УДК 004.93

ПРИМЕНЕНИЕ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ МУЛЬТИМОДАЛЬНОЙ ИДЕНТИФИКАЦИИ

Стефаниди А.Ф., *аспирант Ярославского государственного университета им. П.Г. Демидова, e-mail: antonstefanidi@mail.ru*
Приоров А.Л., *д.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: andcat@yandex.ru*
Топников А.И., *к.т.н., Ярославский государственный университет им. П.Г. Демидова, e-mail: topartgroup@gmail.com*
Хрящев В.В., *к.т.н., доцент Ярославского государственного университета им. П.Г. Демидова, e-mail: vhr@yandex.ru*

THE PROBLEM OF PERSONALITY RECOGNITION USING FACIAL IMAGES AND AUDIO SIGNALS WITH SPEECH RECORDINGS

Stefanidi A.F., Priorov A.L., Topnikov A.I., Hryashev V.V.

Currently, biometric identification systems are often used in mobile applications, banking systems, access control and management systems as well as for the management of mobile robots. In this paper, we consider the problem of personality recognition using facial images and audio signals with speech recordings. The results of the research will be used to create a system of multimodal biometric identification. Since convolutional neural networks demonstrate the highest results regarding the problems of detection, segmentation and classification of objects, this paper also proposes an approach to person identification based on convolutional neural networks. The research was carried out using modern audiovisual database VoxCeleb1. To decrease the computational capability of the experiment, the researchers reduced the number of classes from 1251 to 200. The development results showed the possibility of using the proposed algorithm as a part of a multimodal identity identification system.

Key words: digital speech processing, digital image processing, machine learning, speaker identification, face recognition, convolutional neural network, bimodal biometrics.

Ключевые слова: цифровая обработка речевых сигналов, цифровая обработка изображений, машинное обучение, идентификация диктора, распознавание лиц, сверточная нейронная сеть, бимодальная биометрия.

В настоящее время системы биометрической идентификации личности пользуются высокой популярностью в мобильных приложениях, банкинг-системах, системах контроля и управления доступом, в задачах управления мобильными роботами. В данной работе рассматривается задача распознавания личности с использованием цифровых изображений лиц и речевых сигналов. Результаты исследования планируется использовать в комбинации для создания системы мультимодальной биометрической идентификации. Поскольку сверточные нейронные сети демонстрируют наиболее высокие результаты в задачах детектирования, сегментации и классификации объектов, в данной работы также предложен подход идентификации личности на основе сетей данного типа. Исследования проводились с использованием современной базы аудиовизуальных данных VoxCeleb1. Для снижения вычислительной сложности исследования количество классов уменьшено с 1251 до 200. Результаты моделирования показали возможность применения предлагаемых алгоритмов в составе композитной мультимодальной системы идентификации личности.

Введение

В настоящее время существует множество подходов для идентификации и аутентификации личности, однако методы на основе анализа биометрических признаков являются наиболее эффективными. В частности, они, в отличие от паролей и токенов, не могут быть украдены, потеряны или забыты. Эти важные свойства способствуют развитию и все большему внедрению биометрических технологий [1, 2].

Большинство биометрических систем являются унимодальными, то есть используют в своей работе один источник биометрической информации. Выбор типа биометрических признаков в значительной степени определяет достоинства и недостатки системы идентификации. Так, например, распознавание диктора в шумных реальных условиях является чрезвычайно сложной задачей. Связано это с высокой степенью вариативности внешних и внутренних параметров системы: фоновые разговоры, музыка, смех, фоновые вибрации, эффекты канала передачи информации и микрофона, физиологические особенности говорящего, акцент, эмоции, интонация [3]. В задаче идентификация пользователя по цифровому изображению лица также есть сильная зависимость от внешних и внутренних факторов: степень освещенности, качество светочувствительного датчика, угол наклона и поворота головы, возрастные изменения, появление у человека очков/бороды/усов, эмоциональная активность и мимика. Поэтому для повышения точности и компенсации

Рис. 1. Блох-схема мультимодальной биометрической системы идентификации личности

недостатков унимодальных подходов используют методы комбинирования сильно отличающихся друг от друга признаков. Такие системы называют мультимодальными (число независимых признаков два и более) [4].

Данное исследование посвящено разработке и исследованию методов идентификации личности на основе анализа изображений лиц и аудио сигналов. Полученные решения будет применяться для создания системы мультимодальной биометрического идентификации (рис. 1). Распознавание по лицу и голосу дает возможность получения биометрических параметров в отсутствии физического контакта человека с системой, что расширяет спектр практического использования данной технологии. Использование речи и изображений лиц повышает устойчивость к возможным спуфинг атакам, фальсификации данных, а также попыткам несанкционированного доступа [4].

Сверточные нейронные сети в настоящее время являются стандартным решением в задачах распознавания лиц, именно с их помощью получены наилучшие на сегодня результаты [1]. В задаче текстонезависимой идентификации диктора долгие годы таким стандартным решением являлось использование мел-частотных кепстральных коэффициентов и классификатора на основе моделей гауссовых смесей [4-6]. Для робастности системы к внешним условиям часто использовали универсальную фоновую модель [4, 6, 7], а также совместный факторный анализ [8], метод полной изменчивости [9], вероятностный линейный дискриминантный анализ [10-12]. Однако развитие нейронных сетей и глубокого обучения затронули и область распознавания диктора [13-15]. В последние годы стали появляться работы, демонстрирующие эффективность применения нейросетей в этой задаче. Применение сверточных нейронных сетей в задаче идентификации диктора стало возможным, в том числе и за счет трансформации аудио данных в двумерное представление сигнала. Это достигается в результате перехода из временной области в частотную [5].

Целью работы является разработка алгоритмов идентификации личности с использованием речевой и лицевой модальности. Реализованные алгоритмы будут применяться для проектирования мультимодальной биометрической системы на следующих этапах исследования.

Описание базы аудиовизуальных данных

Для проведения эксперимента использовалась известная тестовая база VoxCeleb1. Это аудиовизуальный набор данных, состоящий из коротких фрагментов человеческой речи и цифровых изображений лиц, извлеченных из видео интервью, загруженных на YouTube [16]. База VoxCeleb1 включает в себя речь спикеров, охватывающих широкий спектр различных национальностей, стиля произношения, профессий и возраста. Она содержит более 150 000 аудио примеров для 1251 класса. Набор голосовых данных является достаточно сбалансированным по гендерному признаку – 55 % составляет речь мужчин и 45 % речь женщин. Длительность аудио записей изменяется от 4 до 145 сек., в среднем 8,2 сек. (рис. 2.). На каждого диктора приходится от 45 до 250 аудио дорожек, в среднем 123.

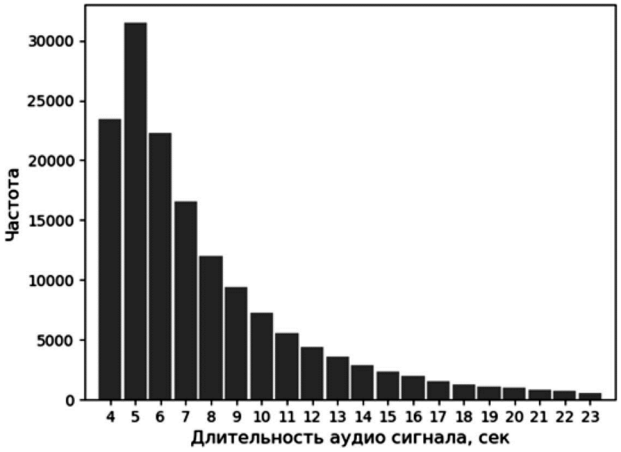

Рис. 2. Гистограмма длительностей звуковых сигналов из базыVoxCeleb1

Важно отметить, что условия записи данных максимально приближены к реальным. Звуковые примеры, включенные в набор данных, собирались в сложных акустических условиях с использованием видеокамер и микрофонов с различными техническими характеристиками. Запись велась в таких местах, как Красная дорожка, открытый стадион, студия телешоу, выступление на сцене перед большой аудиторией, на интервью со съемок кинофильмов и других мест, обладающих неповторимыми и уникальными акустическими свойствами. Во многих примерах присутствуют шумы естественного происхождения: фоновая речь, смех, перекрывающая речь, шумы, вызванные акустическими особенностями помещения [16]. Многие отмечают, что данная база является сложной для задачи идентификации диктора, поскольку звуковые дорожки могут содержать фрагменты с перекрывающейся речью, например, когда двое людей говорят параллельно, тем самым ухудшая процесс обучения нейронной сети и в итоге снижая точность

Рис. 3. Примеры изображений лиц из базы данных VoxCeleb1

Таблица 1. Статистика анализируемой части аудиовизуальной базы VoxCeleb1 на 200 классов

	Обучение	Валидация	Тест	Суммарно	Доля базы VoxCeleb1
Изображения	91 331	11 417	11 314	114 062	9,36%
Речевые сигналы	12 599	1 123	1 343	15 065	9,81%

работы модели. Однако в данной работе принято, что указанное свойство является достаточно распространенным и имеет место быть в повседневной жизни, поэтому важно получить систему, способную работать в такого рода условиях [17, 18].

База VoxCeleb1 также содержит набор цифровых изображений лиц, детектированных и вырезанных в процессе обработки видео роликов с YouTube (рис. 3). Общее количество изображений составляет более 1,2 млн. Данная база лиц имеет следующие особенности: состоит из цветных изображений; лица имеют разный угол поворота/наклона головы, цвет лица/волос различен, наличие/отсутствие очков/бороды, усов; различные сцены и степень освещенности. Это позволяет утверждать, что эксперимент может быть проведен в условиях, приближенных к условиям реальной эксплуатации разрабатываемой системы.

Благодаря наличию хорошо структурированной и размеченной базы данных VoxCeleb1, состоящей из цифровых изображений лиц и речевых сигналов, открывается возможность разработки мультимодальной (бимодальной) системы идентификации на основе двух биометрических признаков: лица и голоса.

Для уменьшения вычислительной сложности и длительности экспериментальных исследований количество определяемых классов уменьшено с 1251 до 200. Это позволило более динамично проводить все этапы исследования и получить высокие результаты для задачи классификации цифровых изображений и аудио сигналов. Также отметим, что практическое применение систем идентификации личности, как правило сводится к задачам определения от нескольких десятков до пары сотен объектов. В табл. 1 представлен объем исследуемой части аудиовизуальной базы VoxCeleb1.

Предобработка данных и архитектуры сетей

Задача распознавания лиц решалась с использованием предобученной сверточной нейронной сети VGGFace. Данная архитектура показывает высокие результаты в задачах классификации изображений [19-22]. Выбрана реализация, обученная на коллекции лиц известных людей VGG-Face в соответствии с базой данных интернет фильмов IMDB. В общей сложности набор состоит из 2622 классов с общим количеством изображений более 2,6 млн.

Сверточная неройнная сеть VGGFace принимает на вход отмасштабированные цветные изображения размером 224×224 пикселей. Для решения задачи верхние слои предобученной сети удалялись и вместо них собирался новый классификатор на основе двух полносвязных слоев, состоящих из 512 нейронов и последующей функцией активации ReLU. На выходе сети использовался 200-мерный softmax-слой (рис. 4). В качестве алгоритма численной оптимизации использовался Adam (adaptive moment estimation) с начальной скоростью сходимости 0,001. В процессе обучения скорость оптимизатора динамически уменьшалось в 0,9 раза при условии попадания в локальный минимум [23].

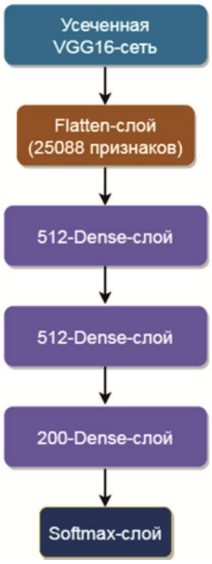

Рис. 4. Архитектура сверточной нейронной сети, используемой для классификации лиц из базы VoxCeleb1

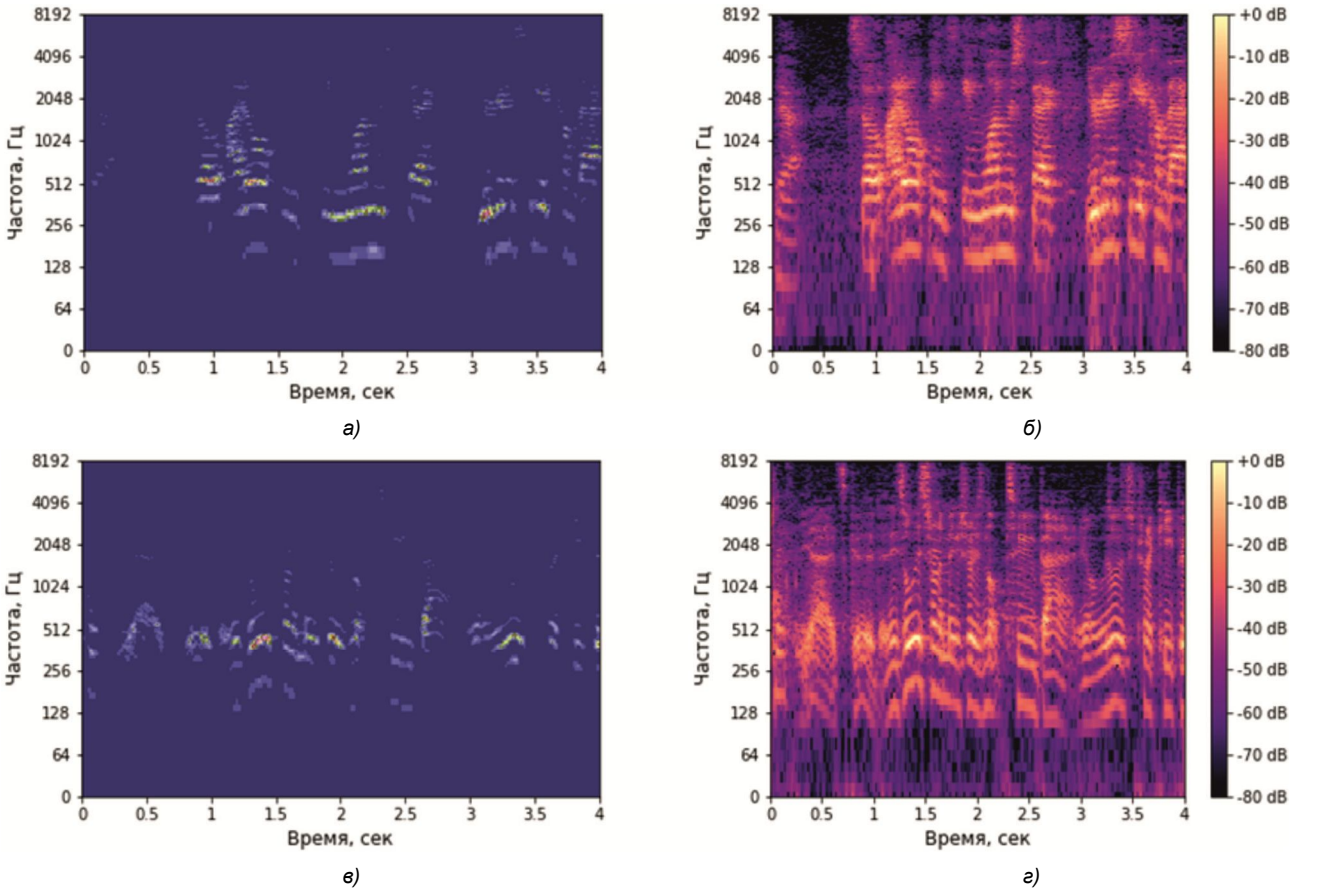

Рис. 5. Примеры частотного спектра аудио сигналов в логарифмическом масштабе: а, в – модуль амплитуды; б, г – логарифм мощности спектрограммы

Для увеличения количества изображений применялся метод искусственной аугментации данных с использованием следующих примитивных преобразований: зумирование, смещение изображения относительно вертикали и горизонтали, изменение ротации. Размер батча составлял 128 изображений. Обучение осуществлялось в течение 50 эпох, что по меркам глубокого обучения считается малым показателем. Связанно это с тем, что сеть не обучалась «с нуля», а брались веса, полученные в ходе анализа данной сетью другой базы изображений лиц – VGG-Face. Это позволило в разы сократить время, необходимое на обучение.

Весь процесс предобработки аудио данных основывался на применение библиотеки с открытым исходным кодом librosa. Звуковые сигналы представлялись в формате wav с частотой дискретизации 16 кГц и уровнем квантования по амплитуде в 16 бит. Для перехода в частотную область использовалось быстрое преобразование Фурье (БПФ) на основе реализации librosa.core.stft со следующими параметрами: 64 мс длина окна, что эквивалентно 1024 временным отчетам, с шагом в 10 мс. Поскольку речевые сигналы имели разную длительность, они приводились к одной длине путем обрезания. Из каждого сигнала производилось случайное извлечение одной звуковой дорожки продолжительностью 4 с. Если изначально звуковая запись была меньше 4 с, то необходимую длину создавали путем дублирования речевых фрагментов этого же сигнала. Таким образом, каждый звуковой сигнал представлялся в виде спектрограммы размерностью 513x401x1. Далее брался модуль FFT и мощность спектрограммы конвертировалась в децибелы (рис. 5). В качестве нормировки использовалась пиковая мощность спектрограммы. Это описывается следующей формулой:

$$S_{dB} = 10 \cdot \log_{10}(S) - 10 \cdot \log_{10}(S_p),$$

где S – мощность спектрограммы сигнала; S_p – пиковая мощность спектрограммы сигнала; S_{dB} – логарифм мощности спектрограммы в децибелах.

Переход из временной в частотную область позволил преобразовать звуковой сигнал в матричное представление. Спектрограмма довольно часто используется в задачах идентификации и аутентификации диктора [15-17]. Двумерный тип объекта хорошо подходит для работы со сверточными нейронными сетями. Для решения задачи распознавания диктора выбрана сверточная нейронная сеть архитектуры VGGM. В качестве оптимизатора использовался Adam со скоростью сходимости 0,001. В процессе обучения скорость оптимизатора динамически уменьшалось в 0,9 раза при условии попадания в локальный минимум. Размер батча составил 32, при этом количество эпох обучения равнялось 200. На рис. 6 представлена архитектура используемой нейронной сети.

Для обработки звуковых и видео данных, а также обучения нейронных сетей применялся суперкомпьютер NVIDIA DGX-1 VOLTA производительностью до 960 Тфлопс, принадлежащий Ярославскому государственному университету им. П.Г. Демидова. В качестве

фреймворка выбран Keras с внутренней реализацией на Tensorflow.

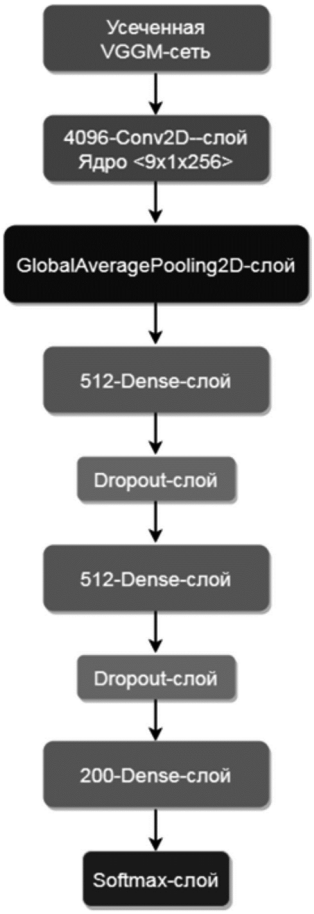

Рис. 6. Архитектура сверточной нейронной сети, используемой для классификации аудио сигналов из базы VoxCeleb1

Результаты исследований

Для анализа процесса обучения сверточных нейронных сетей использовались следующие метрики: оценка доли правильных ответов (accuracy, acc), точность (precision, P), полнота (recall, R), средневзвешенное значение полноты и точности (F-мера, F1-score). На рис. 7 представлен процесс мониторинга обучения сети для классификации лиц из базы цифровых изображений VoxCeleb1.

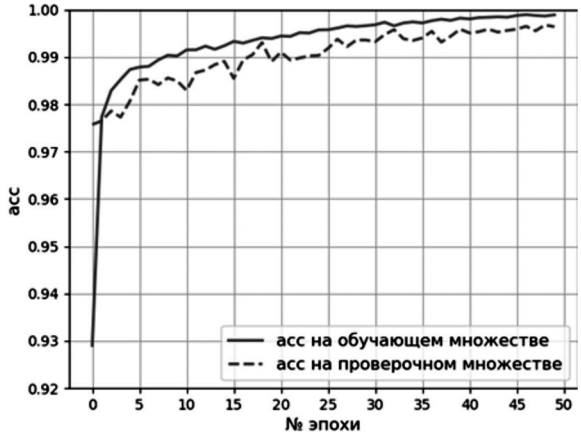

Рис. 7. Анализ доли правильных ответов в процессе обучения сети CNN-VGGFace

Из результатов видно, что уже на первой эпохе обучения точность классификации на проверочном (кросс-валидационном) множестве составляет более 97 %, что является отличным результатом. После прохождения обучения в 50 эпох классификатор имеет точность в 99,64 %. На тестовом наборе данных точность составляет 99,57 %. Важно отметить, что отсутствие признаков переобучения и недообучения, а также высокая точность на тестовом множестве говорят о хорошей обобщающей способности данной нейронной сети.

Для того чтобы качественно провести анализ работы сверточной нейронной сети по метрикам P, R, F1-score важно проверить кросс-валидационную выборку изображений на смещение внутри классов. Для этого построим гистограмму для кросс-валидационного множества, описывающую количество примеров внутри каждого из классов (рис. 8).

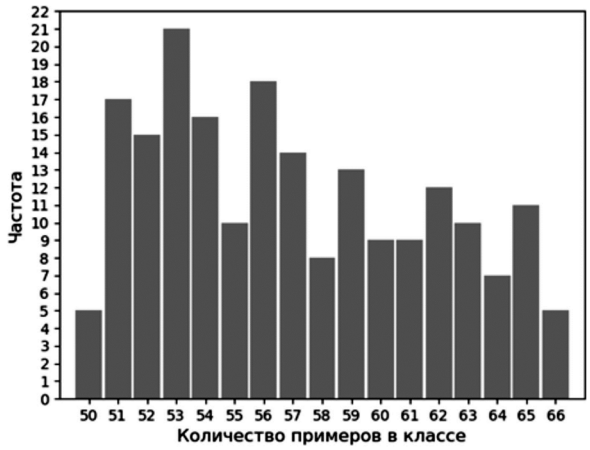

Рис. 8. Гистограмма описания количества примеров в каждом из 200 классов

Из полученных результатов видно, что количество примеров в каждом из классов варьируется от 50 до 66, при этом среднее значение равно 57. Исходя из этого, можно сказать, что перекоса внутри проверочного множества нет, а значит, можно вычислить описанные выше численные показатели с использованием micro- или macro-усредняющего подхода для мультиклассовой классификации. В данной работе используется макро-усреднение. Оно подразумевает расчет количественных показателей P, R, F1-score внутри каждого класса с последующим усреднением. В итоге получены следующие результаты: P = 99,90 %, R = 99,87 %, F1-score = 99,88 % [24].

Ранее уже отмечалось, что работа ведется с выборкой только на 200 классов. Для проведения исследования по идентификации диктора данные делились на три части – обучающую, проверочную и тестовую выборку. Обучающая выборка содержала 12599 аудиозаписей, проверочный и тестовый набор данных содержали 1123 и 1343 аудиозаписи соответственно. Для контроля сходимости модели использовалась категориальная функция потерь $J(Q)$ (Categorical Cross-Entropy Loss), представленная на рис. 9 [25].

Здесь представлен процесс обучения сверточной нейронной сети. Реализация является верной, поскольку $J(Q)$ имеет монотонно убывающую тенденцию, спускаясь в итоге на «плато». Дополнительно во время обуче-

ния проводился мониторинг метрики оценки доли правильных ответов (accuracy, acc) на обучающем и проверочном множестве (рис. 10).

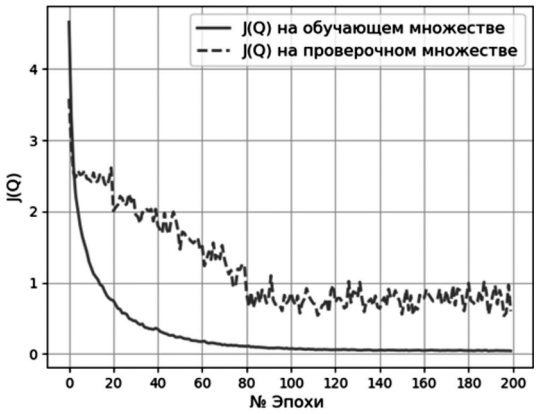

Рис. 9. Анализ функции потерь в процессе обучения сети CNN-VGGM

Из рис. 9 и рис. 10 можно сделать вывод о том, что сеть научилась классифицировать звуковые данные из обучающей выборки, имея уровень accuracy в 98,91 %. Однако если посмотреть на кривую, описывающую точность работы модели на кросс-валидационном множестве, то можно увидеть более низкий уровень accuracy в 78,87 %. Это свидетельствует о том, что модель переобучилась. Одним из самых популярных способов борьбы с переобучением является метод прореживания слоев. К сожалению, такой подход не дал существенных улучшений. Другой способ основывается на синтетическом увеличении речевых данных, но в данной работе он не использовался, поскольку требует проведение дополнительных глубоких исследований. Методы аугментирования речевых сигналов и анализ их влияния на обобщающую способность нейронных сетей будут подробно описаны в следующей работе.

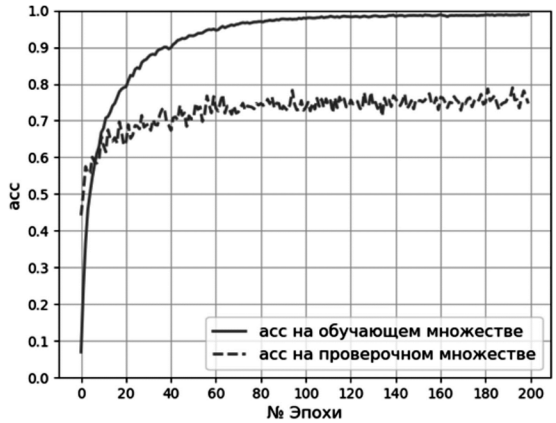

Рис. 10. Анализ доли правильных ответов в процессе обучения сети CNN-VGGM

В табл. 2 отображены результаты работы CNN-VGGM на обучающем, проверочном и тестовом множестве. В качестве оценки точности классификации используется метрика accuracy различного типа – top-1, top-3, top-5.

Видно, что модель показывает на тестовой выборке уровень точности acc-top-5 86,97 % в задаче классификации 200 классов из базы VoxCeleb1.

В качестве рекомендации по улучшению качество работы CNN-VGGM можно использовать аугментирование аудио данных на основе различных преобразований и типов помех. Также увеличение обучающего набора возможно за счет использования полного набора данных VoxCeleb1 или более крупной базы речевых сигналов VoxCeleb2 [16-17]. Стоит отметить, что в данной работе в качестве представления речевого сигнала использовалась его спектрограмма. Такой подход является достаточно универсальным, однако существует ряд альтернативных методов, в частности подход на основе выделения мел-частотных кепстральных коэффициентов. Эти возможности будут учтены при дообучении и тонкой настройке сети CNN-VGGM на следующем этапе исследования.

Заключение

В исследовании рассматривались вопросы классификации личности по таким биометрическим параметрам, как голос и лицо. Для решения поставленных задач использовались сверточные нейронные сети архитектуры VGG-типа различных модификаций. В качестве базы цифровых изображений лиц и речевых сигналов выбран аудиовизуальный набор данных VoxCeleb1. Для уменьшения вычислительной сложности и длительности экспериментальных исследований количество определяемых классов уменьшено с 1251 до 200. Это позволило более динамично проводить все этапы исследования и получить высокие результаты для задачи классификации цифровых изображений и звуковых сигналов.

Для идентификации личности по лицу применялась нейронная сеть CNN-VGGFace, предобученная на крупной базе цифровых изображений VGG-Face. В процессе исследования применялся метод переноса обучения с использованием тонкой настройки. В итоге сеть дообучалась под исследуемый набор данных и показала высокие результаты на тестовом множестве: accuracy = 99,57 %, P = 99,87 %, R = 99,90 %, F1-score = 99,88 %.

Набор звуковых данных представлялся фрагментами голосовой активности. В качестве сети применялась CNN-VGGM. Установлено, что точность классификации дикторов на тестовой выборке по метрики acc-top-5 составила 86,97 %.

Таблица 2. Результаты идентификации дикторов из база речевых сигналов VoxCeleb1

Accuracy	Top-1 (%)	Top-3 (%)	Top-5 (%)
Обучение	93,10	96,73	98,52
Валидация	77,27	83,17	88,32
Тест	74,65	82,40	86,97

На следующем этапе исследования планируется повышение точности распознавания дикторов путем применения мел-частотных кепстральных коэффициентов; модернизации топологии обучаемых сетей; изменения параметров регуляризации; применения методов синтетического аугментирования речевых сигналов; использования полного набора данных VoxCeleb1 или более крупной базы речевых сигналов VoxCeleb2. Дополнительно будут проведены исследования по созданию мультимодального решения на основе речевой и лицевой модальности, а также по разработке алгоритмов объединения биометрических систем.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-37-90158.

Литература

1. Cao Q., Shen L., Xie W., Parkhi O.M., Zisserman A. VGGFace2: A dataset for recognising faces across pose and age, 2018, Web: https://arxiv.org/abs/1710.08092.

2. Lebedev A., Khryashchev V., Priorov A., Stepanova O. Face verification based on convoluional neural network and deep learning, In Proceedings of 15-th IEEE East-West Design and Test Symposium (EWDTS 2017), Novi Sad, Serbia, 2017, pp. 261-265.

3. Stoll L.L. Finding difficult speakers in automatic speaker recognition. Technical Report No. UCB/EECS-2011-152, 2011.

4. Khryashchev V., Topnikov A., Stefanidi A., Priorov A. Bimodal person identification using voice data and face images, In Proceedings SPIE 11041, Eleventh International Conference on Machine Vision, WEB: https://doi.org/10.1117/12.2523138.

5. Reynolds D.A., Quatieri T.F., Dunn R.B. Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, Vol.10, 2000, pp. 19-41.

6. Tupitsin G., Topnikov A., Priorov A. Two-step noise reduction based on soft mask for robust speaker identification, In Proceedings 18th Conference of Open Innovations Association FRUCT, 2016, pp. 351-356.

7. May T., S. van de Par, Kohlrausch A. Noise-Robust speaker recognition combining missing data techniques and universal background modeling, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, 2012, pp. 108-121.

8. Kenny P. Joint factor analysis of speaker and session variability: Theory and algorithms, CRIM, Montreal, (Report) CRIM-06/08-13, 2005.

9. Dehak N., Dehak R., Kenny P., Brümmer N., Ouellet P., Dumouchel P. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification, In Proceedings INTERSPEECH, 2009, pp. 1559-1562.

10. Prince S.J.D., Elder J.H. Probabilistic Linear Discriminant Analysis for Inferences About Identity, In Proceedings IEEE 11th International Conference on Computer Vision ICCV, 2007, pp. 1-8.

11. Garcia-Romero D., Espy-Wilson C.Y. Analysis of i-vector Length Normalization in Speaker Recognition Systems, In Proceedings INTERSPEECH, 2011, pp. 249-252.

12. Kenny P. Bayesian Speaker Verification with Heavy-Tailed Priors, In Proceedings Odyssey, 2010.

13. Ghahabi O., Hernando J. Deep Learning Backend for Single and Multi-session i-Vector Speaker Recognition, IEEE Transactions on Audio, Speech, and Language Processing, vol. 25, no. 4, 2017, pp. 807-817.

14. Ault S.V., Perez R.J., Kimble C.A., Wang J. On Speech Recognition Algorithms, International Journal of Machine Learning and Computing, vol. 8, no. 6, 2018, pp. 518-523.

15. Bunrit S., Inkian T., Kerdprasop N. Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network, International Journal of Machine Learning and Computing, vol. 9, no. 2, 2019, pp. 143-148.

16. Nagrani A., Chung J.S., Zisserman A. VoxCeleb: a large-scale speaker identification dataset, 2017, Web: https://arxiv.org/abs/ 1706.08612v2.

17. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition, In Proceedings Interspeech, 2018, pp. 1086-1090.

18. Xiang X., Wang S., Huang H., Qian Y., Yu K. Margin Matters: Towards More Dicsriminative Deep Neural Network Embeddings for Specker Recognition, 2019, Web: https://arxiv.org/abs/1906. 07317v1.

19. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition, In International Conference on Learning Representations, 2015, Web: https://arxiv.org/abs/ 1409.1556v6.

20. Chatfield K., Simonyan K., Vedaldi A., Zisserman A. Return of the Devil in the Details: Delving Deep into Convolutional Nets, In Proceedings British Machine Vision Conference, 2014, pp. 1-11.

21. Sun Y., Ding L., Wang X., Tang X. DeepID3: Face recognition with very deep neural networks, 2015, Web: https://arxiv.org/ abs/1502.00873.

22. Taigman Y., Yang M., Ranzato M., Wolf L. Deepface: Closing the gap to human-level performance in face verification, In IEEE Conf. on CVPR, 2014.

23. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization, 2017, Web: https://arxiv.org/abs/1412.6980v9.

24. Sokolova M., Japkowicz N., Szpakowicz S. Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation, In Proceeding of National Conference on Artificial Intelligence, 2016, pp. 1-6.

25. Liu W., Wen Y., Yu Z., Yang M. Large-margin softmax loss for convolutional neural networks, In ICML, 2016, pp. 507-516.