**Обзор аналогов.**

Распознавание эмоций в речи (Speech Emotion Recognition, SER) – область исследований, которая использует техники машинного обучения и наборы данных с эмоциональной речью в целях разработки моделей для автоматического распознавания эмоциональной окраски речи. До 2016 года в литературе превалировали традиционные методы, основанные на пофреймовом извлечении локальных низкоуровневых дескрипторов с последующим применением к ним функционалов для получения глобальных признаков изучаемого высказывания или отрывка речи, и использование полученного признакового представления для обучения алгоритмов классификации или регрессии [Schuller, Björn. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM. 61. 90-99. 10.1145/3129340.]. Исследователи изучали многие разработанные вручную признаки и их комбинации. Наиболее эффективными и часто используемыми наборами стали наборы eGeMAPS (88 параметров) [Eyben, Florian & Scherer, Klaus & Schuller, Björn & Sundberg, Johan & Andre, Elisabeth & Busso, Carlos & Devillers, Laurence & Epps, Julien & Laukka, Petri & Narayanan, Shrikanth & Truong, Khiet. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing. 7. 1-1. 10.1109/TAFFC.2015.2457417.] и ComParE (6373 параметра) [Schuller, Björn & Steidl, Stefan & Batliner, Anton & Vinciarelli, Alessandro & Scherer, Klaus & Ringeval, Fabien & Chetouani, Mohamed & Weninger, Felix & Eyben, Florian & Marchi, Erik & Mortillaro, Marcello & Salamin, Hugues & Polychroniou, Anna & Valente, Fabio & Kim, Samuel. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 148-152.]. В качестве классификаторов в литературе наиболее часто встречаются машины опорных векторов, алгоритм k ближайших соседей, скрытые марковские модели, многослойные перцептроны [Akçay, Berkehan & Oguz, Kaya. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication. 116. 10.1016/j.specom.2019.12.001.]. Также, 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 [наша статья]. Самые распространенные задачи распознавания предполагают классификацию различных эмоциональных классов, основанных на теории эмоций Пола Экмана (Гнев, Счастье, Отвращение и др.) или моделирование эмоциональных атрибутов (Возбуждение, Валентность, Доминантность и др.)[Sailunaz, Kashfia & Dhaliwal, Manmeet & Rokne, Jon & Alhajj, Reda. (2018). Emotion Detection from Text and Speech - A Survey. Social Network Analysis and Mining (SNAM), Springer. 8. 10.1007/s13278-018-0505-2.].

Очевидны недостатки традиционных подходов. Во-первых, ручное проектирование признаков требует привлечения экспертов по акустике и психологии, чтобы разработать набор наиболее релевантных параметров[Schuller, Björn & Batliner, Anton & Seppi, D. & Steidl, S. & Vogt, T. & Wagner, J. & Devillers, Laurence & Vidrascu, L. & Amir, Noam & Kessous, Loic & Aharonson, Vered. (2007). The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proc. Interspeech. 2253-2256.]. Только те признаки, которые явно показали высокую степень корреляции с эмоциями, будут отобраны путем обширных и тщательно подготовленных экспериментов, что довольно трудоемко и утомительно. Более того, эффективность выбранных признаков по-прежнему сильно зависит также от реализованной модели распознавания образов, что приводит к снижению эффективности [наша статья]. 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 пайплайна. 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 (CNN) и рекуррентных (RNN) архитектур глубоких нейронных сетей для таких систем. Так, в работе [G. Trigeorgis et al., "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), Shanghai, 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669.] был впервые описан сквозной подход к распознаванию эмоций по голосу. Авторы применили сверточные и рекуррентные с долгой краткосрочной памятью (LSTM) нейронные сети 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.

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, такие как спектрограммы. Спектрограммы – это визуальные представления силы сигнала с течением времени на разных частотах, полученные с помощью кратковременного 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 [A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, ‘‘Sound classification using convolutional neural network and tensor deep stacking network,’’ IEEE Access, vol. 7, pp. 7717–7727, 2019], распознавание речи [N. S. S. Srinivas, N. Sugan, L. S. Kumar, M. K. Nath, and A. Kanhe, ‘‘Speaker-independent japanese isolated speech word recognition using TDRC features,’’ in Proc. Int. CET Conf. Control, Commun., Comput. (IC4), Jul. 2018, pp. 278–283], распознавание человека по голосу [P. Li, Y. Li, D. Luo, and H. Luo, ‘‘Speaker identification using FrFT-based spectrogram and RBF neural network,’’ in Proc. 34th Chin. Control Conf. (CCC), Jul. 2015, pp. 3674–3679], продемонстрировали применимость спектрограмм для извлечения из них скрытых высокоуровневых признаков с помощью сверточных архитектур глубоких нейронных сетей и подтолкнули исследователей на использование спектрограмм в области распознавания эмоций в речи­­­­­­­­.

В работе [Vryzas, Nikolaos & Vrysis, Lazaros & Matsiola, Maria & Kotsakis, Rigas & Dimoulas, Charalampos & Kalliris, George. (2020). Continuous Speech Emotion Recognition with Convolutional Neural Networks. Journal of the Audio Engineering Society. Audio Engineering Society. 68. 14-24. 10.17743/jaes.2019.0043.] продемонстрирована модель распознавания эмоций в речи, основанная на применении к спектрограммам сверточных нейронных сетей. Набор данных, используемый для обучения и тестирования модели - динамическая база данных Acted Emotional Speech Dynamic Database (AESDD). Предлагаемая архитектура сверточной нейронной сети (4 сверточных слоя и 2 полносвязных слоя) превзошла базовую модель машинного обучения (машина опорных векторов в самостоятельно разработанном авторами признаковом пространстве) на 8,4% с точки зрения точности (unweighted accuracy). Авторами [M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.] представлена нейронная сеть, комбинирующая трехмерные сверточные слои, bi-LSTM ячейки и механизм внимания. В качестве входных данных использовались мел-спектрограммы, дополненные первой и второй производной по времени. Получен результат средней невзвешенной полноты (unweighted average recall, UAR) 64,74% на наборе данных IEMOCAP и 82,82% на наборе данных Emo-DB. В статье [Satt, Aharon, S. Rozenberg and R. Hoory. “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms.” INTERSPEECH (2017).] 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 топология, содержащая 3 сверточных и 2 LSTM слоя, точность распознавания на наборе данных IEMOCAP составила 68,8%. Наконец, Mustaqeem и др. в своей работе [Mustaqeem & Sajjad, Muhammad & Kwon, Soonil. (2020). Clustering Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2990405.] для распознавания эмоций по голосу демонстрируют подход с использованием выбора ключевых сегментов высказывания. Выбранные сегменты, представленные в виде спектрограмм, были обработаны предобученной на наборе данных ImageNet глубокой сверточной нейронной сетью ResNet-101, а затем двунаправленной рекуррентной нейронной сетью с долгой краткосрочной памятью (Bi-LSTM-RNN). Система была протестирована на наборах данных IEMOCAP, Emo-DB и RAVDESS, достигнув на этих наборах максимальной точности распознавания 72,25%, 85,87% и 77,02% соответственно.

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 (dropout, weight decay, усложнение модели) 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 так называемое многозадачное обучение (англ. multi-task learning) — одновременное обучение модели группе различных, но взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение». Многозадачное обучение в последнее время было включено во множество моделей глубоких нейронных сетей, решающих проблемы в области компьютерного зрения [Bilen, Hakan & Vedaldi, Andrea. (2017). Universal representations:The missing link between faces, text, planktons, and cat breeds.], обработки речи [Das, A., Hasegawa-Johnson, M., Veselý, K. (2017) Deep Auto-Encoder Based Multi-Task Learning Using Probabilistic Transcriptions. Proc. Interspeech 2017, 2073-2077, DOI: 10.21437/Interspeech.2017-582.] и естественного языка [Sanh, Victor, Thomas Wolf and Sebastian Ruder. “A Hierarchical Multi-task Approach for Learning Embeddings from Semantic Tasks.”], а также обучения с подкреплением [Teh, Yee, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. "Distral: Robust multitask reinforcement learning." In *Advances in Neural Information Processing Systems*, pp. 4496-4506. 2017.]. К примеру, задачи обнаружения лица, распознавания пола и оценки позы человека могут быть одновременно решены с использованием одной сверточной глубокой нейронной сети [R. Ranjan, V. M. Patel, and R. Chellappa, “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, vol. 41, no. 1, pp. 121–135, 2019].

В области распознавания эмоций в речи многозадачное обучение показало хорошие 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. [Parthasarathy, Srinivas and C. Busso. “Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning.” INTERSPEECH (2017).] представили систему для одновременной оценки эмоциональных атрибутов Возбуждение, Валентность, Доминантность, использующую многозадачное обучение глубоких полносвязных нейронных сетей в признаковом пространстве ComParE. Лучшая производительность была достигнута структурой, комбинирующей один общий слой с тремя отдельными слоями для каждой задачи. По сравнению с аналогичной, но однозадачной архитектурой, был продемонстрирован максимальный прирост 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 [Z. Zhang, B. Wu and B. Schuller, "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), Brighton, United Kingdom, 2019, pp. 6705-6709, doi: 10.1109/ICASSP.2019.8682896.] также использовали многозадачное обучение для предсказания значений атрибутов Возбуждение, Валентность и 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 случае 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. [Gideon, John, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. "Progressive neural networks for transfer learning in emotion recognition." arXiv preprint arXiv:1706.03256 (2017).] 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 «эффекта забвения», поскольку сохраняют знания, полученные 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 знаний между задачами распознавания диктора и эмоции: точность распознавания 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. [Latif, Siddique & Rana, Rajib & Khalifa, Sara & Jurdak, Raja & Epps, Julien & Schuller, Bjorn. (2020). Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition. IEEE Transactions on Affective Computing. PP. 1-1. 10.1109/TAFFC.2020.2983669.] представили модель многозадачного обучения для голосового распознавания эмоций, идентификации говорящего и его пола. Для извлечения 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) 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 state-of-the-art решения. На текущий момент данная работа является наиболее широко 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 [R. Cipolla, Y. Gal and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7482-7491, doi: 10.1109/CVPR.2018.00781.] 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), instance segmentation и попиксельной регрессии глубины (depth regression), продемонстрировав улучшение результатов каждой из задач по сравнению с однозадачными моделями. Liebel и Körner [Liebel, L. and M. Körner. “Auxiliary Tasks in Multi-task Learning.” ArXiv abs/1805.06334 (2018)] адаптировали элемент регуляризации в этом методе, предотвратив отрицательные значения регуляризации, что позволило добиться еще лучших результатов на тех же задачах. В работе [T. Gong et al., "A Comparison of Loss Weighting Strategies for Multi task Learning in Deep Neural Networks," in IEEE Access, vol. 7, pp. 141627-141632, 2019, doi: 10.1109/ACCESS.2019.2943604.] проведено сравнение этих и еще нескольких стратегий динамического многозадачного обучения, таких как Dynamic Weighted Average (DWA)[S. Liu, E. Johns and A. J. Davison, "End-To-End Multi-Task Learning With Attention," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 1871-1880, doi: 10.1109/CVPR.2019.00197.] и GradNorm[Chen, Zhao & Badrinarayanan, Vijay & Lee, Chen-Yu & Rabinovich, Andrew. (2017). GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks.] 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 человеческой 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.

**Описание предлагаемого подхода.**

**Модели.**

Были проведены предварительные эксперименты, оценивающие множество архитектур сверточных глубоких нейронных сетей на предмет качества изучения и извлечения высокоуровневых признаков из логарифмированных мел-спектрограмм: AlexNet [Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Communications of the ACM 60, no. 6 (2017): 84-90.]; VGG в конфигурациях A, B, D, E [Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).]; ResNet-50 [He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.]; ???. 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.

**Базовая модель 1.**

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 Рисунке 1, подробное описание её слоёв – в Таблице 1.



Рисунок 1. Схема Базовой модели 1.

|  |  |  |
| --- | --- | --- |
| Слой | Параметры | Размерность на выходе |
| Conv1 | Количество фильтров - 64, размер ядра - 11, шаг - 4, пэддинг - 2 | 64х55х55 |
| MaxPool1 | Размер ядра -3, шаг - 2 | 64х27х27 |
| Conv2 | Количество фильтров - 192, размер ядра - 5, шаг - 1, пэддинг - 2 | 192x27x27 |
| MaxPool2 | Размер ядра - 3, шаг - 2 | 192x13x13 |
| Conv3 | Количество фильтров - 384, размер ядра - 3, шаг - 1, пэддинг - 1 | 384x13x13 |
| Conv4 | Количество фильтров - 256, размер ядра - 3, шаг - 1, пэддинг - 1 | 256x13x13 |
| Conv5 | Количество фильтров - 256, размер ядра - 3, шаг - 1, пэддинг - 1 | 256x12x12 |
| MaxPool3 | Размер ядра - 3, шаг - 2 | 256x5x5 |
| FC1 |  | 6400 |
| Dropout | p = 0.5 |  |
| FC2 |  | 2048 |
| FC3 |  | 512 |
| FC4 |  | <количество классов> |

*Таблица 1. Архитектура Базовой модели 1.*

Модель имеет 17,073,348 параметров, которые занимают 65.13 МБ дискового пространства.

**Базовая модель 2.**

В базовой модели 2 извлечение признаков производится с помощью сверточной нейронной сети, архитектура которой аналогична архитектуре VGG (конфигурация D), кроме количества входных каналов изображения. Далее извлеченные признаки подаются на блок классификатора, архитектура которого аналогична таковой у Базовой модели 1. 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 Рисунке 2, описание её слоёв – в Таблице 2.



Рисунок 2. Схема Базовой модели 2.

|  |  |  |
| --- | --- | --- |
| Слой | Параметры | Размерность на выходе |
| Conv1 | Количество фильтров - 64, размер ядра - 3, шаг - 1, пэддинг - 1 | 64x224x224 |
| MaxPool1 | Размер ядра - 2, шаг - 2, пэддинг - 0 | 64x112x112 |
| Conv2 | Количество фильтров - 128, размер ядра - 3, шаг - 1, пэддинг - 1 | 128x112x112 |
| MaxPool2 | Размер ядра - 2, шаг - 2, пэддинг - 0 | 128x56x56 |
| Conv3 | Количество фильтров - 256, размер ядра - 3, шаг - 1, пэддинг - 1 | 256x56x56 |
| Conv4 | Количество фильтров - 256, размер ядра - 3, шаг - 1, пэддинг - 1 | 256x56x56 |
| MaxPool3 | Размер ядра - 2, шаг - 2, пэддинг - 0 | 256x28x28 |
| Conv5 | Количество фильтров - 512, размер ядра - 3, шаг - 1, пэддинг - 1 | 512x28x28 |
| Conv6 | Количество фильтров - 512, размер ядра - 3, шаг - 1, пэддинг - 1 | 512x28x28 |
| MaxPool4 | Размер ядра - 2, шаг - 2, пэддинг - 0 | 512x14x14 |
| Conv7 | Количество фильтров - 512, размер ядра - 3, шаг - 1, пэддинг - 1 | 512x14x14 |
| Conv8 | Количество фильтров - 512, размер ядра - 3, шаг - 1, пэддинг - 1 | 512x14x14 |
| MaxPool5 | Размер ядра - 2, шаг - 2, пэддинг - 0 | 512x7x7 |
| FC1 |  | 25088 |
| Dropout | p=0.5 |  |
| FC2 |  | 2048 |
| FC3 |  | 512 |
| FC4 |  | <количество классов> |

Таблица 2. Архитектура Базовой модели 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: 1) количество параметров в многозадачной модели будет меньше, чем при построении нескольких моделей, каждая из которых оптимизирована для своих индивидуальных задач; и 2) 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 высокоуровневую структуру, 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 Рисунке 3.

Рисунок 3. Схема предлагаемой модели.



Данная модель имеет 63,762,576 параметров, которые занимают 243.23МБ дискового пространства.

**Эксперименты.**

**Наборы данных.**

Для экспериментального исследования вышеописанных моделей были выбраны наборы данных IEMOCAP и RAMAS.

IEMOCAP (Interactive emotional dyadic motion capture database) [Busso, Carlos & Bulut, Murtaza & Lee, Chi-Chun & Kazemzadeh, Abe & Mower Provost, Emily & Kim, Samuel & Chang, Jeannette & Lee, Sungbok & Narayanan, Shrikanth. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. 10.1007/s10579-008-9076-6.] – это многомодальный набор данных, состоящий из аудио-видеозаписей 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 12ч26мин данных, которые были размечены несколькими аннотаторами как на дискретные эмоциональные классы, так и на непрерывные значения валентности и активации.

RAMAS (The Russian Acted Multimodal Affective Set) ― многомодальный набор данных об испытываемых эмоциях, включающий параллельную запись 12 каналов: аудио, видео, носимые датчики движения и др. В создании датасета приняли участие 10 актеров из ВГИКа (5 мужчин и 5 женщин), воссоздавшие различные ситуации повседневного общения. В наборе данных присутствует две параллельные категориальные разметки: «базовые аффективные состояния» (Anger – гнев, Happiness - радость, Neutral – нейтральная эмоция, Sadness - грусть, Disgust - отвращение, Fear - страх, Surprise - удивление, Shame - стыд, Tiredness – усталость), а также разметка на Domination (доминация) и Submission (подчинение). Разметка произведена по временным меткам. Однако, при сегментации файлов с помощью данных меток и дальнейшем первичном исследовании полученных образцов данных, сделан вывод о низкой репрезентативности этих образцов (пересекающиеся метки, зачастую на записях звучат реплики сразу двух актеров) и принято решение о переразметке этого набора данных не по временным меткам, а по фразам репликам (как это сделано, к примеру, в IEMOCAP). Таким образом, были получены образцы данных, однозначно представленные строго одним актером на записи, а также строго одним эмоциональным состоянием. При этом было принято решение отказаться от класса Tiredness, так как он был представлен крайне мало. Итоговое распределение данных по классам изображено на рисунке 4.

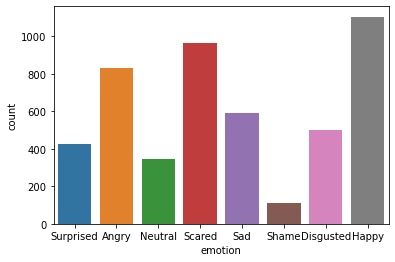


Рисунок 4. Представленность эмоциональных классов

в наборе данных RAMAS после переразметки

**Описание экспериментов.**

Все модели были обучены на трёх выборках:

1. Набор данных IEMOCAP с четырьмя эмоциональными категориями Anger, Happiness, Neutral, Sadness (далее – IEMOCAP-4). Мы мотивируем свой выбор наибольшей популярностью именно данного варианта среди исследователей, что позволит сравнить наши результаты.
2. Набор данных RAMAS, все эмоциональные классы (далее – RAMAS-descrete).
3. Набор данных RAMAS, бинарная классификация «Гнев – Не гнев». Все образцы класса «Anger» и столько же случайно выбранных остальных образцов (далее – RAMAS-binary).

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 Рисунке 5. Все процедуры, подразумевающие использование случайных чисел («Случайно обрезать», «Генерация D», «Случайно обрамить 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 512х512. Данный массив был конвертирован в изображение (значения приведены в диапазон от 0 до 255), которое было уменьшено до размеров 224х224 с применением кубической интерполяции. Наконец, перед непосредственно подачей на вход нейронной сети, проводилась нормализация со средним значением 0.5 и стандартным отклонением 0.225.

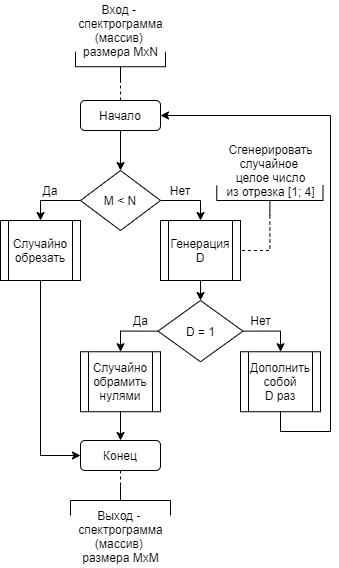


Рисунок 5. Блок-схема алгоритма унификации размера и аугментации спектрограмм.

Модели обучались на тренировочной подвыборке 300 эпох, learning rate – 1e-5, размер мини-батча 32, использован алгоритм оптимизации Adam [Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).], функция потерь – перекрестная энтропия. 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 значений функции потерь отдельных задач на каждом мини-батче (здесь и далее  – 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: ;

б) Усовершенствованный метод на основе гомоскедастичной неопределенности, описанный в [R. Cipolla, Y. Gal and A. Kendall, "Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 7482-7491, doi: 10.1109/CVPR.2018.00781.][Liebel, L. and M. Körner. “Auxiliary Tasks in Multi-task Learning.” ArXiv abs/1805.06334 (2018)];

в) Метод взвешенного среднего: , где .

Лучшие результаты были достигнуты с использованием метода взвешенного среднего.

**Результаты экспериментов.**

Ниже представлены результаты моделей с наименьшим за все время обучения значением функции ошибки на валидационной выборке. В качестве метрик качества выбраны *Accuracy* acc, *Precision* pr, *Recall* rec, а также *F-мера* f1.

Результаты экспериментов представлены в таблицах 3-7. Обозначения эмоциональных классов: ang – Гнев, hap - Радость, neu – Нейтральная эмоция, sad – Грусть, dis – Отвращение, fea – Страх, shm – Стыд, sur – Удивление.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | acc | pr | rec | f1 |
| IEMOCAP-4 | 0,688 | 0,670 | 0,624 | 0,613 |
| RAMAS (descrete) | 0,416 | 0,295 | 0,353 | 0,315 |
| RAMAS (binary) | 0,712 | 0,697 | 0,698 | 0,698 |

Таблица 3. Результаты Базовой модели 1 для классификации эмоций.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | acc | pr | rec | f1 |
| IEMOCAP-4 | 0,694 | 0,674 | 0,631 | 0,631 |
| RAMAS (descrete) | 0,477 | 0,431 | 0,427 | 0,388 |
| RAMAS (binary) | 0,729 | 0,715 | 0,708 | 0,710 |

Таблица 4. Результаты Базовой модели 2 для классификации эмоций.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | acc | pr | rec | f1 |
| Emotion | IEMOCAP-4 | 0,712 | 0,685 | 0,666 | 0,673 |
| RAMAS (descrete) | 0,467 | 0,548 | 0,424 | 0,435 |
| RAMAS (binary) | 0,771 | 0,764 | 0,743 | 0,749 |
| Speaker | IEMOCAP-4 | 0,782 | 0,774 | 0,771 | 0,767 |
| RAMAS (descrete) | 0,818 | 0,811 | 0,804 | 0,802 |
| RAMAS (binary) | 0,839 | 0,831 | 0,838 | 0,833 |
| Gender | IEMOCAP-4 | 0,969 | 0,969 | 0,969 | 0,969 |
| RAMAS (descrete) | 0,936 | 0,937 | 0,936 | 0,936 |
| RAMAS (binary) | 0,928 | 0,931 | 0,927 | 0,928 |

Таблица 5. Результаты многозадачной модели для классификации эмоции (Emotion),

спикера (Speaker) и пола (Gender).

|  |  |  |  |
| --- | --- | --- | --- |
|  | pr | rec | f1 |
| ang | 0,820 | 0,760 | 0,790 |
| hap | 0,500 | 0,390 | 0,440 |
| neu | 0,700 | 0,780 | 0,740 |
| sad | 0,72 | 0,73 | 0,73 |

Таблица 6. Результаты распознавания каждого эмоционального класса

предлагаемой моделью на наборе данных IEMOCAP.

|  |  |  |  |
| --- | --- | --- | --- |
|  | pr | rec | f1 |
| ang | 0,490 | 0,540 | 0,520 |
| hap | 0,420 | 0,840 | 0,560 |
| neu | 0,720 | 0,710 | 0,720 |
| sad | 0,69 | 0,34 | 0,45 |
| dis | 0,310 | 0,340 | 0,330 |
| fea | 0,510 | 0,200 | 0,280 |
| shm | 0,380 | 0,150 | 0,220 |
| sur | 0,860 | 0,270 | 0,410 |

Таблица 7. Результаты распознавания каждого эмоционального класса

предлагаемой моделью на наборе данных RAMAS-descrete.

Сравнение результатов предлагаемой модели с базовыми на задаче классификации эмоций представлено в Таблице 8.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | acc | pr | rec | f1 |
| IEMOCAP | 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** |
| RAMAS-descrete | Базовая модель 1 | 0,416 | 0,295 | 0,353 | 0,315 |
| Базовая модель 2 | 0,457 | 0,431 | **0,426** | 0,388 |
| Предлагаемая модель | **0,467** | **0,548** | 0,424 | **0,435** |
| RAMAS-binary | Базовая модель 1 | 0,712 | 0,697 | 0,698 | 0,698 |
| Базовая модель 2 | 0,729 | 0,715 | 0,708 | 0,710 |
| Предлагаемая модель | **0,770** | **0,760** | **0,838** | **0,749** |

Таблица 8. Сравнение результатов моделей.

**Интерпретация и обсуждение результатов**

На наборе данных IEMOCAP предлагаемая модель достигает лучших результатов в распознавании метки Гнев, худших – в распознавании метки Радость (см. Таблицу 6). 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, f1 – на 0.042 (см. Таблицу 8). Одновременно данная модель успешно решает две смежные паралингвистические задачи: распознавание диктора с точностью 0.782 и распознавание пола с точностью 0.969.

На выборке набора данных RAMAS, содержащей данные для классификации всех представленных в этом наборе эмоциональных категорий (RAMAS-descrete), лучшие результаты получены для классификации Нейтральной эмоции и Радости, худшие – для классификации Стыда (см. Таблицу 6), что объясняется наименьшей представленностью этого класса (см. Рисунок 4). Получен минимальный прирост accuracy на 0.01, precision – на 0.117, f1 – на 0.047 по сравнению с базовыми моделями, а также значение recall, практически равное последнему у Базовой модели 2 – 0.424 против 0.426. На выборке этого набора данных, подразумевающей бинарную классификацию «Гнев – не гнев» (RAMAS-binary) получен минимальный прирост accuracy на 0.041, precision – на 0.045, recall – на 0.13, f1 – на 0.039. Также на наборе данных RAMAS предложенной моделью получена (средняя по двум выборкам) точность распознавания диктора 0.829, пола – 0.932.

В целом, с помощью предлагаемого подхода получено улучшение качества распознавания эмоций по сравнению с базовыми на всех наборах данных, из чего следует эффективность и целесообразность применения концепции многозадачного обучения для увеличения точности моделей распознавания эмоций и одновременного решения смежных паралингвистических задач.