

УДК 004.93'11

10.25209/2079-3316-2024-15-4-79-96



## Нейросетевая классификация видеороликов по малому числу кадров

Александр Владимирович **Смирнов**<sup>1</sup>, Дмитрий Денисович **Парфенов**<sup>2</sup>,  
Игорь Петрович **Тищенко**<sup>3</sup>

<sup>1,3</sup> Институт программных систем им. А. К. Айламазяна РАН, Вельсково, Россия

<sup>2</sup> ГУМРФ имени адмирала С. О. Макарова, Санкт-Петербург, Россия

**Аннотация.** В статье предложен метод нейросетевой классификации коротких видеороликов. Задача классификации рассматривается с точки зрения уменьшения числа требуемых операций для категоризации видеороликов. Предлагаемое решение заключается в использовании небольшого числа кадров (не более 10) для выполнения классификации при помощи самой лёгкой нейросетевой архитектуры семейства моделей ResNet. В ходе исследования создан собственный набор данных для обучения, состоящий из трёх классов: «animals», «cars» и «people». В результате получена точность классификации, равная 79%, а также сформирована база данных классифицируемых видеороликов и разработано приложение с элементами GUI для взаимодействия с классификатором и просмотра результатов.

**Ключевые слова и фразы:** Классификация видео, набор данных, нейронные сети, графический интерфейс пользователя

Для цитирования: Смирнов А. В., Парфенов Д. Д., Тищенко И. П. *Нейросетевая классификация видеороликов по малому числу кадров* // Программные системы: теория и приложения. 2024. Т. 15. № 4(63). С. 79–96.  
[https://psta.psiras.ru/read/psta2024\\_4\\_79-96.pdf](https://psta.psiras.ru/read/psta2024_4_79-96.pdf)

## Введение

В настоящее время большую роль в обработке информации различного типа и происхождения играют нейронные сети. Одной из наиболее востребованных областей применения нейросетевых моделей является обработка графических данных, в частности изображений или видео.

Так, например, в работе [1] задача классификации видео рассматривается как операция присвоения некоторого индекса или метки конкретному видеоролику. Авторы используют собственную модель свёрточной нейронной сети (CNN), а также набор данных UCF11<sup>1</sup>, состоящий из видеороликов с различными действиями людей. В результате была получена точность классификации действий, сравнимая с другими нейросетевыми моделями.

В другой работе [2] основное внимание уделялось поиску ключевых кадров на видео, что в последствии повышало точность классификации. Здесь авторы используют гибридную модель состоящую из свёрточной нейронной сети, технологии кластеризации пиков плотности временных сегментов (TSDPC)<sup>2</sup> и сети с долговременной краткосрочной памятью (LSTM)<sup>3</sup>. В итоге им удалось получить точность на наборах данных HMDB51<sup>4</sup> и UCF101<sup>5</sup> в 82.94% и 91.43% соответственно.

Статья [3] также посвящена поиску ключевых кадров. Разработанный метод основан на использовании шаблонов действий путем определения информативной области каждого кадра. Данный метод был протестирован на наборе данных UCF101 с использованием моделей ConvLSTM<sup>6</sup> и VGG16<sup>7</sup>. В результате была получена точность классификации видео более 80%.

Достаточно важным фактором в задаче классификации видео может стать используемый для обучения набор данных. Именно от используемого набора данных зависит итоговая цель классификации. В статье [4] рассматривается создание набора данных из видеороликов, содержащих действия с разжиганием ненависти. Авторы отмечают, что подобные наборы данных существуют, но содержат в основном текстовую информацию, а

---

<sup>1</sup>UCF11 - Action Recognize<sup>URL</sup>

<sup>2</sup>Кластеризация на основе плотности (Пространственная статистика)<sup>URL</sup>

<sup>3</sup>LSTM – сети долгой краткосрочной памяти<sup>URL</sup>

<sup>4</sup>HMDB51<sup>URL</sup>

<sup>5</sup>UCF101 - Action Recognition<sup>URL</sup>

<sup>6</sup>ConvLSTM<sup>URL</sup>

<sup>7</sup>VGG16 – нейросеть для выделения признаков изображений<sup>URL</sup>

изображения и видео представлены очень редко. Вследствие чего, авторы вручную отобрали 43 часа видео для создания своего набора данных, на котором протестировали нейронную сеть и получили точность около 79%.

В работе [5] решается задача поиска признаков ненормального поведения пассажиров лифта. Эксперименты проводились на наборе видеоданных, содержащем четыре вида ненормального поведения: открывание двери, прыжки, пинки и блокировка двери. При использовании модифицированной модели PP-TSM удалось достичь точности классификации в 95%, что на 10% больше, чем было получено на стандартной модели PP-TSM<sup>8</sup>.

Статья [6] описывает метод обнаружения аномального поведения на видео. В данном случае под таким поведением подразумевается проявление жестокости с использованием оружия. Для обнаружения аномалий поведения авторы используют собственную модель нейронной сети J.QCNN, основанную на квантовой свёрточной нейронной сети (QCNN) [7] и глубокой свёрточной нейронной сети Javeria (DCNN). В результате была получена точность обнаружения более 90% по метрике F1-score<sup>9</sup>.

Следующая работа [8] посвящена распознаванию действий на видео с последующим расставлением временных меток. Здесь предлагается модель на основе метода двухпоточкового слияния информации с механизмами внимания (DSIFAM). При тестировании разработанной модели на наборах данных UCF11 и UCF50<sup>10</sup> была получена точность в 91.2% и 89.1% соответственно.

Помимо анализа готовых видеофайлов, также ведутся работы по обработке потокового видео. Такой подход следует использовать, когда необходимо оперативно принять решение, например, быстро идентифицировать потенциально критические или опасные ситуации. Так в статье [9] рассматривается унифицированная и теоретически обоснованная адаптационная система для решения проблемы онлайн-классификации видеоданных, которая основана на математической модели теории классификации. Благодаря этому авторам удалось получать результат от нейронной сети гораздо быстрее без значительного ущерба для точности по сравнению с классическим подходом.

---

<sup>8</sup>High performance recognition 2D architecture PP-TSM<sup>(RU)</sup>

<sup>9</sup>Что такое F-score и для чего он используется?<sup>(RU)</sup>

<sup>10</sup>UCF50 - Action Recognition Data Set<sup>(RU)</sup>

В статье [10] представлено исследование на тему классификации спортивных видео. Ключевой проблемой является природа данных видеороликов, которая заключается в присутствии большого количества динамических сцен. Для решения поставленных задач, авторы разрабатывают собственную нейросетевую модель, основанную на извлечении ключевых кадров и использующую принципы сиамских сетей<sup>11</sup>. В результате новая модель продемонстрировала точность в 97% на наборе видеоданных высокого разрешения и 87% на наборе видеоданных низкого разрешения.

Анализ мультипликационных видеороликов представлен в работе [11]. Здесь авторы поставили задачу отфильтровать нежелательный контент, содержащий жестокие и откровенно сексуальные сцены. Они использовали предварительно обученную на наборе данных ImageNet<sup>12</sup> сверточную сеть в сочетании с сетью двунаправленной долговременной краткосрочной памяти на основе внимания (BiLSTM<sup>13</sup>). Эксперименты показали, что разработанная модель работает относительно лучше других сетей, достигая точности в 95.3%.

Нейросетевая обработка видеоданных находит применение в различных сферах. Благодаря высокой точности такого анализа некоторые результаты исследований можно применять уже сейчас в различных видеохостингах в качестве инструмента фильтрации и автоматизированной категоризации контента.

В настоящей работе предложен метод нейросетевой классификации коротких видеороликов по небольшому числу кадров. В рамках данной работы был создан собственный обучающий набор данных, состоящий из 3 классов: «animals», «cars» и «people». В результате была получена точность классификации, равная 79% по метрике F1-score, что немного ниже, чем в представленных выше работах. Однако прямое сравнение точности классификации нецелесообразно, так как использовались не только разные нейросетевые модели, но и разные наборы данных.

## 1. Цель и задачи исследования

Классификация изображений или объектов на изображениях предполагает отнесение целевого объекта к одному из рассматриваемых

---

<sup>11</sup> *A Friendly Introduction to Siamese Networks*<sup>URL</sup>

<sup>12</sup> *ImageNet*<sup>URL</sup>

<sup>13</sup> *Bidirectional LSTM*<sup>URL</sup>

классов. При классификации видеороликов задача классификации сводится к определению категории/класса видео. Однако видеоролик по сути является набором изображений/кадров, которые постепенно сменяют друг друга. В таком случае требуется определить класс большого множества отдельно взятых кадров, что может быть ресурсозатратно, так как в одной секунде стандартного видеоролика может быть от 30 кадров. Очевидным выходом из данной ситуации может быть использование лишь части кадров видеоролика.

Отличительной особенностью настоящего исследования является использование малого числа кадров видеоролика для выполнения его классификации. Тесты нейросетевого анализа показали, что для достижения удовлетворительной точности достаточно использовать около  $10 \pm 1$  кадров видео.

Основной целью настоящего исследования является экспериментальная проверка метода нейросетевой классификации/категоризации видеороликов, который основан на анализе малого числа кадров без предварительной обработки. Метод считается условно работоспособным, если точность классификации видеороликов превышает 70%.

## 2. Используемая нейросетевая модель-классификатор

В качестве классификатора рассматривались следующие нейросетевые решения: ResNet [12], Faster R-CNN [13] и EfficientNet [14]. Среди представленных моделей была выбрана 18-слойная<sup>14</sup> нейросетевая архитектура ResNet (рисунок 1), так как она предлагает баланс между глубиной, сложностью и производительностью.

Вероятно, для будущих итераций настоящего исследования будет использована 50-слойная<sup>15</sup> архитектура ResNet. Такое решение было продиктовано тем, что 50-слойная архитектура показывает более высокие результаты точности, но также требует больше времени на обучение и классификацию. Тем не менее, 18-слойная архитектура ResNet способна показывать сопоставимые значения точности<sup>16</sup>. Реализация архитектуры ResNet-18 была выполнена на ЯП Python с использованием инструментария PyTorch<sup>17</sup>.

---

<sup>14</sup>[resnet18](#)<sup>URL</sup>

<sup>15</sup>[resnet50](#)<sup>URL</sup>

<sup>16</sup>[Image Classification on ImageNet](#)<sup>URL</sup>

<sup>17</sup>[PyTorch documentation](#)<sup>URL</sup>

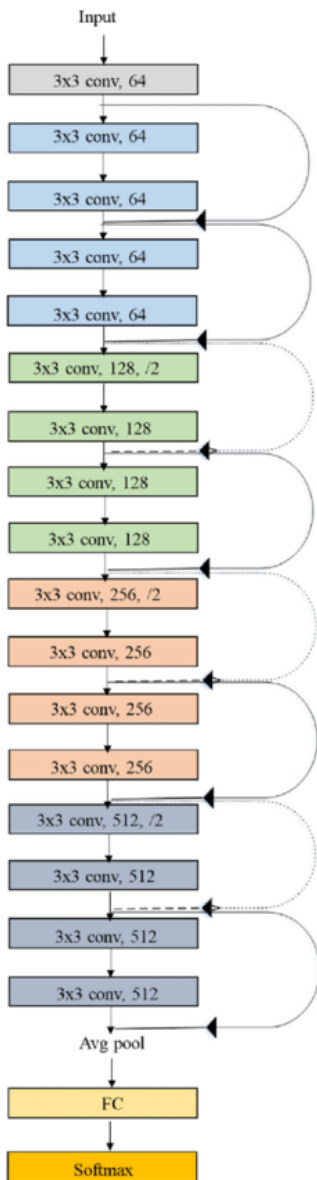


Рисунок 1. Оригинальная архитектура ResNet18

### 3. Создание набора обучающих данных

Несмотря на наличие готовых наборов данных, для решения поставленных задач был создан собственный обучающий набор данных, состоящий из кадров видео с изображением неудач. Так называемое «видео с неудачей» обычно длится менее 10 секунд и содержит не более двух действий-неудач.

По мнению авторов настоящей статьи такой видеоролик с большой долей вероятности может быть успешно классифицирован, так как его содержание достаточно однозначно определено типом отображаемой неудачи. Однако из-за короткой продолжительности «видео с неудачей» часто собирают в тематические подборки, что, с одной стороны, упрощает поиск достаточного количества данных, а с другой стороны, затрудняет извлечение и обработку конкретного ролика.

#### 3.1. Критерии определения классов

Для классификации видеоданных было сформировано три основных класса: «*animals*», «*cars*» и «*people*». Каждый класс определялся наличием на видео объекта конкретного типа, который непосредственно совершал неудачное действие или становился причиной неудач, а именно:

*animals* — класс включал сцены, где причиной неудачи было животное.

Здесь входили различные виды животных, от домашних до диких, которые по своей природе создавали ситуации, подходящие под определение «видео с неудачей»;

*cars* — класс включал сцены, в которых основной причиной возникновения неудачи являлся автомобиль. Подобные видеофрагменты охватывали инциденты с участием автомобилей, такие как аварии, неудачные маневры и другие события, связанные с транспортными средствами;

*people* — класс содержал видеоматериалы, где неудачные ситуации происходили исключительно с участием людей, без присутствия животных или автомобилей. Сцены данного типа включали падения, ошибки и другие инциденты, произошедшие с людьми.

#### 3.2. Формирование набора данных

В первую очередь был осуществлён сбор исходного видеоматериала, сгруппированного по трем ранее описанным категориям. Видеоконтент был извлечён с популярного видеохостинга<sup>18</sup> при помощи специально

---

<sup>18</sup>Видеохостинги: обзор самых популярных площадок<sup>sm</sup>

написанного скрипта на ЯП Python, который позволяет загружать видеофайлы различного формата и качества.

Процесс формирования набора данных состоял из следующих этапов:

*Предварительная обработка видео.* После получения исходных видеофайлов была проведена их предварительная обработка в программе Adobe Premiere Pro<sup>19</sup>. С помощью функции «определение сцен» (Scene Detection) видео было автоматически разделено на отдельные сцены. Этот шаг был необходим, так как скачанный видеоматериал представлял собой подборки «видео с неудачами», которые были смонтированы в единый видеоролик.

*Разделение сцен на кадры.* Полученные ранее сцены были разделены на кадры. В рамках данного процесса из каждой сцены были извлечены последовательности по 10 случайных кадров.

*Верификация и очистка данных.* На этом этапе был проведён тщательный анализ полученных кадров. Кадры проверялись на корректность и соответствие выбранным критериям категорий. Нерелевантные кадры (размытый кадр, чёрный экран и т.п.) были удалены из набора данных.

*Расширение и балансировка набора данных.* Завершающим этапом было расширение и балансировка набора данных. Поскольку изначально количество данных в каждой из категорий могло отличаться, были приняты меры по балансировке классов. Для этого использовались методы дополнения данных, включая операции поворота, изменения масштаба, а также другие трансформации, что позволило создать равномерно распределённый набор данных для последующего обучения нейронной сети.

Таким образом, процесс создания набора данных был тщательно структурирован и включал в себя этапы сбора, обработки, верификации и балансировки данных, что обеспечило получение качественного и надежного материала для дальнейшего обучения модели. Сформированный набор данных содержал по 2939 изображений/кадров для каждого из трёх классов.

#### 4. Обучение нейронной сети

Обучение нейронной сети происходило при помощи трансферного метода. Трансферное обучение (TL)<sup>20</sup> – это метод машинного обучения, при котором модель, предварительно обученная выполнению одной

---

<sup>19</sup> Professional video editing software | Adobe Premiere Pro<sup>®</sup>

<sup>20</sup> Что такое трансферное обучение?<sup>URL</sup>



задачи, перенастраивается для выполнения другой, похожей на предыдущую. Обучение новой модели – это трудоемкий и длительный процесс, требующий большого количества данных, достаточной вычислительной мощности и прохождения нескольких итераций, прежде чем модель будет готова к запуску. Вместо этого применяется метод TL для переобучения существующих моделей, подготавливая их к решению смежных задач с использованием новых данных.

Метод TL был использован на предварительно обученной модели ResNet-18. Переобучение модели выполнялось в течении 25, 35 и 50 эпох. В таблице 1 представлены характеристики модели после обучения.

Таблица 1. Характеристики модели после обучения

Кол-во эпох	Точность на тренировочной выборке, %	Точность на валидационной выборке, %	Ошибка на тренировочной выборке	Ошибка на валидационной выборке
25	78.0	89.7	0.538	0.266
35	77.3	90.4	0.546	0.262
50	81.4	93.6	0.458	0.172

На рисунках 2 и 3 представлены графики точности и ошибки для обучения модели в течение 50 эпох.

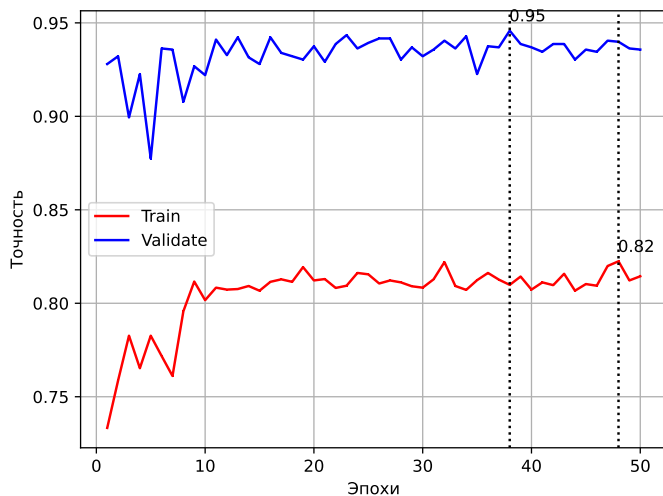


Рисунок 2. График изменения значения точности, полученной на тренировочной и валидационной выборках в процессе обучения

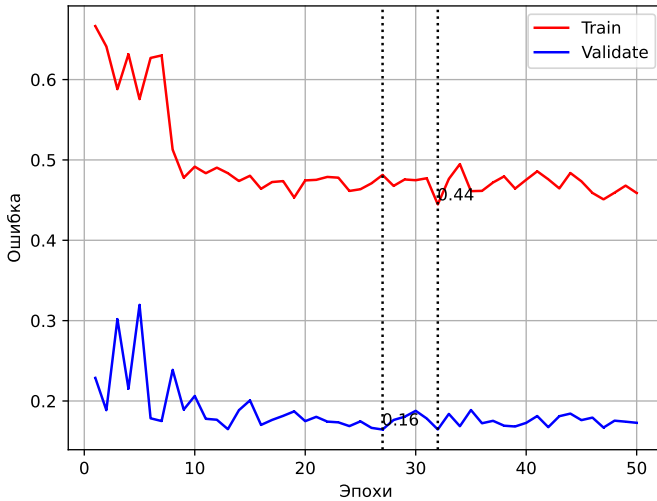


Рисунок 3. График изменения значения ошибки, полученной на тренировочной и валидационной выборках в процессе обучения

## 5. Тестирование нейросети и подсчёт точности

Для тестирования обученной нейронной сети использовалось 35 видеороликов класса «animals», 39 видеороликов класса «cars» и 42 видеоролика класса «people». Из каждого видеоролика извлекалось  $10 \pm 1$  кадров с равными временными промежутками. Принадлежность ролика к какому-либо классу определялась классом большинства его кадров, то есть, если более 50% кадров ролика имели некоторый класс, то и сам видеоролик относился к этому же классу.

Точность рассчитывалась по метрике F1-score с помощью функции `MulticlassF1Score`<sup>21</sup> библиотеки `PyTorch Lightning`<sup>22</sup>. Также были подсчитаны значения Precision (точность)<sup>23</sup> и Recall (полнота)<sup>24</sup> для видеороликов и их кадров. В таблицах 2 и 3 представлены данные, полученные в ходе тестирования нейросети.

<sup>21</sup>`MulticlassF1Score`<sup>URL</sup>

<sup>22</sup>*Turn ideas into AI, Lightning fast*<sup>URL</sup>

<sup>23</sup>`MulticlassPrecision`<sup>URL</sup>

<sup>24</sup>`MulticlassRecall`<sup>URL</sup>

Таблица 2. Данные о количестве используемых кадров и результирующей точности

Класс	Кол-во кадров	Precision	Recall	F1-score
animals	378	0.754	0.576	0.653
cars	427	0.899	0.733	0.807
people	456	0.602	0.824	0.696
Общее	1261	0.752	0.711	0.719

Таблица 3. Данные о количестве используемых видеороликов и результирующей точности

Класс	Кол-во видео	Precision	Recall	F1-score
animals	35	0.884	0.657	0.754
cars	39	0.916	0.846	0.880
people	42	0.666	0.857	0.750
Общее	116	0.822	0.786	0.794

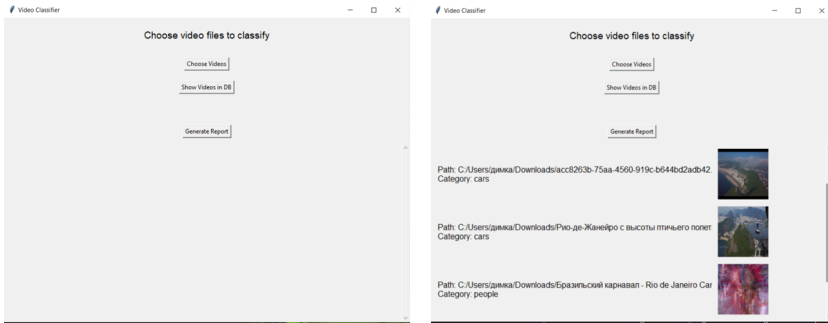
## 6. Приложение для взаимодействия с классификатором и просмотра результатов

В рамках исследования, описанного в настоящей статье, было разработано приложение с графическим интерфейсом пользователя (GUI) для взаимодействия с классификатором, хранения и просмотра результатов классификации видеороликов. На рисунке 4 изображены основные окна приложения.

Разработанное приложение имеет следующий функционал:

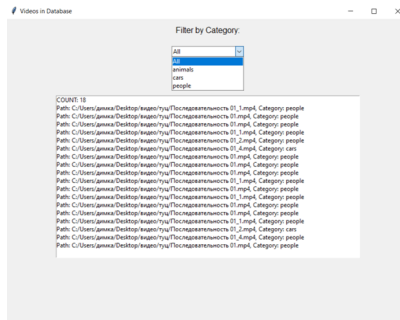
*Выбор видеофайла.* Данная операция позволяет пользователю выбрать один или несколько видеофайлов с локального диска для их классификации. При нажатии на соответствующий элемент GUI открывается диалоговое окно для выбора файлов с расширениями .mp4, .avi и .mov. После завершения выбора выполняется анализ каждого файла. Результатом анализа является определённый класс/категория видео («animals», «cars» или «people»).

*Взаимодействие с базой данных.* При нажатии на соответствующий элемент GUI открывается новое окно, в котором отображается список видеофайлов, ранее классифицированных и сохраненных в базе данных. Далее отправляется запрос к базе данных для извлечения видео на основе выбранного фильтра. Пользователь может выбрать категорию для фильтрации или оставить фильтр на значении «All» для отображения всех видео. В окне предпросмотра отображается список видеофайлов с указанием их пути и категории.



(a) Основное окно приложения

(б) Окно отображения результатов классификации



(в) Окно взаимодействия с базой данных

Рисунок 4. Приложение для взаимодействия с классификатором

*Создание отчёта.* При нажатии на соответствующий элемент GUI приложение генерирует текстовый файл, который содержит сводку о количестве видео в каждой категории и полный список видео с указанием их категорий.

## 7. Анализ полученных результатов

В процессе тестирования нейронной сети была достигнута точность классификации более 79% (0.794), при этом точность классификации отдельного класса «cars» составила 88% (0.880). Данные показатели выше, чем заявленный порог в 70%, при достижении которого предложенный метод считается успешно проверенным и условно работоспособным.





Тем не менее, полученное значение точности меньше, чем в аналогичных работах, рассмотренных ранее. Это может указывать на наличие некоторых недостатков представленного метода классификации. Однако проводить прямое сравнение некорректно, так как использовались не только разные нейросетевые модели, но и разные наборы данных.









Стоит отметить, что в предложенном методе не используются какие-либо алгоритмы предобработки видео/кадров видео. Помимо этого, классификатором выступает относительно простая и лёгкая нейросетевая модель. Основная идея предложенного метода как раз и заключалась в том, чтобы выполнить классификацию/категоризацию видеороликов с использованием минимально возможных инструментария и данных.

## Вывод

В результате проведённого исследования была достигнута точность классификации/категоризации видео более 79%, что доказывает жизнеспособность выбранного метода. Однако следует учесть тот факт, что при анализе видео сложной композиции или при классификации на более конкретные классы, вероятно, потребуются прибегнуть к дополнительным методам извлечения признаков или ключевых кадров, которые были описаны в рассмотренных статьях.

## Список использованных источников

- [1] Duvvuri K., Kanisettypalli H., Jaswanth K., Murali K. *Video classification using CNN and ensemble learning // 2023 9th International Conference on Advanced Computing and Communication Systems.*– V. 1, ICACCS 2023 (17-18 March 2023, Coimbatore, India).– IEEE.– 2023.– ISBN 9798350397383.– Pp. 66–70.  ↑80
- [2] Tang H., Ding L., Wu S., Ren B., Sebe N., Rota P. *Deep unsupervised key frame extraction for efficient video classification // ACM Transactions on Multimedia Computing, Communications and Applications.*– 2023.– Vol. 19.– No. 3.– id. 119.– 17 pp.  ↑80
- [3] Savran K. R., Gan J. Q., Escobar J. J. *A novel keyframe extraction method for video classification using deep neural networks // Neural Computing and Applications.*– 2023.– Vol. 35.– No. 34.– Pp. 24513–24524.  ↑80
- [4] Das M., Raj R., Saha P., Mathew B., Gupta M., Mukherjee A. *HateMM: a multi-modal dataset for hate video classification // Proceedings of the International AAAI Conference on Web and Social Media.*– 2023.– Vol. 17, Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023).– Pp. 1014–1023.  ↑80

- [5] Lei J., Sun W., Fang Y., Ye N., Yang S., Wu J. *A model for detecting abnormal elevator passenger behavior based on video classification* // *Electronics*.– 2024.– Vol. **13**.– No. 13.– id. 2472.– 15 pp.  <sup>↑81</sup>
- [6] Amin J., Anjum M. A., Ibrar K., Sharif M., Kadry S., Crespo R. G. *Detection of anomaly in surveillance videos using quantum convolutional neural networks* // *Image and Vision Computing*.– 2023.– Vol. **135**.– id. 104710.  <sup>↑81</sup>
- [7] Cong I., Choi S., Lukin M. D. *Quantum convolutional neural networks* // *Nature Physics*.– 2019.– Vol. **15**.– Pp. 1273–1278.  <sup>↑81</sup>
- [8] Jianmin H., Jie L. *A video action recognition method via dual-stream feature fusion neural network with attention* // *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*.– 2024.– Vol. **32**.– No. 04.– Pp. 673–694.  <sup>↑81</sup>
- [9] Trędowicz M., Struski Ł., Mazur M., Janusz S., Lewicki A., Tabor J. *PrAViC: probabilistic adaptation framework for real-time video classification*.– 2024.– 12 pp.  arXiv:2406.11443 [cs.CV] <sup>↑81</sup>
- [10] Gao T., Zhang M., Zhu Y., Zhang Y., Pang X., Ying J., Liu W. *Sports video classification method based on improved deep learning* // *Applied Sciences*.– 2024.– Vol. **14**.– No. 2.– id. 948.– 13 pp.  <sup>↑82</sup>
- [11] Kanwal Y., Tabassam N. *An attention mechanism-based CNN-BiLSTM classification model for detection of inappropriate content in cartoon videos* // *Multimedia Tools and Applications*.– 2024.– Vol. **83**.– No. 11.– Pp. 31317–31340.  <sup>↑82</sup>
- [12] He K., Zhang X., Ren S., Sun J. *Deep residual learning for image recognition* // *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016 (27–30 June 2016, Las Vegas, NV, USA)*.– IEEE.– 2016.– ISBN 978-1-4673-8850-4.– Pp. 770–778. <sup>↑83</sup>
- [13] Ren S., He K., Girshick R., Sun J. *Faster R-CNN: towards real-time object detection with region proposal networks* // *Proceedings of the 28th International Conference on Neural Information Processing Systems*.– V. 1, NIPS'15 (December 7–12, 2015, Montreal, Canada), Cambridge: MIT Press.– 2015.– ISBN 9781510825024.– Pp. 91–99.  <sup>↑83</sup>
- [14] Tan M., Le Q. V. *EfficientNet: rethinking model scaling for convolutional neural networks* // *Proceedings of the 36th International Conference on Machine Learning, ICML 2019 (9–15 June 2019, Long Beach, California, USA)*, Proceedings of Machine Learning Research.– vol. **97**.– ICML.– 2019.– ISBN 9781510886988.– Pp. 6105–6114. <sup>↑83</sup>

Поступила в редакцию	01.10.2024;
одобрена после рецензирования	23.10.2024;
принята к публикации	04.11.2024;
опубликована онлайн	20.11.2024.

Рекомендовал к публикации

к.т.н. Е. П. Куршев

## Информация об авторах:



**Александр Владимирович Смирнов**

Младший научный сотрудник Лаборатории методов обработки и анализа изображений, Институт Программных Систем имени А. К. Айламазяна РАН. Научные интересы: компьютерное зрение; нейронные сети; робототехника; автоматизация и управление

ID 0000-0002-7104-1462

e-mail: [asmirnov\\_1991@mail.ru](mailto:asmirnov_1991@mail.ru)



**Дмитрий Денисович Парфенов**

Студент 3 курса по специализации «Прикладная информатика в экономике». Научные интересы: компьютерное зрение; нейронные сети

ID 0000-0002-0369-0524

e-mail: [parfecto@yandex.ru](mailto:parfecto@yandex.ru)



**Игорь Петрович Тищенко**

Кандидат технических наук, ИО директора, Институт Программных Систем имени А. К. Айламазяна РАН. Научные интересы: компьютерное зрение; нейронные сети; робототехника; автоматизация и управление

ID 0000-0002-0369-0524

e-mail: [igor.p.tishchenko@gmail.com](mailto:igor.p.tishchenko@gmail.com)

Вклад авторов: *А. В. Смирнов* – 55% (идея, методология, программное обеспечение, валидация, формальный анализ, расследование, написание черновой версии, доработка и редактирование, визуализация, наставничество, администрирование); *Д. Д. Парфенов* – 35% (сбор материала, курирование данных, написание черновой версии, доработка и редактирование, визуализация); *И. П. Тищенко* – 10% (администрирование, финансирование).

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*



# Neural network classification of videos based on a small number of frames

Alexander Vladimirovich **Smirnov**<sup>1</sup>, Dmitry Denisovich **Parfenov**<sup>2</sup>,  
Igor Petrovich **Tishchenko**<sup>3</sup>

<sup>1,3</sup>Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

<sup>2</sup>Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia

**Abstract.** The article proposes a method for neural network classification of short videos. The classification problem is considered from the point of view of reducing the number of operations required to categorize videos. The proposed solution consists of using a small number of frames (no more than 10) to perform classification using the lightest neural network architecture of the ResNet family of models. As part of the work, a proprietary training dataset was created, consisting of three classes: “animals”, “cars” and “people”. As a result, a classification accuracy of 79% was obtained, a database of classified videos was formed, and an application with GUI elements was developed for interacting with the classifier and viewing the results. (*In Russian*).

**Key words and phrases:** Video classification, dataset, neural networks, graphical user interface


2020 *Mathematics Subject Classification:* 68T10; 68T45

For citation: Alexander V. Smirnov, Dmitry D. Parfenov, Igor P. Tishchenko. *Neural network classification of videos based on a small number of frames*. Program Systems: Theory and Applications, 2024, **15**:4(63), pp. 79–96. (*In Russ.*). [https://psta.pstiras.ru/read/psta2024\\_4\\_79-96.pdf](https://psta.pstiras.ru/read/psta2024_4_79-96.pdf)



## References

- [1] K. Duvvuri, H. Kanisettypalli, K. Jaswanth, K. Murali. “Video classification using CNN and ensemble learning”, *2023 9th International Conference on Advanced Computing and Communication Systems*. V. 1, ICACCS 2023 (17-18 March 2023, Coimbatore, India), IEEE, 2023, ISBN 9798350397383, pp. 66–70. [doi](#)
- [2] H. Tang, L. Ding, S. Wu, B. Ren, N. Sebe, P. Rota. “Deep unsupervised key frame extraction for efficient video classification”, *ACM Transactions on Multimedia Computing, Communications and Applications*, **19**:3 (2023), id. 119, 17 pp. [doi](#)
- [3] K. R. Savran, J. Q. Gan, J. J. Escobar. “A novel keyframe extraction method for video classification using deep neural networks”, *Neural Computing and Applications*, **35**:34 (2023), pp. 24513–24524. [doi](#)
- [4] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, A. Mukherjee. “HateMM: a multi-modal dataset for hate video classification”, *Proceedings of the International AAAI Conference on Web and Social Media*, **17**, Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023) (2023), pp. 1014–1023. [doi](#)
- [5] J. Lei, W. Sun, Y. Fang, N. Ye, S. Yang, J. Wu. “A model for detecting abnormal elevator passenger behavior based on video classification”, *Electronics*, **13**:13 (2024), id. 2472, 15 pp. [doi](#)
- [6] J. Amin, M. A. Anjum, K. Ibrar, M. Sharif, S. Kadry, R. G. Crespo. “Detection of anomaly in surveillance videos using quantum convolutional neural networks”, *Image and Vision Computing*, **135** (2023), id. 104710. [doi](#)
- [7] I. Cong, S. Choi, M. D. Lukin. “Quantum convolutional neural networks”, *Nature Physics*, **15** (2019), pp. 1273–1278. [doi](#)
- [8] H. Jianmin, L. Jie. “A video action recognition method via dual-stream feature fusion neural network with attention”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **32**:04 (2024), pp. 673–694. [doi](#)
- [9] M. Trędowicz, Ł. Struski, M. Mazur, S. Janusz, A. Lewicki, J. Tabor. *PrAViC: probabilistic adaptation framework for real-time video classification*, 2024, 12 pp. [doi](#) arXiv:2406.11443 [cs.CV]
- [10] T. Gao, M. Zhang, Y. Zhu, Y. Zhang, X. Pang, J. Ying, W. Liu. “Sports video classification method based on improved deep learning”, *Applied Sciences*, **14**:2 (2024), id. 948, 13 pp. [doi](#)
- [11] Y. Kanwal, N. Tabassam. “An attention mechanism-based CNN-BiLSTM classification model for detection of inappropriate content in cartoon videos”, *Multimedia Tools and Applications*, **83**:11 (2024), pp. 31317–31340. [doi](#)
- [12] K. He, X. Zhang, S. Ren, J. Sun. “Deep residual learning for image recognition”, *2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2016 (27-30 June 2016, Las Vegas, NV, USA), IEEE, 2016, ISBN 978-1-4673-8850-4, pp. 770–778.

- [13] S. Ren, K. He, R. Girshick, J. Sun. “Faster R-CNN: towards real-time object detection with region proposal networks”, *Proceedings of the 28th International Conference on Neural Information Processing Systems*. V. 1, NIPS’15 (December 7–12, 2015, Montreal, Canada), MIT Press, Cambridge, 2015, ISBN 9781510825024, pp. 91–99. 
- [14] M. Tan, Q. V. Le. “EfficientNet: rethinking model scaling for convolutional neural networks”, *Proceedings of the 36th International Conference on Machine Learning*, ICML 2019 (9-15 June 2019, Long Beach, California, USA), Proceedings of Machine Learning Research, vol. **97**, ICML, 2019, ISBN 9781510886988, pp. 6105–6114.