рисунок 2. Визуализация архитектуры с двумя цветовыми моделями

Для обучения сверточной нейронной сети был использован OpenForensics Dataset. Этот набор данных разделен на тестовую и обучающую выборки. Общее количество изображений в датасете составляет 150904 изображения. Количество изображений в обучающей выборке составляет 140000, в тестовой — 10904. Датасет состоит из реальных и сгенерированных изображений человеческих лиц. Сгенерированные изображения были созданы следующими нейронными сетями: StyleGan3, StyleGan2, BigGan и диффузионными моделями. Размер изображений: 256x256 пикселей. Изображения представлены в цветовом пространстве RGB.

На этом датасете были обучены следующие модели: RGB, CMYK, YCbCr, HSV. А также модели, использующие два цветовых пространства: RGB+YCbCr, RGB+HSV, RGB+CMYK, YCbCr+CMYK, YCbCr+HSV, HSV+CMYK. В таблице представлены значения AUC [10] в результате построения ROC-кривой на тестовой выборке для всех обученных моделей с одним цветовым пространством.

Матрица значений AUC для всех обученных моделей

	RGB	CMYK	YCbCr	HSV
RGB	AUC=0,94	AUC=0,98	AUC=0,98	AUC=0,98
CMYK	AUC=0,98	AUC=0,91	AUC=0,97	AUC=0,98
YCbCr	AUC=0,98	AUC=0,97	AUC=0,95	AUC=0,98
HSV	AUC=0,98	AUC=0,98	AUC=0,98	AUC=0,96

Исходя из значений в таблице можно сделать вывод, что обученные модели с применением нескольких цветовых пространств имеют преимущество в определении сгенерированных изображений, в сравнении с моделями, где используется только одно цветовое пространство.

Также для одной из моделей была создана тепловая карта Grad-CAM (рис. 3).

Рисунок 3. Тепловая карта Grad-CAM сгенерированного изображения

Из изображения на рисунке 3 видно, что обученная модель верно выделяет лицо как сгенерированное (красные и оранжевые области). Модель уделяет значительное внимание глазам и рту, что может указывать на то, что эти области содержат ключевые признаки, которые помогают отличить сгенерированное изображение от реального. Области вокруг контура лица и текстуры кожи также выделяются, что может означать, что модель анализирует детали, такие как текстура кожи и четкость контуров, которые могут отличаться на сгенерированных изображениях. Области с меньшей интенсивностью цвета (синие и зеленые) указывают на части изображения, которые модель считает менее важными. В данном случае, это фон или менее детализированные части изображения.

Далее все эти модели были обучены на еще одном датасете — Dataset Nvidia & StyleGAN. Этот датасет насчитывает 140000 изображений, где 70000 реальных лиц — фотографии из Flickr, собранные Nvidia (реальные люди), 70000 сгенерированных лиц — изображения, созданные с помощью генеративной модели StyleGAN (GAN-сгенерированные лица). На рисунке 4 представлены значения AUC в результате построения ROC-кривой на тестовой выборке на новом датасете.

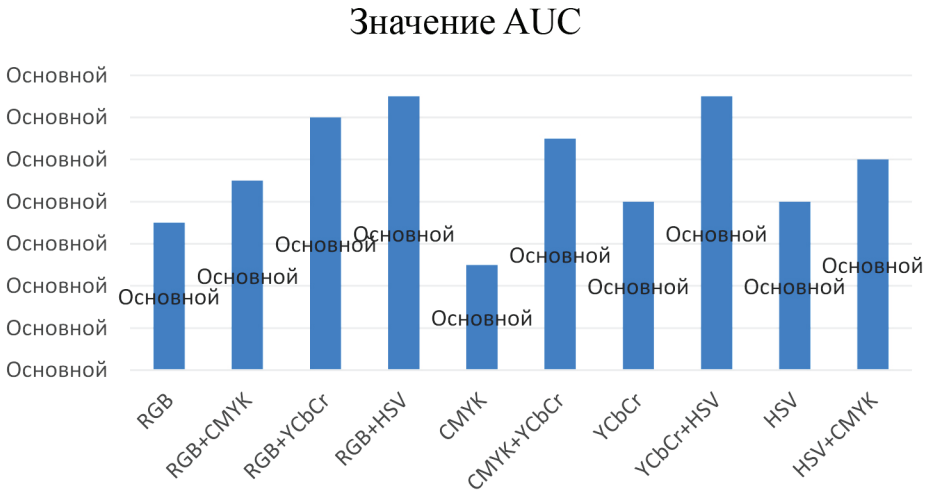

Рисунок 4. Результаты вычисления ROC-кривой для моделей с улучшенной архитектурой на датасете Dataset Nvidia & StyleGAN

Таким образом, объединяя результаты всех проведенных экспериментов, можно сделать вывод, что модели, обученные на датасете OpenForensics Dataset, имеют более высокий процент обнаружения сгенерированных изображений в сравнении с моделями, обученными на датасете Dataset Nvidia & StyleGAN. Кроме того, модели, использующие два входных слоя, успешнее справляются с детектированием сгенерированных изображений, по сравнению с моделями, которые используют одно цветовое пространство (RGB, CMYK, YCbCr, HSV). С помощью демонстрации тепловой карты, были определены ключевые области, на которых фокусируется нейронная сеть при обнаружении сгенерированных изображений. Этими областями являются: глаза, рот, контур лица и текстура кожи. Комбинация цветовых пространств позволяет повысить точность обнаружения сгенерированных изображений.

Библиографический список

1. Mo S., Lu P., Liu X. AI-Generated Face Image Identification with Different Color Space Channel Combinations // Sensors. 2022. Vol. 22(21). P. 8228.

2. Ennehar B.C., Brahim O., Hicham T. An Appropriate Color Space to Improve Human Skin Detection // INFOCOMP Journal of Computer Science. 2010. Vol. 9(4). P. 1–10.

3. Dariusz J.S., W. Miziolek. Human colour skin detection in CMYK colour space // IET Image Processing. 2015. Vol. 9. P. 751–757.

4. Manjare S., Chougule S.R. Skin Detection for Face Recognition Based on HSV Color Space // International Journal of Engineering Sciences & Research Technology. 2013. Vol. 2(7).

5. Khamar Basha Shaika, Ganesan P., Kalist V., Sathish B.S., Merlin J. Mary Jenitha. Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space // Procedia Computer Science 57. 2015. P. 41–48.

6. Dubey S.R., S.K Singh, B.B Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark // Neurocomputing. 2022. Vol. 503. P. 92–108.

7. Szandała T. Review and comparison of commonly used activation functions for deep neural networks // Bio-inspired neurocomputing. 2021. Vol. 1. P.203–224.

8. Bhatt D., C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya and H. Ghayvat. CNN variants for computer vision: History, architecture, application, challenges and future scope // Electronics. 2021. Vol. 10. N. 20. P. 2470.

9. Balamurali K., Chandru S., Razvi M.S., Sathiesh V. Kumar. Face Spoof Detection Using VGG-Face Architecture // Journal of Physics: Conference Series. 2021. Vol. 1917. N. 1. P. 12010.

10. Yang Z, Xu Q, Bao S, Cao X, Huang Q. Learning with multiclass AUC: Theory and algorithms // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. Vol. 44. P. 7747–7763.