ДЕТЕКТИРОВАНИЕ ДЕСТРУКТИВНОГО МУЛЬТИМЕДИА КОНТЕНТА В СОЦИО-КИБЕРФИЗИЧЕСКОЙ СИСТЕМЕ МОНИТОРИНГА СЕТИ ИНТЕРНЕТ

Исхакова А. О.[1], Русаков К. Д.[2], Исхаков А. Ю.[3], Мамченко М. В.[4]

(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Рассматривается проблема мониторинга и фильтрации мультимедиа контента в сети Интернет в части автоматизированного детектирования сцен насилия и агрессии. Задачи обеспечения безопасности индивидуального, группового и массового сознания, включая защиту от деструктивных информационных воздействий и от использования информационных технологий для пропаганды терроризма, являются одним из основных направлений научных исследований в области обеспечения информационной безопасности Российской Федерации. В докладе приводится обзор существующих научно-технических решений в данной области и предлагается подход к детектированию целевого контента. Модель состоит из легких сверточных нейронных сетей для извлечения признаков и модели GRU для кодирования изменений кадров, характеризующих сцены насилия, существующие в видео. Исследованная архитектура имеет потенциал применения на мобильных платформах с низкими вычислительными возможностями, например, Nvidia Jetson Nano. Предложенное методическое обеспечение может быть использовано в реализации модулей распознавания сцен насилия и агрессии в видеопотоке для комплексных социо-киберфизических систем мониторинга.

Ключевые слова: анализ видеопотока, агрессия, насилие, социо-киберфизическая система, искуственная нейронная сеть.

1. Введение

Одним из важных прикладных направлений интеллектуального анализа данных видеопотока является задача автоматического обнаружения насильственных действий и агрессивного

[1] *Анастасия Олеговна Исхакова, к.т.н., с.н.с.* (*iao@ipu.ru*).

[2] *Константин Дмитриевич Русаков, н.с.* (*rusakov.msk@yandex.ru*).

[3] *Андрей Юнусович Исхаков, к.т.н., с.н.с.* (*iay@ipu.ru*).

[4] *Марк Владиславович Марченко, н.с.* (*markmamcha@gmail.com*).

поведения участников. В условиях повсеместного распространения интернет-технологий важной задачей является обеспечение эффективной контент-фильтрации, в том числе оперативного детектирования различных противоправных сцен, содержащих элементы жестокого обращения, убийств и т.д. как в условиях загрузки в социальных сетях, сервисах медиа-хостинга, так и при стриминге (потоковом вещании без сохранения на носители информации). Кроме того, сегодня в общественных местах в рамках проекта «Безопасный город» внедрен внушительный объем средств фото и видеофиксации, требующих разработки средств автоматизированного обнаружения правонарушений из огромных массивов видеоданных, полученных с камер наблюдения. Вышеперечисленные примеры свидетельствуют о высокой степени актуальности развития прикладных систем аналитики для детектирования деструктивного видео-контента.

Большинство известных методов и алгоритмов в данной области направлены на выявление определенных признаков интернет-контента или классификацию/характеристику пользователя. При этом в большей части исследования носят прикладной характер и не находят научного подтверждения эффективности предлагаемых методов. Для решения различных задач, включая аннотирование видео, поиск видео и мониторинг в реальном времени, наиболее сложной составляющей является обнаружение насильственных действий в видеофрагментах. Эта задача включает в себя множество смежных методов компьютерного зрения, в том числе, обнаружение объектов, распознавание действий и классификацию.

В рамках выполняемого проекта по разработке социокиберфизической системы мониторинга разнородного интернет-контента в целях противодействия проявлению агрессии, давления и других форм деструктивного воздействия на индивидуальное и групповое сознание пользователей возникла потребность в реализации модуля детектирования элементов агрессивного поведения и насилия субъектов. Ниже приводится краткое описание используемого алгоритмического обеспечения, применяемого при разработке данного модуля.

2. Анализ текущего состояния исследований

Проблема повышения эффективности алгоритмического обеспечения классификаторов, предоставляющих возможность детектирования агрессивного поведения и сцен насилия, находит свое отражение в работах множества ученых в области машинного обучения, а также разработчиков программного обеспечения видеонаблюдения и видеоаналитики.

В работе [1] предлагается архитектура глубокой нейронной сети для распознавания насильственных видео. Авторы используют сверточную нейронную сеть для извлечения признаков на уровне кадра видео. Далее значения полученных характеристик агрегируются с помощью варианта модуля долговременной краткосрочной памяти. Предложенная архитектура способна улавливать локализованные пространственно-временные особенности, что позволяет анализировать локальное движение, происходящее в видео. Подход авторов предлагает использовать разницу между соседними кадрами в качестве входных данных для модели, тем самым заставляя ее кодировать изменения, происходящие в видео.

Исследование [2] содержит обзор стратегий для определения значимости признаков из различных предварительно обученных моделей для обнаружения насилия в видео. Был создан набор данных, который состоит из видеороликов с насилием и без насилия в различных условиях. В качестве исходных датасетов использованы размеченные фрагменты видеофильмов. Авторы апробировали наиболее известные модели ImageNet; VGG16, VGG19, ResNet50 для извлечения признаков из кадров видео. Согласно представленным в статье результатам, модель ResNet50 оказалась наиболее эффективной в задаче детектирования сцен насилия. Полученные признаки в сочетании с моделью долгой краткосрочной памятью (LSTM) обеспечивают точность 97,06%, что является результатом, превосходящим другие модели, предложенные к тестированию.

Были рассмотрены схожие исследования авторов Kyunghyun, Hochreiter, Junyoung, Gruber, Simonyan, He, Misra и др. На основании результатов апробации, опубликованных

вышеперечисленными группами ученых, была предпринята попытка провести исследование нейросетевой архитектуры управляемого рекуррентного блока GRU [3] для классификации видео на насильственное и ненасильственные. GRU похожа на модель долгой краткосрочной памяти с логическим элементом забывания [4], но имеет меньше параметров, чем LSTM. Производительность GRU по определенным задачам моделирования полифонической музыки, моделирования речевых сигналов и обработки естественного языка аналогична LSTM. Поскольку авторы [5, 6] показали, что GRU демонстрируют лучшую производительность на некоторых наборах данных, то было решено исследовать данную архитектуру в решении задачи распознавания деструктивного поведения.

3. Предлагаемый подход

Блок-схема предложенного подхода показана на рис. 1. Модель состоит из легких сверточных нейронных сетей для извлечения признаков и модели GRU для кодирования изменений кадров, характеризующих сцены насилия, существующие в видео.

Чтобы система могла определить, происходит ли насилие между людьми, присутствующими на видео, она должна быть способна определять местонахождение людей и понимать, как движения этих людей меняются со временем. Сверточные нейронные сети способны получать представление каждого видеокадра как набора некоторых признаков, для кодирования временных изменений которых требуется рекуррентная нейронная сеть. Так как в данной задаче интерес вызывают изменения как в пространственном, так и во временном измерении, в качестве рекуррентной нейронной сети использовалась модель GRU.

Чтобы система могла идентифицировать видео как насильственное или ненасильственное, она должна быть способна кодировать особенности каждого кадра, а также изменение этих кадров во времени. Выход линейных слоев после каждой сверточной нейронной сети представляет собой глобальный дескриптор (эмбеддинг) всего изображения.

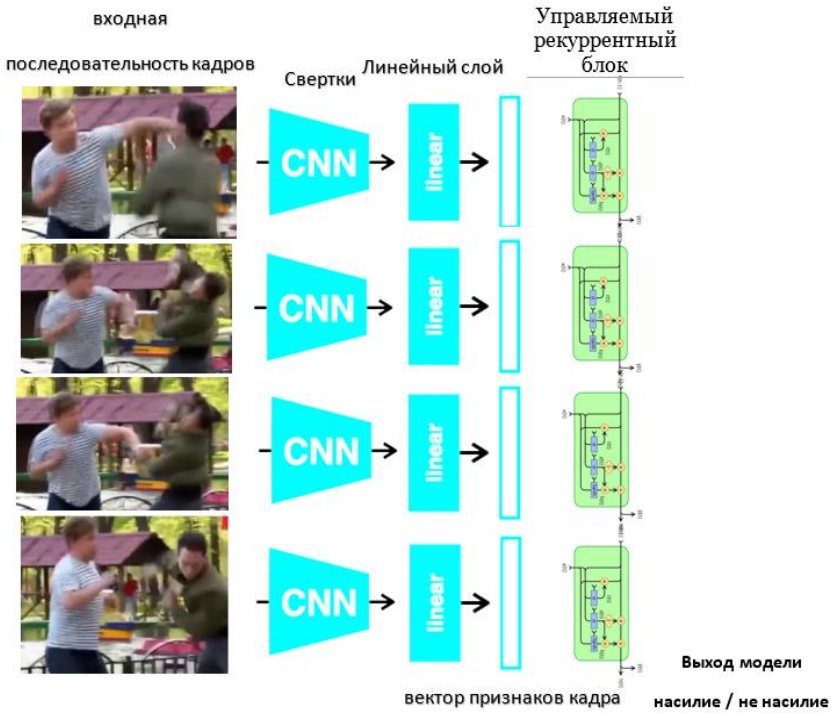

Рис. 1. Архитектура предлагаемой модели

Уравнения модели GRU представлены в выражениях (1), (2), (3):

(1) $z_t = \sigma_g(W_z x_t + U_z h_{t-1} + b_z)$,

(2) $r_t = \sigma_g(W_r x_t + U_r h_{t-1} + b_r)$,

(3) $h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \sigma_g(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h)$,

где x_t – входной вектор; h_t – выходной вектор; z_t – вектор вентиля обновления; r_t – вектор вентиля сброса; W, U, b – матрицы параметров; σ – функции активации.

Сеть работает следующим образом: 10 кадров, идущих последовательно, применяются к модели, при этом под каждым кадром понимается разница между текущим кадром и следую-

щим. Таким образом, сеть вынуждена моделировать изменения, происходящие в соседних кадрах, а не в самих кадрах. Как показано в работе [7], разностное изображение можно рассматривать как грубую и приблизительную версию изображений оптического потока. Таким образом, в предлагаемом методе разница между соседними видеокадрами применяется как вход в сеть. В результате исключается вычислительная сложность, связанная с генерацией изображения оптического потока. После того как все кадры применены, скрытое состояние слоя GRU на этом последнем временном шаге содержит представление примененных входных видеокадров. Это видеопредставление в скрытом состоянии GRU затем применяется к серии полносвязных слоев для классификации. В предложенной подходе в качестве модели CNN для извлечения характеристик уровня кадра была использована модель ResNet34 [8], предварительно обученная в базе данных ImageNet. Проведенные исследования показали, что сети, обученные в базе данных ImageNet, способны к лучшему обобщению и приводят к повышению производительности для таких задач, как распознавание действий [9]. В сверточной сети слой пакетной нормализации добавляется перед линейным слоем, а также после каждого сверточного и линейного слоя применяется нелинейная активация ReLU. Нелинейная активация применяется после каждого сверточного и полносвязного слоев. В сети, вместо того чтобы применять входные кадры как таковые, в качестве входных данных выдается разница между соседними кадрами. Сеть обучена минимизировать потери двоичной кросс-энтропии.

Набор данных

В настоящей работе был использован набор данных, представляющих собой видеофрагменты уличных бунтов и политических беспорядков, полученных из общедоступных ресурсов YouTube, Facebook, Twitter, а также открытых групп месенджера Telegram. Стоит отметить, что полученные видеофрагменты не унифицированы, вследствие чего имеют разное разрешение, длительность и др. параметры. Было собрано 230 видеороликов для каждого класса, т.е. общий объем выборки 460 видеороликов. Поскольку моменты деструктивного поведения на самих

видеороликах занимают меньшую часть времени, во время обучения было принято решение применять балансировку данных во избежание переобучения. Собранный набор данных был разделен на обучающую и тестовую части в соотношении 90% к 10% соответственно. Сформированный датасет в настоящее время дополняется специалистами и вскоре планируется публикация его расширенной версии, что позволит применять его другим исследователям при решении смежных задач.

Результаты эксперимента

В ходе реализации программного решения была использована библиотека глубокого обучения Pytorch для обучения моделей. Обучение проводилось на графическом процессоре NVIDIA Tesla V100 32 Gb. Результаты обучения и валидации модели показаны в таблице 1.

Таблица 1. Результаты обучения и валидации

Эпоха	Точность	
	Обучение	Валидация
1	85,565	80,766
5	91,542	90,756
10	92,785	91,943
20	93,562	93,256
50	94,158	92,549
100	95,489	91,563

Представленные в таблице 1 результаты свидетельствуют о том, что точность модели на валидации начинает падать после 20 эпохи, что говорит о переобучении модели. В результате было принято решение использовать 20 эпох в качестве эталона для данной модели. Неустойчивая точность тестирования и потери при тестировании являются признаками того, что модель не подходит для решения данной задачи. Однако увеличение набора обучающих данных может решить проблему обобщения. Данная гипотеза будет проверена в дальнейшей работе при разработке модуля детектирования деструктивного контента.

В таблице 2 показан сравнительный анализ обученной модели с другими рекуррентными моделями, полученными у различных авторов:

Таблица 2. Сравнительный анализ моделей

Модель	Точность	Количество параметров модели
LSTM	94,6	77,5M
convLSTM	97,1	9,6M
GRU	93,2	0,1M

В таблице 2 помимо точности интерес представляет количество параметров моделей. В связи с тем, что GRU представлен гораздо меньшим количеством параметров, ожидаемыми являются результаты оценки, свидетельствующие о меньшей точность, чем у моделей с более высоким числом параметров.

На рис. 2 представлены примеры результатов успешного распознавания моделью класса деструктивного поведения. На рис. 3 представлены примеры ошибочного распознавания моделью класса деструктивного поведения

Рис. 2. Примеры результаты успешного распознавания

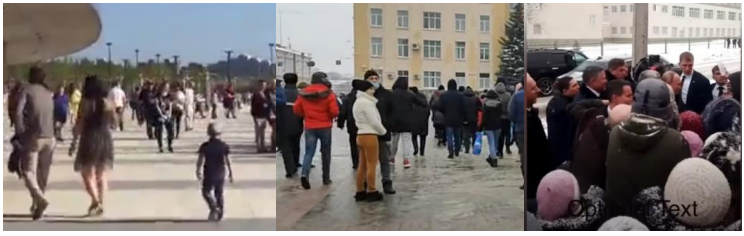

Рис. 3. Примеры ошибочного распознавания деструктивных сцен

4. Заключение

В ходе выполнения данной работы были исследован потенциал извлечения характерных признаков из кадров для распознавания насилия в видеороликах с помощью вычислительно легких решений, основанных на сверточных и рекуррентных нейронных сетях. В ходе выполняемого проекта были проведены эксперименты как с предварительно обученными моделями на базе данных ImageNet, так и на непредобученных моделях. В ходе экспериментов подтвердилось, что предварительно обученная сеть учится гораздо быстрее непредобученной. Также в ходе экспериментов на вход модели подавались кадры с разной последовательностью: от 2 до 30 с шагом 2. При этом для 10 кадров были получены лучшие показатели метрик. В дальнейшем планируется имплементировать эти модели в программное обеспечение для различных мобильных устройств, в том числе на модули Nvidia Jetson.

Исследование выполнено при частичной финансовой поддержке РФФИ в рамках научного проекта № 18-29-22104 (*раздел 3*).

Литература

1. SUDHAKARAN S., LANZ O. *Learning to Detect Violent Videos using Convolutional Long Short-Term Memory* // 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). – IEEE, 2017. – P. 1–6.
2. SUMON S.A., GONI R., HASHEM N.B., SHAHRIA T., RAHMAN R.M. *Violence Detection by Pretrained Modules with Different Deep Learning Approaches* // Vietnam Journal of Computer Science. – 2020. – Vol. 7, No. 1. – P. 19–40.
3. CHO K., VAN MERRIENBOER B., GULCEHRE C., BAHDANAU D., BOUGARES F., SCHWENK H., BENGIO Y. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation* // arXiv. – URL: https://arxiv.org/abs/1406.1078 (дата обращения: 20.07.2021).

4. HOCHREITER S., SCHMIDHUBER J. *Long short-term memory* // Neural computation – 1997. – Vol. 9, No. 8. – P. 1735-1780.
5. CHUNG J., GULCEHRE C., CHO K., BENGIO Y. Empirical *Evaluation of Gated Recurrent Neural Networks on Sequence Modeling* // arXiv. – URL: https://arxiv.org/abs/1412.3555 (дата обращения: 20.07.2021).
6. GRUBER N., JOCKISCH A. *Are GRU cells more specific and LSTM cells more sensitive in motive classification of text?* // Frontiers in Artificial Intelligence. – 2020. – Vol. 3. – P. 1–6.
7. SIMONYAN K., ZISSERMAN A. *Two-stream convolutional networks for action recognition in videos* // arXiv. – URL: https://arxiv.org/abs/1406.2199 (дата обращения: 20.07.2021).
8. HE K., ZHANG X., REN S., SUN J. *Deep Residual Learning for Image Recognition* // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – IEEE, 2016. – P. 770–778.
9. MISRA I., ZITNICK C.L., HEBERT M. *Shuffle and learn: unsupervised learning using temporal order verification* // In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science. Vol. 9905 / Leibe B., Matas J., Sebe N., Welling M. (eds). – Springer, Cham, 2016. – P. 527–544.

DETECTION OF DESTRUCTIVE MULTIMEDIA CONTENT IN A SOCIO-CYBERPHYSICAL INTERNET MONITORING SYSTEM

Anastasia Iskhakova, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, Cand.Sc., senior researcher (iao@ipu.ru).
Konstantin Rusakov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, research assistant (rusakov.msk@yandex.ru).
Andrey Iskhakov, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, research assistant, Cand.Sc., senior researcher (iay@ipu.ru).
Mark Mamchenko, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, research assistant (markmamcha@gmail.com).

Abstract: The research focuses on the problem of monitoring and filtering multimedia content on the Internet in terms of automated detection of scenes of violence and

aggression. The tasks of ensuring security of individual, group and mass consciousness, including protection from destructive information influences and from the use of information technologies for terrorism propaganda, is one of the main directions of scientific research in the field of information security of the Russian Federation. The report provides an overview of existing scientific and technical solutions in this area and offers an approach to target content detection. The model consists of light convolutional neural networks for feature extraction and GRU model for coding frame changes, characterizing the scenes of violence, existing in the video. The proposed methodological support can be used in the implementation of violence and aggression detection modules for complex socio-cyber-physical monitoring systems.

Keywords: video stream analysis, aggression, violence, socio-cyberphysical sys-system, artificial neural network.

УДК 004.89
ББК 32.813.5