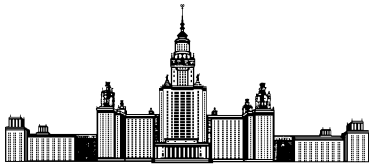
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

Курсовая работа студента 317 группы

«Исследование метода аугментации аудиоданных для задачи распознавания эмоций»

Выполнил:

студент 3 курса 317 группы

Бессонов Аркадий Сергеевич

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Москва, 2022

Содержание

1 Введение **3**

1.1 Обзор литературы 4

2 Постановка задачи **4**

3 Вычислительные эксперименты **8**

3.1 Исходные данные и условия эксперимента 8

3.2 Результаты экспериментов 11

3.3 Дополнительные исследования 14

4 Заключение **16**

Аннотация

Многие задачи, связанные с обработкой и распознаванием речи, требуют понимания речи, которое учитывает паралингвистические особенности. Примером такой задачи является распознавание эмоций. В отличие от задачи распознавания речи, в этой области сравнительно мало размеченных данных, что создает определенные трудности для обучения моделей. В данной работе демонстрируется способ расширения обучающей выборки, характерный для задачи распознавания эмоций. Эффективность предложенного подхода демонстрируется на экспериментах.

1 Введение

Задача распознавания эмоций (*SER – Speech Emotion Recognition*), как одна из задач человеко-машинного взаимодействия (*human computer interaction*), имеет практическую ценность. Например, используется в анализе поведения клиентов в колл-центрах или в более общем случае для оценки психологического состояния человека.

Помимо этого, распознавание эмоций находит широкое применение в играх, личных помощниках. Речь считается простым и эффективным коммуникативным механизмом для передачи человеческих эмоций. С распространением архитектур сверточных нейронных сетей широкое применение получил подход, при котором речь переводилась в спектрограмму (т.н «hand-crafted features»), а затем к полученной спектрограмме применялись подходы из задач компьютерного зрения. В последнее время широкое распространение получили модели, основанные на работе с непосредственным речевым сигналом [1, 2]. Эта особенность моделей позволяет эффективно производить преобразования входного сигнала, например с целью аугментации данных.

Основными сложностями в определении эмоций являются потенциальная неоднозначность и субъективность классификации эмоций как таковых, а также недостаток качественных обучающих данных в нужном количестве. Более того, довольно низкое разнообразие актеров в обучающей выборке способствует быстрому переобучению модели. В то же время, специфика задачи автоматического распознавания эмоций открывает некоторые возможности для искусственного увеличения обучающей выборки. В данной работе предложен довольно универсальный и простой способ снижения эффекта переобучения модели с помощью аугментации обучающей выборки при распознавании эмоций. Результаты работы демонстрируются на популярных датасетах для SER - используется английский и немецкий языки.

1.1 Обзор литературы

Как уже было отмечено, задача распознавания эмоций по аудио ранее решалась преимущественно с помощью спектрограмм аналогично задачам компьютерного зрения (см. например [15]). В этом случае разные способы аугментации из этой области были исследованы на спектрограммах и для некоторых из них нашлось применение для задач ASR (*Automatic Speech Recognition*) [5, 9]. Однако с появлением архитектур, основанных на механизме самовнимания (*self-attention*) [13] и их успехе в задачах обработки естественного языка, все чаще стали появляться решения, использующие модели, обученные на неразмеченных данных (*self-supervised learning*) [7]. Такие решения нацелены на построение представлений низкоуровневых признаков, описывающих речь из аудиосигнала [1, 2]. Успех подобных архитектур привел к тому, что их начали применять также в задаче распознавания эмоций [16, 17].

Последнее время появились статьи, исследующие возможность получения представлений речи, которые бы учитывали её паралингвистические свойства, т.е несмысловые характеристики речи [6, 18]. С увеличением размеров моделей, с одной стороны, улучшается качество работы, однако, с другой стороны, увеличивается риск переобучения. Особенно в задачах SER, где размеченные обучающие данные имеют относительно небольшие размеры, а для редких языков такие данные получить очень сложно. Описанный далее подход является простым и интуитивно понятным, при этом демонстрирует улучшение качества работы.

2 Постановка задачи

Целью данной работы является исследование метода аугментации данных, который представляется естественным для задач SER. Методы для ASR в общем случае имеют больше ограничений на изменение данных - например, в общем случае важен

порядок слов и звуков говорящего, а также целостность каждого произнесенного слова. Напротив, для распознавания эмоциональной окраски речи это не нужно - более того, это потенциально приводит к переобучению и снижению обобщающей способности алгоритмов в условиях сравнительно небольшого количества данных, свойственных этой области - модель может запоминать специфичные для аудиозаписи слова.

В задаче распознавания эмоций традиционно используются модели, показавшие высокое качество для задач распознавания речи. Одной из популярных моделей, которая успешно применяется в подобных задачах, является Wav2Vec2 [2]. Отличительной особенностью этой модели является то, что она работает с аудиосигналом напрямую, без использования спектрограмм. Эти соображения приводят к идее расширения обучающей выборки с помощью перестановки слов или отдельных частей в исходной аудиозаписи, минуя преобразование в спектрограмму. Действительно, интуитивно кажется, что эмоциональный контекст будет сохраняться и дополняться смешивающей аудиозаписью. При таком «смешивании» двух аудиозаписей меняется соотношение эмоциональных оттенков, что может быть учтено соответствующим изменением метки класса для смешанной записи. Эксперименты, проделанные в ходе исследования, показали, что описанная идея действительно улучает качество работы модели. При этом потенциально существует множество вариаций реализации изложенных соображений. В роли критерия качества результатов модели используется точность классификации. Далее в этом и последующих разделах основное внимание будет уделено одной из потенциальных реализаций, которая показала большую гибкость и эффективность. Ниже будет дана более формальная постановка предложенного подхода.

Предложенный подход

Пусть имеется множество $\mathbf{X}$ - выборка, состоящая из векторов разной длины (вообще говоря, произвольной), представляющих набор дискретных отсчетов голосовой записи. Сопоставим каждому объекту $x \in \mathbf{X}$ вектор $y \in \{0,1\}^n$ размерности n, где n - число рассматриваемых классов (эмоциональных категорий). Иначе говоря, y - это one-hot кодирование метки, отвечающей некоторой из n эмоций. Тогда для пар (x_1, y_1) и (x_2, y_2), где $x_1 \in \mathbb{R}^{m_1}, x_2 \in \mathbb{R}^{m_2}$, семплированных из распределения выборки, сопоставим новый объект $(\hat{x}, \hat{y})$, где $\hat{x} = (x_1^{(1)}, x_1^{(2)}, ..., x_1^{(\lfloor m_1 \lambda \rfloor)}, x_2^{(m_2 - \lfloor m_2(1-\lambda) \rfloor)}, ..., x_2^{(m_2)}$, $\hat{y} = \lambda y_1 + (1-\lambda) y_2$, $\lambda \in (0,1), \lambda \sim B(\alpha, \alpha), \alpha \in (0, +\infty)$. Пример применения этой операции показан на рис. 1.

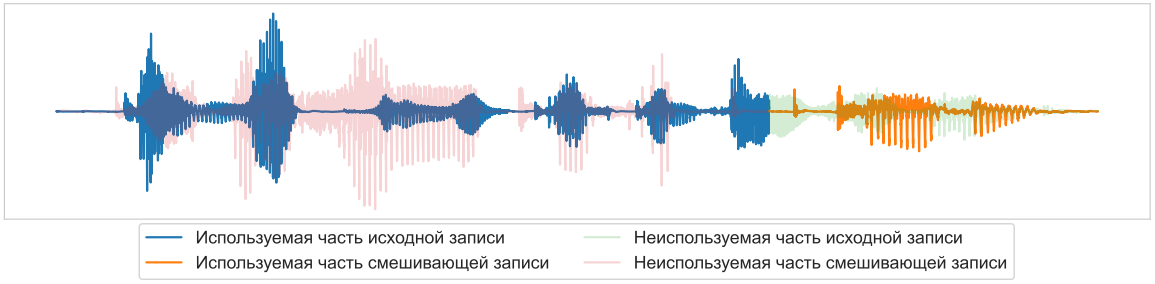


Рис. 1: Пример аугментации сигналов

Т.е, подобно [4], здесь используются предположение о том, что эмоциональная окраска всей записи формируется из экспрессивности каждой эмоции, входящей в запись (предполагается, что экспрессивность определяется размерами частей каждой эмоции, входящей в составную аудиозапись). Как и в [4], в экспериментах используется $B-$распределение, в котором параметр α контролирует силу «смешивания». При $\alpha \to +0$ эффект сводится к семплированию из распределения Бернулли случайной записи данного актера из тренировочной выборки, при $\alpha = 1$ к равномерному распределению пропорций, при $\alpha \to \infty$ к конкатенации половин каждой из записей (см. рис. 2), поскольку распределение вырождается в $\frac{1}{2}$.

Помимо рассмотренного параметра α для настройки доступен параметр p распределения Бернулли ($Bernoulli(p)$), определяющий вероятность для данного объекта проведения аугментации. Также возможно определять вид распределения, используемого при выборе смешивающей записи для данного спикера. В этой работе используется равномерное распределение по аудиозаписям для данного актера. Однако, возможно использовать фильтрацию, например, по схожим эмоциям, или разным эмоциям.

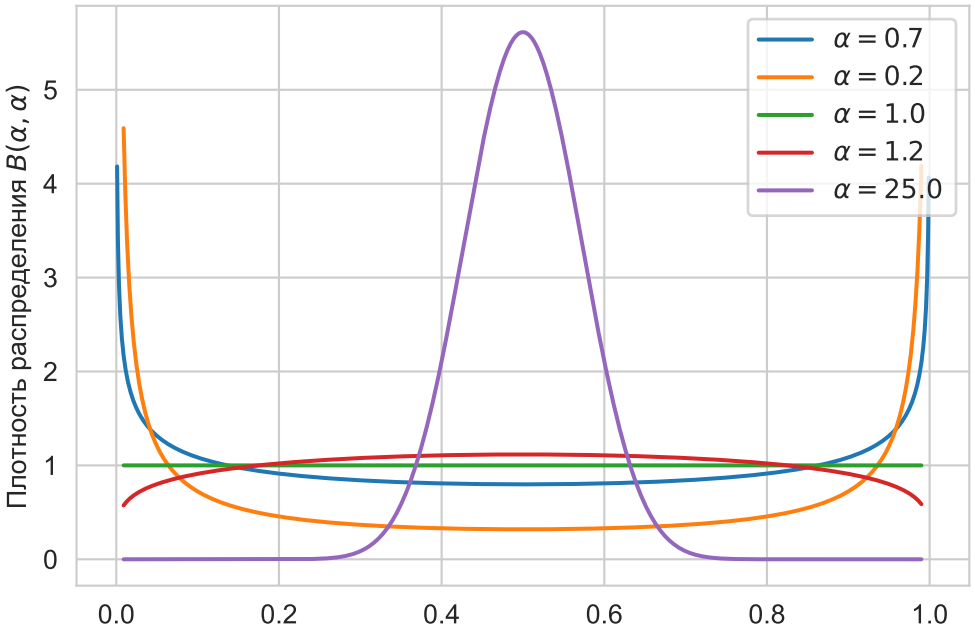


Рис. 2: Вид $B(\alpha, \alpha)$ – распределения в зависимости от α

3 Вычислительные эксперименты

При проведении экспериментов использовалась модель Wav2Vec2-XLSR-53 (Large) [3], предобученная на 53 языках. Выбор в пользу именно такой версии модели сделан потому, что эксперименты проводились на наборах данных из разных языков. Обучение модели производилось на графическом процессоре Nvidia Tesla V100 (16GB). В силу ограниченности вычислительных ресурсов проведенные эксперименты не покрывают все потенциально успешные варианты применения предложенного метода. При сравнении все гиперпараметры оставались неизменными на протяжении всех экспериментов (см. разд. 3.1).

При получении результатов вычисления проводились на разных контрольных выборках независимо, а затем усреднялись для получения более надежных экспериментальных данных (каждая выборка состояла примерно из 30% от общего объема данных). Разбиение проводилось по актерам так, что в валидационной и тренировочной выборках не было одинаковых спикеров.

3.1 Исходные данные и условия эксперимента

В данной работе используется равномерное распределение для выбора смешивающей записи и фиксированное значение $p = 0.5$ – т.е в среднем для каждой второй записи из тренировочной выборки производится смешивание с другой записью того же актера из тренировочной выборки независимо от целевой метки. Само «смешивание» проводилось по схеме, описанной выше (разд. 2).

Как уже отмечалось, в экспериментах используется модель Wav2Vec2-XLSR-53 [3]. При дообучении замораживаются нижние сверточные слои, отвечающие за извлечение признаков из аудио. Затем скрытые состояния на выходе из модели усредняются и полученные векторные представления подаются на вход классификатору,

состоящему из двухслойной полносвязной нейронной сети с гиперболическим тангенсом в качестве функции активации и скрытым слоем размерности 1024.

Темп обучения в начале был выбран равным 10^{-4} и монотонно убывал по квадратичному закону до нуля. Размер батча был выбран равным 3 (при больших размерах происходило переполнение памяти видеокарты) и оставался постоянным для каждого набора данных. Для сравнения полученных результатов в экспериментах были использованы три набора данных для двух языков. В качестве функции потерь использовалась бинарная кросс-энтропия.

RAVDESS

Набор данных Ryerson Audio-Visual Data of Emotional Speech and Song (RAVDESS) [10] представляет собой стандартный набор данных, при записи которых использовался сценарий. Этот набор данных содержит 60 устных предложений и 40 спетых предложений от 24 актеров (12 мужчин, 12 женщин). Представлено восемь категорий эмоций.

В данной работе были использованы только речевые предложения от всех 24 актеров для анализа эмоций, а спетые предложения не использовались. Рассматриваются все восемь категорий эмоций для экспериментов («calm», «happy», «sad», «angry», «fearful», «surprise», «disgust», «neutral»). При проведении экспериментов датасет был разделен на тренировочную и валидационную части так, как было указано ранее (см. разд. 3). Модель обучалась 10 эпох примерно два с половиной часа. Итоговые цифры получены усреднением по пяти фолдам, а при подведении итогов в табл. 1 параметр α был взят равным 0.3 без предварительного подбора.

EMODB

Датасет Berlin Emotional Database (EmoDB) [11] создан Институтом коммуникационных наук Технического университета Берлина. В записи данных участвовали десять профессиональных дикторов (пять мужчин и пять женщин). Набор данных содержит в общей сложности 535 высказываний и включает семь эмоций («anger», «boredom», «anxiety», «happiness», «sadness», «disgust», «neutral»). Разделение на тренировочную и валидационную выборку проводилось по схеме, описанной ранее (см. разд. 3).

При использовании этого датасета модель обучалась 25 эпох в течение примерно полутора часов. Итоговые цифры получены усреднением по двум фолдам. Значение параметра α для данного набора данных при подведении результатов в табл. 1 было взято равным 0.3, так же без подбора.

CREMA-D

CREMA-D [12] — это набор данных из 7442 оригинальных записей с 91 актером. Среди актеров 48 мужчин и 43 женщин в возрасте от 20 до 74 лет, принадлежащими к разным расам и этническим группам (афроамериканцам, азиатам, кавказцам, латиноамериканцам и неуказанным).

Актеры говорили на выбор одно из 12 предложений. Предложения были представлены с использованием одной из шести различных эмоций («anger», «disgust», «fear», «happy», «neutral», «sad») и четырех различных уровней выразительности эмоций (низкий, средний, высокий и неопределенный).

При проведении экспериментов уровни выразительности были объединены. Схема разделения выборки на тренировочную и валидационную та же (см. разд. 3) - для валидации использовано 14 спикеров. Итоговые цифры получены усреднением по двум фолдам. Обучение модели проводилось 7 эпох. При этом обучение модели

занимает около 12 часов. Значение параметра α для данного датасета в при подсчете результатов табл. 1 было взято равным 0.4.

3.2 Результаты экспериментов

По результатам обучения модели в соответствии с описанным способом получены значения точности, достигаемой моделью в конце обучения на каждом наборе данных в зависимости от применения аугментации. Как видно из таблицы, (см. табл. 1) аугментация действительно оказывает положительное влияние на качество работы модели. Для набора данных CREMA-D этот эффект оказался почти не заметен (отличие было в тысячных знаках). Но с другой стороны, как будет указано далее, результат аугментации для этого набора данных тем не менее виден на графиках функции потерь (см. рис. 5).

	RAVDESS	EMODB	CREMA-D
Точность с аугментацией	**0.81**	**0.96**	**0.77**
Точность без аугментации	0.79	0.92	0.77

Таблица 1: Сравнение достигаемой точности (*accuracy*) с использованием аугментации и без

Помимо увеличения точности классификации наблюдается также более плавное и монотонное убывание функции потерь в процессе обучения. Результат виден на всех наборах данных (см. рис. 3, рис. 4, рис. 5). На графиках полупрозрачной пунктирной линией показано значение функции потерь между эпохами, сплошной линией соединены значения по числу эпох.

Исследования показывают, что наибольший эффект от аугментации наблюдается на небольших наборах данных (RAVDESS, EMODB). Действительно, видно, что в обоих случаях (рис. 3, рис. 4) выход на плато происходит на меньших значениях функции потерь. Интересно отметить, что в случае использования RAVDESS без

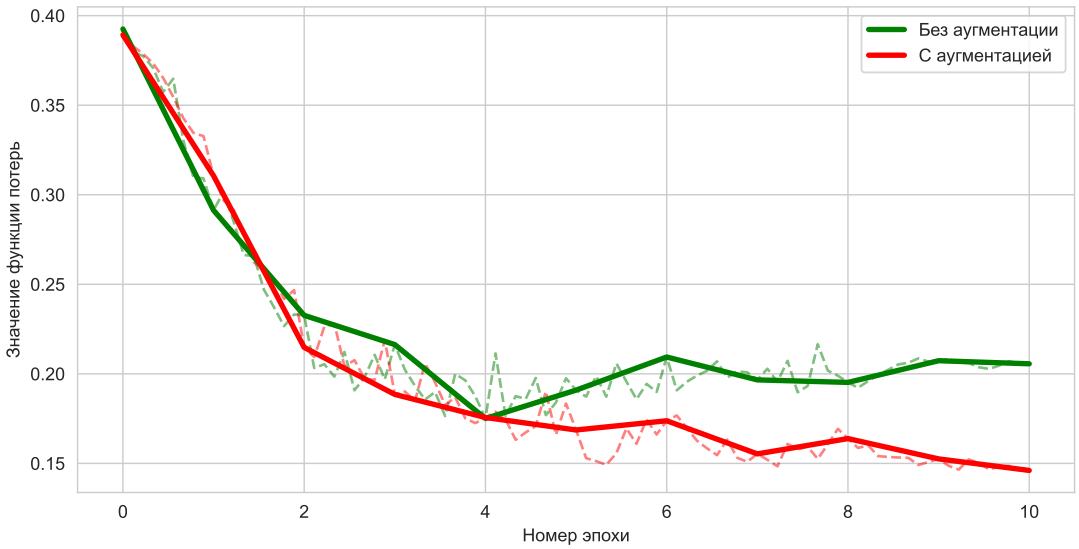


Рис. 3: График зависимости функции потерь от эпохи при использовании датасета RAVDESS

использования аугментации начинает происходить переобучение на четвертой эпохе, в то время как с аугментацией функция потерь продолжает монотонно убывать.

В то же время на больших наборах данных несмотря на то, что аугментация предотвращает переобучение, положительный эффект в её использовании ниже, поскольку данных достаточно для обучения (у датасета CREMA-D значительно больший размер, чем у всех остальных исследованных примерно в 10-12 раз). Из рис. 5 видно, что модель начинает переобучаться после второй эпохи. При этом с аугментацией функция потерь выходит на асимптоту, в то время как без неё происходит дальнейшее переобучение модели. При этом достигаемая точность почти не изменилась и монотонно выходила на асимптоту в процессе обучения. Важно также отметить, что применение аугментации почти не повлияло на время обучения модели во всех случаях.

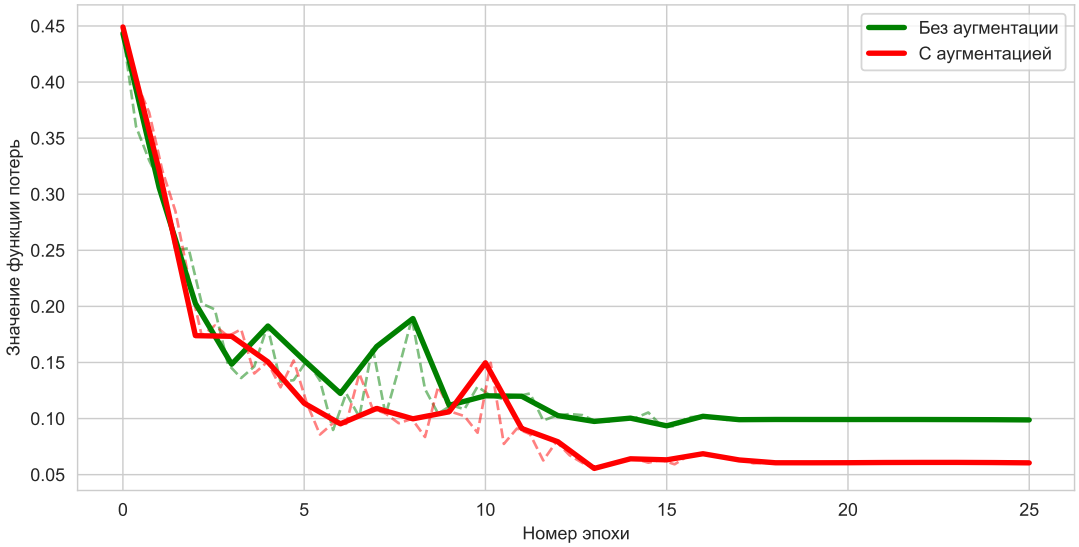


Рис. 4: График зависимости функции потерь от эпохи при использовании датасета EMODB

Также были проведены эксперименты по исследованию влияния параметра α на достигаемую точность при фиксированном количестве эпох (равном 6) на датасете EMODB (поскольку он самый маленький из рассмотренных). Аналогично тому, как было написано ранее, было зафиксировано две группы актеров (примерно по 30% от общего количества каждая). Далее поочередно каждая группа использовалась в качестве валидационного набора данных. Результаты получены усреднением результатов по этим двум группам и представлены в табл. 2. В эксперименте было перебрано порядка 10 значений параметра от 0 до 1 с шагом 0.1, а также достаточно большие значения α для того, чтобы рассмотреть как можно больше характерных значений параметра. Для облегчения восприятия данных были оставлены только наиболее представительные значения параметра α.

Как видно из табл. 2, значения из интервала $(0, 1)$ оказываются хорошим выбором. При достаточно больших значениях параметра наибольшее достигаемая точ-

α	0.1	0.3	0.6	0.8	5.0	10.0	15.0	25.0
Точность	0.95	**0.96**	**0.96**	0.95	**0.96**	0.93	0.94	0.95

Таблица 2: Сравнение достигаемой точности в зависимости от параметра α

ность немного снижается, хотя и не значительно. Таким образом, наблюдается некоторая устойчивость результатов по параметру α, поскольку для всех рассмотренных значений достигается большая точность, чем без применения аугментации.

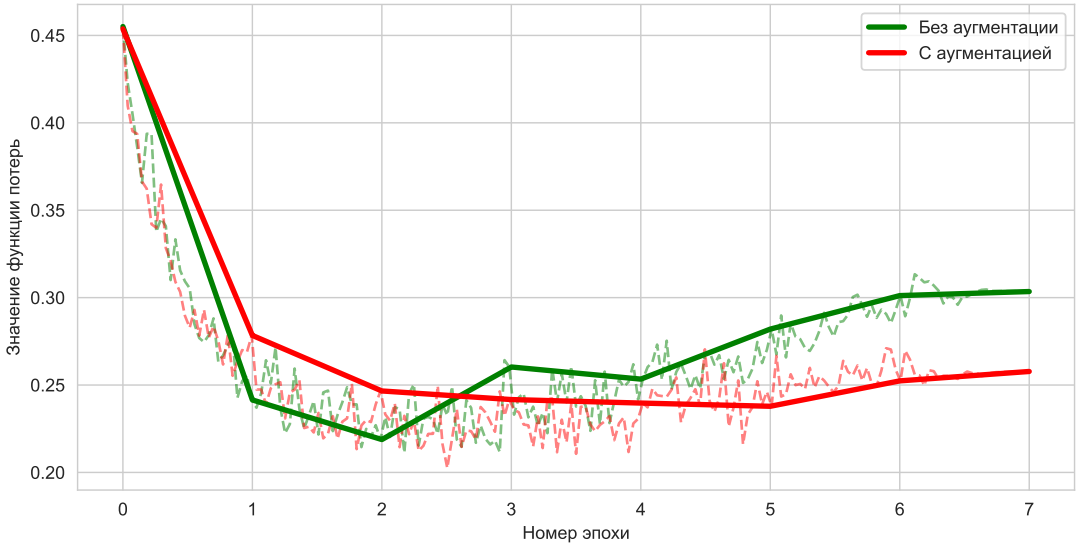


Рис. 5: График зависимости функции потерь от эпохи при использовании датасета CREMA-D

3.3 Дополнительные исследования

Как следует из экспериментов, описанный способ аугментации действительно улучшает качество работы модели. Во всех проведенных экспериментах модели, использующие данный подход, давали результаты лучше. Было также замечено, что

выбор оптимального параметра α должен проводиться для каждого набора данных отдельно. Типичные значения лежат в диапазоне $[0.1, 1.0]$. В целом, значение $\alpha = 0.3$ давало хорошие результаты на всех наборах данных.

Поскольку обучение используемой в экспериментах модели требует значительных вычислительных ресурсов, не было возможности рассмотреть все возможные способы смешивания записей и зависимости качества работы моделей от параметра α. Помимо указанной стратегии также были исследованы две другие основные стратегии «смешивания» - 1) Замена слов в записи словами из других записей того же актера. 2) Вставка фрагмента случайной длины (выбиралась из равномерного распределения) в данную аудиозапись.

Для реализации первого подхода использовалась разметка записи, полученная с помощью Wav2Vec2. С помощью разметки можно определить границы слов, а далее заменить выделенные слова словами из других записей. Для экспериментов использовался набор данных RAVDESS, поскольку в нем небольшое число уникальных текстов предложений, поэтому интуитивно казалось, что данный подход должен хорошо сработать. Однако на практике данный способ не показал ощутимого улучшения результатов.

Во втором способе из равномерного распределения выбиралось место в исходной записи (индекс). Затем, также из равномерного распределения, производился выбор другой аудиозаписи того же актера и выбор длины фрагмента этой записи (аналогично, из равномерного распределения). После этого производилась вставка этого фрагмента и изменение метки аналогично описанному в разд. 2 подходу. Этот подход также не оказал существенного влияния на качество работы модели. Возможно это связано с тем, что при вставке фрагмент аудио часто попадал в окрестность середины записи, тем самым могла теряться наиболее эмоциональная часть исходной

записи (которая в среднем приходилась на середину из-за того, что велась запись коротких предложений).

4 Заключение

Предложенный подход отличается от существующих главным образом тем, что меняются метки, т.е аугментация приводит не только к устойчивости модели к изменениям в данных, но и в некотором смысле «калибрует» модель, заставляя ее запоминать характерные особенности эмоциональной речи.

Работы, посвященные аугментации в задаче SER не так распространены, как для задач классификации изображений. В похожих работах по аугментации аудио, во-первых, используется подход со спектрограммами (т.е в точности применяются методы для работы с изображениями), а во-вторых небольшие части спектрограмм вставляются в исходную в большей степени лишь для эффекта «маскирования» спектрограммы (*spectrogram masking*) [5].

Также близким подходом является миксап (*mixup*) [4] для спектрограмм [5]. Однако в таком подходе спектрограммы накладываются друг на друга (т.е берется линейная комбинация спектрограмм), что может интерпретироваться как сложение аудиосигналов. Описанный подход интуитивно понятен - эмоциональная характеристика речи зависит от соотношения эмоциональной экспрессии каждой её части и конечная метка может определяться пропорционально каждой из компонент, а также прост и эффективен в реализации.

После проведения экспериментов и исследования предложенного подхода было обнаружено, что ранее уже применялась схожая с изложенной идея [8]. Однако она применялась при исследовании классификации изображений на примере датасета

CIFAR [14], и этому подходу не было уделено достаточное внимание. Также он не получил применения в задачах ASR/SER.

Итак, можно считать, что описанный подход в некотором смысле дополняет предыдущие идеи, ранее используемые в задачах классификации изображений, и может применяться наряду с ними для задачи распознавания эмоций, особенно с небольшими датасетами (например, редкими языками). Также описанный подход демонстрирует потенциальную полезность дальнейших, более сложных, экспериментов с преобразованием аудиосигналов для увеличения обучающей выборки в задачах SER.

Список литературы

[1] Schneider S. et al. wav2vec: Unsupervised pre-training for speech recognition //arXiv preprint arXiv:1904.05862. – 2019.

[2] Baevski A. et al. wav2vec 2.0: A framework for self-supervised learning of speech representations //Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 12449-12460.

[3] Conneau A. et al. Unsupervised cross-lingual representation learning for speech recognition //arXiv preprint arXiv:2006.13979. – 2020.

[4] Zhang H. et al. mixup: Beyond empirical risk minimization //arXiv preprint arXiv:1710.09412. – 2017.

[5] Gharib S. et al. Acoustic scene classification: A competition review //2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP). – IEEE, 2018. – С. 1-6.

[6] Shor J., Venugopalan S. TRILLsson: Distilled Universal Paralinguistic Speech Representations //arXiv preprint arXiv:2203.00236. – 2022.

[7] Morais E. et al. Speech emotion recognition using self-supervised features //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2022. – С. 6922-6926.

[8] Summers C., Dinneen M. J. Improved mixed-example data augmentation //2019 IEEE Winter Conference on Applications of Computer Vision (WACV). – IEEE, 2019. – С. 1262-1270.

[9] Park D. S. et al. Specaugment: A simple data augmentation method for automatic speech recognition //arXiv preprint arXiv:1904.08779. – 2019.

[10] Livingstone S. R., Russo F. A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English //PloS one. – 2018. – Т. 13. – №. 5. – С. e0196391.

[11] Burkhardt F. et al. A database of German emotional speech //Interspeech. – 2005. – Т. 5. – С. 1517-1520.

[12] Cao H. et al. Crema-d: Crowd-sourced emotional multimodal actors dataset //IEEE transactions on affective computing. – 2014. – Т. 5. – №. 4. – С. 377-390.

[13] Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

[14] Krizhevsky A. et al. Learning multiple layers of features from tiny images. – 2009.

[15] Prasomphan S. Detecting human emotion via speech recognition by using speech spectrogram //2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). – IEEE, 2015. – С. 1-10.

[16] Sharma M. Multi-Lingual Multi-Task Speech Emotion Recognition Using wav2vec 2.0 //ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2022. – С. 6907-6911.

[17] Pepino L., Riera P., Ferrer L. Emotion recognition from speech using wav2vec 2.0 embeddings //arXiv preprint arXiv:2104.03502. – 2021.

[18] Chen S. et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing //arXiv preprint arXiv:2110.13900. – 2021.