**Архитектуры моделей.**

**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, 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, 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.**

Многозадачное обучение (англ. multi-task learning) — одновременное обучение группе взаимосвязанных задач, для каждой из которых задаются свои пары «ситуация, требуемое решение». Обучая одну модель нескольким задачам одновременно, мы вносим регуляризацию, которая, гипотетически, должна улучшить качество классификации. 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. 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.

**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.

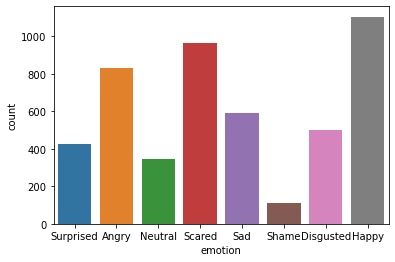


Рисунок 4. Представленность эмоциональных классов в наборе данных RAMAS после переразметки

**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, использован алгоритм оптимизации Adam. Для предотвращения переобучения, производилась остановка обучения модели, если значение функции ошибки на валидационной выборке не уменьшалось в течение 30 эпох. 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 моделей представлены в таблицах 3-5.

Таблица 3. Результаты Базовой модели 1.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | 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).

Сравнение результатов предлагаемой модели с базовыми на задаче бинарной классификации «Гнев – не гнев» набора данных RAMAS представлено в Таблице 6.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | acc | pr | rec | f1 |
| Базовая модель 1 | 0,712 | 0,697 | 0,698 | 0,698 |
| Базовая модель 2 | 0,729 | 0,715 | 0,708 | 0,710 |
| Предлагаемая модель | **0,771** | **0,764** | **0,743** | **0,749** |

Таблица 6. Сравнение результатов базовых и предлагаемой модели

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%.