Научная визуализация, 2024, том 16, номер 2, страницы 67 - 80, DOI: 10.26583/sv.16.2.06

Визуализация и классификация движений человека на основе скелетной структуры: Нейросетевой подход к анализу спортивных упражнений и сравнение методологий

В.О. Кузеванов[1], Д. В. Тихомирова[2]

Национальный исследовательский ядерный университет "МИФИ", Москва, Россия

[1] ORCID: 0009-0004-5415-1477, vl.kuzevanov@gmail.com
[2] ORCID: 0000-0002-0812-2331, dvsulim@mail.ru

Аннотация

Авторы статьи рассматривают и сравнивают различные существующие подходы к HAR, анализируют преимущества и недостатки платформ извлечения скелетной структуры человека из видеопотока, а также оценивают значение визуального представления в процессе анализа движений. В данной статье представлен пример реализации одного из подходов к распознаванию движений человека (Human Action Recognition – HAR), основанный на использовании интерпретируемости и визуальной выразительности, присущих скелетным структурам. В работе разработана и реализована специальная сеть с долговременной памятью (Long Short-Term Memory – LSTM), предназначенная для классификации человеческой деятельности, которая была обучена и протестирована в домене спортивных упражнений. Включение в состав LSTM ячеек памяти и механизмов управления не только снимает проблему затухающего градиента, но и позволяет слою LSTM избирательно сохранять и использовать релевантную информацию в длинных последовательностях, что делает их весьма эффективными в задачах со сложными временными зависимостями. Проблема с затухающим градиентом достаточна распространена в глубоких нейронных сетях и заключается в том, что при обратном распространении ошибки во время обучения сети градиент может сильно уменьшаться по мере прохождения через слои сети к начальным слоям. Это может привести к тому, что веса в начальных слоях практически не обновляются, что делает обучение этих слоев невозможным или замедляет его процесс. Полученное решение может использоваться для создания виртуального фитнес-ассистента, работающего в режиме реального времени. Кроме того, данный подход позволит создавать интерактивные обучающие приложения с визуализацией скелетной структуры человека, системы анализа и мониторинга движений в области медицины и реабилитации, а также для разработки систем безопасности с контролем доступа, основанных на анализе визуальных данных о движении частей тела человека.

Ключевые слова: компьютерное зрение, нейронная сеть, машинное обучение, скелетная структура.

1. Введение. Обзор работ в области распознавания действий

Проблема идентификации паттернов поведения человека по видеопотоку представляет для вычислительных устройств непростую задачу, поэтому одним из важнейших объектов исследований в научных областях компьютерного зрения и машинного обучения является способность компьютерных систем идентифицировать, сегментировать и классифицировать деятельность человека на основе данных, собранных различными

датчиками. Информационные системы распознавания активности находят применение во многих сферах, включая системы видеонаблюдения, человеко-компьютерные интерфейсы, робототехнику и здравоохранение. В частности, такие технологии могут быть применены и в спорте не только в трансляциях спортивных мероприятий, но и в персональных помощниках, фитнес-ассистентах для улучшения качества выполнения упражнений спортсменом. Обширная возможность применения подобного рода систем и их польза обуславливают актуальность исследований в данном направлении. Одним из самых популярных подходов к вычислению сложных задач, позволяющих превзойти возможности человека является концепция глубокого обучения, подраздел машинного о бучения (Deep Learning – DL) [1].

На сегодняшний день существует большое количество исследований и методик распознавания движений человека в видео. Так М. Вригкас, Х. Никоу и И. А. Какадиарис [2] предлагают следующую декомпозицию действий человека (см. рис. 1) и иерархию методов распознавания (см. рис. 2).

Рис. 1. Декомпозиция человеческих действий

Целью HAR является исследование действий из видеопоследовательностей или неподвижных изображений. Системы распознавания стремятся правильно классифицировать входные данные в лежащую в их основе категорию деятельности. В зависимости от сложности деятельность человека подразделяется на: 1) жесты, 2) атомарные действия, 3) взаимодействия человека с объектом или человека с человеком, 4) групповые действия, 5) поведение и 6) события.

В области исследований по распознаванию деятельности человека предложены несколько подходов. Разделяют исследования на 2D (с явными моделями формы и без них) и 3D подходы [3]. Также была представлена новая таксономия, посвященная анализу движения человека, отслеживанию с одно- и много-ракурсных камер и распознаванию человеческой деятельности [4].

Моделирование 3D-данных также является новой тенденцией, появившейся вместе со специальными камерами, способными определять глубину объектов, которую можно использовать для 3D-реконструкции. Человеческое тело состоит из костей и соединяющих суставов, что позволяет моделировать эту структуру в 3D пространстве с помощью камер глубины, получая более сильные признаки, в сравнении с моделированием структуры человека в 2D пространстве. Исследование Аггарвал Дж.К. и Ся Л. [5] представляет созданную классификацию методов распознавания человеческой деятельности HAR с использованием 3D-стереосистем и систем захвата движения, уделив

основное внимание методам, которые используют в своих расчетах данные о глубине. Система на подобие Microsoft Kinect [6] сыграли важную роль в захвате движений, идентифицируя скелетную структуру и движение сочленений с использованием датчиков глубины.

В исследованиях выделяются две основные категории: (1) одномодальные и (2) мультимодальные методы распознавания в соответствии с характером информации, приходящей с датчиков, которые они используют. Далее они разделяются на подкатегории в зависимости от того, как они моделируют деятельность человека.

Унимодальные методы представляют деятельность человека на основе данных одной модальности, таких как изображения, и далее они подразделяются на: 1) пространственно-временные, 2) стохастические, 3) методы, основанные на правилах и 4) методы, основанные на форме.

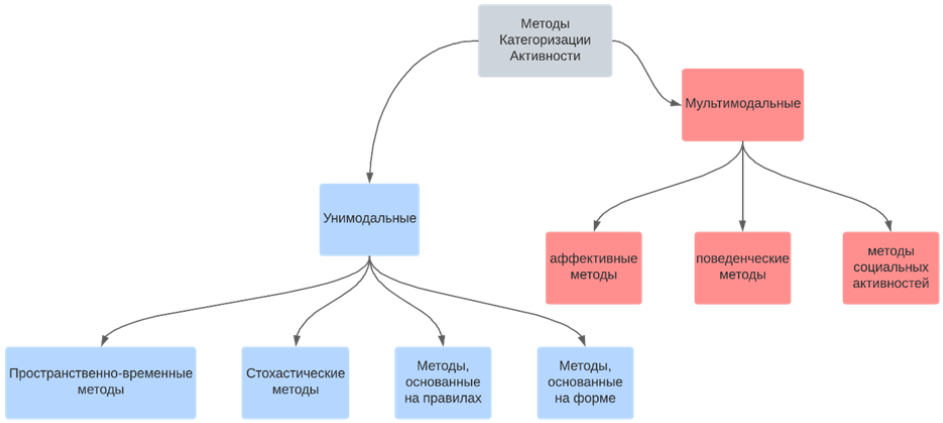

Рис. 2. Иерархическая категоризация методов

Пространственно-временные методы основываются на концепте представления движений в виде набора пространственно-временных характеристик или траекторий [7]. Стохастические методы анализируют активность человека на основе статистических моделей (например, скрытые марковские модели) [8]. Методы, основанные на правилах, используют описательный набор правил [9]. Методы, основанные на форме, используют в анализе движений человека смоделированные формы в пространстве [10].

Мультимодальные методы объединяют используют информацию сразу из нескольких модальностей и подразделяются на следующие категории: 1) аффективные, 2) поведенческие и 3) методы социальных сетей [11].

Аффективные методы представляют активность человека через эмоциональные коммуникации и аффективные состояния [12]. Поведенческие методы распознают различные поведенческие признаки, невербальных мультимодальных сигналов, таких как жесты, мимика и слуховые сигналы [13]. Методы социальных сетей моделируют характеристики и поведение людей на нескольких уровнях взаимодействия между людьми в социальных событиях, начиная с жестов, движений тела и речи [14].

1.1 Методы, основанные на форме

Части человеческого тела могут быть описаны различными способами в 2D-пространстве и в 3D-пространстве как прямоугольные участки, наборы координат определенных точек и сочленений, как объемные фигуры (см. рис. 3). Хорошо известно, что алгоритмы распознавания деятельности по силуэту человека (основанные на форме) становятся все популярнее с приходом нейронных сетей, однако из-за использования данных только одной модальности необходимо с большой точностью распознавать части тела человека.

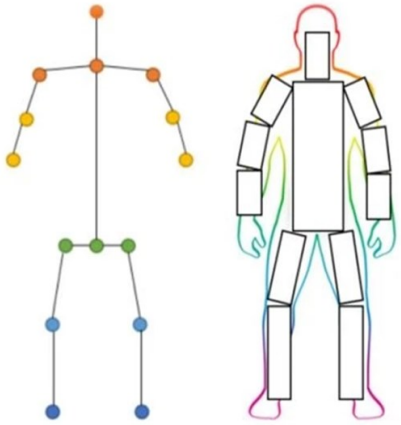

Рис. 3. Представление человека в 2D и 3D пространстве

Одной из ключевых особенностей данной методики, основанной на извлечении скелетной структуры, является представление тела человека и его скелета набором координат в пространстве, что представляет собой отход от традиционных представлений на основе пикселей. Такой подход позволяет не только снизить размерность исходных данных, но и уловить существенные пространственные и временные особенности, необходимые для распознавания действий.

Действие могут классифицироваться достаточно большим количеством различных способов, но наиболее часто встречающимися методами являются: 1) метод голосования по кадрам, 2) метод глобальной гистограммы, 3) метод классификации SVM и 4) метод динамической временной деформации.

Графические модели широко используются в трехмерном моделировании позы человека. Сочетание дискриминационных и генеративных моделей улучшает оценку позы человека.

Процедуры распознавания могут быть реализованы в режиме реального времени с использованием пошагового обновления ковариации и методов классификации ближайших соседей. Оценка позы человека очень чувствительна к различным обстоятельствам, включая изменения освещения, точки обзора, окклюзии, беспорядок на фоне и одежду человека. Недорогие технологии, такие как Microsoft Kinect и другие датчики RGB-D, могут эффективно бороться с этими ограничениями и обеспечивать достаточно точную оценку.

1.2 Структура системы распознавания на основе скелетного представления

Исходя из технологических возможностей и эффективности рассмотренных методов категоризации человеческой активности было решено остановиться на подходе, использующем представление скелетной структуры. Эффективная реализация системы распознавания действий предполагает анализ видеопотока с последующим извлечением признаковых данных. На рисунке 4 представлена предполагаемая авторами структура системы распознавания.

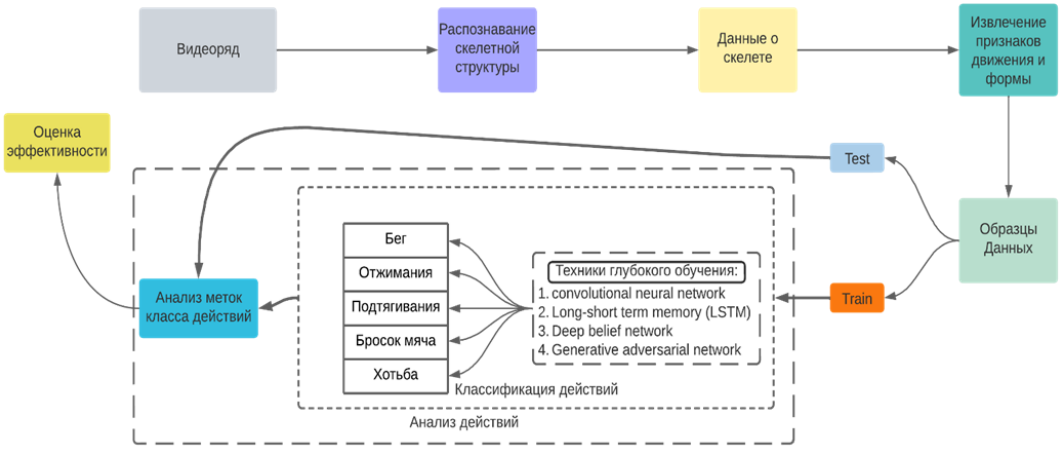

Рис. 4. Концепт системы распознавания действий

Предлагаемая архитектура системы состоит из трех основных этапов: 1) извлечение скелета, 2) извлечение пространственных и временных характеристик и 3) распознавание активности.

При этом следует отметить, что признаки извлекаются путем отслеживания ключевых суставов тела человека.

2. Платформы извлечения признаков скелетной структуры

С развитием технологий искусственного интеллекта появилось большое количество предобученных моделей и платформ, позволяющих быстро и качественно выполнять часто встречающие подзадачи машинного обучения. К одной из таких подзадач относится анализ изображения или видеопотока и выявление скелетной структуры человека.

На данный момент существует достаточное количество платформ, доступных в открытом доступе для пользования, решающих проблему извлечения координат. К ним относятся такие платформы как OpenPose [15], Detectron2 [16], MediaPipe [17] и YOLOv7 [18]. Наибольшей популярностью пользуются инструменты OpenPose, MediaPipe и YOLOv7 Pose.

MediaPipe – это система с открытым исходным кодом от Google для создания кроссплатформенных настраиваемых решений машинного обучения для прямой и потоковой передачи мультимедиа. MediaPipe в настоящее время находится в активной разработке и содержит обширную документацию, включая демонстрации и примеры использования встроенных функций. В системе используется топология ориентиров BlazePose 33. BlazePose – это набор из 3 топологий: COCO keypoints, Blaze Palm и Blaze Face. Она работает в два этапа: обнаружение и отслеживание. Поскольку обнаружение не производится в каждом кадре, MediaPipe может быстрее выполнять вывод. Для оценки позы в MediaPipe используются три модели.

OpenPose – это система обнаружения нескольких человек в режиме реального времени с открытым исходным кодом для совместного обнаружения ключевых точек человеческого тела, ладоней, лица и ступней. Этот проект сильно зависит от набора данных CMU Panoptic Studio. OpenPose также включает демонстрации и примеры использования встроенных функций.

Модель YOLO (You Only Look Once) v7 является последней в семействе моделей YOLO. Модели YOLO представляют собой одноступенчатые детекторы объектов. В YOLO кадры изображений представлены через извлечение признаков. Эти функции объединяются и смешиваются, а затем передаются в голову сети. Эта модель предсказывает местоположения и классы объектов, вокруг которых должны быть нарисованы ограничивающие рамки.

2.1 Сравнительный анализ платформ

Авторы данной статьи производят сравнение платформ в контексте применимости для анализа видео спортивных упражнений, выполненных одним спортсменом в кадре.

Согласно исследованию Радзки Р. [19], все решения обладают хорошей точностью определения сочленений человеческого тела при рендеринге относительно статичных изображений; при плохом освещении или независимо от того, смотрит ли человек прямо в камеру.

Самая большая проблема в задаче распознавания скелетной структуры заключается в размытости движения, что, наряду с увеличением скорости движения, приводит к большим ошибкам в представлении положения ориентиров, вплоть до полной потери обнаружения. В этой области MediaPipe оказался гораздо более эффективным в борьбе со сбоями. В этом тесте MediaPipe показал значительно большую устойчивость к размытию, чем OpenPose. На рисунке 5 показано влияние размытия движения на обнаружение ориентира.

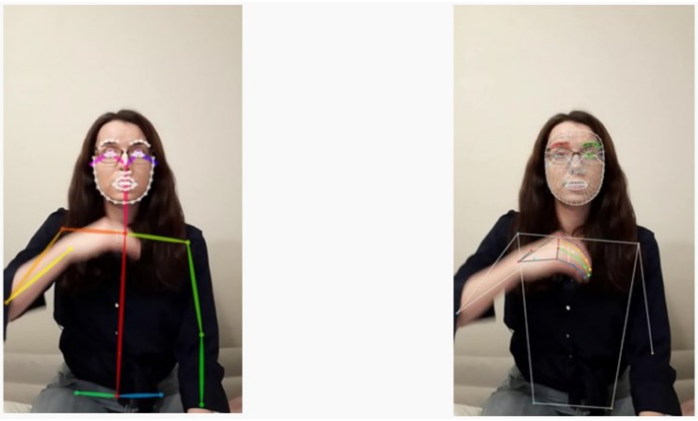

Рис. 5. Пример работы OpenPose и MediPipe при размытии

Кукил и В. Гупта [20] отмечают, что YOLOv7 Pose хуже справляется с детекцией человеческой фигуры в малом масштабе по сравнению с MediaPipe. На рисунке 6 заметно, что модель не оказалась способна обнаружить человека на всех кадрах в видеопотоке потеряв его из виду при уменьшении размера человеческой фигуры в кадре. MediaPipe в свою очередь смог определить человека в значительно меньшем масштабе.

Рис. 6. Пример работы YOLOv7 и MediaPipe с малым масштабом

Данная проблема обуславливается различием в используемых фреймворками методах оценки позы. MediaPipe отслеживает человека после подтверждения обнаружения объекта с использованием вспомогательной сети. С другой стороны, YOLOv7 выполняет обнаружение для каждого кадра в отдельности, что уменьшает ее качество, но значительно увеличивает скорость вычислений.

Все модели имеют свои преимущества и недостатки. Модель OpenPose требует больших вычислительных мощностей и специальных платформ для работы, однако ее точность высока за счет возможности определения позы человека в каждом кадре сети. Кроме того, по сравнению с моделью MediaPipe, ее обучение занимает много времени. Модель OpenPose выводит только двумерные точки, поэтому она не может точно определить углы наклона некоторых суставов, поэтому OpenPose подходит для высокоточных задач, где пренебрегают отслеживанием в реальном времени. С другой стороны, модель MediaPipe является легкой и быстрой и не требует установки на какие-либо платформы. Хотя точность ниже, чем у OpenPose, обучение и развертывание занимает меньше времени: используя вспомогательный детектор, конвейер определяет местоположение области интереса (в данном случае это человеческая фигура) (Region of Interest – ROI) в кадре, затем трекер предсказывает ориентиры позы и маску сегментации в пределах области интереса ROI, используя на входе изображение обрезанное по области интереса, тем самым не учитывая лишнюю информацию. MediaPipe выдает 3D-точки, поэтому точно определяет углы между суставами и отображает их на экране. По этой причине MediaPipe подходит для приложений, работающих в реальном времени. По сравнению с YOLOv7, MediaPipe показывает хорошие результаты на входных данных низкого разрешения. MediaPipe быстрее, чем YOLOv7, работает с процессорными выводами, также сравнительно хорошо справляется с обнаружением удаленных объектов (в нашем случае – людей). Однако, когда речь заходит об окклюзии, YOLOv7 показывает более эффективный результат. YOLOv7 также лучше справляется с анализом быстрых движений, в случае высокого разрешения входных изображений.

YOLOv7 и OpenPose – это фреймворки для обнаружения нескольких человек. MediaPipe не обладает таким функционалом, однако применительно к задаче распознавания одного единственного спортсмена в кадре это ограничение не является проблемой.

Исходя из проанализированных характеристик каждой платформы был сделан вывод, что наиболее подходящей платформой для моделирования скелетной структуры одного спортсмена из видеопотока является MediaPipe ввиду ее представления скелетной структуры в 3D пространстве, что позволяет более точно рассчитывать углы между точками скелета и оценивать качество выполнения упражнений.

2.2 Извлекаемые признаки скелетной структуры с помощью MediPipe

Модель ориентиров в MediaPipe Pose прогнозирует расположение 33 ориентиров позы. Детектируемые точки указаны на рисунке 7.

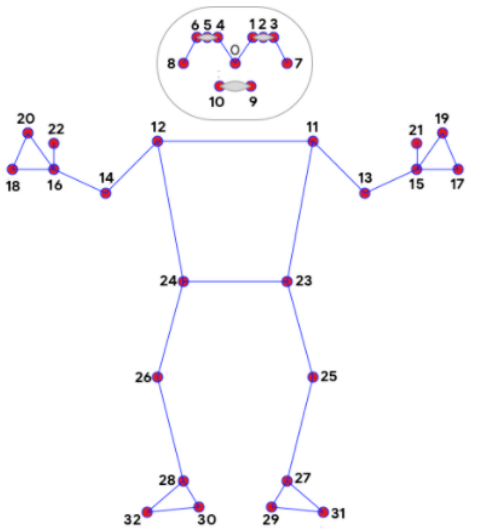

Рис. 7. Распознаваемые точки скелетной структуры

Модель возвращает следующие данные для каждого ориентира:

- x и y: координаты ориентира, нормализованные [0.0, 1.0] по ширине и высоте изображения соответственно.
- z: представляет глубину ориентира с глубиной в середине бедер, являющейся исходной точкой, и чем меньше значение, тем ближе ориентир находится к камере. Величина z использует примерно тот же масштаб, что и x.
- visibility: значение [0.0, 1.0], указывающее вероятность того, что ориентир будет виден (присутствует и не перекрыт) на изображении.

Обработка видео проходила с помощью языка программирования Python и библиотек OpenCV[21], Numpy [22], MediaPipe.

3. Предлагаемая методология распознавания действий спортсмена на видео

Авторы предлагают методологию, основанную на фреймворке глубокого обучения на базе слоев с долговременной памятью LSTM [23].

LSTM – это специализированный рекуррентный нейросетевой слой, который был специально создан в замену обычному рекуррентному слою для работы с последовательным набором данных, где важны долгосрочные темпоральные зависимости между отдельными пакетами данных. Примерами таких данных может быть текст, звук, последовательности изображений (видео) или временные ряды.

Такая архитектура нейронной сети позволяет обрабатывать входной пакет информации, основываясь на «знании» о предыдущих обработанных пакетах. Модуль LSTM углубляет эту концепцию и представляет новые части модуля, именуемые как «состояние ячейки», которое может сохранять информацию в течение длительного времени. Это состояние ячейки контролируется тремя вентилями памяти (memory gates): входным (input gate), забывания (forget gate) и вывода (output gate). Данные «вентили» определяют, какую информацию следует сохранить или отбросить из состояния ячейки.

Вентиль ввода отвечает за определение информацию, которую необходимо добавить в состояние ячейки, тогда как вентиль забывания отвечает за удаление участков информации из состояния ячейки. Вентиль вывода решает какую информацию вывести из состояния ячейки.

Архитектура LSTM выбрана за ее способность улавливать временные зависимости в последовательных данных, что делает ее особенно подходящей для динамической природы человеческих движений. Данная нейронная сеть оптимизирована как по точности, так и по интерпретируемости, что соответствует общим целям улучшения понимания сложной человеческой деятельности.

Тренировочный датасет состоит из наборов видео, соответствующих одному из 15 категорий движений, выбранных в качестве примеров для тестирования подхода. Внутри видео присутствует только один человек выполняющий действия. В качестве основы был использован датасет UCF101 – Action Recognition Data Set [24], который находится в открытом доступе.

Алгоритм обработки видео был следующим:

1) Провести итерирование по всем доступным видео покадрово;
2) Применить модель распознавания скелетной структуры человека;
3) Извлечь координаты сочленений;
4) Добавить к кадру информацию о том, к какому классу действий он принадлежит (1 из 15);
5) Сохранить извлеченную информацию в файл специального расширения «.npy».

3.1 Сбор и обработка датасета

UCF101 – это набор данных (dataset), предназначенный специально для задачи распознавания действий человека, состоящий из реалистичных коротких видеороликов, собранных с платформы YouTube и имеющих 101 категорию различных действий, относящихся к разным сферам жизнедеятельности человека. В данном датасете содержится 13 320 видеороликов, обеспечивающих наибольшее разнообразие с точки зрения действий, вариаций движения камеры, внешнего вида и поз объекта, масштаба объекта, точки обзора, загроможденного фона, условий освещения и т. д.

Видео в датасете сгруппированы в маленькие кластеры по 4-7 штук. Видео из одной группы могут иметь некоторые общие черты, такие как фон, угол обзора, освещение, удаленность объектов и т. д.

В данной работе обработка проводилась на 15 классах: 1) Archery, 2) BenchPress, 3) Biking, 4) BodyWeightSquats, 5) BoxingPunchingBag, 6) Drumming, 7) HandstandPushups, 8) HandstandWalking, 9) HighJump, 10) HorseRiding, 11) JavelinThrow, 12) JumpingJack, 13) PullUps, 14) PushUps, 15) WalkingWithDog.

Для каждого видео собирается отдельный «.npy» файл с соответствующим названием, в котором хранятся построчно информация о *x, y, z* координатах скелетной структуры, распознанных на каждом отдельном кадре.

3.2 LSTM нейронная сеть классификации

Анализ движения тела во времени и прогнозирование выполняются с использованием сети LSTM. Итак, ключевые точки из последовательности кадров отправляются в LSTM для классификации движений. Модель LSTM, которая используется для классификации действий на основе ключевых точек, обучается с помощью библиотеки машинного обучения Tensorflow [25].

Входные данные для обучения содержат последовательность ключевых точек (33 ключевых точек на кадр) и соответствующие метки действий. Непрерывная последовательность из 24 кадров используется для идентификации конкретного действия. Пример последовательности из 24 кадров будет представлять собой многомерный массив размером 24×99 следующим образом:

$$[[x_0 y_0 z_0 \dots x_{98} y_{98} z_{98}]\ [x_0 \dots [x_0 y_0 \dots y_0 z_0 \dots z_0 \dots\ \dots\ \dots x_{98} \dots x_{98} y_{98} \dots y_{98} z_{98}]\ ] \quad (1)$$

Каждая строка содержит 33 значения ключевых точек. Каждая ключевая точка представлена значениями (x, y, z), следовательно, всего 99 значений в строке.

Для обучения нейронная сеть была сконфигурирована следующим образом:

Таблица 1 – последовательность слоев нейронной сети

Слой	Выходящая размерность	Количество параметров
LSTM	(24, 128)	116736
Dropout	(24, 128)	0
LSTM	(24, 256)	394240
Dropout	(24, 256)	0
LSTM	256	525312
Batch_normalization	256	1024
Dense	256	65792
Dense	128	32896
Dense	64	8256
Dense	15	975

Общее количество параметров: 1 145 231

Обучающее количество параметров: 1 144 719

Не обучающиеся параметры: 512

Во время обучения использовался оптимизатор Adam, функция ошибки – categorical cross – entropy loss и метрика categorical accuracy. Согласно Кингма Д. П. и Б. Дж. Адам [26], этот метод «эффективен в вычислительном отношении, требует мало памяти, ин-

вариантен к диагональному масштабированию градиентов и хорошо подходит для задач, больших с точки зрения данных / параметров».

Категориальная кросс-энтропия в данной работе используется в качестве функции потерь для модели классификации с несколькими классами, где есть две или более выходных метки [27]. Выходной метке присваивается значение кодирования одной из категорий в форме 0 или 1. Выходная метка, если она присутствует в целочисленной форме, преобразуется в категориальное кодирование с использованием метода.

Категориальная точность считает частоту соответствия прогнозов реальным меткам.

Тренировка модели велась в течение 56 эпох и обучалась около часа.

Наблюдение за этапами обучения модели проводилось посредством инструмента визуализации TensorBoard [28]. TensorBoard — это набор веб-приложений для проверки и понимания ваших запусков и графиков моделей TensorFlow. TensorBoard предназначен для работы полностью в автономном режиме, не требует доступа к сети Интернет. TensorBoard предоставляет пользователю необходимый функционал для визуализации своих экспериментов, а также инструменты, необходимые для экспериментов с машинным обучением:

- Отслеживание и визуализация метрик качества и ошибок в зависимости от итерации/эпохи алгоритма;
- Визуализация динамического/статического графа модели (операции и слои);
- Отображение медиаданных (изображений, текста и аудиоданных);
- Профилирование программ TensorFlow.

Во время обучения модели были построены графики зависимости функции ошибки от эпохи выполнения алгоритма на тренировочных данных (красный) и график функции ошибки на валидации (синий) (см. рис. 8).

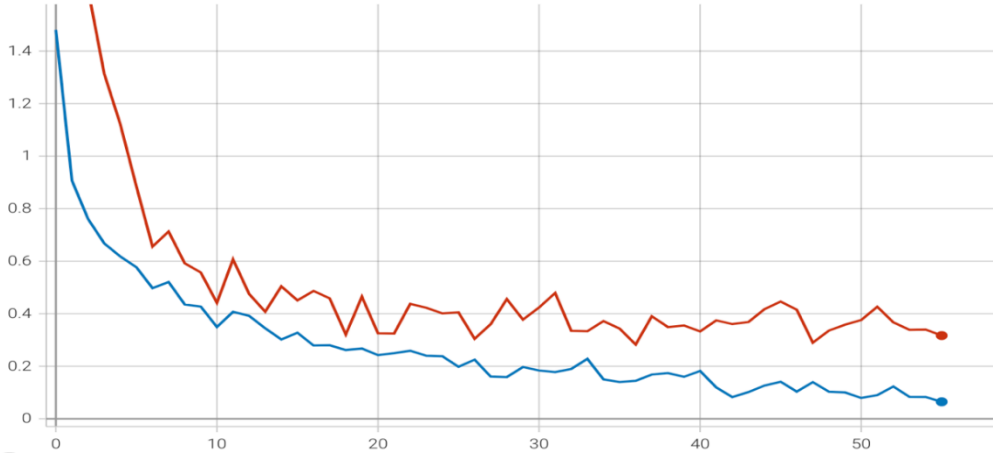

Рис. 8. График зависимости функции ошибки от эпохи выполнения алгоритма

Построение двух таких графиков позволяет найти оптимальную точку остановки обучения, то есть момент, когда дальнейшее обучение модели машинного обучения перестает приносить значимое улучшение метрик качества (высчитываются по выбранным и подходящим под конкретную задачу) на тестовом наборе данных, и начинается переобучение на тренировочном наборе данных. Одним из признаков переобучения модели является уменьшение ошибки на тренировочном наборе и увеличение/выход на плато на. Место, где ошибка на тестовом наборе данных начинает увеличиваться, обычно считается оптимальным местом для остановки обучения.

Также были построены графики, демонстрирующие изменения качества предсказаний с течением итерации обучения, по метрике категориальной точности (см. рис. 9).

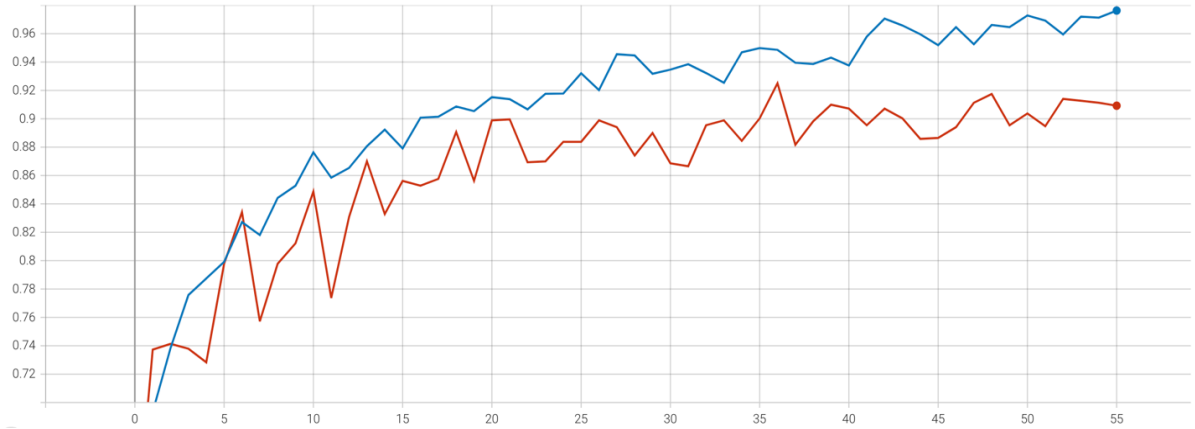

Рис. 9. График зависимости точности предсказаний от эпохи выполнения алгоритма

3.3 Результаты обучения

В процессе обучения сети рассчитываются метрики precision (2), recall (3) и функция потерь при обучении, тестировании и валидации, чтобы получить представление о том, насколько хорошо обучена сеть. Наконец, для проверки работы модели подается около 1000 тестовых видео, на которых рассчитываются метрики, в конце представлена матрица ошибок для наглядного представления качества распознавания.

$$precision(a, X) = \frac{TP}{TP + FP} \tag{2}$$

$$recall(a, X) = \frac{TP}{TP + FN} \tag{3}$$

Результаты обучения:

- Training loss: 0.0647
- Training categorical accuracy: 0.9763
- Validation loss: 0.3169
- Validation categorical accuracy: 0.9092

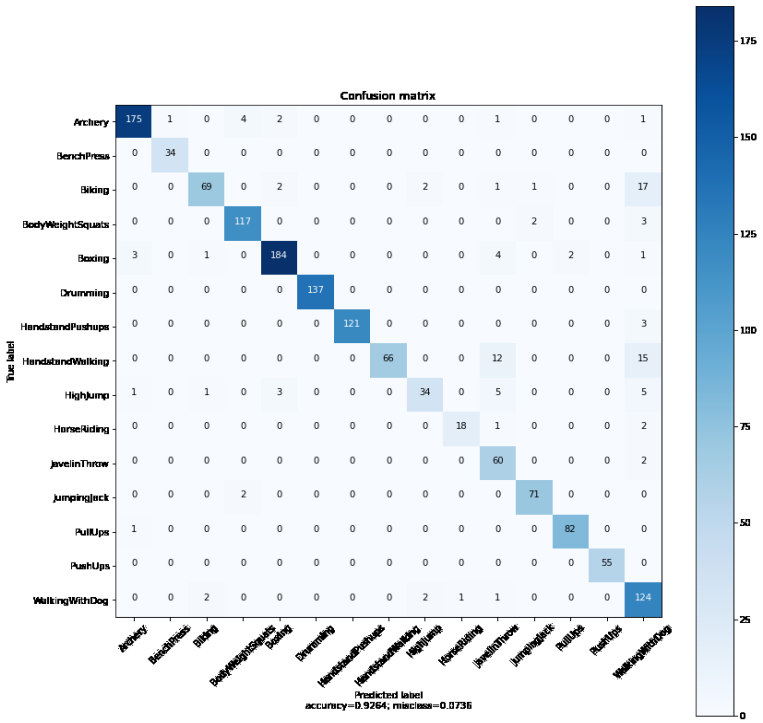

Рис. 10. Матрица ошибок

Совместно с построением матрицы ошибок (см. рис. 10), точность и полнота позволяют хорошо интерпретировать результаты, проследить соотношение предсказаний алгоритма и истинный ответ.

На рисунке 11 представлен результаты работы на видео, принадлежащим различным классам действий.

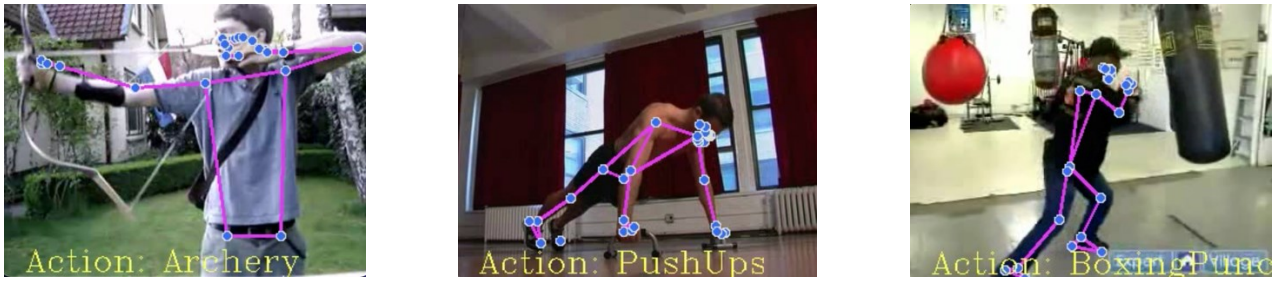

Рис. 11. Пример работы модели по видеопотоку на разных классах

4. Заключение

В ходе данной работы был произведен анализ методологий распознавания движений человека на видео. По мнению многих исследователей, одной из самых многообещающих является методология, основанная на форме. В результате сравнительного анализа платформ извлечения скелетной структуры, наиболее подходящей для задачи распознавания спортивных упражнений, выполненных единственным человеком в кадре, является платформа MediaPipe. Она обладает преимуществами в скорости работы на CPU, обработке артефактов на видео и дает трехмерные координаты точек, что позволяет более точно рассчитывать углы между сочленениями скелета, тем самым позволяет более точно оценивать качество выполнения упражнений спортсменом.

Авторы статьи предлагают собственную реализацию подхода, основанного на анализе формы человеческого тела и расположения в пространстве, используя платформу извлечения скелетной структуры MediaPipe и нейронную сеть, основанную на слоях с долговременной памятью LSTM. Данная реализация показывает достаточное качество распознавания действий на датасете UCF101 в режиме реального времени. Данная методология может быть применима, в частности, для моделирования поз спортсменов в 3D пространстве, визуализации их скелетной структуры, что позволит интерпретировать результаты и отслеживать качество выполняемых упражнений в автоматическом режиме без присутствия тренера.

Другие способы применения технологий распознавания человеческих движений включают использование в медицинской сфере для анализа движений пациентов с целью реабилитации и диагностики положений, в которых находится пациент, в игровой индустрии для создания виртуальных персонажей, отслеживания движений игроков, улучшения и создание уникального игрового опыта, а также в области безопасности для мониторинга и обнаружения подозрительных действий на видео посредством внедрения данной технологии в камеры видеонаблюдения.

Благодаря возможности визуализации скелетной структуры, исследователи и специалисты могут более наглядно представлять и анализировать данные о движениях, что существенно облегчает процесс интерпретации результатов и повышает их понятность, а также упрощает взаимодействие пользователей с любыми системами, в которых будет применяться данная технологии, в виду более просто зрительного восприятия.

5. Благодарности

Работа выполнена при финансовой поддержке РНФ (грант № 22-11-00213, https://rscf.ru/project/22-11-00213/).

Список используемой литературы:

1. Laith A., Jinglan Zhang. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. №8 (53). 2021.
2. M. Vrigkas, C. Nikou, I. A. Kakadiaris. A Review of Human Activity Recognition Methods. Robot. AI. Vol. 2. 2015.
3. Gavrila, D. M. The Visual Analysis of Human Movement: A Survey // Computer Vision and Image Understanding. 1999. №1. P. 82–98.
4. Aggarwal J.K., Cai Q. Human Motion Analysis: A Review // Computer Vision and Image Understanding. 1999. №3. P. 428–440.
5. Aggarwal J.K., Xia L. Human activity recognition from 3D data: A review // Pattern Recognition Letters. 2014. №48. P. 70–80.
6. Ghotkar A., Vidap P., Deo K. Dynamic Hand Gesture Recognition using Hidden Markov Model by Microsoft Kinect Sensor // International Journal of Computer Applications. 2016. №5. P. 5–9.
7. Li R., Zickler T. Discriminative virtual views for cross-view action recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2012. P. 2855–2862.
8. Iosifidis A., Tefas A., Pitas I. Activity-based person identification using fuzzy representation and discriminant learning // EEE Transactions on Information Forensics and Security. 2012. Vol. 7. №2. P. 530–542.
9. Chen C. Y. Grauman, K. Efficient activity detection with max-subgraph search. EEE Conference on Computer Vision and Pattern Recognition. 2012. P. 1274–1281.
10. Tran K. N., Kakadiaris I. A., Shah S. K. Part-based motion descriptor image for human action recognition // Pattern Recognition. 2012. №7. P. 2562–2572.
11. Wu Q., Wang Z., Deng F., Chi Z., Feng D. D. Realistic human action recognition with multimodal feature selection and fusion // EEE Transactions on Systems, Man, and Cybernetics: Systems. 2013. №4. P. 875–885.
12. Martinez H. P., Yannakakis G. N., Hallam J. (2014). Don't Classify Ratings of Affect; Rank Them! // IEEE Transactions on Affective Computing. 2014. № 5. P. 314–326.
13. Vrigkas M., Nikou C., Kakadiaris I. A. Classifying Behavioral Attributes Using Conditional Random Fields // Artificial Intelligence: Methods and Applications. 2014. №8445. P. 95–104.
14. Marín-Jiménez M. J., etc. Human interaction categorization by using audio-visual cues // Machine Vision and Applications volume. 2014. №25. P. 71–84.
15. OpenPose Documentation. OpenPose. 2022. URL: https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/index.html (accessed: 01.12.2022).
16. Detectron2. Meta AI. 2022. URL: https://ai.facebook.com/tools/detectron2/ (accessed: 03.12.2022).
17. MediaPipe Pose. MediaPipe. 2022. URL: https://google.github.io/mediapipe/solutions/pose.html (accessed: 05.12.2022).
18. Wang C.Y., Bochkovskiy A., Liao H. Y. YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors // Computer Vision and Pattern Recognition. 2022. 15 p. (https://arxiv.org/abs/2207.02696)
19. Radzki P. Detection of Human Body Landmarks - Mediapipe and Openpose Comparison // HearAI. 03.04.2022. URL: https://www.hearai.pl/post/14-openpose/ (accessed: 05.12.2022).
20. Kukil, V. Gupta. YOLOv7 Pose vs MediaPipe in Human Pose Estimation. LearnOpenCV. 2022. URL: https://learnopencv.com/yolov7-pose-vs-mediapipe-in-human-pose-estimation/ (accessed: 11.10.2022).
21. OpenCV. Intel. 2021. URL: https://opencv.org/ (accessed: 11.10.2022).
22. NumPy Documentation. NumPy. 2022. URL: https://numpy.org/doc/stable/ (accessed: 23.10.2022).

23. Shipra S. What is LSTM? Introduction to Long Short-Term Memory // Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/ (accessed: 16.03.2024)
24. UCF101 - Action Recognition Data Set. UCF: Center for Research in Computer Vision. 2013. URL: https://www.crcv.ucf.edu/data/UCF101.php (accessed: 22.10.2022).
25. TensorFlow Documentation. TensorFlow. 2022. URL: https://www.tensorflow.org/api_docs (accessed: 10.12.2022).
26. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // Machine Learning. Conference Paper at ICLR. 2015. 15 p. (https://arxiv.org/abs/1412.6980)
27. Kumar A. Keras – Categorical Cross Entropy Loss Function // Data Analytics. 28.10.2020. URL: https://vitalflux.com/keras-categorical-cross-entropy-loss-function/ (accessed: 15.12.2022).
28. TensorBoard. TensorFlow. 2022. URL: https://www.tensorflow.org/tensorboard (accessed: 17.12.2022).

Scientific Visualization, 2024, volume 16, number 2, pages 67 - 80, DOI: 10.26583/sv.16.2.06

Visualization and Classification of Human Movements Based on Skeletal Structure: A Neural Network Approach to Sport Exercise Analysis and Comparison of Methodologies

V.O. Kuzevanov[1], D. V. Tikhomirova[2]

National Research Nuclear University “MEPhI”, Moscow, Russia

[1] ORCID: 0009-0004-5415-1477, vl.kuzevanov@gmail.com
[2] ORCID: 0000-0002-0812-2331, dvsulim@mail.ru

Abstract

The authors of the paper review and compare different existing approaches to Human Action Recognition (HAR), analyze the advantages and disadvantages of platforms for extracting human skeletal structure from video stream, and evaluate the importance of visual representation in the motion analysis process. This paper presents an example implementation of one of the approaches to HAR based on the use of interpretability and visual expressiveness inherent in skeletal structures. In this work, an ad hoc network with Long Short-Term Memory (LSTM) for human activity classification is designed and implemented, which has been trained and tested in the domain of sports exercises. LSTM incorporation of memory cells and gating mechanisms not only mitigates the vanishing gradient problem but also enables LSTMs to selectively retain and utilize relevant information over extended sequences, making them highly effective in tasks with complex temporal dependencies. The problem with a fading gradient is quite common in deep neural networks and is that if the error is back propagated during the training of the network, the gradient can decrease strongly as it travels through the layers of the network to the initial layers. This can lead to the fact that the weights in the initial layers are practically not updated, which makes training of these layers impossible or slows down its process. The resulting solution can be used to create a real-time virtual fitness assistant. The resulting solution can be used to create a real-time virtual fitness assistant. In addition, this approach will make it possible to create interactive training applications with visualization of human skeletal structure, motion analysis and monitoring systems in the field of medicine and rehabilitation, as well as for the development of security systems with access control based on the analysis of visual data on the movement of human body parts.

Keywords: computer vision, neural network, machine learning, skeletal structure.

References

1. Laith A., Jinglan Zhang. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data. №8 (53). 2021.
2. M. Vrigkas, C. Nikou, I. A. Kakadiaris. A Review of Human Activity Recognition Methods. Robot. AI. Vol. 2. 2015.
3. Gavrila, D. M. The Visual Analysis of Human Movement: A Survey // Computer Vision and Image Understanding. 1999. №1. P. 82–98.
4. Aggarwal J.K., Cai Q. Human Motion Analysis: A Review // Computer Vision and Image Understanding. 1999. №3. P. 428–440.
5. Aggarwal J.K., Xia L. Human activity recognition from 3D data: A review // Pattern Recognition Letters. 2014. №48. P. 70–80.
6. Ghotkar A., Vidap P., Deo K. Dynamic Hand Gesture Recognition using Hidden Markov Model by Microsoft Kinect Sensor // International Journal of Computer Applications. 2016. №5. P. 5–9.

7. Li R., Zickler T. Discriminative virtual views for cross-view action recognition. IEEE Conference on Computer Vision and Pattern Recognition. 2012. P. 2855–2862.
8. Iosifidis A., Tefas A., Pitas I. Activity-based person identification using fuzzy representation and discriminant learning // EEE Transactions on Information Forensics and Security. 2012. Vol. 7. №2. P. 530–542.
9. Chen C. Y. Grauman, K. Efficient activity detection with max-subgraph search. EEE Conference on Computer Vision and Pattern Recognition. 2012. P. 1274–1281.
10. Tran K. N., Kakadiaris I. A., Shah S. K. Part-based motion descriptor image for human action recognition // Pattern Recognition. 2012. №7. P. 2562–2572.
11. Wu Q., Wang Z., Deng F., Chi Z., Feng D. D. Realistic human action recognition with multimodal feature selection and fusion // EEE Transactions on Systems, Man, and Cybernetics: Systems. 2013. №4. P. 875–885.
12. Martinez H. P., Yannakakis G. N., Hallam J. (2014). Don't Classify Ratings of Affect; Rank Them! // IEEE Transactions on Affective Computing. 2014. № 5. P. 314–326.
13. Vrigkas M., Nikou C., Kakadiaris I. A. Classifying Behavioral Attributes Using Conditional Random Fields // Artificial Intelligence: Methods and Applications. 2014. №8445. P. 95–104.
14. Marín-Jiménez M. J., etc. Human interaction categorization by using audio-visual cues // Machine Vision and Applications volume. 2014. №25. P. 71–84.
15. OpenPose Documentation. OpenPose. 2022. URL: https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/index.html (accessed: 01.12.2022).
16. Detectron2. Meta AI. 2022. URL: https://ai.facebook.com/tools/detectron2/ (accessed: 03.12.2022).
17. MediaPipe Pose. MediaPipe. 2022. URL: https://google.github.io/mediapipe/solutions/pose.html (accessed: 05.12.2022).
18. Wang C.Y., Bochkovskiy A., Liao H. Y. YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors // Computer Vision and Pattern Recognition. 2022. 15 p. (https://arxiv.org/abs/2207.02696)
19. Radzki P. Detection of Human Body Landmarks - Mediapipe and Openpose Comparison // HearAI. 03.04.2022. URL: https://www.hearai.pl/post/14-openpose/ (accessed: 05.12.2022).
20. Kukil, V. Gupta. YOLOv7 Pose vs MediaPipe in Human Pose Estimation. LearnOpenCV. 2022. URL: https://learnopencv.com/yolov7-pose-vs-mediapipe-in-human-pose-estimation/ (accessed: 11.10.2022).
21. OpenCV. Intel. 2021. URL: https://opencv.org/ (accessed: 11.10.2022).
22. NumPy Documentation. NumPy. 2022. URL: https://numpy.org/doc/stable/ (accessed: 23.10.2022).
23. Shipra S. What is LSTM? Introduction to Long Short-Term Memory // Analytics Vidhya. URL: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/ (accessed: 16.03.2024)
24. UCF101 - Action Recognition Data Set. UCF: Center for Research in Computer Vision. 2013. URL: https://www.crcv.ucf.edu/data/UCF101.php (accessed: 22.10.2022).
25. TensorFlow Documentation. TensorFlow. 2022. URL: https://www.tensorflow.org/api_docs (accessed: 10.12.2022).
26. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization // Machine Learning. Conference Paper at ICLR. 2015. 15 p. (https://arxiv.org/abs/1412.6980)
27. Kumar A. Keras – Categorical Cross Entropy Loss Function // Data Analytics. 28.10.2020. URL: https://vitalflux.com/keras-categorical-cross-entropy-loss-function/ (accessed: 15.12.2022).
28. TensorBoard. TensorFlow. 2022. URL: https://www.tensorflow.org/tensorboard (accessed: 17.12.2022).