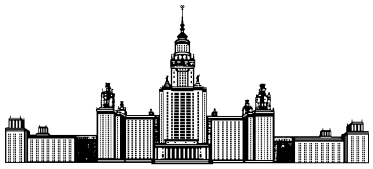

Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Лукьянов Павел Александрович

Методы аугментации аудиоданных

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Научный руководитель:
д.ф-м.н., профессор
Дьяконов Александр Геннадьевич

Москва, 2022

Содержание

Аннотация

В данной работе предлагается метод аугментации аудиоданных SwapVerticalStripes, основанный на перестановке вертикальных полос в мел-спектрограмме, и его модификации SwapNeighboringStripes, SwapSeveralStripes. Также предлагается алгоритм применения методов аугментации с выбором конкретного метода аугментации после каждой эпохи обучения. Проведенные вычислительные эксперименты показывают возможную применимость предлагаемых подходов в задаче аудиоклассификации.

1 Введение

Понятию аугментации сложно дать точное определение, в данной работе под аугментацией понимается создание новых данных с помощью модификации уже имеющихся. Использование аугментации может быть особенно полезно для небольшой обучающей выборки и может улучшить обобщающую способность модели, являясь мощным инструментом в борьбе с переобучением.

Исследование методов аугментации данных актуально в настоящее время. Аугментация успешно используется при решении многих задач глубинного обучения, связанных с обработкой изображений, звуковых данных, текстов.

В данной работе рассматриваются методы аугментации аудиоданных, а именно мел-спектрограмм. Мел-спектрограмма получается после применения оконного преобразования Фурье [1] и мел-фильтров [2]. Мел-спектрограммы представляют собой двумерные матрицы, поэтому их можно рассматривать как изображения и многие подходы к аугментации картинок применимы к аудиоданным. Например, метод Random Erasing [3], сводящийся к вырезанию случайных прямоугольников из изображения, может быть использован в задаче аудиоклассификации [4]. Также в задаче классификации звуковых данных применяются такие методы аугментации, как Shift Augmentation [5] — сдвиг мел-спектрограммы влево или вправо, Noise Augmentation [5] — добавление Гауссовского шума, Loudness Augmentation [5] — регулирование громкости, Speed augmentation [5] — ускорение или замедление аудиозаписи.

SpecAugment [6] — один из наиболее известных методов аугментации аудиоданных, который показал свою эффективность в задаче автоматического распознавания речи. Политика аугментации SpecAugment определяется 3 возможными преобразованиями:

1. Time warping [6] (искривление времени)
2. Frequency masking [6] (зануление значений мел-спектрограммы внутри горизонтальной полосы)
3. Time masking [6] (зануление значений мел-спектрограммы внутри вертикальной полосы)

В настоящее время известны некоторые модификации SpecAugment: SpliceOut [7], SpecAugment++ [8].

В данной работе рассматриваются методы аугментации, которые могут применяться "на лету"(онлайн-аугментация), т.е. преобразования мел-спектрограмм, соответствующие этим аугментациям, должны выполняться достаточно быстро.

На практике выбирается некоторый набор заранее заданных методов аугментации. Пусть N - число выбранных методов ($N \geq 1$). В процессе обучения чаще всего используются следующие стратегии применения методов аугментации:

1. к каждому объекту обучающей выборки применяются изначально или во время обучения все N аугментаций. Таким образом, число используемых данных на каждой эпохе увеличивается в N раз.
2. преобразование, которое будет применено к конкретной мел-спектрограмме, выбирается случайным образом с вероятностью $\frac{1}{N}$ [9].

Однако возможны и другие стратегии использования методов аугментации. В работе [10] оптимальная политика применения методов аугментации ищется с помощью методов обучения с подкреплением. В работе [11] предлагается идея минимизации максимальных потерь среди аугментированных данных:
$\min_\theta E_{x\sim D} \max_i L(Augment_i(x), \theta)$, где
D — датасет,
θ — параметры нейронной сети,
L — функция потерь,
$\{Augment_1, Augment_2, ..., Augment_n\}$ — методы аугментации..

В данной работе предлагается метод аугментации SwapVerticalStripes, основанный на перестановке вертикальных полос в мел-спектрограмме, его модификации SwapNeighboringStripes, SwapSeveralStripes и алгоритм применения методов аугментации с выбором конкретного метода аугментации после каждой эпохи обучения.

2 Существующие методы аугментации

Ниже представлены известные подходы к аугментации аудиоданных, используемые в работе.
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] [6]
   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, T - параметр аугментации.
   В результате применения аугментации:
   $S \to S \cdot M(\{0, \ldots, \text{FreqSize} - 1\}, \{t_0, \ldots, t_0 + t - 1\})$

2. FreqMasking[2] [6]
   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, 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] [5]
   К каждому значению в мел-спектрограмме добавляется $g \sim N(0, \sigma)$ (для каждого значения мел-спектрограммы генерируется свое g), где σ - параметр аугментации (в данной работе $\sigma = 0.01$).

4. TimeShift[4] [5]
   Сдвигаем все значения мел-спектрограммы относительно временной оси влево

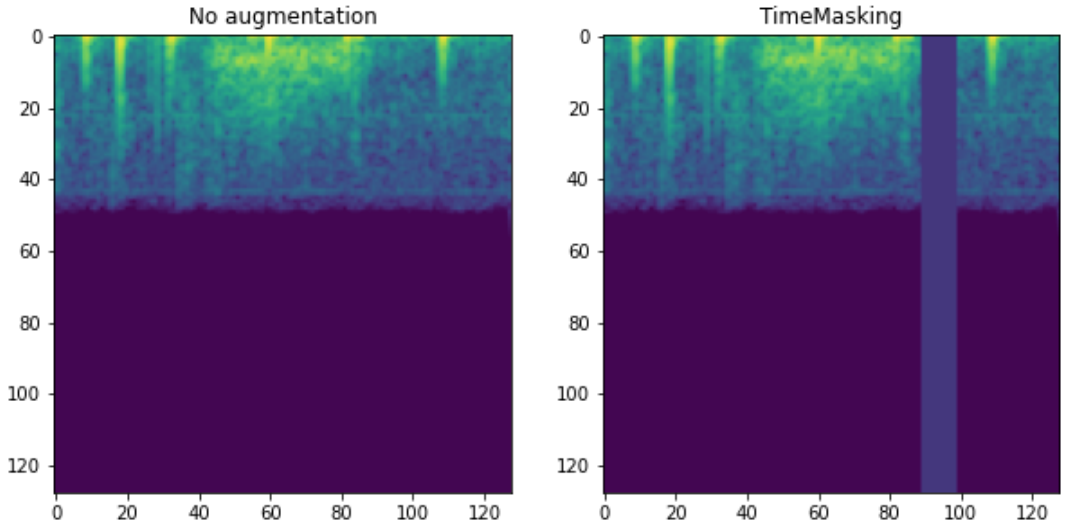

Рис. 1: TimeMasking

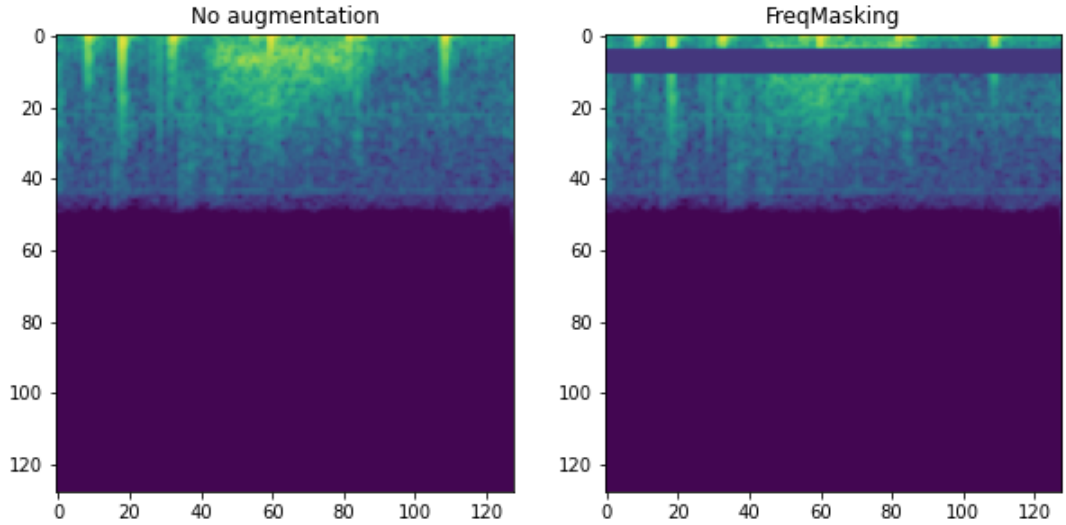

Рис. 2: FreqMasking

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.

В данной работе мел-спектрограммы нормализуются следующим образом: $value = \frac{value - mean}{std}$, где mean — математическое ожидание значений мел-спектрограммы, std — стандартное отклонение.

Поэтому замена некоторых значений мел-спектрограммы на 0 в результате применения аугментации — это замена на математическое ожидание.

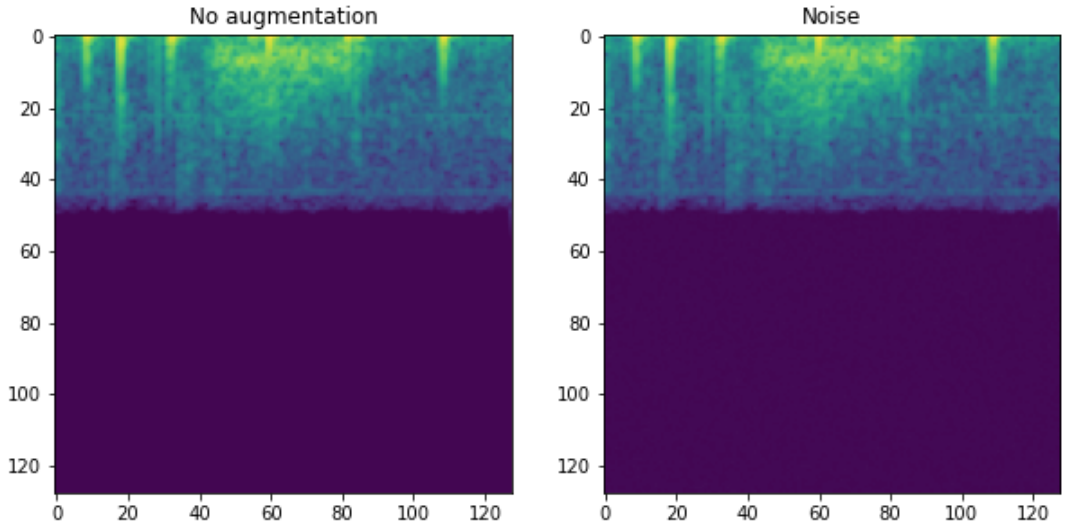

Рис. 3: Noise

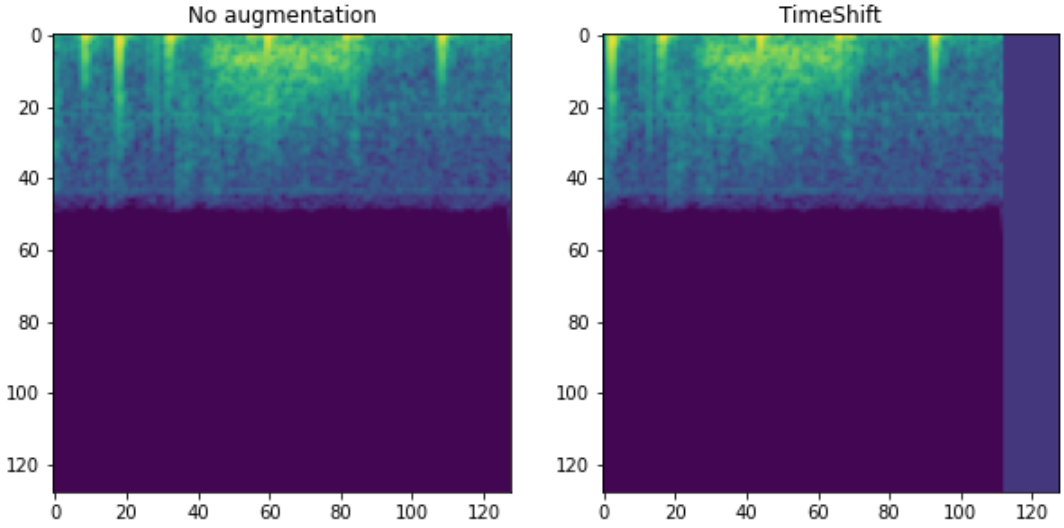

Рис. 4: TimeShift

3 Предлагаемые подходы

3.1 Методы аугментации, основанные на перестановке вертикальных полос в мел-спектрограмме

Перестановка слов — метод аугментации, используемый в задачах, связанных с обработкой текстов [12]. Подобная интуиция может быть применима и к звуковым данным.

В данной работе предлагается метод **SwapVerticalStripes**: [5]
$t \sim U\{0, T\}, t_1 \sim U\{t, \text{TimeSize} - 1 - t\}, t_2 \sim U\{t, \text{TimeSize} - 1 - t\}, |t_1 - t_2| >= t,$
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 вертикальных полос в мел-спектрограмме.

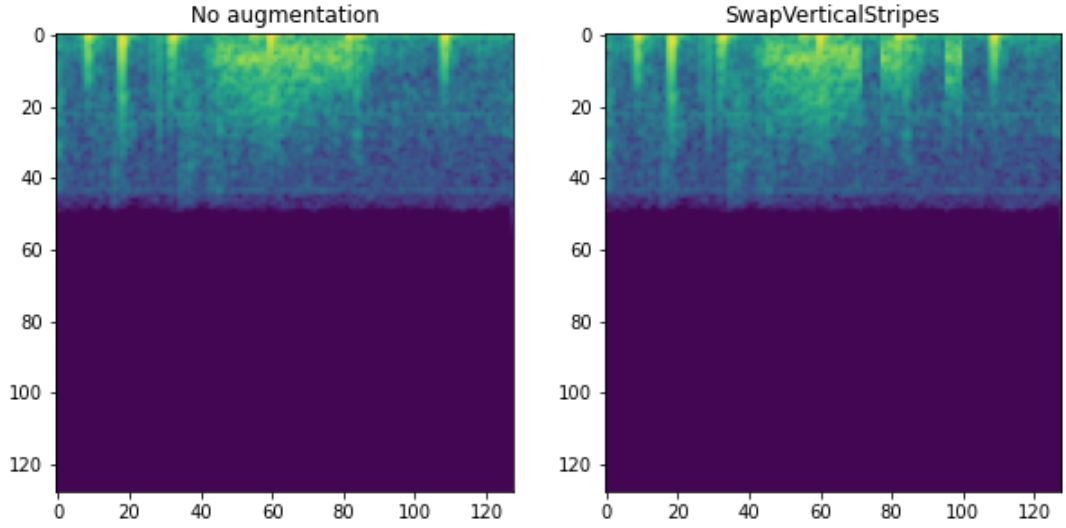

Рис. 5: SwapVerticalStripes

Также в данной работе предлагаются следующие модификации метода SwapVerticalStripes:

1. **SwapNeighboringStripes** [6]
   $t \sim U\{0, T\}, t_0 \sim U\{t, \text{TimeSize} - 1 - t\}$, 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 предлагаемого метода SwapNeighboringStripes заключается в перестановке соседних вертикальных полос.

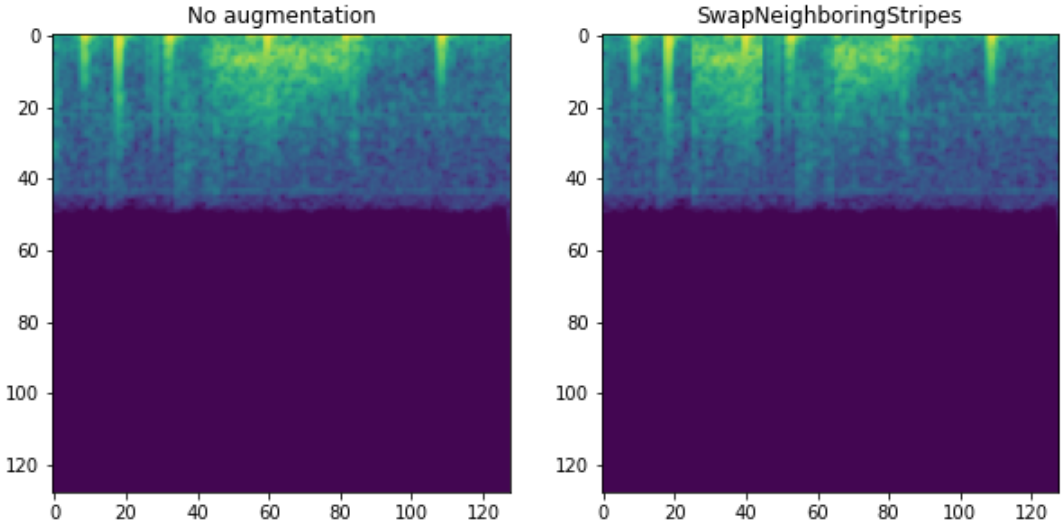

Рис. 6: SwapNeighboringStripes

2. **SwapSeveralStripes** [7]
   T, N — параметры аугментации
   $n \sim U\{0, N\}$
   В результате применения аугментации (процедура повторяется n раз):
   $T_0 = \lfloor \frac{T}{n} \rfloor$
   $t \sim U\{0, T\}, t_1 \sim U\{t, \text{TimeSize} - 1 - t\}, t_2 \sim U\{t, \text{TimeSize} - 1 - t\}, |t_1 - t_2| >= t$
   Идея предлагаемого метода SwapSeveralStripes заключается в перестановке нескольких вертикальных полос.

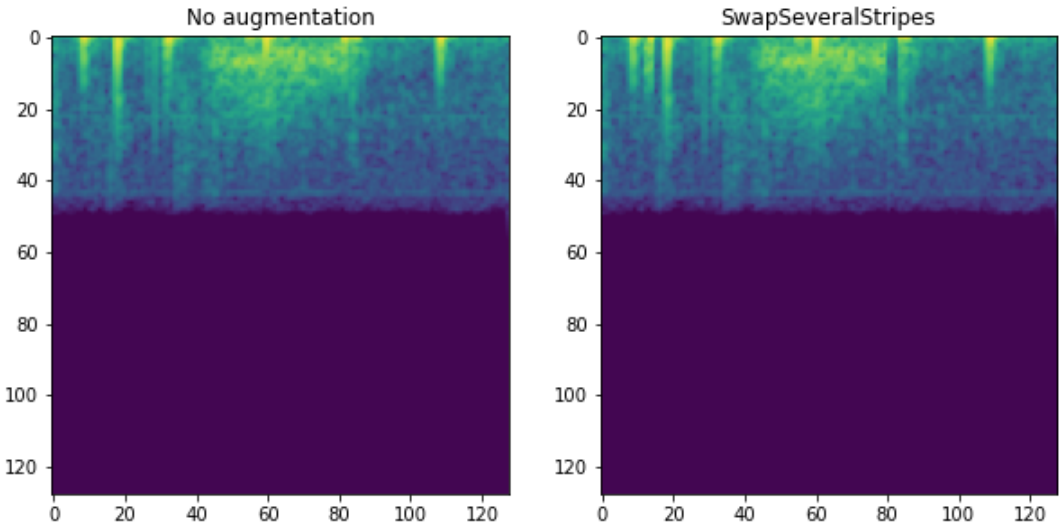

Рис. 7: SwapSeveralStripes

3.2 Алгоритм применения методов аугментации с выбором конкретного метода аугментации после каждой эпохи обучения

Введем следующую операцию:
Augmentation$(X) = \{\text{Augmentation}(x)\ \forall x \in X\}$, где X — датасет,
Augmentation — метод аугментации.

В данной работе предлагается алгоритм 1 применения методов аугментации.

Идея предлагаемого подхода заключается в следующем: в конце n-ой эпохи обучения выбирается метод аугментации, на котором модель нейронной сети работает хуже всего, и далее выбранный метод используется в процессе обучения на n + 1 эпохе.

Стоит отметить, что в работе [13] подобная идея используется для нахождения "худшего" с точки зрения метрика качества на валидационной выборке параметра аугментации, используемого в процессе обучения.

Algorithm 1 Предлагаемый алгоритм

Augmentations = $\{Augment_1, Augment_2, ..., Augment_n\}$ — заданный набор аугментаций,
$Augment$ — случайно выбранная аугментация из Augmentations,
(X_{val}, y_{val}) — валидационный датасет,
(X_{train}, y_{train}) — обучающая выборка,
f — метрика качества,
M — число эпох обучения нейронной сети
Цикл от $j = 1$ **до** M **выполнять**
    train-шаг с применением $Augment$
    вычисление $F_i = f(Augment_i(X_{val}), y_{val}), i = \overline{1, n}$
    $Augment = Augment_k$, где $k = argmin_k(F_k)$
Конец цикла

4 Вычислительные эксперименты

Для исследования применимости предложенных подходов в задаче классификации вычислительные эксперименты проведены с использованием трех датасетов: HeartBeatSounds [14] [15] (звуки сердцебиения), GTZAN [16] [17] (классификация музыкальных жанров), Audio MNIST [18] [19] (классификация произнесенных человеком цифр).

Датасет HeartBeatSounds [14] [15] представляет собой записи звуков сердцебиения (656 файлов формата .wav). Задача — определить, к какому из 3 типов относятся звуки на записи: normal, murmur, extrastole.

Датасет GTZAN [16] [17] состоит из 1000 музыкальных записей (файлов формата .wav). Задача классификации заключается в определении музыкального жанра. Всего в датасете представлено 10 музыкальных жанров: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, rock.

Датасет Audio MNIST [18] [19] 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. В данной работе использовалась версия датасета на kaggle [19].

В этих датасетах оставлены только корректно считываемые записи, длина которых больше некоторого порогового значения. Из оставшихся файлов в каждом датасете были извлечены фиксированные по длине куски записи (это необходимо для того, чтобы мел-спектрограммы были одного размера), при этом в случае датасета HeartBeatSounds [14] [15] из одного файла в зависимости от длины записи могло быть извлечено несколько непересекающихся кусков.

Ниже представлено количество элементов в каждом классе в каждом из трех датасетов после предобработки данных:

- HeartBeatSounds [14] [15]
    1. normal — 1296
    2. murmur — 500
    3. extrastole — 172

- GTZAN [16] [17]
    1. blues — 100
    2. classical — 100
    3. country — 100
    4. disco — 100
    5. hip hop — 100
    6. jazz — 99
    7. metal — 100
    8. pop — 100
    9. reggae — 100
    10. rock — 100
- Audio MNIST [18] [19]
    1. 0 — 300
    2. 1 — 289
    3. 2 — 284
    4. 3 — 278
    5. 4 — 294
    6. 5 — 299
    7. 6 — 263
    8. 7 — 300
    9. 8 — 297
    10. 9 — 299

В случае датасетов GTZAN [16] [17] и Audio MNIST [18] [19] нет явного дисбаланса классов, поэтому в задачах классификации с этими датасетами использовалась простая в интерпретации метрика качества — процент верно классифицированных объектов. В случае датасета HeartBeatSounds [14] [15] присутствует дисбаланс классов, поэтому дополнительно к указанной выше метрике использовалась учитывающая дисбаланс классов метрика качества — сбалансированная точность. В данной работе использовались модели нейронных сетей resnet18 [20], resnet50 [20] и алгоритм оптимизации Adam [21]. В рамках экспериментов нейронная сеть обучается 100 эпох. Функция потерь — кросс-энтропия.

В данной работе значения параметров для всех типов аугментаций, где используются эти параметры, считаем равными:

- $F = \lfloor 0.2 \cdot \text{FreqSize} \rfloor$
- $T = \lfloor 0.2 \cdot \text{TimeSize} \rfloor$
- $\text{max\_shift} = \lfloor 0.2 \cdot \text{TimeSize} \rfloor$

- $N = 4$

Датасеты разбиваются на train_valid и test в отношении 4 : 1. train_valid, в свою очередь, разбивается на train и valid в том же отношении. Обучение происходит на выборке train. После обучения берется лучший по метрике результат на валидационной выборке valid и считается метрика на тестовой выборке test. Именно по метрике качества на тестовой выборке оценивается эффективность методов аугментации.

Датасеты разбиваются на train, valid и test при 5 разных фиксированных random_seed. Результаты, соответственно, усредняются. В процессе обучения аугментация применяется с вероятностью $\frac{1}{2}$ к каждому сэмплу в каждом батче.

В качестве методов аугментации для исследования применимости предлагаемого алгоритма был выбран набор из 5 методов: TimeMasking[1] [6], FreqMasking[2] [6], Noise[3] [5], TimeShift[4] [5], SwapVerticalStripes[5]. В рамках экспериментов проводится сравнение предлагаемого алгоритма с RandAugment [9].

4.1 Результаты экспериментов

Результаты экспериментов представлены в таблицах ниже.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	81.98 ± 2.34	82.23 ± 2.4
SwapVerticalStripes	**83.2 ± 1.3**	**83.65 ± 1.07**
SwapNeighboringStripes	81.62 ± 0.69	**83.4 ± 1.71**
SwapSeveralStripes	**83.55 ± 0.49**	**84.42 ± 1.92**

Таблица 1: Результаты экспериментов (Heartbeat Sounds [14] [15]) с предлагаемыми методами аугментации SwapVerticalStripes, SwapNeighboringStripes, SwapSeveralStripes. Метрика качества — процент верно классифицированных объектов.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	0.66 ± 0.034	0.692 ± 0.04
SwapVerticalStripes	**0.699 ± 0.029**	0.681 ± 0.038
SwapNeighboringStripes	**0.69 ± 0.029**	0.7 ± 0.027
SwapSeveralStripes	**0.687 ± 0.026**	**0.709 ± 0.029**

Таблица 2: Результаты экспериментов (Heartbeat Sounds [14] [15]) с предлагаемыми методами аугментации SwapVerticalStripes, SwapNeighboringStripes, SwapSeveralStripes. Метрика качества — сбалансированная точность.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	74.3 ± 3.03	73.0 ± 3.24
SwapVerticalStripes	**76.6 ± 2.67**	**75.6 ± 3.68**
SwapNeighboringStripes	**75.6 ± 2.75**	71.4 ± 4.91
SwapSeveralStripes	**75.4 ± 2.18**	72.7 ± 3.4

Таблица 3: Результаты экспериментов (GTZAN [16] [17]) с предлагаемыми методами аугментации SwapVerticalStripes, SwapNeighboringStripes, SwapSeveralStripes. Метрика качества — процент верно классифицированных объектов.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	95.66 ± 0.81	94.49 ± 0.42
SwapVerticalStripes	95.42 ± 0.88	95.46 ± 1.05
SwapNeighboringStripes	95.63 ± 0.85	94.53 ± 0.4
SwapSeveralStripes	95.7 ± 0.52	94.39 ± 0.99

Таблица 4: Результаты экспериментов (Audio MNIST [18] [19]) с предлагаемыми методами аугментации SwapVerticalStripes, SwapNeighboringStripes, SwapSeveralStripes. Метрика качества — процент верно классифицированных объектов.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	81.98 ± 2.34	82.23 ± 2.4
RandAugment [9]	83.1 ± 0.92	84.57 ± 1.3
Предлагаемый алгоритм	**86.65 ± 0.67**	**86.75 ± 0.76**

Таблица 5: Результаты экспериментов (Heartbeat Sounds [14] [15]) с предлагаемым алгоритмом применения методов аугментации. Метрика качества — процент верно классифицированных объектов.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	0.66 ± 0.034	0.692 ± 0.04
RandAugment [9]	0.713 ± 0.031	0.677 ± 0.036
Предлагаемый алгоритм	**0.762 ± 0.023**	**0.753 ± 0.02**

Таблица 6: Результаты экспериментов (Heartbeat Sounds [14] [15]) с предлагаемым алгоритмом применения методов аугментации. Метрика качества — сбалансированная точность.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	74.3 ± 3.03	73.0 ± 3.24
RandAugment [9]	75.0 ± 2.61	74.9 ± 2.63
Предлагаемый алгоритм	**76.8 ± 1.75**	72.2 ± 2.8

Таблица 7: Результаты экспериментов (GTZAN [16] [17]) с предлагаемым алгоритмом применения методов аугментации. Метрика качества — процент верно классифицированных объектов.

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	95.66 ± 0.81	94.49 ± 0.42
RandAugment [9]	95.8 ± 0.67	95.49 ± 0.77
Предлагаемый алгоритм	96.04 ± 0.76	94.84 ± 1.43

Таблица 8: Результаты экспериментов (Audio MNIST [18] [19]) с предлагаемым алгоритмом применения методов аугментации. Метрика качества — процент верно классифицированных объектов.

4.2 Анализ полученных результатов

Результаты экспериментов показывают:

- В случае датасета Audio MNIST [18] [19] с помощью предлагаемых методов SwapVerticalStripes, SwapNeighboringStripes, SwapSeveralStripes и предлагаемого алгоритма применения методов аугментации не удалось получить улучшения в качестве. Стоит отметить, что это может быть связано с особенностью данных или с тем, что и без использования аугментации удается достичь хорошего качества

- Использование предлагаемого метода SwapVerticalStripes позволило получить прирост в качестве в задачах аудиоклассификации Heartbeat Sounds Classification [14] [15] (за исключением случая использования resnet50 в качестве нейронной сети и сбалансированной точности в качестве метрики качества) и GTZAN Classification [16] [17]

- Использование предлагаемого метода SwapSeveralStripes позволило получить прирост в качестве в задаче аудиоклассификации Heartbeat Sounds Classification [14] [15], а также в задаче аудиоклассификации GTZAN Classification [16] [17] при использовании resnet18 в качестве нейронной сети

- Предлагаемый метод SwapNeighboringStripes показал менее стабильные результаты, чем SwapVerticalStripes и SwapSeveralStripes, однако с его помощью в некоторых случаях можно получить прирост в качестве

- В задаче аудиоклассификации Heartbeat Sounds Classification [14] [15] показано существенное преимущество предлагаемого алгоритма применения методов аугментации над RandAugment [9]

- В случае датасета GTZAN [16] [17] предлагаемый алгоритм позволил получить прирост в качестве относительно RandAugment [9] при использовании модели нейронной сети resnet18, однако в случае resnet50 наблюдается снижение качества не только по сравнению с RandAugment [9], но и по сравнению с тем случаем, когда обучение нейронной сети происходит без аугментации

5 Заключение

В процессе выполнения работы получены следующие результаты:

- Предложен и реализован метод аугментации аудиоданных SwapVerticalStripes, основанный на перестановке вертикальных полос в мел-спектрограмме, а также его модификации SwapNeighboringStripes, SwapSeveralStripes
- Проведены вычислительные эксперименты, показывающие возможную применимость предложенного метода SwapVerticalStripes и его модификаций в задаче аудиоклассификации
- Предложен и реализован алгоритм применения методов аугментации аудиоданных с выбором конкретного метода аугментации после каждой эпохи обучения
- Проведены вычислительные, показывающие преимущество предложенного алгоритма над алгоритмом RandAugment [9] в задаче аудиоклассификации Heartbeat Sounds Classification [14] [15]

Список литературы

[1] *Harris F.* On the Use of Windows for Harmonic Analysis With the Discrete Fourier Transform // *In Proceedings of the IEEE, Jan. 1978, Vol. 66, Num. 1, 51-83.*

[2] `https://librosa.org/doc/main/generated/librosa.filters.mel.html`

[3] *Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, Yi Yang.* Random Erasing Data Augmentation // *arXiv preprint arXiv:1708.04896.* - 2017.

[4] *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 // *arXiv preprint arXiv:2008.05175.* - 2020.

[5] *Steffen Illium, Robert Muller, Andreas Sedlmeier and Claudia Linnhoff-Popien.* Surgical Mask Detection with Convolutional Neural Networks and Data Augmentations on Spectrograms // *arXiv preprint arXiv:2008.04590.* - 2020.

[6] *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 // *arXiv preprint arXiv:1904.08779.* - 2019.

[7] *Arjit Jain, Pranay Reddy Samala, Deepak Mittal, Preethi Jyoti, Maneesh Singh.* SpliceOut: A Simple and Efficient Audio Augmentation Method // *arXiv preprint arXiv:2110.00046.* - 2021.

[8] *Helin Wang, Yuexian Zou, Wenwu Wang.* SpecAugment++: A Hidden Space Data Augmentation Method for Acoustic Scene Classification // *arXiv preprint arXiv:2103.16858.* - 2021.

[9] *Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le.* RandAugment: Practical automated data augmentation with a reduced search space // *arXiv preprint arXiv:1909.13719.* - 2019.

[10] *Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, Quoc V. Le.* AutoAugment: Learning Augmentation Policies from Data // *arXiv preprint arXiv:1805.09501.* - 2018.

[11] *Chengyue Gong, Tongzheng Ren, Mao Ye, Qiang Liu.* MaxUp: A Simple Way to Improve Generalization of Neural Network Training // *arXiv preprint arXiv:2002.09024.* - 2020.

[12] *Jason Wei, Kai Zou.* EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks // *arXiv preprint arXiv:1901.11196.* - 2019.

[13] *Yu Shen, Laura Zheng, Manli Shu, Weizi Li, Tom Goldstein, Ming C. Lin.* Improving Robustness of Learning-based Autonomous Steering Using Adversarial Images // *arXiv preprint arXiv:2102.13262.* - 2021.

[14] *Bentley, P. and Nordehn, G. and Coimbra, M. and Mannor, S.* The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. — 2011. `http://www.peterjbentley.com/heartchallenge/index.html`

[15] Kaggle-датасет Heartbeat Sounds

`https://www.kaggle.com/kinguistics/heartbeat-sounds`

[16] *G. Tzanetakis and P. Cook. Musical genre classification of audio signals.* // IEEE Transactions on Speech and Audio Processing. — 2002.

[17] GTZAN Dataset — Music Genre Classification

`https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification`

[18] *Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, Wojciech Samek.* Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals // *arXiv preprint arXiv:1807.03418.* — 2018.

[19] Kaggle-датасет Audio MNIST

`https://www.kaggle.com/datasets/alanchn31/free-spoken-digits`

[20] *Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.* Deep Residual Learning for Image Recognition // In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[21] *Diederik P. Kingma, Jimmy Ba.* Adam: A Method for Stochastic Optimization // In the 3rd International Conference for Learning Representations, San Diego, 2015.