**Архитектуры моделей.**

Были проведены предварительные эксперименты, оценивающие множество архитектур сверточных глубоких нейронных сетей на предмет качества изучения и извлечения высокоуровневых признаков из логарифмированных мел-спектрограмм: AlexNet [ссылка]; VGG в конфигурациях A, B, D, E [ссылка]; ResNet-50 [ссылка]; ???. 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 классификации, были выбраны в качестве базовых.

**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 [ссылка]. В базовой модели 1 извлечение признаков производится с помощью сверточной нейронной сети, архитектура которой аналогична архитектуре AlexNet [ссылка], 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-16 (конфигурация D) [ссылка], 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. Многозадачное обучение (англ. multi-task learning) — одновременное обучение модели группе различных, но взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение». Преимущества многозадачного обучения: 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), от 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.

Волновая форма – STFT с параметрами

В ходе экспериментов были протестированы различные стратегии взвешивания значений функции потерь отдельных задач (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) Метод на основе гомоскедастичной неопределенности, описанный в 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 .



Рисунок 3. Схема предлагаемой модели.

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

**Наборы данных.**

**IEMOCAP** [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.]

IEMOCAP (Interactive emotional dyadic motion capture database) – 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. Существует несколько вариантов дискретной разметки этого набора данных, нами была выбрана разметка на четыре эмоциональные категории (Anger – гнев, Happiness – радость, Neutral – нейтральная эмоция, Sadness – грусть). Выбор разметки мотивирован наибольшей популярностью именно данного варианта среди исследователей, что позволит сравнить наши результаты.

**RAMAS** [Perepelkina, Olga & Kazimirova, Evdokia & Konstantinova, Maria. (2018). RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Studying Emotion Recognition. 10.7287/peerj.preprints.26688v1.]

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), воссоздающие 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 данной разметки и дальнейшем первичном исследовании полученных образцов данных, сделан вывод о низкой репрезентативности этих данных (пересекающиеся метки, зачастую на записях слышно сразу двух актеров) и принято решение о переразметке этого набора данных не по временным меткам, а по фразам актеров (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, в ходе разметки, было принято решение отказаться от класса Tiredness, так как он был представлен крайне мало. Итоговое распределение данных по классам изображено на рисунке 4.

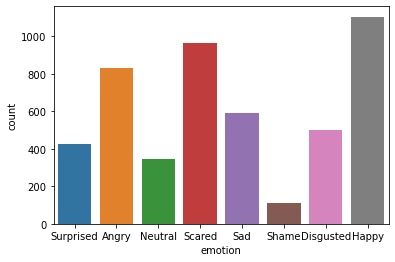


Рисунок 4. Представленность эмоциональных классов в наборе данных RAMAS после переразметки

**Описание и результаты экспериментов.**

Все модели были обучены на трёх выборках данных:

1. Набор данных IEMOCAP с четырьмя эмоциональными категориями Anger, Happiness, Neutral, Sadness (далее – IEMOCAP-4)
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).

Наборы данных были разделены на тренировочную и валидационную выборки в пропорции 4:1.

Модели обучались на тренировочной выборке 300 эпох, learning rate – 1e-5, размер мини-батча 32, использован алгоритм оптимизации Adam [ссылка], функция потерь – Cross Entropy Loss. Для предотвращения переобучения, производилась остановка обучения модели, если значение функции ошибки на валидационной выборке не уменьшалось в течение 30 эпох.

В ходе экспериментов были протестированы различные стратегии взвешивания значений функции потерь отдельных задач (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) Метод на основе гомоскедастичной неопределенности, описанный в 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 обучения были достигнуты с использованием взвешенного среднего.

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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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 |

Результаты экспериментов представлены в таблицах 3-7.

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.

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.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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.

Таким образом, предлагаемой моделью получено улучшение точности классификации гнева в сравнении с базовыми в среднем на 5%, при этом одновременно этой же моделью получена возможность классифицировать пол говорящего с точностью 92% и идентифицировать личность говорящего с точностью 83%.