Оригинальная статья / Original article

https://doi.org/10.21869/2223-1560-2021-25-1-82-109

Применение многозадачного глубокого обучения в задаче распознавания эмоций в речи

А.В. Рябинов [1], М.Ю. Уздяев [1] ✉, И.В. Ватаманюк [1]

[1] Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Санкт-Петербургский институт информатики и автоматизации Российской академии наук 14-я линия В.О. 39, г. Санкт-Петербург 199178, Российская Федерация

✉ e-mail: uzdyaev.m@iias.spb.su

Резюме

Цель исследования. *Эмоции играют одну из ключевых ролей в регуляции поведения человека. Решение задачи автоматического распознавания эмоций позволяет повысить эффективность функционирования целого ряда цифровых систем: систем обеспечения безопасности, человеко-машинных интерфейсов, систем электронной коммерции и т.д. При этом отмечается низкая эффективность современных подходов распознавания эмоций в речи. Данная работа посвящена исследованию автоматического распознавания эмоций в речи с помощью методов машинного обучения.*

Методы*. В статье описан и протестирован подход к автоматическому распознаванию эмоций в речи на основе многозадачного обучения глубоких сверточных нейронных сетей архитектур AlexNet и VGG с применением автоматического подбора коэффициентов весов каждой задачи при вычислении итогового значения потери в процессе обучения. Все модели были обучены на выборке набора данных IEMOCAP с четырьмя эмоциональными категориями «гнев», «счастье», «нейтральная эмоция», «грусть». В качестве входных данных используются обработанные специализированным алгоритмом лог-мел спектрограммы высказываний.*

Результаты. *Рассмотренные модели были протестированы на основе численных метрик: доля верно распознанных экземпляров, точность, полнота, f-мера. По всем вышеперечисленным метрикам получено улучшение качества распознавания эмоций предлагаемой моделью по сравнению с двумя базовыми однозадачными моделями, а также с известными решениями. Это достигается благодаря применению автоматического взвешивания значений функций потерь от отдельных задач при формировании итогового значения ошибки в процессе обучения.*

Заключение. *Полученное улучшение качества распознавания эмоций по сравнению с известными решениями подтверждает целесообразность применения концепции многозадачного обучения для увеличения точности моделей распознавания эмоций. Разработанный подход позволяет достичь равномерного и одновременного снижения ошибок отдельных задач и используется в области распознавания эмоций в речи впервые.*

Ключевые слова: *многозадачное обучение; сверточные нейронные сети; речевые технологии; автоматическое распознавание эмоций; анализ аудиосигналов речи.*

Конфликт интересов: *Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.*

Финансирование: Работа выполнена при поддержке РФФИ (18-29-22061_мк).

Для цитирования: Рябинов А.В., Уздяев М.Ю., Ватаманюк И.В. Применение многозадачного глубокого обучения в задаче распознавания эмоций в речи // Известия Юго-Западного государственного университета. 2021; 25(1): 82-109. https:// doi.org/10.21869/2223-1560-2021-25-1-82-109.

Поступила в редакцию 23.12.2020 *Подписана в печать 17.02.2021* *Опубликована 31.03.2021*

Applying Multitask Deep Learning to Emotion Recognition in Speech

Artem V. Ryabinov [1], Mikhail Yu. Uzdiaev [1] ✉, Irina V. Vatamaniuk [1]

[1] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS),
St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
39, 14th Line, St. Petersburg 199178, Russian Federation

✉ e-mail: uzdyaev.m@iias.spb.su

Abstract

Purpose of research. *Emotions play one of the key roles in the regulation of human behaviour. Solving the problem of automatic recognition of emotions makes it possible to increase the effectiveness of operation of a whole range of digital systems such as security systems, human-machine interfaces, e-commerce systems, etc. At the same time, the low efficiency of modern approaches to recognizing emotions in speech can be noted. This work studies automatic recognition of emotions in speech applying machine learning methods.*

Methods*. The article describes and tests an approach to automatic recognition of emotions in speech based on multitask learning of deep convolution neural networks of AlexNet and VGG architectures using automatic selection of the weight coefficients for each task when calculating the final loss value during learning. All the models were trained on a sample of the IEMOCAP dataset with four emotional categories of 'anger', 'happiness', 'neutral emotion', 'sadness'. The log-mel spectrograms of statements processed by a specialized algorithm are used as input data.*

Results*. The considered models were tested on the basis of numerical metrics: the share of correctly recognized instances, accuracy, completeness, f-measure. For all of the above metrics, an improvement in the quality of emotion recognition by the proposed model was obtained in comparison with the two basic single-task models as well as with known solutions. This result is achieved through the use of automatic weighting of the values of the loss functions from individual tasks when forming the final value of the error in the learning process.*

Conclusion*. The resulting improvement in the quality of emotion recognition in comparison with the known solutions confirms the feasibility of applying multitask learning to increase the accuracy of emotion recognition models. The developed approach makes it possible to achieve a uniform and simultaneous reduction of errors of individual tasks, and is used in the field of emotions recognition in speech for the first time.*

Keywords: *multitask learning; convolution neural networks; speech technologies, automatic emotion recognition; analysis of audio signals of speech.*

Conflict of interest. *The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.*

Funding: *The work was supported by the Russian Foundation for Basic Research (18-29-22061_mk).*

For citation: Ryabinov A. V., Uzdiaev M. Yu., Vatamaniuk I.A. Applying Multitask Deep Learning to Emotion Recognition in Speech. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University*. 2021; 25(1): 82-109 (In Russ.). https://doi.org/10.21869/2223-1560-2021-25-1-82-109.

Received 23.12.2020 *Accepted 17.02.2021* *Published 31.03.2021*

***

Введение

Развитие моделей, методов и систем автоматического распознавания и интерпретации эмоционального состояния, является важной и актуальной задачей искусственного интеллекта. Особо стоит выделить область распознавания эмоций в речи (Speech Emotion Recognition, SER), которая рассматривает модели машинного обучения, обученные на наборах данных, которые содержат невербальные проявления эмоций в речи. Такие модели получили широкое распространение в интерфейсах человек-компьютер, в целом, и голосовых пользовательских интерфейсов (Alexa, Cortana, Siri, Алиса), в частности. Кроме того, модели распознавания эмоций получили широкое распространение в следующих областях: в приложениях речевого анализа в области медицины [1], безопасности [2], робототехники [3], автоматизированных систем [4]. Тем не менее, на текущем этапе своего развития, модели автоматического распознавания эмоций в речи не могут обеспечить должную производительность на реальных данных [5], что определяет необходимость разработки новых подходов к распознаванию эмоций человека в речи.

Эмоции являются «сложными психическими процессами и состояниями, связанными с инстинктами, потребностями, мотивами и отражающих в форме непосредственного переживания (удовлетворения, страха, радости и т.д.) значимость действующих на индивида явлений и ситуаций для осуществления его жизнедеятельности. Сопровождая практически любые проявления активности субъекта, эмоции служат одним из главных механизмов внутренней регуляции психической деятельности и поведения, направленных на удовлетворение актуальных потребностей» [6]. Среди прочих структурных компонент (импрессивная, когнитивная, физиологическая) [7], для задач автоматического распознавания эмоций наиболее важной является поведенческая или моторная компонента, т.е. внешне наблюдаемая специфическая двигательная активность, которая связана с тем или иным эмоциональным состоянием индивида. Именно возможность внешнего наблюдения и регистрации поведения на расстоянии без непосредственного участия испытуемого делает эту компоненту ведущей в задачах распознавания эмоций. В задачах распознавания эмоций в речи используются невербальные паралингвистические акустические признаки аудиосигнала речи.

Решая задачу автоматического распознавания эмоций, необходимо определить адекватное представление последних с точки зрения возможности обработки этой информации на компьютере и согласованности с теоретическими положениями психологии эмоций. В этой связи, чаще всего на практике используют две эмоциональные модели [8]. Первая – это дискретные

классы, например, «Большая шестерка» эмоциональных категорий, основанная на теории эмоций Пола Экмана (Гнев, Счастье, Нейтральная, Грусть, Отвращение, Страх) [9]. Вторая модель – непрерывный подход, заключающийся в представлении каждой эмоции как базиса в многомерном пространстве, чьими измерениями являются эмоциональные атрибуты Знак (Valence), Интенсивность (Arousal), Доминантность (Dominance) [10, 11].

Активное применение нейронных сетей и моделей глубокого обучения в задачах распознавания эмоций началось около десяти лет назад. Однако лишь в последние годы были предложены сквозные (end-to-end) модели [12-17], использующие для обучения непосредственно запись человеческого голоса. Узким местом таких моделей является отсутствие репрезентативных выборок данных, а также отсутствие формализованных методик сбора и разметки данных проявления эмоций в аудиосигнале речи. Поэтому основным общим недостатком современных сквозных систем является переобучение и низкая способность к обобщению. В данной работе мы рассматриваем подход к решению данных проблем с помощью многозадачного обучения (англ. multi-task learning) – одновременного обучения модели глубокого обучения группе различных, но взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение» [18]. Мы рассматриваем три смежные паралингвистические задачи – распознавание эмоций по голосу (классификация одного из дискретных эмоциональных классов), распознавание диктора, распознавание пола диктора. Также мы впервые применяем в данной области методы автоматического подбора коэффициентов при вычислении итогового значения ошибки, что способствует дальнейшему улучшению результата. Мы обучаем и проверяем разработанную модель на наборе данных IEMOCAP [19]. Предложенная модель превосходит результаты распознавания известных решений в области распознавания эмоций в речи.

Обзор литературы

До 2016 года в литературе превалировали традиционные методы, основанные на покадровом извлечении локальных низкоуровневых дескрипторов с последующим их комбинированием для получения глобальных признаков изучаемого высказывания или отрывка речи, и использование полученного признакового представления для обучения алгоритмов классификации или регрессии. Исследователи изучали многие низкоуровневые сконструированные вручную дескрипторы и их комбинации. Наиболее эффективными и часто используемыми наборами стали наборы eGeMAPS (88 параметров) [20] и ComParE (6373 параметра) [21]. В качестве классификаторов в литературе наиболее часто встречаются метод опорных векторов (Support Vector Machines, SVM), алгоритм k-ближайших соседей (k-

Nearest Neighbors, k-NN), скрытые марковские модели (Hidden Markov Model, HMM), многослойные перцептроны [22]. Так же, как и многие другие задачи машинного обучения, речевое распознавание эмоций сильно зависит от набора данных, используемого для обучения. Отличия между наборами данных, вызванные различными постановками задачи распознавания эмоций в речи, включают в себя: наличие искусственно и/или натурально выраженных эмоций, язык, половозрастной состав дикторов и их количество, разметка.

Очевидны недостатки традиционных подходов. Во-первых, использование сконструированных вручную дескрипторов признаков требует привлечения экспертов по акустике и психологии, чтобы разработать набор наиболее релевантных параметров [23]. Помимо выбираемого пространства признаков, эффективность систем распознавания также сильно зависит от реализованной модели распознавания образов, что может привести к снижению результатов распознавания. В этом отношении перспективной альтернативой являются так называемые сквозные (end-to-end) системы. Они направлены на автоматическое изучение наиболее надежных представлений, связанных с определенной задачей, используя различные топологии нейронных сетей для обучения как процессу извлечения признаков, так и классификации, исключая таким образом процедуру ручного проектирования признаков из процесса распознавания эмоций в речи. Недавние достижения в области глубокого обучения в целом и его применения к таким задачам, как распознавание речи и идентификация по голосу, указали на перспективность использования различных сверточных (Convolutional Neural Network, CNN) и рекуррентных (Recurrent Neural Network, RNN) архитектур глубоких нейронных сетей для таких систем. Так, в работе [12] был впервые описан сквозной подход к распознаванию эмоций по голосу. Авторы применили сверточные и рекуррентные с долгой краткосрочной памятью нейронные сети (Long Short-Time Memory Recurrent Neural Network, LSTM-RNN) для обработки «сырого» дискретизированного сигнала в формате wav. Было показано, что использование этого подхода значительно превосходит традиционные подходы, связанные с техниками цифровой обработки сигналов (в качестве базовых методов применялись признаковые представления eGeMAPS и ComParE, классификаторы SVM и BiLSTM-DRNN) в задаче распознавания эмоций на наборе данных RECOLA [24].

Однако представление аудиосигнала в виде волновой формы достаточно полно передает лишь амплитудную характеристику, в то время, как важнейшая частотная характеристика может остаться без внимания. В этой связи широкое распространение получили подходы, основанные на обработке отображений аудиосигнала через различные частотно-временные представления, та-

кие, как спектрограммы. Спектрограммы – это визуальные представления амплитуды сигнала с течением времени на разных частотах, полученные с помощью оконного преобразования Фурье (Short-Time Fourier Transform, STFT) и представляющие собой двухмерный график, по горизонтальной оси которого отложено время, по вертикальной – частота, а интенсивность или цвет точки отображает амплитуду отдельной частоты в конкретный момент времени. Последние исследования в различных сферах анализа звука, таких, как: классификация событий по звуку [25], распознавание речи [26], распознавание человека по голосу [27], продемонстрировали применимость спектрограмм для извлечения из них скрытых признаков с помощью сверточных архитектур глубоких нейронных сетей и подтолкнули исследователей на использование спектрограмм в области распознавания эмоций в речи.

В работе [13] продемонстрирована модель распознавания эмоций в речи, основанная на обработке спектрограмм сверточными нейронными сетями. Набор данных, используемый для обучения и тестирования модели – динамическая база данных Acted Emotional Speech Dynamic Database (AESDD) [28]. Предлагаемая архитектура сверточной нейронной сети (4 сверточных слоя и 2 полносвязных слоя) превзошла базовую модель машинного обучения (метод опорных векторов в самостоятельно разработанном авторами признаковом пространстве) на 8,4% с точки зрения доли верно распознанных экземпляров. Авторами [14] представлена нейронная сеть, комбинирующая трехмерные сверточные слои, двунаправленные LSTM ячейки и механизм внимания [29]. В качестве входных данных использовались мел-спектрограммы, дополненные первой и второй производной по времени. Получен результат среднего значения невзвешенной полноты 64,74% на наборе данных IEMOCAP [19] и 82,82% на наборе данных Emo-DB [30]. В статье [15] предложен метод распознавания эмоций по логарифмированным спектрограммам с помощью сверточной нейронной сети и LSTM. Авторы рассмотрели десятки комбинаций топологий нейронных сетей и их параметров. Были протестированы как исключительно сверточные топологии (от двух до восьми сверточных слоев с различными комбинациями размеров окон свертки), так и топологии с одним-двумя сверточными слоями и одним-двумя слоями LSTM. Лучшие результаты показала архитектура, содержащая 3 сверточных и 2 LSTM слоя, точность распознавания на наборе данных IEMOCAP составила 68,8%.

Описанные выше решения имеют один главный общий недостаток: они страдают от переобучения, что ведет к серьезному снижению производительности в условиях несоответствия между тренировочными и тестовыми данными. Данная проблема решается, в общем случае, регуляризацией модели с помо-

щью таких техник, как прореживание (dropout) [31], сокращение веса (weight decay) или добавлением новых тренировочных данных, в том числе, с помощью техник аугментации. Однако переобучение может быть связано не только с ограниченным размером обучающих данных или недостаточной сложностью модели. Общепринятая методология оптимизации описанных выше моделей глубокого обучения только в рамках одной задачи игнорирует потенциальную богатую информацию в тренировочном сигнале. В этой связи альтернативным эффективным подходом к улучшению результата является многозадачное обучение. В последнее время оно было включено во множество моделей глубоких нейронных сетей, решающих проблемы в области компьютерного зрения [32], обработки речи [33] и естественного языка [34], а также обучения с подкреплением [35]. К примеру, задачи обнаружения лица, распознавания пола и оценки позы человека могут быть одновременно решены с использованием одной сверточной глубокой нейронной сети [36].

В области распознавания эмоций в речи многозадачное обучение показало высокие результаты для моделей, обучаемых по прецедентам. Большинство из существующих в области распознавания эмоций в речи подходов совместно обучаются определенным эмоциональным атрибутам для улучшения как точности, так и генерализации. Так, Parthasarathy и др. [37] представили систему для одновременной оценки эмоциональных атрибутов Возбуждение, Валентность, Доминантность, использующую многозадачное обучение глубоких полносвязных нейронных сетей в признаковом пространстве ComParE. Лучшие результаты были достигнуты структурой, комбинирующей один общий слой с тремя отдельными слоями для каждой задачи. По сравнению с аналогичной, но однозадачной архитектурой, был продемонстрирован максимальный прирост корреляционного коэффициента согласованности (concordance correlation coefficient, CCC) на 4,7% для однокорпусных и 14,0% для кросс-корпусных экспериментов, а полученные с помощью t-SNE визуализации активаций последних скрытых слоев нейронной сети проиллюстрировали, что многозадачное обучение создает лучшие высокоуровневые представления. Zhang и Schuller [16] также использовали многозадачное обучение для предсказания значений атрибутов Знак, Интенсивность, Доминантность. В качестве исходного представления был использован дискретизированный сигнал в формате wav. Дополнительно авторами был реализован механизм внимания с целью зафиксировать распределение вклада различных отрезков записи для каждой отдельной задачи. Для оценки эффективности системы была проведена серия экспериментов на базе данных IEMOCAP. Каждый эмоциональный атрибут был дискретизирован как имеющий в каждом отдельном слу-

чае Высокое, Среднее, или Низкое значение; таким образом, предсказание значения каждого атрибута рассматривалось как задача трехклассовой классификации. Получены результаты точности предсказания: 48,7% для возбуждения, 63,8% для валентности и 51,6% для доминантности, что незначительно превосходит как рассмотренные в той же статье базовые системы (eGeMAPS + SVM, eGeMAPS + RNN), так и однозадачный подход к классификации каждого атрибута с использованием аналогичной архитектуры нейронной сети.

Обе описанные выше работы, однако, не используют спектрограммы в качестве представления аудиосигнала. Также очевидно, что помимо информации, кодирующей эмоциональное состояние говорящего, речь и ее представление в виде спектрограммы содержит большое количество не относящейся к эмоциям информации, поэтому вместо использования в качестве задач моделирование эмоциональных атрибутов, перспективным выглядит создание системы для одновременного решения смежных паралингвистических задач. Например, Gideon и др. [38] исследовали перенос обучения между тремя паралингвистическими задачами: распознавание диктора, пола и эмоции, применяя для этого прогрессивные нейронные сети. В то время как классическая стратегия переноса обучения предполагает предварительное обучение глубокой нейронной сети на исходном наборе данных и дальнейшую тонкую настройку на целевом наборе данных из другой задачи и/или домена, прогрессивные нейронные сети представляют альтернативный способ, позволяющий избежать т.н. «эффекта забывания», выражающегося в невозможности выделения значимой информации из данных нейронной сетью. Это происходит вследствие архитектурных особенностей, обеспечивающих возможность сохранения информации, полученной при обучении решению исходной задачи. В статье предложена архитектура прогрессивной нейронной сети с пятью скрытыми полносвязными слоями. Результаты этого подхода значительно превзошли как стандартное обучение глубокой нейронной сети, так и классическую стратегию переноса обучения между задачами распознавания диктора и эмоции: среднее значение невзвешенной полноты 65,7% на наборе данных IEMOCAP. Однако авторами было использовано признаковое представление eGeMAPS и простая полносвязная топология, а прогрессивные нейронные сети при своем расширении и углублении начинают требовать огромного количества параметров для настройки (для параллельного решения новой задачи требуется увеличение количества параметров модели в 2 раза), что делает их применение нецелесообразным для обработки спектрограмм. В своей недавней работе Latuf и др. [17] представили модель многозадачного обучения для голосового распознавания эмоций, идентификации говорящего и его пола. Для извле

чения высокоуровневых признаков авторами использован состязательный автоэнкодер, а для каждой задачи используется свой блок-классификатор, состоящий из сверточных и полносвязных слоев. Также используется стратегия предварительного обучения модели: не задействуя задачу распознавания эмоций, авторы используют большой набор данных LibriSpeech, созданный для решения задач в области распознавания речи и дикторов. Таким образом, модель первично обучается извлечению признаков на значительно большем количестве данных, чем доступно для задачи распознавания эмоций. После предобучения проводится тонкая настройка модели одновременно по трем задачам на наборах данных с эмоциональной речью. Полученные результаты (68,8% на наборе данных IEMOCAP и 63,6% на наборе данных MSP-IMPROV [39]) превосходят как таковые у этой же модели без предобучения автоэнкодера, так и результаты аналогичной архитектуры при однозадачном обучении, а также известные авторам на тот момент лучшие решения. На текущий момент данная работа является наиболее полно раскрывающей возможности как обработки спектрограмм, так и многозадачного обучения в области распознавания эмоций в речи. Однако и у нее есть недостаток: при вычислении итогового значения ошибки для обратного распространения, авторами была использована формула со статическими коэффициентами, которые в ходе экспериментов выбирались путем множества проб и ошибок.

В недавних исследованиях в области многозадачного обучения было продемонстрировано, что очень важно найти подходящие стратегии взвешивания значений функции потерь каждой задачи, чтобы минимизировать общие эмпирические потери без приоритета в обучении одной задачи над другими. В то же время, именно динамические методы подбора коэффициентов имеют решающее значение в многозадачном обучении, поскольку проблемы с конфликтующими градиентными сигналами, исходящими от отдельных задач в разные моменты обучения, могут ухудшить производительность модели. Kendall et al. в [40] предложил метод взвешивания на основе гомоскедастичной неопределенности и применил его к сверточным нейронным сетям для одновременного решения трех задач компьютерного зрения, а именно семантической сегментации (semantic segmentation), пообъектной сегментации (instance segmentation) и попиксельной регрессии карты глубины (depth regression), продемонстрировав улучшение результатов каждой из задач по сравнению с однозадачными моделями. Liebel и Körner [41] адаптировали элемент регуляризации в этом методе, предотвратив отрицательные значения регуляризации, что позволило еще сильнее улучшить результаты на тех же задачах. В работе [42] проведено сравнение этих и еще нескольких стратегий динамического многозадачного обучения, таких как Dynamic Weighted Average (DWA)

[43] и GradNorm [44] на наборах данных Multi-MNIST, NYU v2 и IMDB-WIKI. Продемонстрировано небольшое превосходство усовершенствованного метода на основе неопределенности.

Таким образом, мы делаем вывод, что в современной литературе не освещено применение метода автоматического динамического взвешивания функции потерь в многозадачном обучении глубоких сверточных нейронных сетей спектрограммам речи для одновременного решения паралингвистических задач распознавания эмоций, распознавания диктора и распознавания пола диктора. Разработка соответствующей системы для улучшения точности распознавания эмоций в речи является целью данного исследования.

Материалы и методы

Были проведены предварительные эксперименты, оценивающие множество архитектур сверточных глубоких нейронных сетей на предмет качества изучения и извлечения высокоуровневых признаков из логарифмированных мел-спектрограмм: AlexNet [45]; VGG [46]; ResNet-50 [47]. Мы не вносили никаких изменений в архитектуры данных нейронных сетей, кроме изменения количества нейронов их последних полносвязных слоев для соответствия количеству эмоциональных классов, а также изменения количества каналов исходного изображения с 3 до 1. Две модели, показавшие на этом этапе лучшие результаты, были выбраны в качестве базовых, ниже приведено их описание, а также описание предлагаемого подхода.

На вход этой и всех описанных далее моделей подается одноканальное нормализованное изображение логарифмированной мел-спектрограммы речевого сигнала. В этой и во всех описанных далее моделях используется функция активации ReLU. В базовой модели 1 извлечение признаков производится с помощью сверточной нейронной сети, архитектура которой аналогична архитектуре AlexNet, кроме количества входных каналов изображения. Далее извлеченные признаки подаются на блок классификатора, состоящего из 4 полносвязных слоев. В целях регуляризации, после первого слоя производится dropout 50% нейронов этого слоя. Схематическое изображение модели представлено на рис. 1, подробное описание её слоёв – в табл. 1.

Модель имеет 17,073,348 параметров, которые занимают 65,13 МБ дискового пространства.

В базовой модели 2 извлечение признаков производится с помощью сверточной нейронной сети, архитектура которой аналогична архитектуре 11-слойной нейронной сети VGG, кроме количества входных каналов изображения. Далее извлеченные признаки подаются на блок классификатора, архитектура которого аналогична таковой у Базовой модели 1.

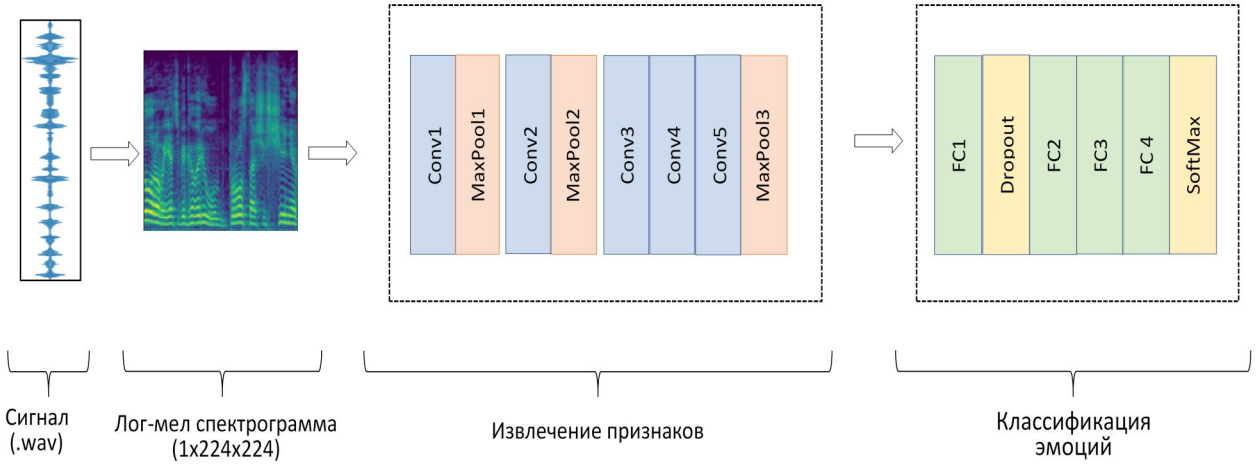

Рис. 1. Схема Базовой модели 1

Fig. 1. Diagram of Reference model 1

Таблица 1. Архитектура Базовой модели 1

Table 1. Architecture of Reference model 1

Слой / Layer	Параметры / Parameters	Размерность на выходе / Output dimension
Conv1	Количество фильтров – 64, размер ядра – 11x11, шаг – 4, дополнение (padding) – 2	64x55x55
MaxPool1	Размер ядра –3x3, шаг – 2	64x27x27
Conv2	Количество фильтров – 192, размер ядра – 5x5, шаг – 1, дополнение – 2	192x27x27
MaxPool2	Размер ядра – 3x3, шаг – 2	192x13x13
Conv3	Количество фильтров – 384, размер ядра – 3x3, шаг – 1, дополнение – 1	384x13x13
Conv4	Количество фильтров – 256, размер ядра – 3x3, шаг – 1, дополнение – 1	256x13x13
Conv5	Количество фильтров – 256, размер ядра – 3x3, шаг – 1, дополнение – 1	256x12x12
MaxPool3	Размер ядра – 3x3, шаг – 2	256x5x5
FC1		6400
Dropout	p = 0.5	
FC2		2048
FC3		512
FC4		<количество классов>

Таким образом, при сравнении результатов Базовой модели 1 и Базовой модели 2, сравниваются между собой блоки извлечения признаков этих моделей. Схематическое изображение модели представлено на рис. 2, описание её слоёв приведено в табл. 2.

Модель имеет 61,652,740 параметров, которые занимают 235,19 МБ дискового пространства.

Предлагаемая модель

Для повышения точности распознавания был выбран подход, основанный на многозадачном обучении. Преимущества данного подхода можно выразить в следующих положениях: во-первых, количество параметров в многозадачной модели будет меньше, чем при построении нескольких моделей, каждая из которых оптимизирована для своих индивидуальных задач; и, во-вторых, что более важно, модели, обученные выполнять множество задач одновременно, должны иметь возможность путем индуктивного переноса знаний между задачами извлекать из представлений исходного сигнала более общую информацию, обеспечивая таким образом регуляризацию модели и лучшую производительность каждой задачи с меньшими объемами тренировочных данных. Предлагаемая модель для многозадачного обучения представляет собой блок извлечения признаков, идентичный Базовой модели 2, и блок классификации, состоящий из одного общего полносвязного слоя (25088 нейронов), после которого происходит разделение нейронной сети на независимые друг от друга классификаторы, архитектуры которых идентичны таковым у Базовой модели 2. В качестве задач выбраны паралингвистические задачи классификации эмоций, классификации спикера и классификации пола. Таким образом, сравнивая результаты Базовой модели 2 и Предлагаемой модели, можно делать выводы о работоспособности концепции многозадачного обучения в контексте нашей задачи. Схематическое изображение модели представлено на рис. 3.

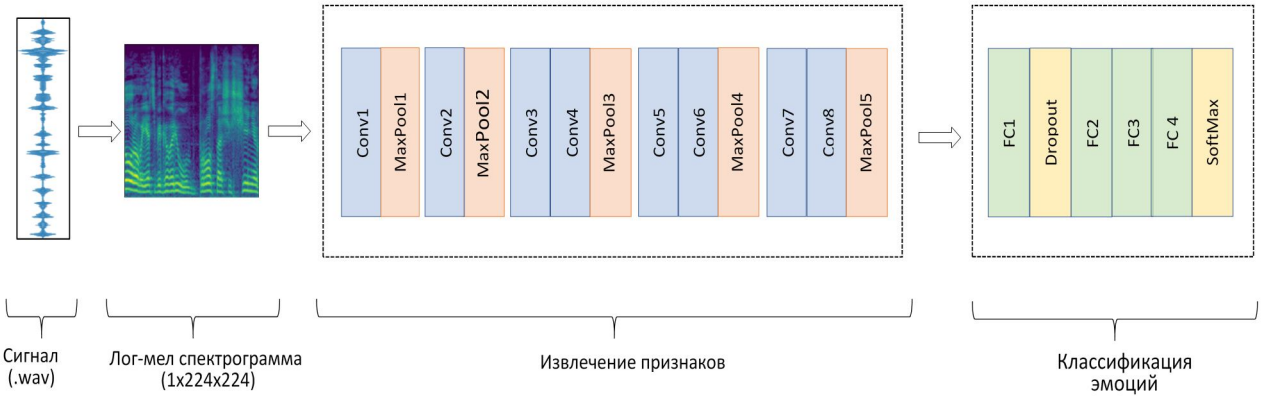

Рис. 2. Схема Базовой модели 2

Fig. 2. Diagram of Reference model 2

Таблица 2. Архитектура Базовой модели 2

Table 2. Architecture of Reference model 2

Слой / Layer	Параметры / Parameters	Размерность на выходе / Output dimension
Conv1	Количество фильтров – 64, размер ядра – 3x3, шаг – 1, дополнение – 1	64x224x224
MaxPool1	Размер ядра – 2, шаг – 2	64x112x112
Conv2	Количество фильтров – 128, размер ядра – 3x3, шаг – 1, дополнение – 1	128x112x112
MaxPool2	Размер ядра – 2, шаг – 2	128x56x56
Conv3	Количество фильтров – 256, размер ядра – 3x3, шаг – 1, дополнение – 1	256x56x56
Conv4	Количество фильтров – 256, размер ядра – 3x3, шаг – 1, дополнение – 1	256x56x56
MaxPool3	Размер ядра – 2, шаг – 2	256x28x28
Conv5	Количество фильтров – 512, размер ядра – 3x3, шаг – 1, дополнение – 1	512x28x28
Conv6	Количество фильтров – 512, размер ядра – 3x3, шаг – 1, дополнение – 1	512x28x28
MaxPool4	Размер ядра – 2, шаг – 2	512x14x14
Conv7	Количество фильтров – 512, размер ядра – 3x3, шаг – 1, дополнение – 1	512x14x14
Conv8	Количество фильтров – 512, размер ядра – 3x3, шаг – 1, дополнение – 1	512x14x14
MaxPool5	Размер ядра – 2, шаг – 2	512x7x7
FC1		25088
Dropout	p=0.5	
FC2		2048
FC3		512
FC4		<количество классов>

Данная модель имеет 63,762,576 параметров, которые занимают 243,23МБ дискового пространства.

Эксперименты

Для экспериментального исследования вышеописанных моделей был выбран набор данных IEMOCAP [19] – многомодальный набор данных, состоящий из аудиовидеозаписей диалогов полупрофессиональных актеров на английском языке, в ситуациях, стимулирующих различные эмоциональные реакции (как сценарных, так и импровизированных). В записи участвовало 10 актеров (5 мужчин и 5 женщин), в ходе записи было получено в общей сложности 12 ч 26 мин данных, которые были размечены несколькими аннотаторами как на дискретные эмоциональные классы, так и на непрерывные значения валентности и активации.

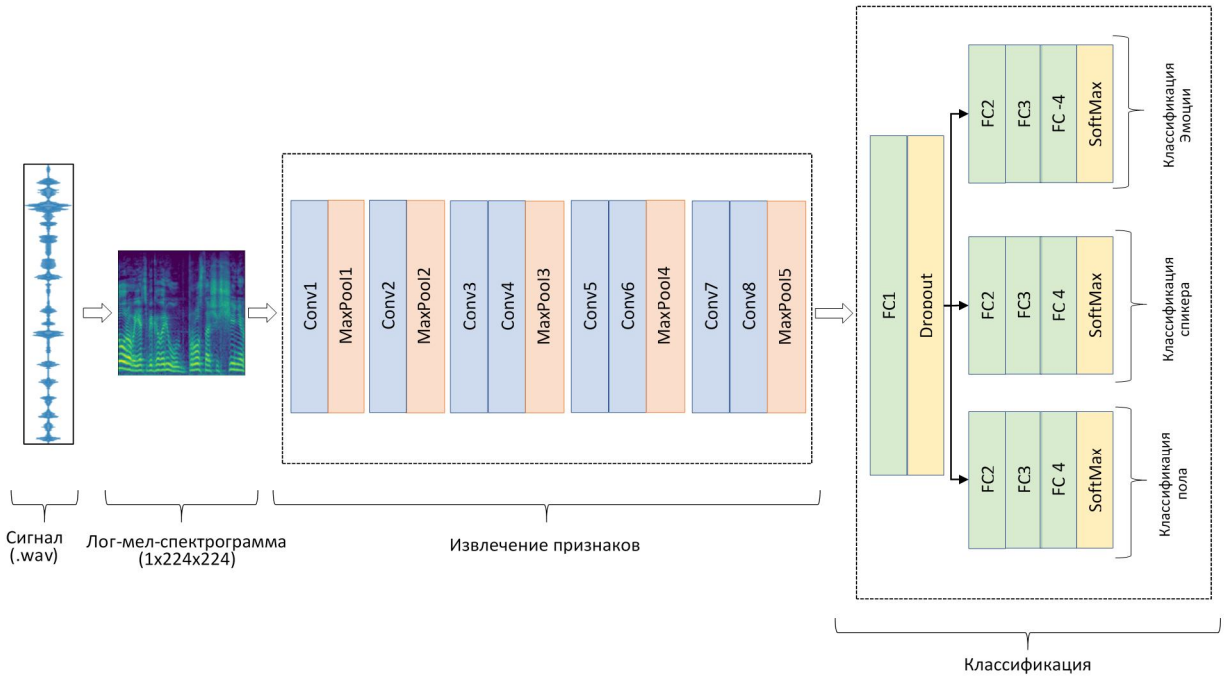

Рис. 3. Схема предлагаемой модели

Fig. 3. Diagram of the proposed model

Все модели были обучены на выборке набора данных IEMOCAP с четырьмя эмоциональными категориями Гнев, Счастье, Нейтральная эмоция, Грусть (далее – IEMOCAP-4). Мы мотивируем свой выбор наибольшей популярностью именно данного варианта среди исследователей, что позволит сравнить наши результаты.

Были сгенерированы логарифмированные мел-спектрограммы высказываний с помощью алгоритма STFT (количество компонент разложения – 2048, длина окна – 2048 фреймов, ширина шага окна – 512 фреймов, количество мел-фильтров – 512) и произведено разделение на тренировочную и валидационную подвыборки в пропорции 4:1. Поскольку модель требует от входных данных единого размера, а также в целях осуществления простейшей аугментации данных, на каждой эпохе полученные спектрограммы подвергались обработке алгоритмом, блок-схема которого изображена на рис. 5. Все процедуры, подразумевающие использование случайных чисел («Случайно обрезать», «Генерация D», «Случайно заполнить нулями»), выполнялись на случайном зерне генератора для данных из тренировочной выборки, и на фиксированном зерне генератора для данных из валидационной выборки. Таким образом, случайные изменения вносились каждую эпоху только в тренировочные данные, в то время как валидационные данные из эпохи в эпоху не изменялись. На выходе этого алгоритма мы получали массив данных размером 512x512. Данный массив был конвертирован в изображение (значения приведены в диапазон от 0 до 255), которое было

уменьшено до размеров 224x224 с применением кубической интерполяции. Наконец, перед непосредственно подачей на вход нейронной сети, проводилась нормализация со средним значением 0,5 и стандартным отклонением 0,225.

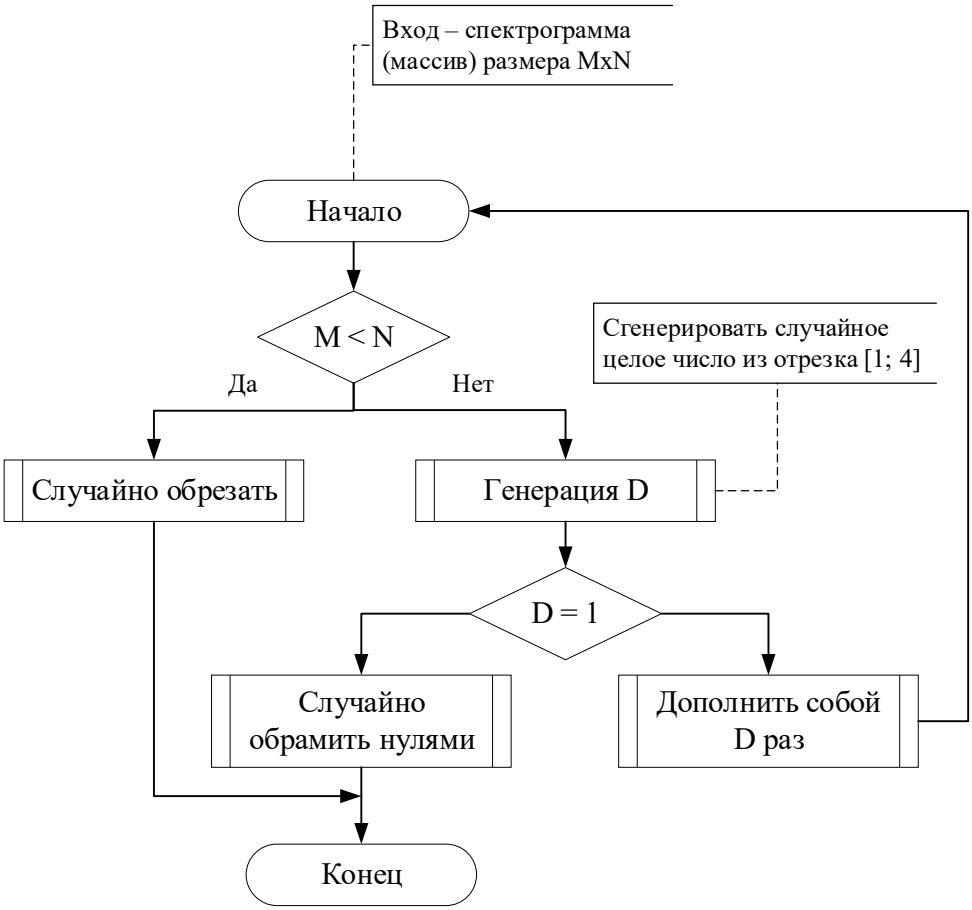

Рис. 4. Блок-схема алгоритма унификации размера и аугментации спектрограмм

Fig. 4. Flowchart of the size unification and spectrogram augmentation algorithm

Модели обучались методом стохастического пакетного градиентного спуска с малым размером пакета (Mini-Batch Stochastic Gradient Descent) на тренировочной подвыборке 300 эпох, параметр скорости обучения равнялся 1e-5, с размером пакета, равным 32; использован алгоритм оптимизации Adam [48], функция потерь – перекрестная энтропия: $L_{CE} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$, где N – количество классов; y_i – истинное значение класса i, принимающее значения 1 (верно) или 0 (неверно); $\hat{y}_i$ – сгенерированное нейронной сетью значение вероятности класса i.

Для предотвращения переобучения производилась остановка обучения модели, если значение функции ошибки на валидационной подвыборке не уменьшалось в течение 30 эпох.

В ходе экспериментов были также имплементированы и протестированы следующие стратегии взвешивания зна-

чений функции потерь отдельных задач на каждом пакете обрабатываемых данных (здесь и далее L_{total}; $L_{emotion}, L_{speaker}$, L_{gender} – соответственно значение итоговой потери для обратного распространения на текущем мини-батче; значения потери при классификации эмоции, спикера и пола на текущем мини-батче):

а) Невзвешенная сумма:

$$L_{total} = L_{emotion} + L_{speaker} + L_{gender};$$

б) Усовершенствованный метод на основе гомоскедастичной неопределенности, описанный в [40; 41];

в) Метод взвешенного среднего: $L_{total} = \alpha L_{emotion} + \beta L_{speaker} + \gamma L_{gender}$,

где $\alpha = \frac{L_{emotion}}{L_{emotion} + L_{speaker} + L_{gender}}$;

$$\beta = \frac{L_{speaker}}{L_{emotion} + L_{speaker} + L_{gender}};$$

$$\gamma = \frac{L_{gender}}{L_{emotion} + L_{speaker} + L_{gender}}.$$

Лучшие результаты в ходе экспериментов были достигнуты с использованием метода взвешенного среднего.

Результаты и их обсуждение

Ниже представлены результаты моделей с наименьшим за все время обучения значением функции ошибки на валидационной подвыборке. В качестве метрик качества выбраны доля верно распознанных экземпляров (Accuracy) $acc = \frac{TP + TN}{TP + TN + FP + FN}$; мера точности (Precision) $pr = \frac{TP}{TP + FP}$; мера полноты (Recall) $rec = \frac{TP}{TP + FN}$; а также *F-мера* (F_1) $F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = 2 \cdot \frac{pr \cdot rec}{pr + rec}$.

Здесь *TP* (True Positive), *TN* (True Negative), *FP* (False Positive), *FN* (False Negative) – соответственно количество истинноположительных, истинноотрицательных, ложноположительных и ложноотрицательных экземпляров данных, распознанных тестируемой моделью. Результаты экспериментов представлены в табл. 3 и 4.

Сравнение результатов предлагаемой модели с базовыми и известными state-of-the-art решениями на задаче классификации эмоций набора данных IEMOCAP представлено в табл. 5.

Таблица 3. Результаты предлагаемой модели для распознавания эмоции, диктора и пола диктора

Table 3. Results of the proposed model for recognition of an emotion, a speaker and his/her gender

	acc	*pr*	*rec*	F_1
Эмоция	0,712	0,685	0,666	0,673
Диктор	0,782	0,774	0,771	0,767
Пол диктора	0,969	0,969	0,969	0,969

Таблица 4. Результаты распознавания каждого эмоционального класса предлагаемой моделью на наборе данных IEMOCAP

Table 4. Results of recognition of every individual emotional class by the proposed model for the IEMOCAP dataset

	pr	*rec*	F_1
Гнев	0,820	0,760	0,790
Радость	0,500	0,390	0,440
Нейтральная эмоция	0,700	0,780	0,740
Грусть	0,72	0,73	0,73

Таблица 5. Сравнение результатов моделей

Table 5. Comparison of the results of the models

	acc	pr	rec	f1
Chen и др. [14]	-	-	0,647	-
Gideon и др. [38]	-	-	0,657	-
Satt и др. [15]	0,688	-	-	-
Latuf и др. [17]	0,688	-	-	-
Базовая модель 1	0,688	0,670	0,624	0,613
Базовая модель 2	0,695	0,674	0,631	0,630
Предлагаемая модель	**0,712**	**0,685**	**0,666**	**0,672**

На наборе данных IEMOCAP предлагаемая модель достигает лучших результатов в распознавании метки Гнев, худших – в распознавании метки Радость (см. табл. 4). Подобная картина совпадает с результатами экспериментов других исследователей на этом наборе данных, и обусловлена его особенностями, а именно высокой степенью разнообразия данных и наименьшей представленностью экземпляров класса Радость. Предлагаемой моделью по сравнению с базовыми получен минимальный прирост accuracy на 0,017, precision – на 0,011, recall – на 0,035, F_1 – на 0,042 (см. табл. 4). Одновременно данная модель успешно решает две смежные паралингвистические задачи: распознавание диктора с точностью 0,782 и распознавание пола с точностью 0,969.

Выводы и дальнейшее развитие

С помощью предложенного в данной работе подхода получено улучшение качества распознавания эмоций по сравнению с известными на наборе данных IEMOCAP. Это обусловливает целесообразность применения концепции многозадачного обучения для увеличения точности моделей распознавания эмоций и одновременного решения смежных паралингвистических задач.

При этом, модель достигает полученных результатов благодаря применению автоматического взвешивания значений функций потерь от отдельных задач при формировании итогового значения ошибки в процессе обучения. Такая стратегия позволяет достичь равномерного и одновременного снижения ошибок отдельных задач, и используется в области распознавания эмоций в речи впервые. Модель превышает результаты распознавания эмоций по сравнению с известными подходами (71,2% верно распознанных экземпляров), а также показывает высокие результаты распознавания диктора и пола (78,2% и 96,9% верно распознанных экземпляров соответственно).

В качестве дальнейших исследований планируется выполнить сравнение результатов распознавания на других наборах данных (RAVDESS [49], Emo-DB [30], MSP-IMPROV [39], MSP-PODCAST [50]); провести кросс-корпусные и кросс-языковые эксперименты (обучение на исходном наборе данных/языке, валидация на целевом наборе данных/языке); проверить производительность модели при использовании другого представления эмоций (параметры Знак, Интенсивность, Доминация); исследовать внутреннее устройство модели с помощью различных техник визуализации (t-SNE [51], Grad-CAM [52]); использовать имеющиеся в наличии большие речевые наборы данных для включения в процесс распознавания эмоций предварительное обучение модели на задачах распознавания диктора и пола.

Список литературы

1. Tokuno S., Tsumatori, G., Shono S., Takei E., Yamamoto T., Suzuki G., Mituyoshi S., Shimura M. Usage of emotion recognition in military health care // *Defense Science Research Conference and Expo* (*DSR*). IEEE, 2011, P. 1-5. https://doi.org/10.1109/DSR.2011.6026823

2. Saste S.T., Jagdale S.M. Emotion recognition from speech using MFCC and DWT for security system // *2017 international conference of electronics, communication and aerospace technology* (*ICECA*). IEEE, 2017. 1. P. 701-704. https://doi.org/10.1109/ICECA.2017.8203631

3. Rázuri J.G., Sundgren D., Rahmani R., Moran A., Bonet I., Larsson A. Speech emotion recognition in emotional feedbackfor human-robot interaction // *International Journal of Advanced Research in Artificial Intelligence* (*IJARAI*). 2015. No. 4(2). P. 20-27. https://doi.org/10.14569/IJARAI.2015.040204

4. Bojanić M., Delić V., Karpov A. Call redistribution for a call center based on speech emotion recognition // *Applied Sciences*. 2020. 10(13). P. 4653. https://doi.org/10.3390/app10134653

5. Björn W., Schuller L. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends // *Communications of the Acm*. 2018. 61(5). P. 90-99. https://doi.org/10.1145/3129340

6. Вилюнас В.К. *Эмоции* // Большой психологический словарь/под общ. ред. Б.Г. Мещерякова, В.П. Зинченко. URL: https://psychological.slovaronline.com/2078-EMOTSII

7. Ильин Е.П. *Эмоции и чувства.* СПб.: Издательский дом "Питер", 2011.

8. Sailunaz K., Dhaliwal M., Rokne J., Alhajj R. Emotion detection from text and speech: a survey // *Social Network Analysis and Mining*. 2018. 8(1). P. 28. https://doi.org/10.1007/s13278-018-0505-2

9. Ekman P. Facial expression and emotion // *American psychologist.* 1993. 48 (4). P. 384. https://doi.org/10.1037/0003-066X.48.4.384

10. Russell J.A. Affective space is bipolar // *Journal of personality and social psychology.* 1979. 37 (3). P. 345. https://doi.org/10.1037/0022-3514.37.3.345

11. Russell J.A. Culture and the categorization of emotions // *Psychological bulletin*. – 1991. 110 (3). P. 426. https://doi.org/10.1037/0033-2909.110.3.426

12. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network / G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou // *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*). IEEE, 2016. P. 5200-5204. https://doi.org/10.1109/ ICASSP.2016.7472669

13. Continuous Speech Emotion Recognition with Convolutional Neural Networks / N. Vryzas, L. Vrysis, M. Matsiola, R. Kotsakis, C. Dimoulas, G. Kalliris // *Journal of the Audio Engineering Society*. 2020. 68 (1/2). P. 14-24. https://doi.org/10.17743/jaes.2019.0043

14. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition / M. Chen, X. He, J. Yang, H. Zhang // *IEEE Signal Processing Letters.* 2018. 25(10). P. 1440-1444. https://doi.org/10.1109/LSP.2018.2860246

15. Satt A., Rozenberg S., Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms // *Interspeech.* 2017. P. 1089-1093. https://doi.org/ 10.21437/Interspeech.2017-200

16. Zhang Z., Wu B., Schuller B. Attention-augmented end-to-end multi-task learning for emotion prediction from speech // *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2019. P. 6705-6709. https://doi.org/10.1109/ICASSP.2019.8682896

17. Affective video content analysis: A multidisciplinary insight / Y. Baveye, C. Chamaret, E. Dellandréa, L. Chen // *IEEE Transactions on Affective Computing*. 2017. 9(4). P. 396-409. https://doi.org/1-1. 10.1109/TAFFC.2020.2983669

18. Caruana R. Multitask learning // *Machine learning*. 1997. 28(1). P. 41-75. https://doi.org/10.1023/A:1007379606734

19. IEMOCAP: Interactive emotional dyadic motion capture database / C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S.S. Narayanan // *Language resources and evaluation.* 2008. 42(4). P. 335. https://doi.org/10.1007/s10579-008-9076-6

20. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing / F. Eyben, K.R. Scherer, B.W. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, K. Truong // *IEEE transactions on affective computing.* 2015. 7(2). P. 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

21. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism / B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, S. Kim // *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France. 2013. URL: https://mediatum.ub.tum.de/doc/1189705/file.pdf

22. Akçay M.B., Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers // *Speech Communication.* 2020. 116. P. 56-76. https://doi.org/10.1016/j.specom.2019.12.001

23. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals / B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. V. Kessous Aharonson // *Eighth Annual Conference of the International Speech Communication Association*. 2007. P. 2253-2256. URL: https://www.isca-speech.org/archive/interspeech_2007/i07_2253.html

24. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions / F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne // *2013 10th IEEE international conference and workshops on automatic face and gesture recognition* (*FG*). IEEE, 2013. P. 1-8. https://doi.org/10.1109/FG.2013.6553805

25. Sound classification using convolutional neural network and tensor deep stacking network / A. Khamparia, D. Gupta, N.G. Nguyen, A. Khanna, B. Pandey, P. Tiwari // *IEEE Access*. 2019. 7. P. 7717-7727. https://doi.org/10.1109/ACCESS.2018.2888882

26. Speaker-independent Japanese isolated speech word recognition using TDRC features / N.S.S. Srinivas, N. Sugan, L.S. Kumar, M.K. Nath, A. Kanhe // *2018 International CET Conference on Control, Communication, and Computing* (*IC4*). IEEE, 2018. P. 278-283. https://doi.org/10.1109/CETIC4.2018.8530947

27. Speaker identification using FrFT-based spectrogram and RBF neural network / P. Li, Y. Li, D. Luo, H. Luo // *2015 34th Chinese Control Conference* (*CCC*). IEEE, 2015. P. 3674-3679. https://doi.org/10.1109/ChiCC.2015.7260207

28. Speech emotion recognition for performance interaction / N. Vryzas, R. Kotsakis, A. Liatsou, C.A. Dimoulas, G. Kalliris // *Journal of the Audio Engineering Society.* 2018. 66(6). P. 457-467. https://doi.org/10.17743/jaes.2018.0036

29. Attention-based models for speech recognition / J.K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio // *Advances in neural information processing systems.* 2015. 28. P. 577-585. URL: https://papers.nips.cc/paper/2015/hash/ 1068c6e4c8051cfd4e9ea8072e3189e2-Abstract.html

30. A database of German emotional speech / F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss // *Ninth European Conference on Speech Communication and Technology.* 2005. URL: https://www.isca-speech.org/archive/archive_papers/ interspeech_2005/i05_1517.pdf

31. Dropout: a simple way to prevent neural networks from overfitting / N. Srivastava, G. Hinton, A., Krizhevsky I. Sutskever, R. Salakhutdinov // *The journal of machine learning research.* 2014. 15(1). P. 1929-1958. https://dl.acm.org/doi/abs/10.5555/2627435.2670313

32. Bilen H., Vedaldi A. Universal representations: The missing link between faces, text, planktons, and cat breeds //arXiv preprint arXiv:1701.07275. 2017.

33. Das A., Hasegawa-Johnson M., Veselý K. Deep Auto-Encoder Based Multi-Task Learning Using Probabilistic Transcriptions // I*NTERSPEECH*. 2017. P. 2073-2077. https://doi.org/10.21437/Interspeech.2017-582

34. Sanh V., Wolf T., Ruder S. A hierarchical multi-task approach for learning embeddings from semantic tasks // *Proceedings of the AAAI Conference on Artificial Intelligence.* – 2019. 33. P. 6949-6956. https://doi.org/10.1609/aaai.v33i01.33016949

35. Distral: Robust multitask reinforcement learning / Y. Teh, V. Bapst, W.M. Czarnecki, J. Quan, J. Kirkpatrick, R. Hadsell, N. Heess, R. Pascanu // *Advances in Neural Information Processing Systems.* 2017. 30. P. 4496-4506. URL: https://proceedings.neurips.cc/ paper/2017/hash/0abdc563a06105aee3c6136871c9f4d1-Abstract.html

36. Ranjan R., Patel V.M., Chellappa R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017. 41(1). P. 121-135. https://doi.org/10.1109/TPAMI.2017.2781233

37. Parthasarathy S., Busso C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning // *Interspeech.* 2017. P. 1103-1107. URL: https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1494.PDF

38. Progressive neural networks for transfer learning in emotion recognition / J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, E.M. Provost // arXiv preprint arXiv:1706.03256. 2017.

39. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception / C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E.M. Provost //

IEEE Transactions on Affective Computing. 2016. 8(1). P. 67-80. https://doi.org/10.1109/TAFFC.2016.2515617

40. Kendall A., Gal Y., Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. P. 7482-7491. https://doi.org/10.1109/CVPR.2018.00781

41. iebel L., Körner M. Auxiliary tasks in multi-task learning // arXiv preprint arXiv:1805.06334. 2018.

42. A comparison of loss weighting strategies for multi task learning in deep neural networks / T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, O.H. Elibol // *IEEE Access*. 2019. 7. P. 141627-141632. https://doi.org/10.1109/ACCESS.2019.2943604

43. Liu S., Johns E., Davison A. J. End-to-end multi-task learning with attention // *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019. P. 1871-1880. https://doi.org/10.1109/CVPR.2019.00197

44. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks / Z. Chen, V. Badrinarayanan, C.Y. Lee, A. Rabinovich // *International Conference on Machine Learning*. PMLR, 2018. P. 794-803. URL: http://proceedings.mlr.press/v80/chen18a.html

45. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks // *Communications of the ACM*. 2017. 60(6). P. 84-90. URL: https://dl.acm.org/doi/abs/10.1145/3065386

46. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition //arXiv preprint arXiv:1409.1556. 2014.

47. He K. et al. Deep residual learning for image recognition // *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016. P. 770-778. URL: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

48. Kingma D.P., Ba J. Adam: A method for stochastic optimization // arXiv preprint arXiv:1412.6980. 2014.

49. Livingstone S.R., Russo F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English // *PloS one*. 2018. 13(5). P. e0196391. https://doi.org/10.1371/journal.pone.0196391

50. Mariooryad S., Lotfian R., Busso C. Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora // *Fifteenth Annual Conference of the International Speech Communication Association*. 2014. URL: https://www.isca-speech.org/archive/interspeech_2014/i14_0238.html

51. Maaten L., Hinton G. Visualizing data using t-SNE // Journal of machine learning research. 2008. 9(Nov). P. 2579-2605. URL: https://www.jmlr.org/papers/v9/vandermaaten08a.html

52. Grad-cam: Visual explanations from deep networks via gradient-based localization / R.R. Sel-varaju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra // *Proceedings of the IEEE international conference on computer vision*. 2017. P. 618-626. URL: https:// openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_ Explanations_ ICCV_2017_paper.html

References

1. Tokuno S., Tsumatori, G., Shono S., Takei E., Yamamoto T., Suzuki G., Mituyoshi S., Shimura M. Usage of emotion recognition in military health care. *Defense Science Research Conference and Expo* (*DSR*). IEEE, 2011:1-5. https://doi.org/10.1109/DSR.2011.6026823

2. Saste S.T., Jagdale S.M. Emotion recognition from speech using MFCC and DWT for security system. *2017 international conference of electronics, communication and aerospace technology* (*ICECA*). IEEE, 2017; 1:701-704. https://doi.org/10.1109/ICECA.2017.8203631

3. Rázuri J.G., Sundgren D., Rahmani R., Moran A., Bonet I., Larsson A. Speech emotion recognition in emotional feedbackfor human-robot interaction. *International Journal of Advanced Research in Artificial Intelligence* (*IJARAI*), 2015, 4(2), pp. 20-27. https://doi.org/10.14569/IJARAI.2015.040204

4. Bojanić M., Delić V., Karpov A. Call redistribution for a call center based on speech emotion recognition. *Applied Sciences*, 2020, no. 10(13), pp. 46-53. https://doi.org/10.3390/ app10134653

5. Björn W., Schuller L. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the Acm*, 2018, no. 61(5), pp.90-99. https://doi.org/10.1145/3129340

6. Vilyunas V.K. [Emotions]. *Bol'shoj psihologicheskij slovar'* [Big psychological dictionary] /pod obshch. red. B.G. Meshcheryakova, V.P. Zinchenko (In Russ.). Available at: https://psychological.slovaronline.com/2078-EMOTSII

7. Il'in E.P., *Emocii i chuvstva* [Emotions and feelings]. Saint-Petersburg, Piter Publ., 2011 (In Russ.)

8. Sailunaz K., Dhaliwal M., Rokne J., Alhajj R. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 2018, no. 8(1), p. 28. https://doi.org/10.1007/s13278-018-0505-2

9. Ekman P. Facial expression and emotion. *American psychologist*, 1993. 48(4), 384 p. https://doi.org/10.1037/0003-066X.48.4.384

10. Russell J.A. Affective space is bipolar. *Journal of personality and social psychology,* 1979, no. 37 (3), 345 p. https://doi.org/10.1037/0022-3514.37.3.345

11. Russell J.A. Culture and the categorization of emotions. *Psychological bulletin*, 1991, no. 110 (3), 426 p. https://doi.org/10.1037/0033-2909.110.3.426

12. Trigeorgis G., Ringeval F., Brueckner R., Marchi E., Nicolaou M.A., Schuller B., Zafeiriou S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. *2016 IEEE international conference on acoustics, speech and signal processing* (*ICASSP*). IEEE, 2016:5200-5204. https://doi.org/10.1109/ICASSP.2016.7472669

13. Vryzas N., Vrysis L., Matsiola M., Kotsakis R., Dimoulas C., Kalliris G. Continuous Speech Emotion Recognition with Convolutional Neural Networks. *Journal of the Audio Engineering Society*, 2020, no. 68(1/2), pp. 14-24. https://doi.org/10.17743/jaes.2019.0043

14. Chen M., He X., Yang J., Zhang H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters,* 2018, no. 25(10), pp.1440-1444. https://doi.org/10.1109/LSP.2018.2860246

15. Satt A., Rozenberg S., Hoory R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Interspeech,* 2017, pp. 1089-1093. https://doi.org/10.21437/Interspeech.2017-200

16. Zhang Z., Wu B., Schuller B. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). IEEE, 2019, pp. 6705-6709. https://doi.org/10.1109/ICASSP.2019.8682896

17. Baveye Y., Chamaret C., Dellandréa E., Chen L. Affective video content analysis: A multidisciplinary insight. *IEEE Transactions on Affective Computing*, 2017, no. 9(4), pp. 396-409. https://doi.org/1-1. 10.1109/TAFFC.2020.2983669

18. Caruana R. Multitask learning. *Machine learning*, 1997, no. 28(1), pp. 41-75. https://doi.org/10.1023/A:1007379606734

19. Busso C., Bulut M., Lee C.C., Kazemzadeh A., Mower E., Kim S., Chang J., Lee S., Narayanan S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation,* 2008, no. 42(4), 335 p. https://doi.org/10.1007/s10579-008-9076-6

20. Eyben F., Scherer K.R., Schuller B.W., Sundberg J., André E., Busso C., Devillers L., Epps J., Laukka P., Narayanan S., Truong K. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing,* 2015, no. 7(2), pp. 190-202. https://doi.org/10.1109/TAFFC.2015.2457417

21. Schuller B., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Mortillaro M., Salamin H., Polychroniou A., Valente F., Kim S. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013. Available at: https://mediatum.ub.tum.de/doc/1189705/file.pdf

22. Akçay M.B., Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication.* 2020, no. 116, pp. 56-76. Available at: https://doi.org/10.1016/j.specom.2019.12.001

23. Schuller B., Batliner A., Seppi D., Steidl S., Vogt T., Wagner J., Devillers L., Vidrascu L., Amir N., Kessous L. Aharonson V. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 2253-2256. Available at: https://www.isca-speech.org/archive/interspeech_2007/i07_2253.html

24. Ringeval F., Sonderegger A., Sauer J., Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE international conference and workshops on automatic face and gesture recognition* (*FG*). IEEE, 2013, pp. 1-8. https://doi.org/10.1109/FG.2013.6553805

25. Khamparia A., Gupta D., Nguyen N.G., Khanna A., Pandey B., Tiwari P. Sound classification using convolutional neural network and tensor deep stacking network. *IEEE Access*, 2019; 7:7717-7727. https://doi.org/10.1109/ACCESS.2018.2888882

26. Srinivas N.S.S., Sugan N., Kumar L.S., Nath M.K., Kanhe A. Speaker-independent Japanese isolated speech word recognition using TDRC features. *2018 International CET Conference on Control, Communication, and Computing* (*IC4*). IEEE, 2018, pp. 278-283. https://doi.org/10.1109/CETIC4.2018.8530947

27. Li P., Li Y., Luo D., Luo H. Speaker identification using FrFT-based spectrogram and RBF neural network. *2015 34th Chinese Control Conference* (*CCC*). IEEE, 2015, pp. 3674-3679. https://doi.org/10.1109/ChiCC.2015.7260207

28. Vryzas N., Kotsakis R., Liatsou A., Dimoulas C.A., Kalliris G. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society,* 2018, 66(6), pp.457-467. https://doi.org/10.17743/jaes.2018.0036

29. Chorowski J.K., Bahdanau D., Serdyuk D., Cho K., Bengio Y. Attention-based models for speech recognition. *Advances in neural information processing systems,* 2015, 28, pp. 577-585. Available at: https://papers.nips.cc/paper/2015/hash/1068c6e4c8051cfd4e9ea8072e3189e2-Abstract.html

30. Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W.F., Weiss B. A database of German emotional speech. *Ninth European Conference on Speech Communication and Technology,* 2005. Available at: https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_1517.pdf

31. Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research.* 2014, no. 15(1), pp.1929-1958. Available at: https://dl.acm.org/doi/abs/10.5555/2627435.2670313

32. Bilen H., Vedaldi A. Universal representations: The missing link between faces, text, planktons, and cat breeds. arXiv preprint arXiv:1701.07275. 2017.

33. Das A., Hasegawa-Johnson M., Veselý K. Deep Auto-Encoder Based Multi-Task Learning Using Probabilistic Transcriptions. *INTERSPEECH*, 2017, pp. 2073-2077. https://doi.org/10.21437/Interspeech.2017-582

34. Sanh V., Wolf T., Ruder S. A hierarchical multi-task approach for learning embeddings from semantic tasks. *Proceedings of the AAAI Conference on Artificial Intelligence,* 2019, no. 33. pp. 6949-6956. https://doi.org/10.1609/aaai.v33i01.33016949

35. Teh Y., Bapst V., Czarnecki W.M., Quan J., Kirkpatrick J., Hadsell R., Heess N., Pascanu R. Distral: Robust multitask reinforcement learning. *Advances in Neural Information Processing Systems,* 2017, no. 30, pp.4496-4506. Available at: https://proceedings.neurips.cc/paper/2017/hash/0abdc563a06105aee3c6136871c9f4d1-Abstract.html

36. Ranjan R., Patel V.M., Chellappa R. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, no. 41(1), pp. 121-135. https://doi.org/10.1109/TPAMI.2017.2781233

37. Parthasarathy S., Busso C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. *Interspeech.* 2017:1103-1107. Available at: https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1494.PDF

38. Gideon J., Khorram S., Aldeneh Z., Dimitriadis D., Provost E.M. Progressive neural networks for transfer learning in emotion recognition. arXiv preprint arXiv:1706.03256. 2017.

39. Busso C., Parthasarathy S., Burmania A., AbdelWahab M., Sadoughi N., Provost E.M. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing,*. 2016, no. 8(1), pp.67-80. https://doi.org/10.1109/TAFFC.2016.2515617

40. Kendall A., Gal Y., Cipolla R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp.7482-7491. https://doi.org/10.1109/CVPR.2018.00781

41. Liebel L., Körner M. Auxiliary tasks in multi-task learning. arXiv preprint arXiv:1805.06334. 2018.

42. Gong T., Lee, T., Stephenson C., Renduchintala V., Padhy S., Ndirango A., Keskin G., Elibol O.H. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*. 2019; 7:141627-141632. https://doi.org/10.1109/ACCESS.2019.2943604

43. Liu S., Johns E., Davison A. J. End-to-end multi-task learning with attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871-1880. https://doi.org/10.1109/CVPR.2019.00197

44. Chen Z., Badrinarayanan V., Lee C.Y., Rabinovich A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. *International Conference on Machine Learning*. PMLR, 2018. pp.794-803. http://proceedings.mlr.press/v80/chen18a.html

45. Krizhevsky A., Sutskever I., Hinton G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, no. 60(6), pp.84-90. https://dl.acm.org/doi/abs/10.1145/3065386

46. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

47. He K. et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2016, pp. 770-778. Available at: https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

48. Kingma D.P., Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.

49. Livingstone S.R., Russo F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 2018, no. 13(5):e0196391. https://doi.org/ 10.1371/ journal.pone.0196391

50. Mariooryad S., Lotfian R., Busso C. Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. *Fifteenth Annual Conference of the International Speech Communication Association*. 2014. Available at: https://www.isca-speech.org/archive/interspeech_2014/i14_0238.html

51. Maaten L., Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*, 2008, 9(Nov), pp. 2579-2605. Available at: https://www.jmlr.org/papers/v9/ vandermaaten08a.html

52. Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Gradcam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. 2017:618-626. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

Информация об авторах / Information about the Authors

Рябинов Артем Валерьевич, программист лаборатории автономных робототехнических систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, г. Санкт-Петербург, Российская Федерация, e-mail: iamryabinov@gmail.com, ORCID: http://orcid.org/0000-0002-3572-4493

Artem V. Ryabinov, Software Engineer of Laboratory of Autonomous Robotic Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation, e-mail: iamryabinov@gmail.com, ORCID: http://orcid.org/0000-0002-3572-4493

Уздяев Михаил Юрьевич, младший научный сотрудник лаборатории технологий больших данных социокиберфизических систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, г. Санкт-Петербург, Российская Федерация, e-mail: m.y.uzdiaev@gmail.com, ORCID: http://orcid.org/0000-0002-7032-0291

Mikhail Yu. Uzdiaev, Junior Researcher of Laboratory of Big Data In Socio-Cyberphysical Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation, e-mail: m.y.uzdiaev@gmail.com, ORCID: http://orcid.org/0000-0002-7032-0291

Ватаманюк Ирина Валерьевна, младший научный сотрудник лаборатории автономных робототехнических систем, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербургский институт информатики и автоматизации Российской академии наук, г. Санкт-Петербург, Российская Федерация, e-mail: vatamaniuk.i.v@gmail.com, ORCID: http://orcid.org/0000-0001-5388-8152

Irina V. Vatamaniuk, Junior Researcher of Laboratory of Autonomous Robotic Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russian Federation, e-mail: vatamaniuk.i.v@gmail.com, ORCID: http://orcid.org/0000-0001-5388-8152