Применение методов глубокого обучения для распознавания эмоционального состояния человека на видеоизображении

А. А. Москвин, А. Г. Шишкин

Московский государственный университет им. М. В. Ломоносова, г. Москва, Россия

Аннотация. Представлены метод и архитектура разработанной нейроннной сети, позволяющие определять в режиме реального времени при ограниченных вычислительных ресурсах эмоциональное состояние человека по видеопоследовательности в которой присутствует как речевой сигнал, относящийся к источнику для которого нужно определить состояние, так и его лицо. Визуальная информация показана с помощью 16 последовательных кадров размером 96x96 пикселей, а аудиоинформация - с помощью 140 характерных признаков для последовательности из 37 окон. На основе экспериментальных исследований разработана архитектура нейросетевой модели с использованием сверхточных и рекуррентных нейронных сетей. Использование аудиоинформации совместно с визуальной информацией позволяет увеличить точность распознавания на 12%.

Ключевые слова: искусственные нейронные сети, глубокое обучение, распознавание эмоций, видеоизображение, речевой сигнал.

DOI 10.14357/20718594190201

Введение

В настоящее время активно применяются адаптивные технологии, включающие широкий спектр программно-аппаратных решений, позволяющих изменять свое поведение во время работы. К числу областей, где адаптивные технологии играют существенную роль, относится распознавание специфических признаков объектов в заданном потоке данных.

Выделение признаков, основанное на глубоком обучении, зачастую решает поставленные задачи даже эффективнее, чем человек, особенно при больших объемах данных, где необходимо найти нелинейные зависимости между ними. В качестве одного из примеров можно привести распознавание лиц и их отдельных деталей на изображении.

Интерес к данным задачам значителен в связи с разнообразием их применения в таких областях, как охранные системы, верификация пользователей, телеконференции, компьютерные игры, медицина и т.д. Реализация технологии распознавания эмоционального состояния человека по видеопоследовательности поможет людям, страдающим психоневрологическими расстройствами, в определении интонационной окраски речи, предоставит возможность отделения наигранной эмоции от настоящей.

Процесс определения эмоций по видеопоследовательности состоит из двух больших этапов: поиск и выделение лиц на отдельных изображениях, и анализ входных данных с целью установления типа эмоции. При поиске лиц на видеоизображениях на данный момент используются в основном следующие методы: метод главных компонент [1], алгоритм AdaBoost [2], сравнение шаблонов [3].

✉ *Москвин Алексей Алексеевич. E-mail: alalmoskvin@gmail.com*

В области распознавания различных эмоций на лицах людей, как на отдельных снимках, так и на видеоизображениях, в последние годы был получен ряд интересных результатов. Например, программное обеспечение, представленное на конференции Future Decoded 2014 в Великобритании, способно анализировать эмоции человека по выражению его лица [4]. Алгоритм определяет наличие на отдельном снимке таких базовых эмоций, как гнев, презрение, отвращение, страх, счастье, спокойствие, печаль и удивление. Результаты, представленные в числовом формате, принадлежат промежутку от 0 до 1 и представляют собой вероятность наличия определенной эмоции, где 0 - отсутствие эмоции, а 1 - явно выраженная эмоция. Распознавание эмоций также является частью более крупного проекта Microsoft Project Oxford [5]. Помимо идентификации эмоций на изображении, в него включены орфографическая проверка текста на английском языке и распознавание речи.

Итоги конкурса INTERSPEECH 2009 Emotion Challenge [6] показали, что точность распознавания эмоционального человека с помощью одного только аудиоканала далека от идеальной. Так, были получены следующие результаты: средняя эффективность на двух классах 71,86%, на пяти классах - 58,83%. Следствием этого стало появление работ, в которых используются различные типы сигналов, описывающих эмоции. В работе [7] рассмотрены мультимодальные (изображение, речевой и тактильный сигналы) ограниченные машины Больцмана (ОМБ), способные как порождать, так и распознавать шесть эмоциональных состояний, которые представляются как активации нейронов верхнего слоя. Построенная модель позволяет воспроизводить эмоцию, соответствующую отсутствующей модальности на основе двух других.

Для классификации эмоциональных состояний в реальном времени по видеопоследовательности, как правило, используются сверточные нейронные сети. В работе [8] для обучения нейронной сети использовалось несколько наборов данных. Достигнута эффективность распознавания – 57,1%. Существенным недостатком работы являются небольшие размеры использованных наборов данных, в которых представлены весьма схожие между собой видеопоследовательности.

В работе [9] при помощи сверточных сетей и классификатора Байеса все изображения разбивались на 3 класса. Для анализа использовалось 6470 изображений. База данных была собрана из фотографий различных социальных мероприятий, таких как свадьба - для позитивных, собрания - для нейтральных и протесты - для негативных эмоциональных состояний. Точность на тестовой выборке составила 64,68%. Анализ результатов показал, что для одних типов эмоций предпочтительнее метод Байеса, а для других - сверточная нейросеть, однако, их совместное использование всегда показывает более высокую эффективность, чем каждый метод по отдельности.

Гибридный подход к решению проблемы, в котором за распознавание эмоций на основе аудио- и видеосигналов отвечают различные методы, представлен в [10]. Так, визуальные данные из видеоизображения обрабатываются сверточными нейросетями, а аудиопоток анализируется на основе глубоких сетей доверия. Помимо этого, авторы выделяют характерные эмоциональные признаки вокруг рта с помощью метода К-средних и применяют автокодировщик для описания пространственно-временной информации. На престижном конкурсе EmotiW 2014 разработанная модель показала общую эффективность распознавания семи базовых эмоций, равную 47,67%. В статье 2018 года “Context-aware Cascade Attention-based RNN for Video Emotion Recognition” использовался только видеоканал для получения информации о представленном классе, как результат – итоговая точность распознавания составляла 45,51% [11]. Отличный результат – свыше 99% показала работа с использованием анализа электроэнцефалографии головного мозга [12].

В данной работе с использованием технологии глубоких нейронных сетей разработана и программно-реализована модель, позволяющая распознавать в режиме реального времени при ограниченных вычислительных ресурсах эмоции человека по видеопоследовательности, в которой присутствует как визуальная, так и речевая информация.

1.Предложенный метод

В основе разработанного метода лежит применение глубоких нейронных сетей к видео- и аудио каналам источника информации. Затем по результату анализа информации из каждого канала происходит принятие решения о принадлеж-

ности видеозаписи к одному из семи классов - нейтральное состояние, злость, грусть, испуг, разочарование, радость, удивление. Обработка каждого из каналов данных представляет собой отдельную большую подзадачу. Для видеоканала можно выделить следующие основные этапы:

• разбиение видеопоследовательности на отдельные кадры;

• поиск и выделение лиц из изображений;

• предварительная обработка изображений;

• использование полученных изображений в качестве входных данных для нейронной сети.

Первый этап представляет собой разбиение каждой видеозаписи на одинаковое число, а именно, 16 изображений, каждое из которых должно содержать лицо. На данном шаге основной проблемой является выбор изображения, на котором объект не находится в движении, т. е. не запечатлен момент, когда объект имеет расплывчатые очертания. Предпочтительнее использовать один из кадров видеопоследовательности на котором границы объекта являются более четкими. Далее с помощью метода AdaBoost [2] выполняется поиск и выделение лица на изображении. На этапе предобработки все лица приводятся к унифицированному формату, а также исключаются изображения, которые не могут быть использованы по тем или иным причинам, например, их размер составляет менее 50 пикселей по высоте и ширине. Последний этап заключается в передаче изображения на вход ранее предобученной нейронной сети, которая имеет 7 выходов, соответствующих вероятностям принадлежности данного ряда изображений каждому из рассматриваемых классов.

С точки зрения логики работы этапы обработки аудиоканала, приведенные ниже, аналогичны его анализу, но внутренняя составляющая этапов отличается:

• разбиение аудиосигнала на отдельные окна;

• выделение характерных признаков из окна;

• использование полученных характерных признаков в качестве входных данных для рекуррентной сети.

В нашем случае исследуемый аудиосигнал разделялся на отдельные окна на основе метода периодограмм Уэлча [13] – вектор отсчетов сигнала делился на перекрывающиеся сегменты (как правило, использовалось 50%-е перекрытие), после чего каждое окно умножалось на весовую функцию и для него вычислялось дискретное преобразование Фурье (ДПФ). Так как в нашей задаче важна локализация параметров, то, помимо ДПФ, также использовалось дискретное косинусное преобразование. Исходя из практических соображений, длина окна выбиралась равной 20 мс. Для каждого окна вычислялось большое число характерных признаков как во временной, так и в частотной областях. Далее из выделенных характерных признаков выбирались наиболее существенные, устойчивые к внешним шумам и позволяющие адекватно описывать данный отрезок сигнала для анализа в реальном времени. Заключительный этап аналогичен последнему этапу обработки видеоканала, отличие состоит лишь в архитектуре использованной нейронной сети.

Данные. Одной из важных задач работы были поиск и получение наборов данных для последующего анализа. Почти все существующие базы данных состоят из файлов, на которых присутствуют видеоизображения с лицами людей, но их принадлежность к одному из рассматриваемых эмоциональных состояний неизвестна, а звуковая «дорожка» или отсутствует, или состоит из внешнего шума. Вследствие этого были рассмотрены различные варианты создания набора данных, а именно, видеохостинги, художественные фильмы, готовые базы данных, видеозаписи из социальных сетей.

В результате нами было принято решение использовать фрагменты художественных фильмов, дополнив их наиболее подходящим к поставленной задаче набором данных TR-CS-11-02 Acted Facial Expressions In The Wild Database [14], для которого была выполнена необходимая обработка. В этой базе данных представлена выборка, разделенная на 7 искомых классов, однако, некоторые видеопоследовательности были неудовлетворительно короткими (менее 1 с), а на некоторых были только лица в профиль. Окончательное число использованных нами видеопоследовательностей: нейтральное – 207, злость – 197, грусть – 178, испуг – 127, разочарование – 114, счастье – 113, удивление - 120.

2.Метод выбора отдельных изображений из видеопоследовательности

Для использования видеоканала необходимо выбрать последовательность изображений, удовлетворяющую следующим свойствам:

наличие лица на каждом изображении, хорошее качество каждого изображения.

После этого можно разбить все видеопоследовательности на равные части и в каждой точке разбиения взять отдельный кадр. Однако в этом случае вследствие движения объектов существует вероятность того, что выбранное изображение будет содержать нечеткое («размытое») лицо, что в дальнейшем скажется на качестве обучения. Соответственно, возникает необходимость выбора кадра, не имеющего размытия, для распознавания.

2.1. Создание базы данных размытых и четких изображений

Для определения того, является ли анализируемый объект на рассматриваемом изображении четким или размытым, была создана отдельная база данных, куда были включены изображения с объектами обоих типов. За основу базы были взяты случайные кадры из видеопоследовательностей. Так как ручное разделение случайных изображений на четкие и размытые является продолжительным по времени процессом, то к каждому изображению был применен один из матричных фильтров: низкочастотный $L = \frac{1}{14}\begin{bmatrix}1 & 2 & 1\\ 2 & 2 & 2\\ 1 & 2 & 1\end{bmatrix}$ (записывался в категорию размытых изображений) и высокочастотный $H = \begin{bmatrix}0 & -1 & 0\\ -1 & 5 & -1\\ 0 & -1 & 0\end{bmatrix}$ (соответствует четкому изображению) [15]. Так как размытым является движущийся объект, то следует применять фильтр только к той части изображения, которая содержит лицо, а не ко всему изображению. Границы применимости фильтра выбирались опытным путем, на основе того, какую часть кадра занимает лицо. Перед применением каждого фильтра с помощью метода AdaBoost [2] находился центр лица, координаты которого являлись центром квадратной области применимости фильтра. Размер области задавался случайным образом в пределах от 30 до 60% от всего изображения.

2.2. Определение типа изображения

Для определения типа изображения - размытое или четкое - была построена нейронная сеть прямого распространения с шестью полносвязными слоями. Для каждого изображения в градациях серого из обучающей выборки были построены гистограммы, представляющие собой вектора значений от 0 до 255, которые использовались в качестве входов нейронной сети. Достаточно простая архитектура нейронной сети объяснялась требованиями решения исходной задачи в реальном времени. Однако она не позволила получить удовлетворительные результаты. Поэтому были добавлены еще два слоя с целью внесения во входные данные шумов. Результаты данной модели также не были удовлетворительными в полной мере. Для визуализации данных были построены гистограммы изображений. На Рис. 1 на каждом графике представлены сравнения количества пикселей определенного значения у размытого изображения (светло-серый), выбранных случайным образом.

Затем к анализируемым изображениям были применены фильтры Собеля по разным координатам и фильтр Лапласа. Найдены гистограммы каждого из этих изображений, представляющие собой вектора, размерностью 3×256, и поданы на

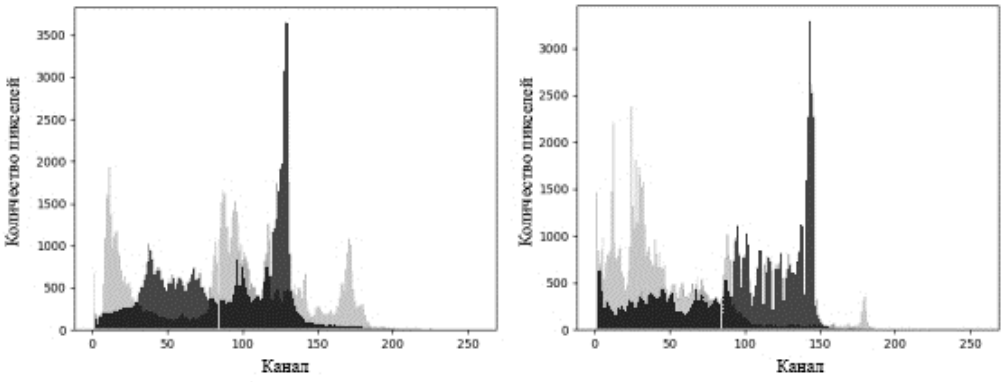

Рис. 1. Сравнение количества пикселей определенного значения выбранных случайным образом у размытого (светло-серый) и четкого (темно-серый) изображений

вход нейронной сети. Точность определения составила свыше 98%. Данная модель использовалась в дальнейшем при выборе необходимого кадра из видеопоследовательности.

3. Выделение характерных признаков из аудиоканала

Амплитуда речевого сигнала существенно изменяется во времени. В частности, амплитуда невокализованных сегментов речевого сигнала значительно меньше амплитуды вокализованных сегментов. Как результат, необходимо выделять достаточно короткие сегменты речевого сигнала и уже в них определять разлиные характерные признаки.

Для решения поставленной задачи речевой сигнал разбивался на отдельные окна длительностью 20 мс. Затем для каждого окна при помощи программы OpenSmile [16] были выделены следующие характерные признаки.

1. Кепстральные коэффициенты нелинейного масштаба или иначе мел-частотные кепстральные коэффициенты (15 первых коэффициентов). Как известно, речеобразование зависит, в первую очередь, от формы голосового тракта, определяющей, какие звуки произносятся в данный момент. Форма голосового тракта отражается, например, в огибающей кратковременного спектра мощности речевого сигнала. Для представления этой огибающей с помощью небольшого объема данных предназначены мел-частотные кепстральные коэффициенты. Мел – единица высоты звука, основанная на восприятии этого звука органами слуха. Зависимость между частотой звука и мелами можно описать простой формулой: $B(f)=1125\ln(1+f/700)$.
2. Коэффициенты линейного предсказания (8 первых коэффициентов).
3. Частота нулевых пересечений. Она низка для вокализованных фрагментов речи и высока для невокализованных [17].
4. Частота основного тона.
5. Среднее значение спектра анализируемого речевого сигнала.
6. Нормализованные средние значения спектра.
7. Интенсивность, амплитуда, энергия, номер окна.

Также были добавлены статистические характеристики и моменты для вышеописанных параметров: среднее арифметическое огибающей сигнала, стандартное отклонение значений огибающей, максимальное значение, минимальное значение, дельта коэффициенты, момент третьего порядка, момент четвертого порядка, доля времени, когда значение величины превышает 75% от максимума, доля времени, когда значение величины превышает 90% от максимума, первый квартиль, второй квартиль, третий квартиль.

Таким образом, для каждого окна было получено 562 параметра. Вследствие большого числа характерных признаков необходимо было уменьшить их количество, оставив только наиболее существенные.

3.1. Ковариация параметров

Для вычисления оптимальной размерности итогового количества характерных признаков аудиоканала был применен метод ковариации параметров, заключающийся в следующем:

1. Выделить все пары параметров аудиоканала, коэффициент ковариации которых больше некоторого порогового значения.
2. Из всех пар выбрать только те, которые встречаются в более чем 50% аудиодорожках, которые были исследованы.
3. Построить граф связности и выделить в нем компоненты связности.
4. Из каждой компоненты выбрать любой параметр, который в дальнейшем окажется в результирующем множестве.

В аудиодорожках для всех семи классов были найдены пары характерных признаков, коэффициент ковариации которых больше, чем величина C, где

$C \in [0.50,0.55,060,065,0.70,0.75,0.80,0.85,0.90]$.

Были проанализированы данные количества различных пар в аудиодорожках определенного класса, на основе которых можно сделать вывод, что рассматриваемые классы не обладают специфическими свойствами наличия слишком большого или слишком малого количества пар ковариации по отношению к другим, т. е. описанный выше алгоритм не зависит от класса эмоции.

Всего было проанализировано 1136 аудиодорожек, из которых были выбраны такие пары, которые встречаются более чем в 568 сигналах. Найдены компоненты связности для каждого значения границы C, после чего по-

строены зависимости размера получившегося результирующего множества пар ковариаций и компонент связности от коэффициента ковариации, которые показаны на Рис. 2.

Для минимизации размерности и максимизации ковариации было выбрано значение коэффициента ковариации, равное 0,90. Данное решение принято на основе Рис. 2, из которого следует, что при 90%-ном совпадении более половины признаков исчезает, что является достаточным для решения поставленной задачи условием.

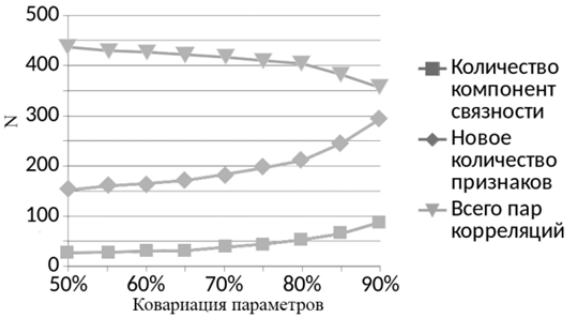

Рис. 2. Зависимость N от процента ковариации параметров

N – количество компонент связности (показано квадратами), новое количество признаков (ромбами) или общее количество пар корреляций (треугольниками)

3.2. Автокодировщик

Автокодировщик представляет собой нейронную сеть, состоящую из двух логических частей – кодирование и декодирование. Кодирование представляет собой некоторую функцию $g: g(x) = x', dim(x) > dim(x')$, а декодирование - функцию $f: f(x') = x'', dim(x) = dim(x'')$, где x' - вектор новых полученных параметров. Требуется выбрать функции f и g так, чтобы функция ошибки $L(x, x'') \to 0$. Для поставленной задачи уменьшения количества параметров на входной слой была подана последовательность из 30 векторов, каждый из которых содержал 562 параметра. При кодировании исходных параметров временное следование параметров не влияет на их сжатие, поэтому автокодировщик не содержит рекуррентных слоев, а только полносвязные, и имеет форму, представленную в Табл. 1.

Табл. 1. Структура автокодировщика

Тип слоя	Количество входов	Количество выходов
Полносвязный слой 1	562	400
Полносвязный слой 2	400	256
Полносвязный слой 3	256	400
Полносвязный слой 4	400	562

После анализа различных метрик, стало понятно, что из разницы в единицах измерения данных и необходимости понимания, насколько каждый полученный признак отличается от оригинального, была выбрана следующая функция ошибки: $\frac{1}{N}\sqrt{\sum_i |\log(x_i'') - \log(x_i)|^2}$.

3.3. Виды слоев нейронных сетей

Полносвязный слой. Слой, в котором каждый нейрон соединен со всеми нейронами на предыдущем уровне, причем каждая связь имеет свой весовой коэффициент.

Сверточный слой. В отличие от полносвязного, в сверточном слое нейрон соединен лишь с ограниченным количеством нейронов предыдущего уровня, т. е. сверточный слой аналогичен применению операции свертки, где используется лишь матрица весов небольшого размера (ядро свертки), которую «двигают» по всему обрабатываемому слою. Еще одна особенность сверточного слоя в том, что он немного уменьшает изображение за счет краевых эффектов.

Слой подвыборки. Слои этого типа выполняют уменьшение размерности (обычно в несколько раз). Это можно делать разными способами, но зачастую используется метод выбора максимального элемента (max-pooling) – вся карта признаков разделяется на ячейки, из которых выбираются максимальные по значению.

Слой прореживания. Это способ борьбы с переобучением в нейронных сетях, обучение которых обычно производят стохастическим градиентным спуском, случайно выбирая некоторые объекты из выборки. Регуляризация заключается в изменение структуры сети: каждый нейрон выбрасывается с некоторой вероятностью p. По такой прореженной сети производится обучение, для оставшихся весов делается градиентный шаг, после чего все выброшенные нейроны возвращаются в нейронную сеть. Таким образом, на каждом шаге стохастического градиента мы настраиваем одну из возможных 2N-архитектур сети, где под архитектурой мы понимаем структуру связей

между нейронами, а через N обозначаем суммарное число нейронов. При тестировании нейронной сети нейроны уже не выбрасываются, но выход каждого умножается на (1 − p). Благодаря этому на выходе нейрона мы будем получать математическое ожидание его ответа по всем 2N-архитектурам. Таким образом, обученную с помощью регуляризации нейронную сеть можно рассматривать как результат усреднения 2N-сетей.

Рекурентный слой. Чтобы перейти от полносвязных сетей к рекуррентным достаточно связать нейроны не только со следующим и предыдущим слоями, но и между собой. Такая реализация позволяет уйти от четко заданного размера входных данных - те нейроны, на которые информация не пришла, будут обрабатывать информацию своих соседей. Также система становится независимой от позиции параметра в исходной последовательности.

4. Структура нейронных сетей и их обучение

Для решения задачи распознавания эмоционального состояния человека с помощью аудиовизуальной информации были использованы различные типы нейронных сетей, а именно, сверточные и рекуррентные сети. Для обработки изображений применялись сверточные сети, а для анализа аудиоканала – рекуррентные.

Для аудиоканала в качестве итоговой модели была реализована рекуррентная нейронная сеть, состоящая из девяти слоев и принимающая на вход матрицу 37 x 256, где 256 – количество параметров одного фрейма аудиодорожки, а 37 – количество окон. На выходе получается вектор из семи действительных чисел. Каждое число, лежащее на отрезке от 0 до 1, показывает, какова вероятность принадлежности исходных параметров одному из классов. Структура рекуррентной сети представлена на Рис. 3.

С помощью данной модели уже к 120 эпохе была достигнута точность 39%, которая в дальнейшем не изменялась. В то же самое время, как видно из Рис. 4, точность на обучающей выборке равна 100% . Это говорит о наличии переобучения – состояния системы, когда модель использует признаки, характерные только для обучающей выборки, и не способна к обобщению [18]. Чтобы решить проблему переобучения был использован метод перекрестной проверки (cross-validation), который заключается в разбиении имеющихся данных на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге тестирование проводится на каждой из k частей данных. Также в модель нейронной сети были добавлены два слоя прореживания, в которых часть нейронов при обучении *исключается* из сети с вероятностью p. Таким образом, вероятность того, что нейрон останется в сети, составляет q=1−p. Таким образом, вероятность того, что нейрон останется в сети, составляет q=1−p. «Исключение» нейрона означает, что при любых входных данных или параметрах он

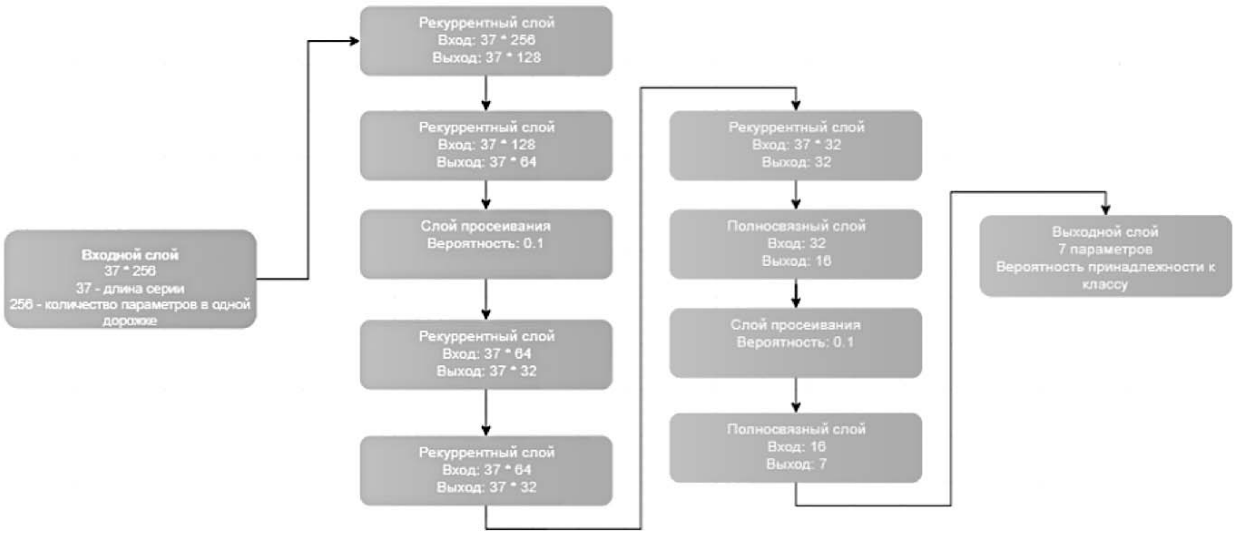

Рис. 3. Графическое представление структуры нейронной сети аудиоканала

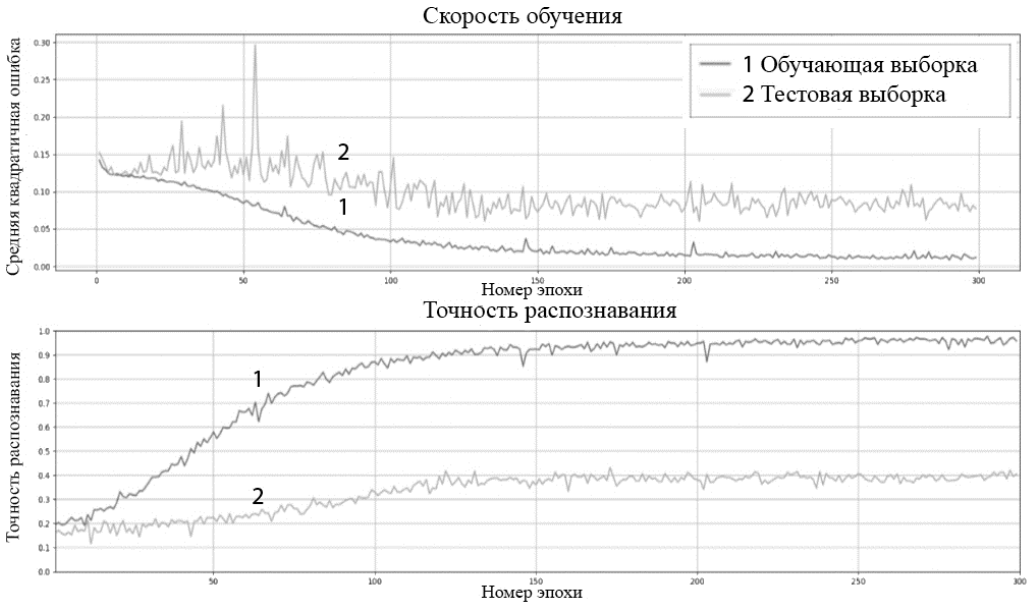

Рис. 4. Результаты обучения и тестирования нейронной сети для аудиоканала

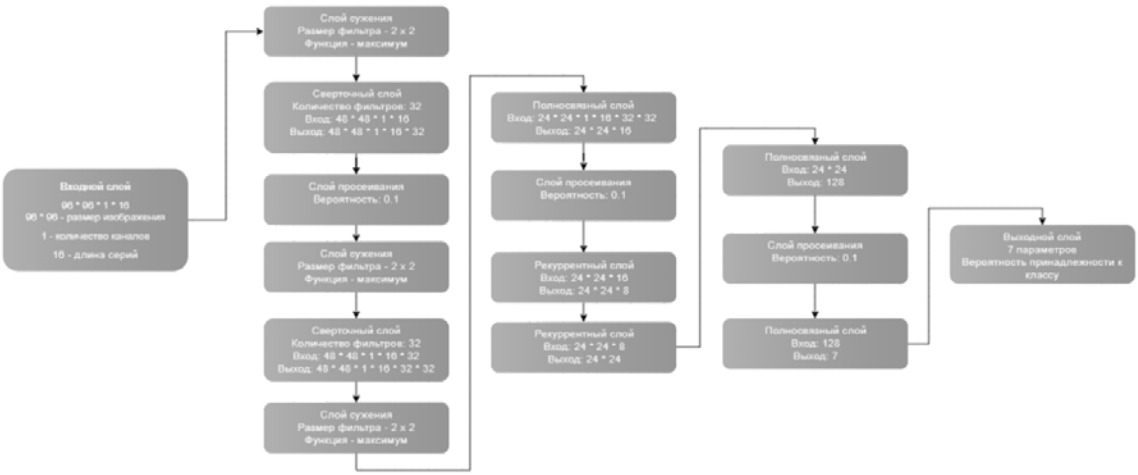

Рис. 5. Графическое представление структуры нейронной сети для видеоканала

возвращает нулевое значение. Исключенные нейроны *не вносят свой вклад* в процесс обучения ни на одном из этапов алгоритма обратного распространения ошибки; поэтому исключение хотя бы одного из нейронов равносильно обучению новой нейронной сети.

При анализе видеоизображения использовалась более сложная, представленная на Рис. 5 модель, которая принимала на вход 16 изображений 96x96 пикселей в градациях серого и включала в себя 16 слоев, среди которых 2 сверточных слоя, 3 слоя подвыборки, 3 слоя прореживания, 3 рекурентных слоя и 4 полносвязных слоя.

Обучение модели происходило быстрей, чем в случае с аудиоканалом, и к 70 эпохе была достигнута максимальная точность на обучающей выборке (Рис. 6). После этого наблюдались небольшие колебания на тестовой выборке в пределах от 45 до 50%. Как результат, был выбран пик с 49%-ной точностью, которую и было принято считать итоговой.

5. Анализ полученных результатов

Результаты работы предложенного метода на фиксированной тестовой выборке можно иллюстрировать с помощью кривой ошибок, а качество оценить как площадь под этой кривой. Кривая ошибок представляет собой зависимость доли верных положительных классификаций от доли ложных при изменении порогового значения решающего правила и является

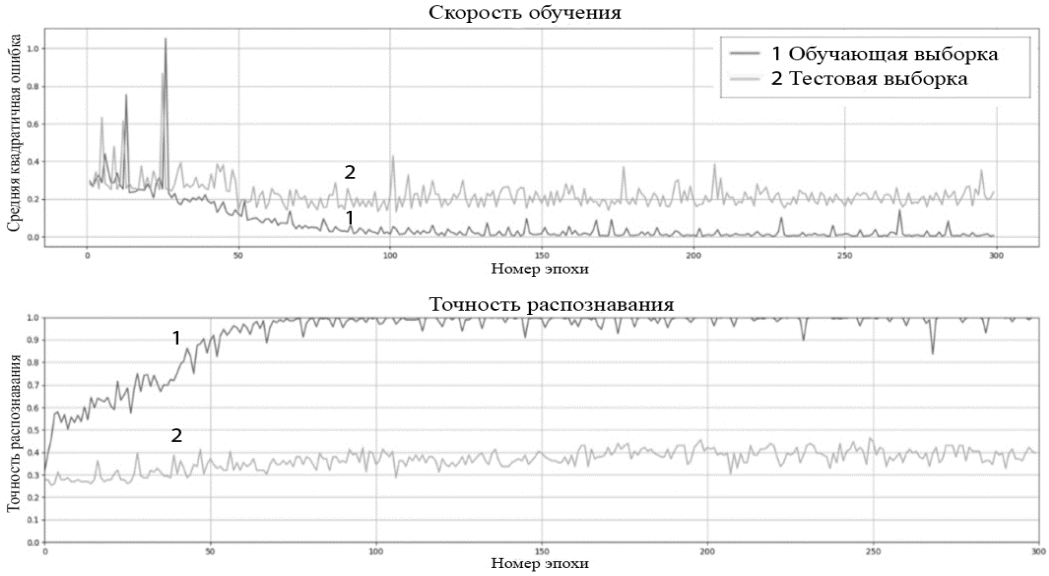

Рис. 6. Результаты обучения и тестирования нейронной сети для видеоканала

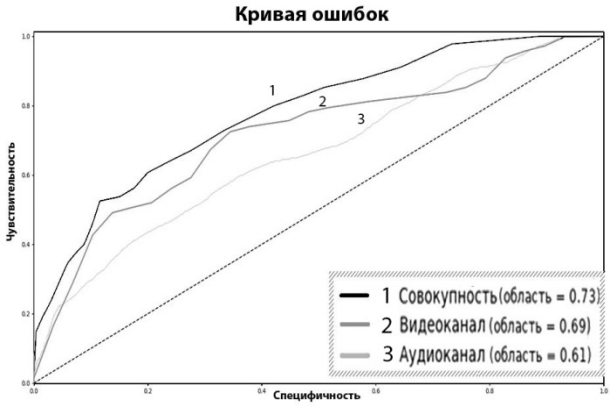

Рис. 7. Кривая ошибок

одним из самых популярных функционалов эффективности решения задач бинарной классификации [19]. Для того чтобы расширить эти понятия на задачи небинарной классификации можно использовать парное сравнение, т.е. рассмотреть множество задач бинарной классификации по типу один-против-всех. На Рис. 7 представлена общая кривая ошибок для задачи распознавания эмоций при использовании видео- и аудиоканалов, а также кривые ошибок при использовании только изображений или только аудиоканала. Видно, что распознавание эмоционального состояния в том случае, когда задействована вся доступная информация, является более эффективным по сравнению с использованием отдельных каналов.

Для совмещения результатов двух нейронных сетей использовалась элементарная нейронная сеть из двух нейронов, принимающей на вход два параметра – результат нейронной сети аудиоканала и результат нейронной сети видеоканала, а на выходе получалось принадлежность к классу.

Из Рис. 8, где представлены результаты обучения нейронной сети на совокупности видео- и аудиоканалов, следует, что после 20 эпох была получена точность, равная 69%. На тестовой выборке точность достигла максимального значения–59%.

Анализ матрицы ошибок, представленной на Рис. 9, показал, что чаще всего ошибки возникали в случае эмоций, которые физиологически похожи. Так, например, в 61% случаев злость распознавалась, как разочарование, и в 42% страх определялся как разочарование.

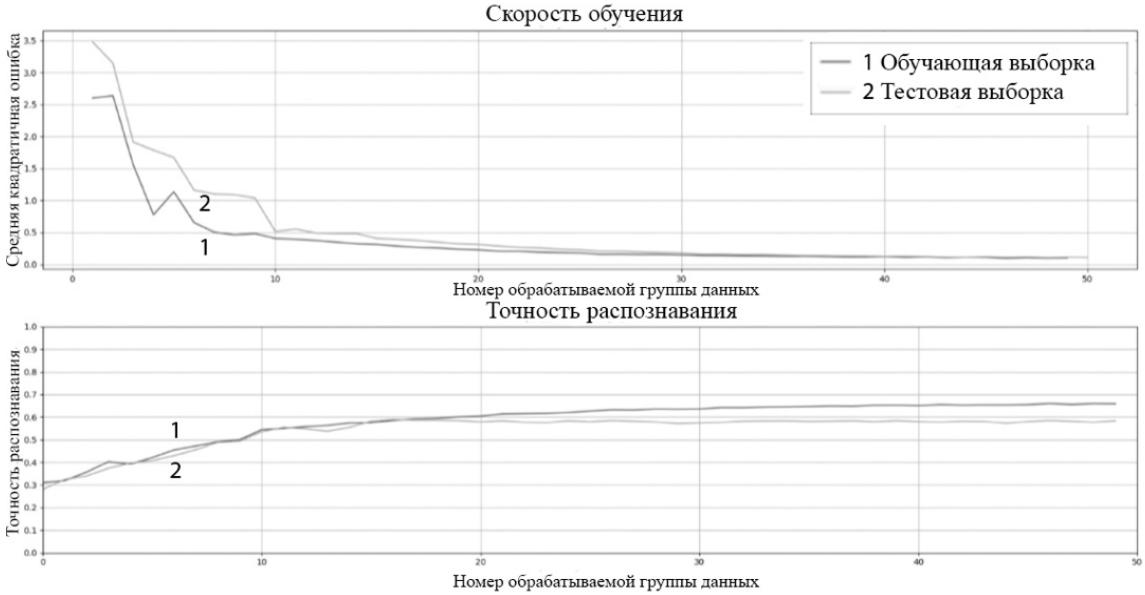

Рис. 8. График обучения и тестирования нейронной сети для совокупности каналов

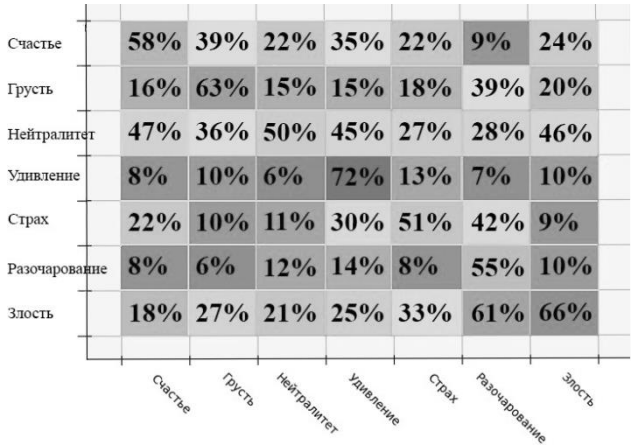

Рис. 9. Матрица ошибок для семи классов эмоций

Заключение

В данной работе рассмотрена задача определения эмоционального состояния человека по видеоизображению. Для классификации одной из семи заданных эмоций предложено использовать глубокие нейронные сети, а именно, сверточные и рекуррентные нейронные. Дополнительную сложность задаче придает тот факт, что эмоциональное состояние человека может изменяться на протяжении одной и той же видеопоследовательности.

Для обучения нейронных сетей была собрана база данных из фрагментов художественных фильмов, а также использована свободно распространяемая база TR-CS-11-02 Acted Facial Expressions In The Wild Database [14]. Вследствие передвижения людей их лица на отдельных кадрах могут быть размытыми. Для выбора четких изображений, необходимых для обучения модели, была создана специальная нейросеть со скромными требованиями к вычислительным ресурсам.

Реализовано автоматическое структурирование данных и приведение их к унифицированному формату, включающее в себя создание выборки для обучения, создание тестовой выборки, обработка данной выборки путем приведения к одинаковому размеру и удаление частей изображения, не влияющих на классификацию. Для определения вероятности принадлежности эмоционального состояния к одному из семи возможных классов разработана архитектура и программно реализована глубо-

кая нейросеть, состоящая, в том числе, из сверточных и рекуррентных слоев. Для решения проблемы переобучения нейронной сети использовался метод оценки аналитической модели и ее поведения на независимых данных, а также были добавлены прореживающие слои в глубокую нейросеть.

На основе ряда проведенных экспериментов были выбраны оптимальные параметры нейронной сети, а также данные для обработки изображений, благодаря которым точность классификации эмоций составила 59%. При этом на компьютере с процессором Intel(R) Core(TM) i5-6200U @ 2.30GHz и 8 ГБ ОЗУ скорость обработки запроса была менее 0,5с, что дает возможность решать задачу в реальном времени.

Следует отметить, что использование аудиоинформации, помимо визуальной, позволило значительно увеличить вероятность корректного распознавания эмоционального состояния человека.

В будущем планируется использовать для обучения модели предобученную на статических изображениях лиц нейронную сеть, реализовать возможность распознавания эмоции в случае различных поворотов головы.

Литература

1. Спицын, В. Г., Болотова, Ю. А., Шабалдина, Н. В., Тхи, Б., Чанг, Т., Хоанг, Ф. Н. Распознование лиц на основе методы главных компонент с применением вейвлет-дескрипторов Хаара и Добеши. Ba Ria – Vung Tau University, Ba Ria – Vung Tau, Vietnam.
2. Huang X., Acero A., Hon H.-W. Spoken language processing: a guide to theory, algorith, and system development // Prentice Hall PTR. 2001.
3. Ashok V., Balakumaran T., Gowrishankar C., Vennila I. L. A., Kumar, A. N. The Fast Haar Wavelet Transform for Signal & Image Processing. Arxiv Preprint ArXiv, 2010. *7*(1), P.126–130.
4. Chronaki G., Hadwin J. A., Garner M., Maurage P., Sonuga-Barke E. J. S., The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. Br. J. Dev. Psychol., 2015, vol. 33, no. 2, P. 218–236.
5. Jorda M., Instructor N. M., Ng A., CS229: Machine Learning techniques - Emotion Classification on face images. 2015.
6. Fan Y., Lam J. C. K., Li V.O.K. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. Retrieved, 2017.
7. Horii T., Nagai Y., Asada M. Emotion Recognition and Generation through Multimodal Restricted Boltzmann Machines // Proceedings of the IROS 2015 Workshop on Grounding robot autonomy: Emotional and social interaction in robot behavior. October 2, 2015, Hamburg, Germany.
8. Duncan D., Shine G., English Ch., Facial Emotion Recognition in Real Time. 2017.
9. Fan Y., Lam J. C. K., Li V.O.K. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. Retrieved, 2017.
10. Kahou S. E., Bouthillier X., Lamblin P., EmoNets: Multimodal deep learning approaches for emotion recognition in video // Journal on Multimodal User Interfaces. June 2016, Volume 10, Issue 2, P. 99–111.
11. Sun M. C., Hsu S. H., Yang M. C., Chien J. H. Context-aware Cascade Attention-based RNN for Video Emotion Recognition. In 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018.
12. Moon S. E., Jang S., Lee J. S. Convolutional Neural Network Approach for EEG-Based Emotion Recognition Using Brain Connectivity and its Spatial Information. In *ICASSP,* IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. 2018.
13. Rabiner L.R., Juang B.-H. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.
14. McDuff D., El Kaliouby R., Senechal T., Amr M., Cohn J. F., Picard R., Affectiva-mit facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected „in-the-wild // IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2013, P. 881–888.
15. Bolon P., Two-Dimensional Linear Filtering // Digital Filters Design for Signal and Image Processing. 2010, P. 233–260.
16. Lukianitsa A. A., Shishkin A. G., Automatic detection of changes in emotional States via speech signal // Speech technologies. 2009, №3, № 60-76.
17. Eyben F., Weninger F., Wöllmer M., Schuller B., München T., Open-Source Media Interpretation by Large feature-space Extraction // TU Munchen, MMK. November 2016.
18. Goodfellow I., Bengio Y., Courville A., Deep Learning, 2016, P. 499-522.
19. Ferri C., Hernández-Orallo J., Salido M. A., Volume under the ROC Surface for Multi-class Problems. 2013, P. 108–120.

Application of deep learning methods to recognize the emotional state of a person in a video image

A. A. Moskvin, A. G. Shishkin

Moscow state University M. V. Lomonosov, Moscow, Russia

Abstract. In this paper, using the use of deep neural networks developed and implemented a model that allows you to determine in real time with limited computing resources emotional state of a person

by video sequence, which is present as a voice signal related to the source for which you want to determine the state, and his face full face. Visual information is represented by 16 consecutive frames with a size of 96x96 pixels, and voice - with 140 of characteristics for a sequence of the 37 Windows. On the basis of experimental studies, the architecture of the model using convolutional and recurrent neural networks is developed. For 7 classes that meet different emotional States - neutral state, anger, sadness, fright, joy, disappointment and surprise - the recognition efficiency is 59%. Studies have shown that the use of audio information in conjunction with the visual can increase the accuracy of recognition by 12%. The created system is dynamic in terms of selection of parameters, narrowing or expanding the number of classes, as well as the ability to easily add, accumulate and use information from other external devices for further development and improvement of classification accuracy.
Keyword: artificial neural networks, deep learning, emotion recognition, video, speech signal.

DOI 10.14357/20718594190201

References

1. Спицын, В. Г., Болотова, Ю. А., Шабалдина, Н. В., Тхи, Б., Чанг, Т., & Хоанг, Ф. Н. (n.d.). Распознование лиц на основе методы главных компонент с применением вейвлет-дескрипторов Хаара и Добеши. Ba Ria – Vung Tau University, Ba Ria – Vung Tau, Vietnam.
2. Huang X., Acero A., Hon H.-W., Spoken language processing: a guide to theory, algorith, and system development // Prentice Hall PTR. 2001.
3. Ashok, V., Balakumaran, T., Gowrishankar, C., Vennila, I. L. A., & Kumar, A. N. (2010). The Fast Haar Wavelet Transform for Signal & Image Processing. Arxiv Preprint ArXiv, *7*(1), 126–130.
4. Chronaki G., Hadwin J. A., Garner M., Maurage P., Sonuga-Barke E. J. S., The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. Br. J. Dev. Psychol., 2015, vol. 33, no. 2, P. 218–236.
5. Jorda M., Instructor N. M., Ng A., CS229: Machine Learning techniques - Emotion Classification on face images. 2015.
6. Fan Y., Lam J. C. K., Li V.O.K. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. Retrieved, 2017.
7. Horii T., Nagai Y., Asada M., Emotion Recognition and Generation through Multimodal Restricted Boltzmann Machines // Proceedings of the IROS 2015 Workshop on Grounding robot autonomy: Emotional and social interaction in robot behavior. October 2, 2015, Hamburg, Germany.
8. Duncan D., Shine G., English Ch., Facial Emotion Recognition in Real Time. 2017.
9. Fan Y., Lam J. C. K., Li V.O.K. Multi-Region Ensemble Convolutional Neural Network for Facial Expression Recognition. Retrieved, 2017.
10. Kahou S. E., Bouthillier X., Lamblin P., EmoNets: Multimodal deep learning approaches for emotion recognition in video // Journal on Multimodal User Interfaces. June 2016, Volume 10, Issue 2, P. 99–111.
11. Sun M. C., Hsu S. H., Yang M. C., Chien J. H. Context-aware Cascade Attention-based RNN for Video Emotion Recognition. In 2018 1st Asian Conference on Affective Computing and Intelligent Interaction, ACII Asia 2018.
12. Moon S. E., Jang S., Lee J. S. (2018). Convolutional Neural Network Approach for EEG-Based Emotion Recognition Using Brain Connectivity and its Spatial Information. In ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings.
13. Rabiner L.R., Juang B.-H. Fundamentals of speech recognition. Prentice Hall, Englewood Cliffs, NJ, 1993.
14. McDuff D., El Kaliouby R., Senechal T., Amr M., Cohn J. F., Picard R., Affectiva-mit facial expression dataset (AM-FED): Naturalistic and spontaneous facial expressions collected „in-the-wild // IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2013, P. 881–888.
15. Bolon P., Two-Dimensional Linear Filtering // Digital Filters Design for Signal and Image Processing. 2010, P. 233–260.
16. Lukianitsa A. A., Shishkin A. G., Automatic detection of changes in emotional States via speech signal // Speech technologies. 2009, №3, № 60-76.
17. Eyben F., Weninger F., Wöllmer M., Schuller B., München T., Open-Source Media Interpretation by Large feature-space Extraction // TU Munchen, MMK. November 2016.
18. Goodfellow I., Bengio Y., Courville A., Deep Learning, 2016, pp. 499-522.
19. Ferri C., Hernández-Orallo J., Salido M. A., Volume under the ROC Surface for Multi-class Problems. 2013, P. 108–120.