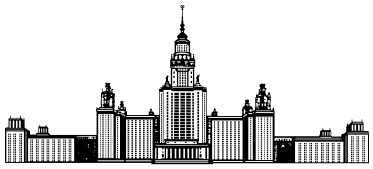

Московский государственный университет имени М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

Лукьянов Павел Александрович

Методы аугментации аудиоданных

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

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

Москва, 2022

Содержание

1 Введение **3**

2 Существующие методы аугментации **4**

3 Предлагаемые подходы **6**
3.1 SwapVerticalStripes 6
3.2 Алгоритм применения методов аугментации с выбором конкретного метода аугментации после каждой эпохи 7

4 Вычислительные эксперименты **8**
4.1 Результаты экспериментов 8
4.2 Анализ полученных результатов 9

5 Заключение **9**

Литература **10**

Аннотация

В данной работе предлагается метод аугментации аудиоданных SwapVerticalStripes, основанный на перестановке вертикальных полос в мел-спектрограмме. 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-спектрограммы представляют собой изображения, поэтому многие подходы к аугментации картинок применимы и к аудиоданным. Например, метод Random Erasing [1], сводящийся к вырезанию случайных прямоугольников из изображения, может быть использован в задаче аудиоклассификации [2]. Также в задаче классификации звуковых данных применяются такие методы аугментации, как Shift Augmentation [3] — сдвиг мел-спектрограммы влево или вправо, Noise Augmentation [3] — добавление Гауссовского шума, Loudness Augmentation [3] — регулирование громкости, Speed augmentation [3] — ускорение или замедление аудиозаписи.

SpecAugment [4] — 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 [4]
2. Frequency masking [4]
3. Time masking [4]

В настоящее время известны некоторые модификации SpecAugment: SpliceOut [5], SpecAugment++ [6].

В данной работе рассматриваются методы аугментации, которые могут применяться на лету, 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 $\frac{1}{N}$ [7].

Однако возможны и другие стратегии использования методов аугментации. В работе [8] оптимальная политика применения методов аугментации ищется с помощью методов обучения с подкреплением. В работе [9] 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 столбцов в мел-спектрограмме, 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] [4]
   $t \sim U\{0, T\}, t_0 \sim U\{0, \text{TimeSize} - 1 - t\}$, 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] [4]
   $f \sim U\{0, F\}, f_0 \sim U\{0, \text{FreqSize} - 1 - f\}$, 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] [3]
   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] [3]
   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.

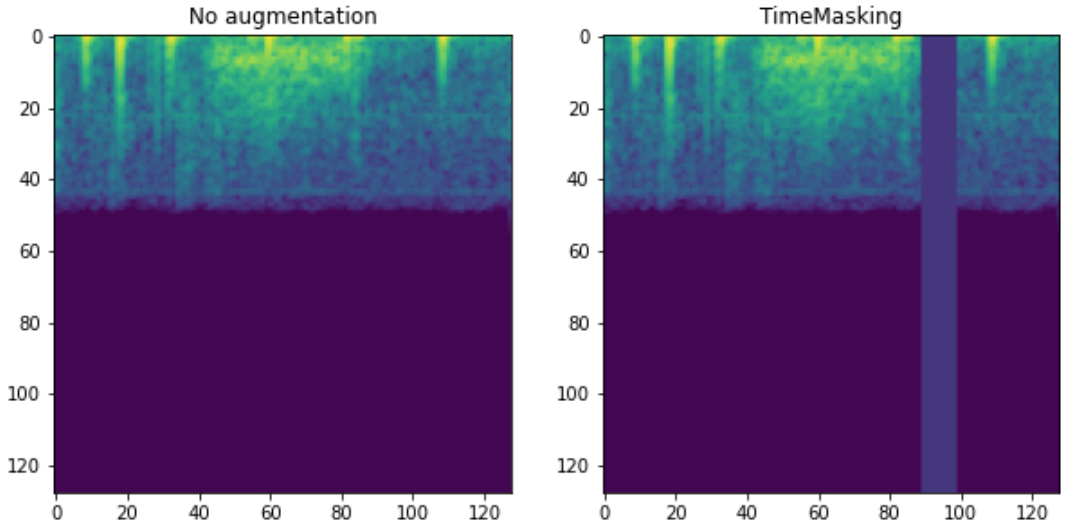

Рис. 1: TimeMasking

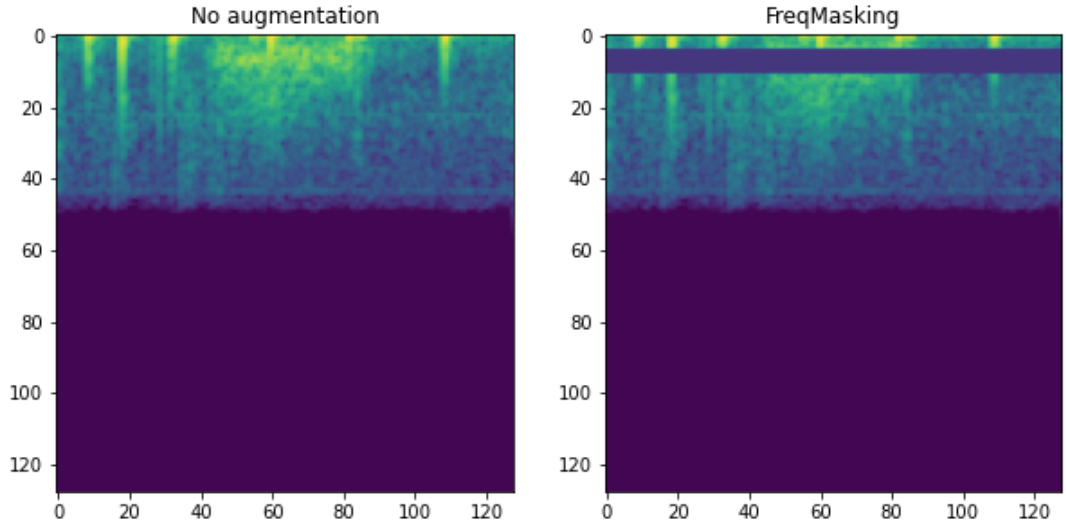

Рис. 2: FreqMasking

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.

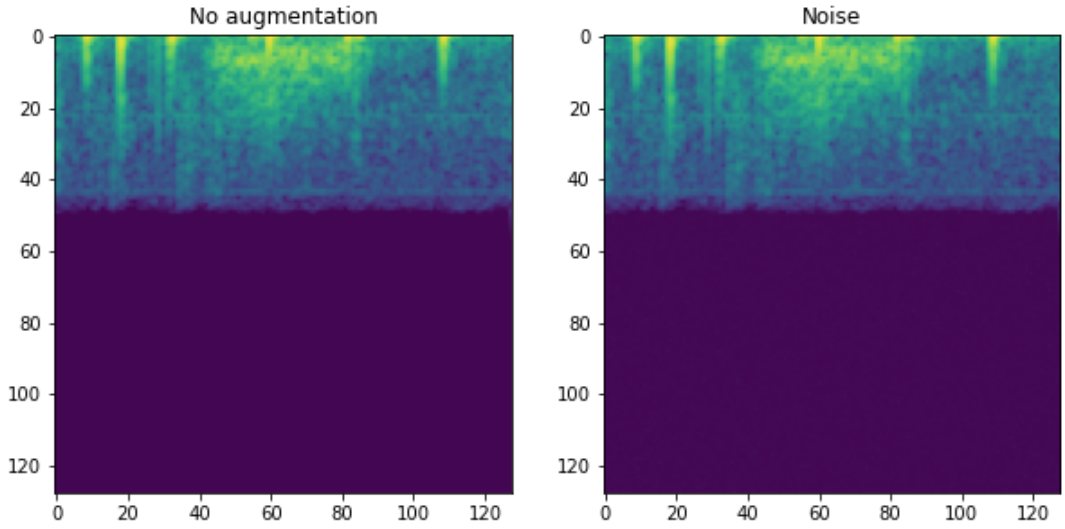

Рис. 3: Noise

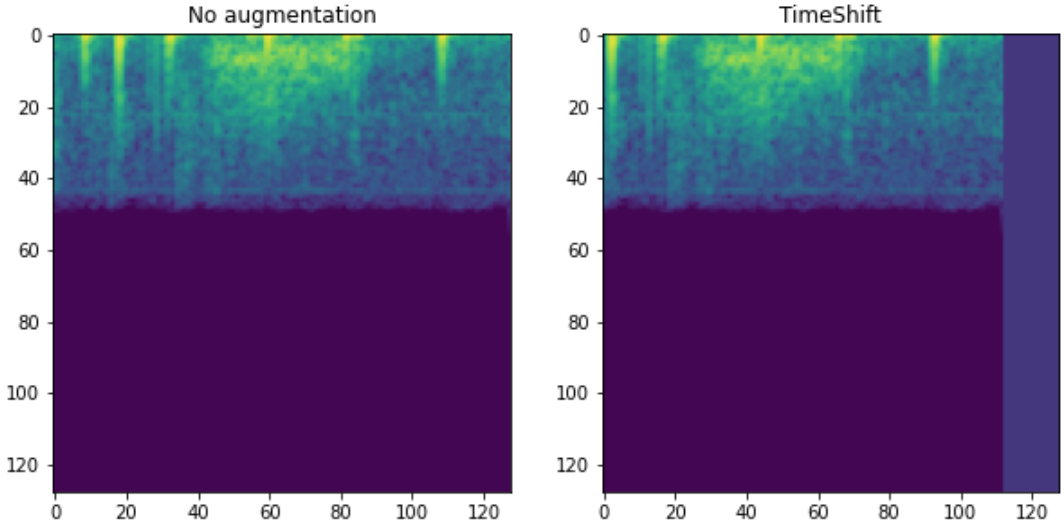

Рис. 4: TimeShift

3 Предлагаемые подходы

3.1 SwapVerticalStripes

Перестановка слов — метод аугментации, используемый в задачах, связанных с обработкой текстов [10]. 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.

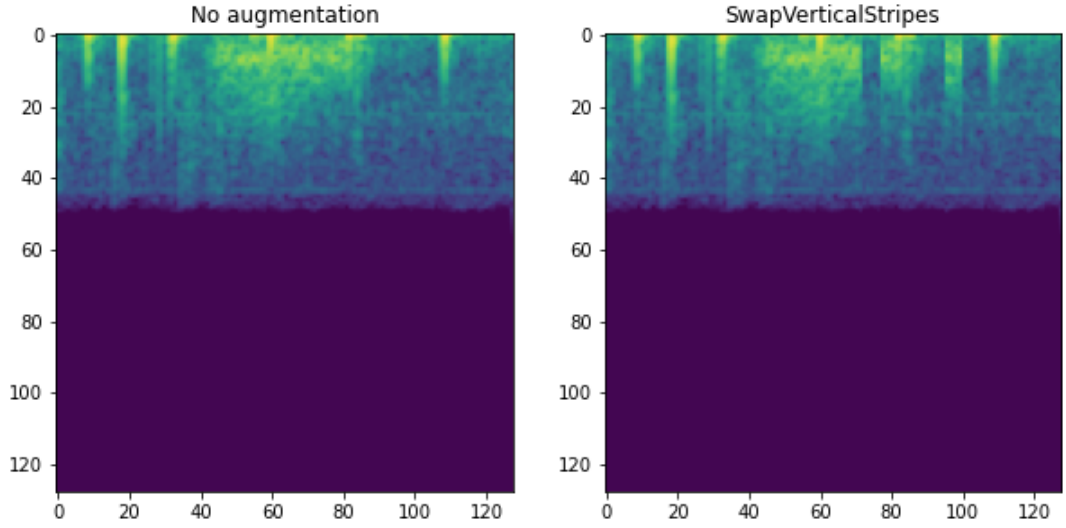

Рис. 5: SwapVerticalStripes

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

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

В данной работе предлагается алгоритм применения методов аугментации:

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 $j = 0$ **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

Идея предлагаемого метода 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 [11] подобная идея используется для нахождения "худшего" с точки зрения метрика качества на валидационной выборке параметра аугментации, используемого в процессе обучения.

4 Вычислительные эксперименты

Для исследования применимости предложенных подходов в задаче классификации вычислительные эксперименты проведены с использованием двух датасетов: Heartbeat Sounds [12] [13] (звуки сердцебиения) и GTZAN [14] [15] (классификация музыкальных жанров). Метрика качества — процент верно классифицированных объектов. В рамках экспериментов использовались модели нейронных сетей resnet18 [16], resnet50 [16] и алгоритм оптимизации Adam [17].

Датасет HeartBeatSounds [12] [13] 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 [14] [15] состоит из 1000 музыкальных записей (файлов формата .wav). Задача классификации заключается в определении музыкального жанра.

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.

В рамках экспериментов нейронная сеть обучается 100 эпох. Функция потерь — кросс-энтропия.

В данной работе значения параметров F и T для всех типов аугментаций, где используются эти параметры, считаем равными $\lfloor 0.2 \cdot \text{FreqSize} \rfloor$ и $\lfloor 0.2 \cdot \text{TimeSize} \rfloor$ соответственно. Параметр max_shift считаем равным $\lfloor 0.2 \cdot \text{TimeSize} \rfloor$.

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 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] [4], FreqMasking[2] [4], Noise[3] [3], TimeShift[4] [3], SwapVerticalStripes[5]. В рамках экспериментов проводится сравнение предлагаемого алгоритма с RandAugment [7].

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

Таблица 1: Результаты экспериментов (Heartbeat Sounds [12] [13]) с предложенным методом аугментации SwapVerticalStripes[5]

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	74.3 ± 3.03	73.0 ± 3.24
SwapVerticalStripes	76.6 ± 2.67	75.6 ± 3.68

Таблица 2: Результаты экспериментов (GTZAN [14] [15]) с предлагаемым методом аугментации SwapVerticalStripes[5]

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	81.98 ± 2.34	82.23 ± 2.4
RandAugment [7]	83.1 ± 0.92	84.57 ± 1.3
Предлагаемый алгоритм	86.65 ± 0.67	86.75 ± 0.76

Таблица 3: Результаты экспериментов (Heartbeat Sounds [12] [13]) с предлагаемым алгоритмом применения методов аугментации

Метод аугментации	resnet18	resnet50
Аугментация отсутствует	74.3 ± 3.03	73.0 ± 3.24
RandAugment [7]	75.0 ± 2.61	74.9 ± 2.63
Предлагаемый алгоритм	76.8 ± 1.75	72.2 ± 2.8

Таблица 4: Результаты экспериментов (GTZAN [14] [15]) с предлагаемым алгоритмом применения методов аугментации

4.2 Анализ полученных результатов

Результаты экспериментов показывают, что использование предлагаемого метода SwapVerticalStripes позволяет получить прирост в качестве в задачах аудиоклассификации Heartbeat Sounds Classification [12] [13] и GTZAN Classification [14] [15]. Также результаты показывают преимущество предлагаемого алгоритма применения методов аугментации над RandAugment [7] в задаче аудиоклассификации Heartbeat Sounds Classification [12] [13]. В случае датасета GTZAN [14] [15] предлагаемый алгоритм позволяет получить прирост в качестве относительно RandAugment [7] при использовании модели нейронной сети resnet18, однако в случае resnet50 наблюдается снижение качества не только по сравнению с RandAugment [7], 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

- Проведены вычислительные эксперименты, показывающие возможную применимость предложенного метода SwapVerticalStripes 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 [7] в задаче аудиоклассификации Heartbeat Sounds Classification [12] [13]

Список литературы

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

[2] *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.

[3] *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.

[4] *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.

[5] *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.

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

[7] *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.

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

[9] *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.

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

[11] *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.

[12] *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`

[13] Kaggle-датасет Heartbeat Sounds

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

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

[15] GTZAN Dataset — Music Genre Classification

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

[16] *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.

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