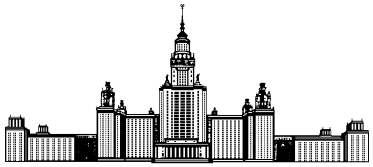
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА

«Исследование методов аугментации аудиоданных»
«Research of audio data augmentation methods»

Выполнил:

студент 5 курса 517 группы

Лукьянов Павел Александрович

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Москва, 2021

Содержание

Аннотация

В данной работе исследуются методы аугментации звуковых данных. На двух задачах аудиоклассификации проверяется эффективность как давно известных методов аугментации, так и предложенных в данной работе подходов. Результаты экспериментов показывают перспективность некоторых из предложенных методов.

1 Введение

Исследование методов аугментации очень актуально в настоящее время. Понятию аугментации сложно дать точное определение, в данной работе под ней будем понимать создание новых данных на основе уже имеющихся. С помощью аугментации можно существенно расширить объем обучающей выборки, что особенно хорошо в тех случаях, когда исходных данных не очень много. Применение аугментации способно повысить обобщающую способность модели.

Аугментация данных используется при решении многих задач глубинного обучения, связанных с обработкой изображений, текстов, звуковых данных. Классическим примером задачи, в которой используется аугментация, является задача классификации [2], [3]. Применение аугментации позволяет добиться лучшего качества в задаче автоматического распознавания речи [1].

В данной работе будем исследовать применение методов аугментации в задаче аудиоклассификации. Соответственно, аугментация будет применяться к звуковым данным, а именно к мел-спектрограммам [1], которые, в свою очередь, будут подаваться на вход нейронной сети как изображения.

Обычная спектрограмма получается после применения оконного преобразования Фурье к коротким кускам речевого сигнала. После же применения мел-фильтров к этой спектрограмме и получается мел-спектрограмма [8], в которой частота выражена в мелах.

В данной работе мел-спектрограммы нормализуются следующим образом:

$value = \frac{value - mean}{std}$, где mean – математическое ожидание значений мел-спектрограммы, std – стандартное отклонение.

Такая нормализация, в частности, важна для корректного обоснования некоторых методов аугментации, описанных в следующих разделах.

2 Существующие методы аугментации

В этом разделе представлены известные методы аугментации аудиоданных, которые будем исследовать и на основе которых будут предложены некоторые из новых подходов.

Здесь и далее считаем, что FreqSize – размерность мел-спектрограммы по частотной оси, TimeSize – размерность мел-спектрограммы по временной оси. S – матрица значений мел-спектрограммы.

Также введем матрицу $M(I, J)$, где I, J - множества индексов:

$$M(I,J) = \{M(i,j)\} = \begin{cases} 0, & (i,j) \in I \times J, \\ 1, & \text{иначе.} \end{cases}$$

Стоит рассматривать только случаи, когда в представленных ниже аугментациях значения t, f, shift ненулевые. В противном случае ($t = 0$, или $f = 0$, или shift $= 0$) мел-спектрограмма никак не изменяется.

1. TimeMasking 1. [1]

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.

   В результате применения аугментации:

   $S \to S \cdot M(\{0, \ldots, \text{FreqSize} - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

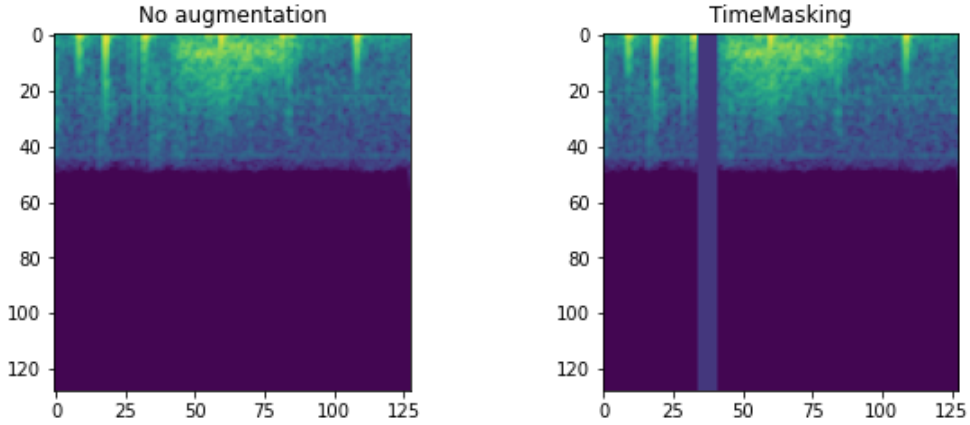


Рис. 1: TimeMasking

2. FreqMasking 2. [1]

   $f \sim U\{0, F\} f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, F - параметр аугментации.

   В результате применения аугментации:

   $S \to S \cdot M(\{f_0, \ldots, f_0 + f - 1\}, \{0, \ldots, \text{TimeSize} - 1\})$

3. Noise 3. [2]

   К каждому значению в мел-спектрограмме добавляется $g \sim N(0, \sigma)$ (для каж-

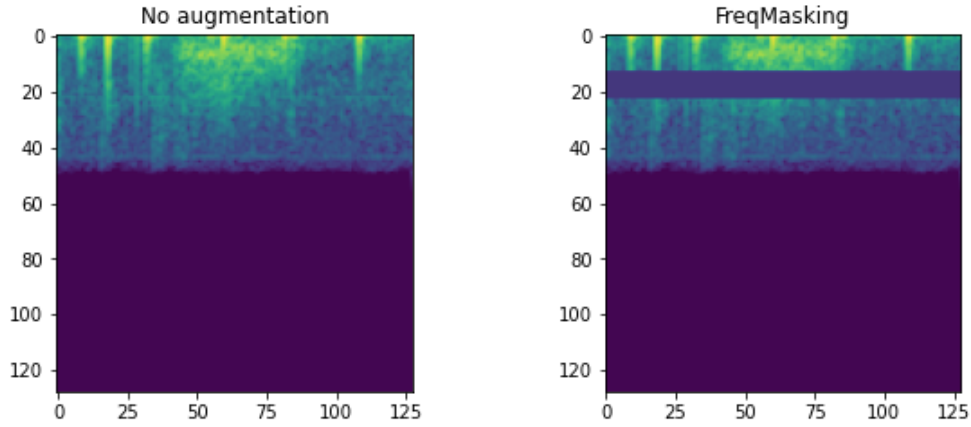


Рис. 2: FreqMasking

дого значения мел-спектрограммы генерируется свое g), где σ - параметр аугментации (в данной работе $\sigma = 0.01$).

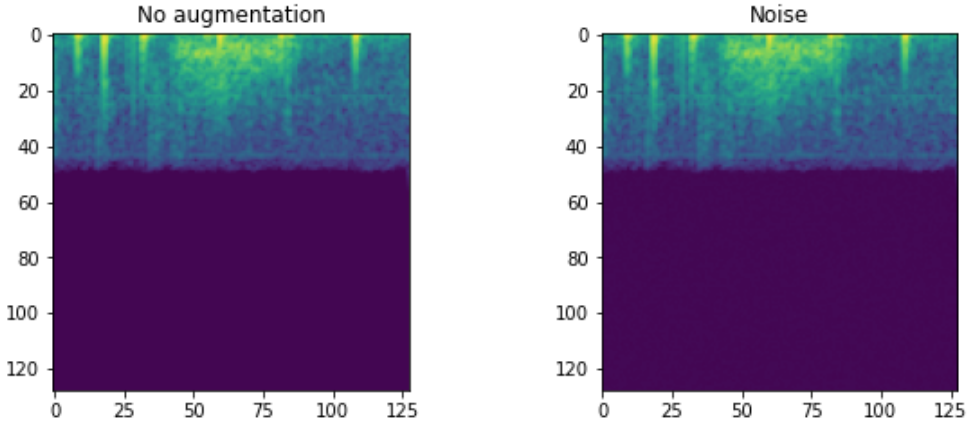


Рис. 3: Noise

4. TimeShift 4. [2]

   Сдвигаем все значения мел-спектрограммы относительно временной оси влево или вправо на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига определяется знаком shift: если shift > 0, происходит сдвиг вправо, если shift < 0 - влево. Пустая область, образующаяся в результате сдвига, заполняется нулями.

5. RandomErasing 5. [3] Случайное вырезание прямоугольника в мел-спектрограмме. $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - t - 1\}$,

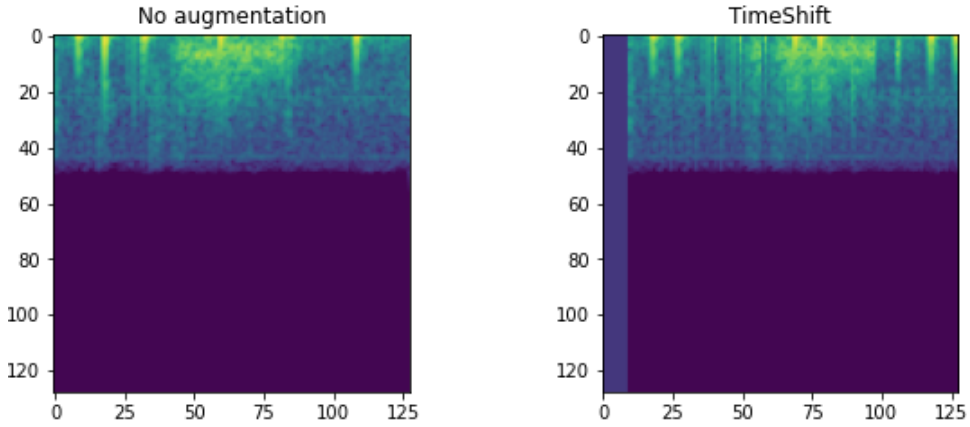


Рис. 4: TimeShift

$f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - f - 1\}$, T, F - параметры аугментации.

В результате применения аугментации:

$S \to S \cdot M(\{f_0, \ldots, f_0 + f - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

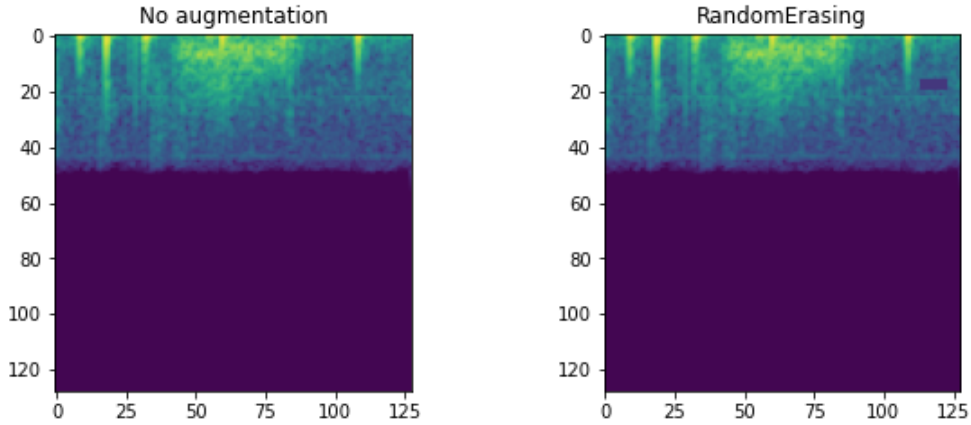


Рис. 5: RandomErasing

Стоит отметить, что в связи с нормализацией мел-спектрограммы замена некоторых значений на 0 в результате применения аугментации – это замена на математическое ожидание [1].

3 Предлагаемые подходы к аугментации

Ниже представлены предлагаемые возможные подходы к аугментации звуковых данных:

1. FreqShift 6.

   Аналог TimeShift, только теперь сдвиг происходит относительно частотной оси. shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига определяется знаком shift: если shift > 0, происходит сдвиг вниз, если shift < 0 - вверх. Пустая область, образующаяся в результате сдвига, заполняется нулями.

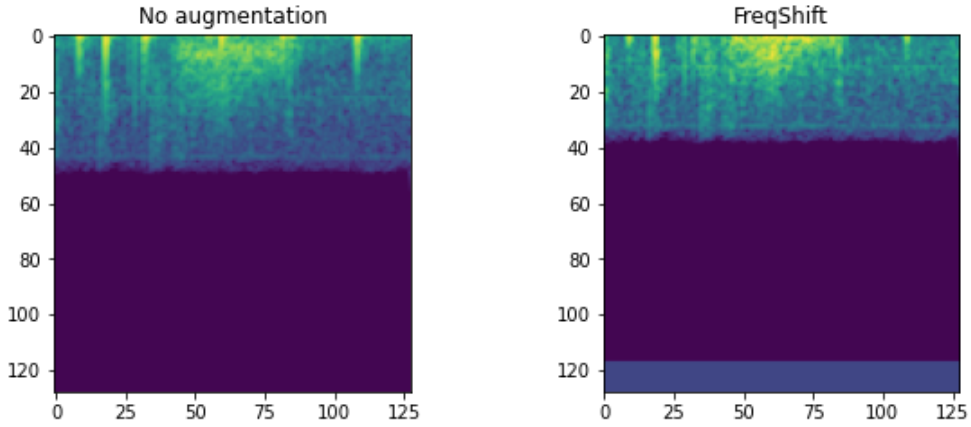


Рис. 6: FreqShift

2. TimeNoising 7.

   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - t - 1\}$, T - параметр аугментации.

   Ко всем значениями мел-спектрограммы, индексы которых принадлежат множеству $\{0, \ldots, \text{FreqSize}-1\} \times \{t_0, \ldots, t_0+t-1\}$, добавляется значение $g \sim N(0, \sigma)$ (для каждого значения мел-спектрограммы генерируется свое g), где σ - параметр аугментации (в данной работе $\sigma = 0.1$). Идея метода заключается в том, чтобы зашумлять не всю мел-спектрограмму, а только отдельные ее участки.

3. FreqNoising 8.

   $f \sim U\{0, F\}, f_0 \sim U\{0, FreqSize - f - 1\}$, F - параметр аугментации.

   Ко всем значениями мел-спектрограммы, индексы которых принадлежат множеству $\{f_0, \ldots, f_0 + f - 1\} \times \{0, \ldots, \text{TimeSize} - 1\}$, добавляется значение $g \sim N(0, \sigma)$ (для каждого значения мел-спектрограммы генерируется свое g), где σ - параметр аугментации (в данной работе $\sigma = 0.1$).

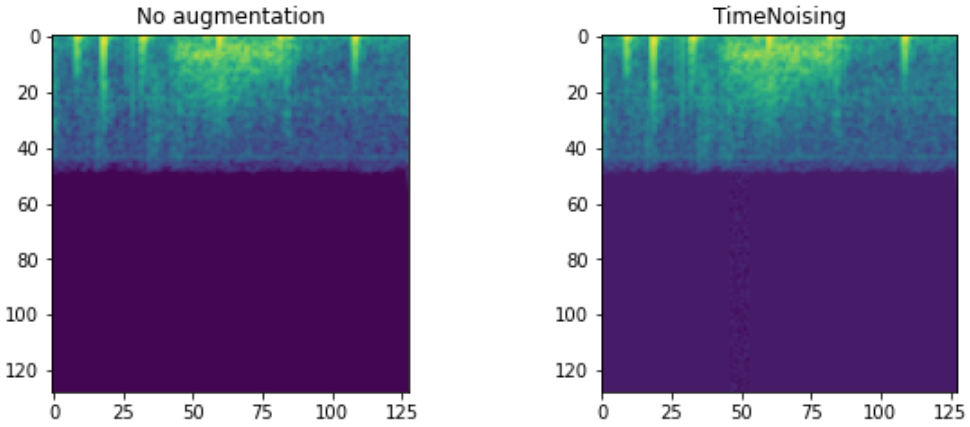


Рис. 7: TimeNoising

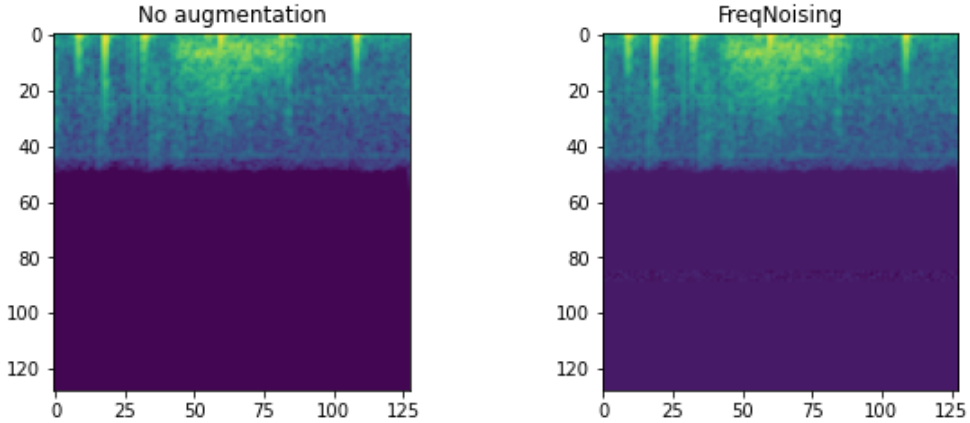


Рис. 8: FreqNoising

4. TimeCycleShift 9.

   Циклический сдвиг всех значений мел-спектрограммы относительно временной оси влево или вправо на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига выбирается так же, как и в TimeShift.

5. FreqCycleShift 10.

   Циклический сдвиг всех значений мел-спектрограммы относительно частотной оси вверх или вниз на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига выбирается так же, как и в FreqShift.

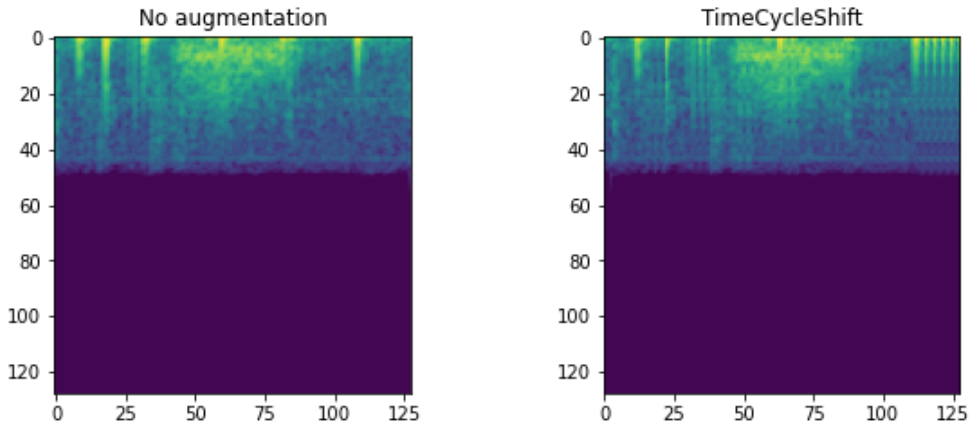


Рис. 9: TimeCycleShift

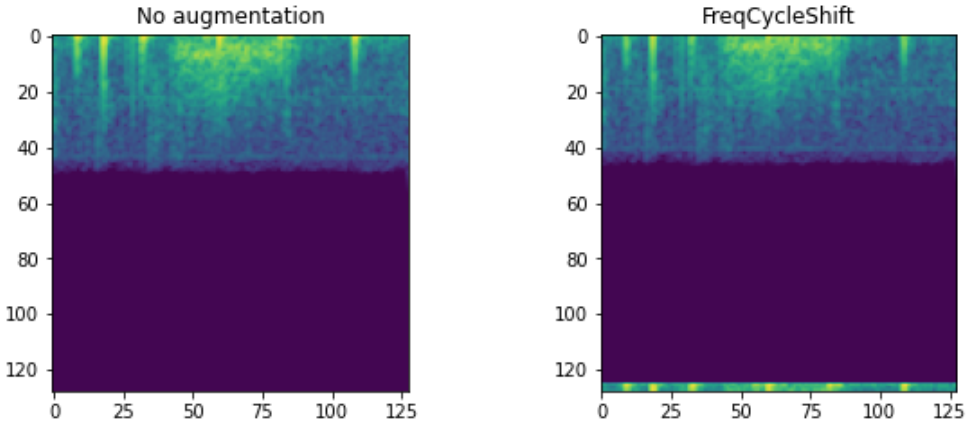


Рис. 10: FreqCycleShift

6. TimeSpecialShift 11.

   Сдвиг всех значений мел-спектрограммы относительно временной оси влево или вправо на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига выбирается так же, как и в TimeShift. Пустая область, образующаяся в результате сдвига, заполняется значениями из исходной спектрограммы $S[0 : \text{FreqSize} - 1; 0 : |\text{shift}| - 1]$ в случае сдвига вправо или $S[0 : \text{FreqSize} - 1; \text{TimeSize} - |\text{shift}| : \text{TimeSize} - 1]$ в противном случае. Идея метода заключается в том, чтобы пустой участок, образующийся в результате сдвига, заполнять не нулями, а значениями из соседнего участка.

7. FreqSpecialShift 12.

   Сдвиг всех значений мел-спектрограммы относительно частотной оси вверх или

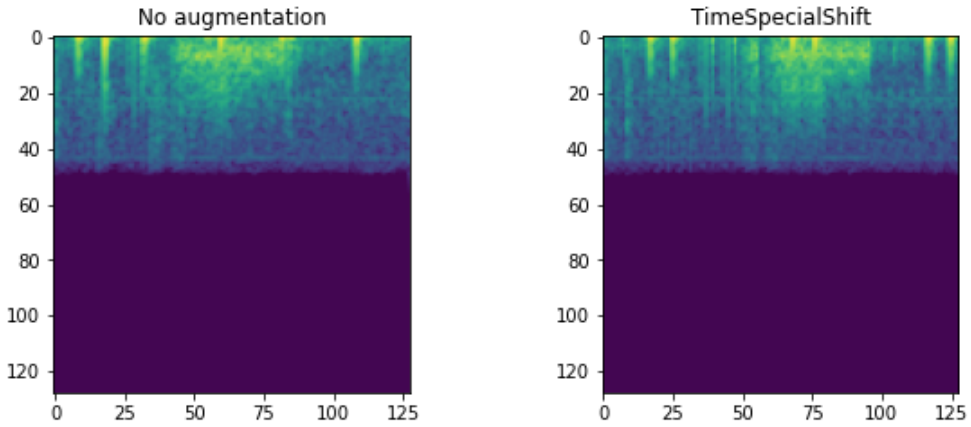


Рис. 11: TimeSpecialShift

вниз на |shift|, где shift $\sim U\{-\text{max\_shift}, \text{max\_shift}\}$, max_shift - параметр аугментации. Направление сдвига выбирается так же, как и в FreqShift. Пустая область, образующаяся в результате сдвига, заполняется значениями из исходной спектрограммы $S[0 : |\text{shift}| - 1; 0 : \text{TimeSize} - 1]$ в случае сдвига вниз или $S[\text{FreqSize} - |\text{shift}| : \text{FreqSize} - 1; 0 : \text{TimeSize} - 1]$ в противном случае.

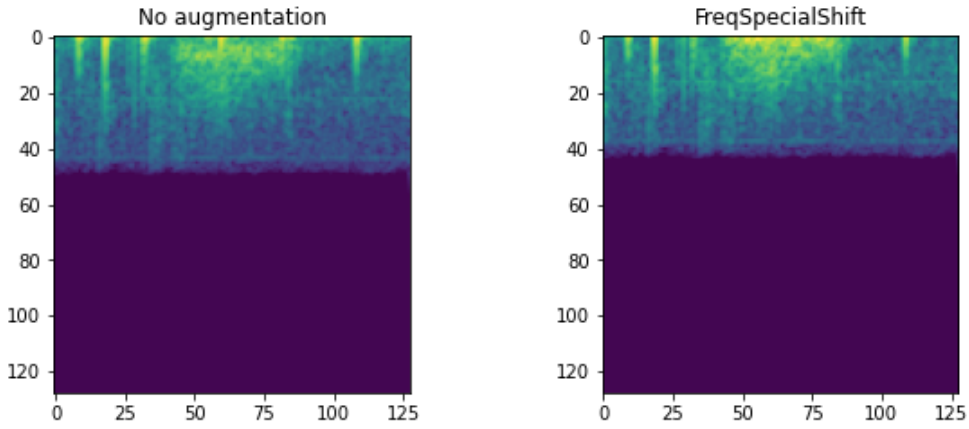


Рис. 12: FreqSpecialShift

8. TimeSwapAugmentation 13.

   $t \sim U\{0, T\}, t_0 \sim U\{t, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.

   В результате применения аугментации:

   $S[0 : \text{FreqSize} - 1; t_0 : t_0 + t - 1] \leftrightarrow S[0 : \text{FreqSize} - 1; t_0 - t : t_0 - 1]$

   Идея метода заключается в перестановке соседних участков.

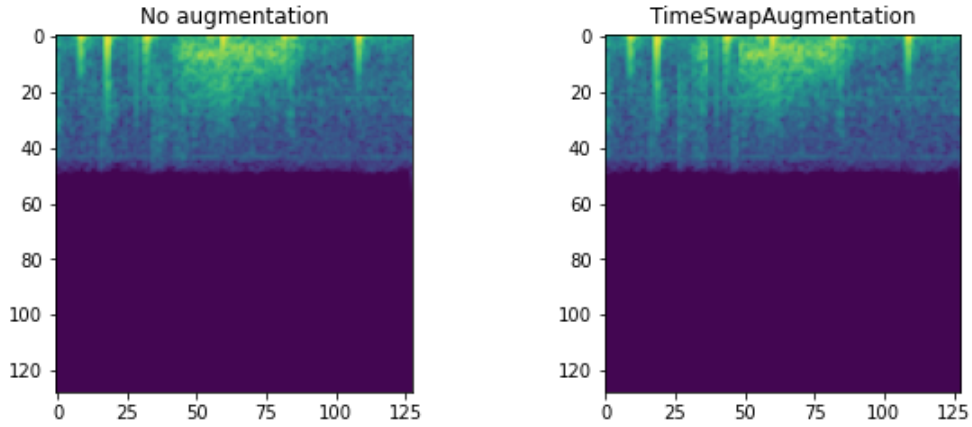


Рис. 13: TimeSwapAugmentation

9. FreqSwapAugmentation 14. $f \sim U\{0, F\}$, $f_0 \sim U\{f, \text{FreqSize}-1-f\}$, F - параметр аугментации.

   В результате применения аугментации:

   $S[f_0 : f_0 + f - 1; 0 : \text{TimeSize} - 1] \leftrightarrow S[f_0 - f : f_0 - 1; 0 : \text{TimeSize} - 1]$

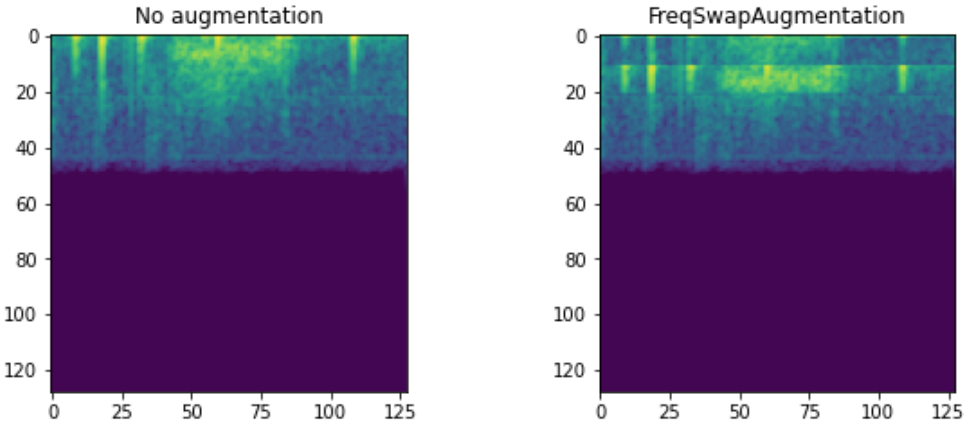


Рис. 14: FreqSwapAugmentation

10. TimeReplyMasking 15.

    $t \sim U\{0, T\}, t_0 \sim U\{t, \text{TimeSize} - 1 - 2t\}$, T - параметр аугментации.

    Участок $S[0 : \text{FreqSize} - 1; t_0 : t_0 + t - 1]$ заменяется на один из участков $S[0 : \text{FreqSize}-1, t_0-t : t_0-1]$, $S[0 : \text{FreqSize}-1; t_0+t : t_0+2t-1]$ в зависимости от направления. Направление выбирается с вероятностью 0.5. Идея метода похожа на идею в TimeSpecialShift.

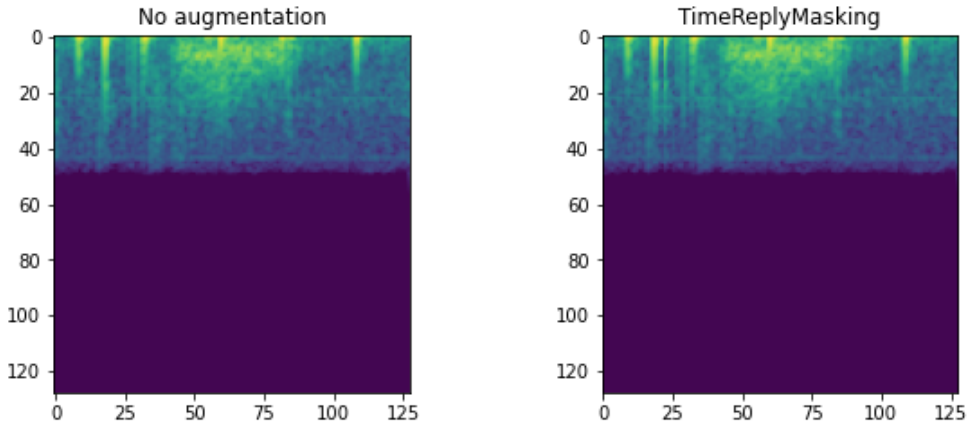


Рис. 15: TimeReplyMasking

11. FreqReplyMasking 16.

    $f \sim U\{0, F\}, f_0 \sim U\{f, \text{FreqSize} - 1 - 2f\}$, F - параметр аугментации.

    Участок $[f_0 : f_0 + f - 1; 0 : \text{TimeSize} - 1]$ заменяется на один из участков $[f_0 - f : f_0 - 1; 0 : \text{TimeSize} - 1]$, $[f_0 + f : f_0 + 2f - 1; 0 : \text{TimeSize} - 1]$ в зависимости от направления. Направление выбирается с вероятностью 0.5.

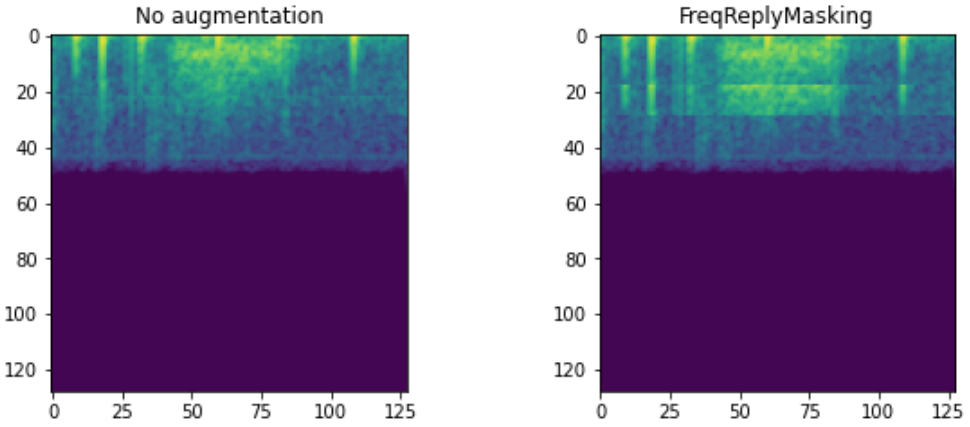


Рис. 16: FreqReplyMasking

12. TimeRandomSwap 17.

    $t \sim U\{0, T\}, t_1 \sim U\{t, \text{TimeSize} - 1 - t\}, t_2 \sim U\{t, \text{TimeSize} - 1 - t\}, |t_1 - t_2| >= t$,

    T - параметр аугментации.

    В результате применения аугментации:

    $S[0 : \text{FreqSize} - 1; t_1 : t_1 + t - 1] \leftrightarrow S[0 : \text{FreqSize} - 1; t_2 : t_2 + t - 1]$

    Идея метода заключается в перестановке произвольных участков.

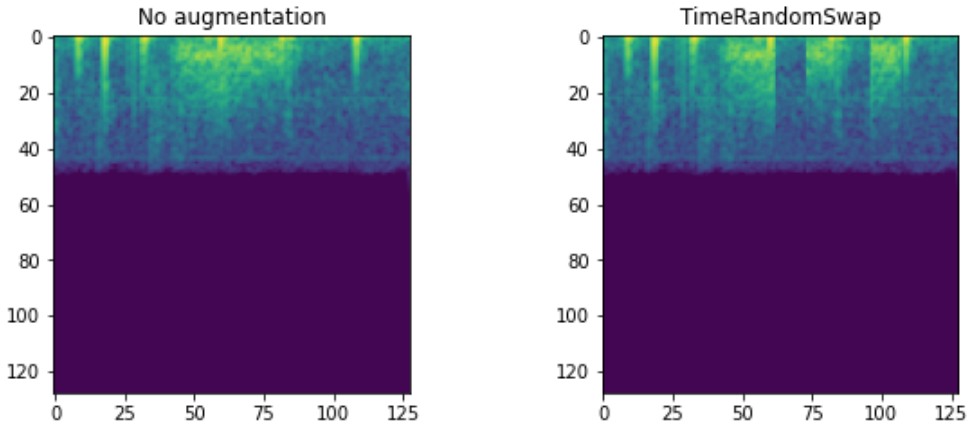


Рис. 17: TimeRandomSwap

13. FreqRandomSwap 18.

    $f \sim U\{0, F\}, f_1 \sim U\{f, \text{FreqSize}-1-f\}, f_2 \sim U\{f, \text{FreqSize}-1-f\}, |f_1-f_2| >= f,$

    F - параметр аугментации.

    В результате применения аугментации:

    $S[f_1 : f_1 + f - 1; 0 : \text{TimeSize} - 1] \leftrightarrow S[f_2 : f_2 + f - 1; 0 : \text{TimeSize} - 1].$

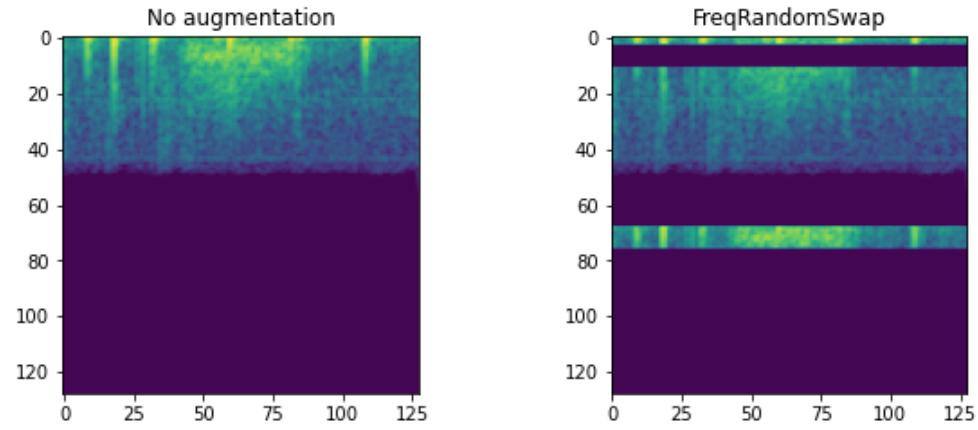


Рис. 18: FreqRandomSwap

4 Вычислительные эксперименты

Методы аугментации будем анализировать применительно к задачам аудиоклассификации. Для этой цели будем использовать 2 набора данных: AudioMnist [7] и HeartBeatSounds [6].

Датасет AudioMnist состоит из 3000 записей (файлов формата .wav), на которых некоторый человек произносит одну из 10 цифр. Соответственно, задача классификации заключается в том, чтобы определить какую конкретно цифру произносит человек на записи.

Датасет HeartBeatSounds представляет собой записи звуков сердцебиения (656 файлов формата .wav). Задача - определить, к какому из 3 типов относятся звуки на записи: normal, murmur, extrastole.

В этих двух датасетах оставим только те записи, длина которых больше некоторого порогового значения. Из оставшихся файлов в каждом датасете извлекаем фиксированные по длине куски записи. Это необходимо для того, чтобы мелспектрограммы были одного размера.

Для решения задач классификации будем использовать нейронные сети архитектур resnet18 и resnet50 и алгоритм оптимизации Adam. Нейронная сеть будет обучаться 75 эпох в случае AudioMnist и 70 эпох в случае HeartBeatSounds. Функция потерь - кросс-энтропия.

В данной работе значения параметров F и T для всех типов аугментаций, где используются эти параметры, считаем равными $0.1 \cdot \text{FreqSize} - 1$ и $0.1 \cdot \text{TimeSize} - 1$ соответственно. Параметр max_shift будем считать равным $0.1 \cdot \text{FreqSize} - 1$ в случае сдвигов относительно частотной оси и $0.1 \cdot \text{TimeSize} - 1$ в случае сдвигов относительно временной оси.

Для оценивания качества классификации будем использовать Accuracy – долю верно классифицированных объектов.

Датасеты разбиваются на train_0 и test в отношении $4 : 1$. train_0, в свою очередь, разбивается на train и valid в том же отношении. Обучение происходит на выборке train. После обучения берется лучший по Accuracy результат на валидационной выборке valid и считается Accuracy на тестовой выборке test. Именно по Accuracy на тестовой выборке будем оценивать эффективность методов аугментации.

Датасеты разбиваются на train, valid и test при 5 разных фиксированных random_seed. Результаты, соответственно, усредняются. В процессе обучения аугментация применяется к каждому сэмплу в каждом батче.

4.1 Результаты экспериментов

Результаты экспериментов представлены в таблице 1. В ней используются следующие сокращения:

R18 = resnet18

R50 = resnet50

AM = AudioMnist

HB = HeartBeatSounds

Метод аугментации	R18 + AM	R18 + HB	R50 + AM	R50 + HB
No Augmentation	**0.954 ± 0.01**	**0.83 ± 0.016**	**0.953 ± 0.009**	**0.824 ± 0.018**
TimeMasking	0.952 ± 0.006	0.829 ± 0.014	0.956 ± 0.006	0.826 ± 0.007
FreqMasking	0.952 ± 0.004	0.829 ± 0.014	0.957 ± 0.004	0.825 ± 0.013
Noise	0.958 ± 0.006	**0.837 ± 0.009**	0.951 ± 0.009	0.821 ± 0.015
RandomErasing	**0.962 ± 0.005**	0.823 ± 0.01	0.951 ± 0.01	0.817 ± 0.013
TimeShift	**0.961 ± 0.006**	**0.866 ± 0.014**	0.957 ± 0.003	**0.863 ± 0.015**
FreqShift	0.937 ± 0.013	0.818 ± 0.01	0.939 ± 0.013	0.821 ± 0.01
TimeNoising	**0.96 ± 0.003**	0.818 ± 0.01	0.956 ± 0.008	0.816 ± 0.018
FreqNoising	0.957 ± 0.004	0.829 ± 0.013	0.952 ± 0.01	0.823 ± 0.019
TimeCycleShift	**0.962 ± 0.006**	**0.87 ± 0.01**	0.956 ± 0.014	**0.872 ± 0.017**
FreqCycleShift	0.937 ± 0.013	0.819 ± 0.004	0.929 ± 0.018	0.818 ± 0.017
TimeSpecialShift	0.953 ± 0.011	**0.865 ± 0.011**	0.952 ± 0.006	**0.858 ± 0.015**
FreqSpecialShift	0.942 ± 0.006	0.821 ± 0.011	0.943 ± 0.007	**0.835 ± 0.007**
TimeSwapAugmentation	0.957 ± 0.009	**0.835 ± 0.015**	0.95 ± 0.011	**0.833 ± 0.015**
FreqSwapAugnentation	0.952 ± 0.006	0.812 ± 0.016	0.953 ± 0.008	0.828 ± 0.019
TimeReplyMasking	0.957 ± 0.006	**0.834 ± 0.017**	0.953 ± 0.006	0.826 ± 0.022
FreqReplyMasking	**0.959 ± 0.008**	0.819 ± 0.016	0.953 ± 0.006	0.815 ± 0.01
TimeRandomSwap	**0.96 ± 0.003**	**0.835 ± 0.008**	**0.96 ± 0.005**	0.823 ± 0.021
FreqRandomSwap	0.957 ± 0.005	0.822 ± 0.009	0.947 ± 0.007	0.828 ± 0.007

Таблица 1: Результаты экспериментов

4.2 Анализ полученных результатов

Стоит отметить, что в данных экспериментах аугментация применяется не очень агрессивно. Вероятно, увеличение значений параметров аугментаций позволило бы улучшить результаты. Также стоит сказать, что в некоторых случаях с применением аугментаций моделе требуется большее количество эпох, в данной работе число эпох фиксировано.

Полученные результаты показывают:

- Сдвиги относительно частотной оси (FreqShift, FreqCycleShift, FreqSpecialShift) показали плохие результаты, приводящие почти во всех случаях к ухудшению качества.

- Некоторые аугментации (например, TimeNoising) привели к улучшению качества на одном датасете (AudioMnist), но в то же время привели к значительному снижению качества на другом датасете (HeartBeatSounds).

- Сдвиги относительно временной оси (TimeShift, TimeCycleShift, TimeSpecialShift) показали хорошие результаты, особенно на датасете HeartBeatSounds. Однако TimeSpecialShift проигрывает в результатах обычному TimeShift, поэтому TimeSpecialShift, на мой взгляд, не является достаточно перспективным для дальнейших исследований. Метод же TimeCycleShift проявил себя лучше остальных в данных экспериментах.

- Методы TimeSwapAugmentation, TimeReplyMasking, TimeRandomSwap оказались достаточно стабильными в данных экспериментах. Поэтому они являются перспективными для дальнейших исследований.

- Остальные методы аугментации, которые не упоминались выше, не привели к стабильным результатам, что, впрочем, не ставит крест на них, поскольку среди этих методов есть известные методы аугментации, активно использующиеся в настоящее время. Все известные подходы к аугментации, за исключением TimeShift, не показали хороших результатов в данных экспериментах.

5 Заключение

В процессе выполнения работы получены следующие результаты:

- Предложены и реализованы подходы к аугментации аудиоданных.
- Проведены вычислительные эксперименты, показывающие эффективность некоторых из предложенных подходов.
- На основании результатов экспериментов сделан вывод о том, что методы TimeCycleShift, TimeSwapAugmentation, TimeReplyMasking, TimeRandomSwap являются наиболее перспективными для дальнейших исследований.

Список литературы

[1] Daniel S. Park , William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. 2019.

[2] Steffen Illium, Robert Muller, Andreas Sedlmeier and Claudia Linnhoff-Popien. Surgical Mask Detection with Convolutional Neural Networks and Data Augmentations on Spectrograms. 2020.

[3] Haiwei Wu, Lin Zhang, Lin Yang, Xuyang Wang, Junjie Wang, Dong Zhang, Ming Li. Mask Detection and Breath Monitoring from Speech: on Data Augmentation, Feature Representation and Modeling. 2020.

[4] Sarala Padi, Dinesh Manocha, Ram D.Sriram. Multi-Window Data Augmentation Approach for Speech Emotion Recognition. 2020.

[5] Yeongtae Hwang, Hyemin Cho, Hongsun Yang, Dong-Ok Won, Insoo Oh, and Seong-Whan Lee. Mel-spectrogram augmentation for sequence-to-sequence voice conversion. 2020.

[6] Heartbeat Sounds. Classifying heartbeat anomalies from stethoscope audio. https://www.kaggle.com/kinguistics/heartbeat-sounds

[7] Audio MNIST. https://www.kaggle.com/alanchn31/free-spoken-digits

[8] https://librosa.org/doc/main/generated/librosa.feature.melspectrogram.html