УДК 004.891.3

ОБЗОР МЕТОДОВ КЛАССИФИКАЦИИ ЗВУКОВ ГОРОДСКОЙ СРЕДЫ

Г. М. Мкртчян, аспирант, ассистент кафедры МКиИТ МТУСИ, Москва, Россия;
orcid.org/0000-0002-5802-5513, e-mail: g.m.mkrtchyan@mtuci.ru
Н. А. Кравченко, студент МТУСИ, Москва, Россия;
orcid.org/0009-0006-8897-2331, e-mail: n.a.kravchenko@mtuci.ru

Классификация звуков городской среды – сложная задача, которая имеет некоторые общие грани и с задачей классификации изображений, и с задачей обработки естественного языка. Описываются методы подготовки аудиоданных, представлены некоторые типы архитектур глубоких нейронных сетей, используемых для классификации звуков городской среды, такие как 1DCNN, EsResNet, AST, PaSST. Обсуждаются преимущества и недостатки таких архитектур. Рассмотрены методы дистилляции и переноса знаний, используемые для повышения эффективности используемых методов. ***Целью работы*** *является сравнение результатов обучения моделей на нескольких наборах данных, включая ESC-50, UrbanSound8K и FSD50K, на основе метрик mAP и Accuracy.*

Ключевые слова*: свёрточная нейронная сеть, end-to-end, 1D-CNN, ESResNet, AST, PaSST, перенос знаний, дистилляция знаний, UrbanSound8k, ESC-50, FSD50K, классификация аудиосигналов, извлечение признаков, спектрограмма, наборы данных, оценочные метрики.*

DOI: 10.21667/1995-4565-2024-88-53-64

Введение

В контексте урбанизации растет потребность в детектировании множества городских событий для немедленного вмешательства или сбора данных в исследовательских целях. Многие из таких ситуаций могут быть определены по звуковому сопровождению, а следовательно, после решения задачи классификации звуков городской среды становится возможным проектировать системы, автоматически реагирующие на подобные события. К примеру, систему оповещения аварийных служб, систему, выявляющую стаи беспризорных животных или нарушения в режиме ведения строительных работ.

Методы классификации звуков многочисленны, однако последние несколько лет большинство применяемых алгоритмов основано на использовании глубоких нейронных сетей разных архитектур.

Для применения подобных методов классификации аудиозаписей используют подходы к подготовке аудиоданных, перечисленных в следующем разделе.

Теоретическая часть

Подготовка данных

Формально аудиозапись, закодированная импульсно-кодовой модуляцией [1] длиной n_i секунд, записанная с частотой дискретизации f_i, является вектором длиной $m_i = n_i f_i$.

Большинство классификаторов требует фиксированной длины вектора входных признаков M, и для использования таких моделей на аудиозаписях разной длительности существует два подхода.

1. Векторы больше M обрезаются до размера M. Векторы меньше M дополняются нулями до размера M.

2. Аудиосигнал разделяется на несколько кадров фиксированной длины с использованием скользящего окна размером M, классификатор применяется на каждый кадр, результаты объединяются, чаще всего суммируются.

Для понижения размерности, чаще всего по этому вектору, с помощью оконного преобразования Фурье [2] рассчитывается спектрограмма, применив к которой треугольные банки фильтров шкалы Мел [3], можно получить Мел-спектрограмму. Эту Мел-спектрограмму используют как входной двумерный тензор признаков для модели классификатора.

Другой стандартный подход – из Мел-спектрограммы с помощью дискретного косинусного преобразования рассчитать Мел-кепстральные коэффициенты и уже их передавать в модель классификатор [4].

Нейросетевые методы классификации звуков окружающей среды

End-to-end 1DCNN [5] (1D Convolution Neural Network, одномерная свёрточная нейронная сеть без предварительного процесса выделения признаков) – метод, основанный на одномерной свёрточной нейронной сети, для классификации аудио. Для получения фиксированной длины аудио используется подход кадрирования аудиосигнала. Каждый кадр в отличие от большинства методов классификации звуков подается в нейронную сеть напрямую:

$$\mathrm{T} = \mathrm{G}(\mathrm{X} \mid \Theta) = g_L(\ldots \mathrm{g}_2(\mathrm{g}_1(\mathrm{X} \mid \Theta) \mid \Theta) \mid \Theta_L), \quad (1)$$

где g_i – функция, обозначающий применение i-го слоя нейронной сети с параметром Θ_i; L – количество слоев в сети; G – композиция всех функций g_i; Θ – набор всех параметров Θ_i. Для свёрточных слоев работа i-го слоя может быть выражена следующим образом:

$$T_i = g_i(X_i \mid \Theta_i) = h(W \otimes X_i + b), \Theta_i = [W, b], \quad (2)$$

где $\otimes$ обозначает операцию свертки; X_i – двумерная входная матрица из N наборов признаков; W – набор из N одномерных ядер, используемых для извлечения нового набора признаков из входного вектора; b – вектор смещения; $h(x)$ – функция активации. Размеры X_i, W, T_i равны $(\mathrm{N}, \mathrm{d}), (\mathrm{N}, \mathrm{m})$ и $(\mathrm{N}, \mathrm{d} - \mathrm{m} + 1)$ соответственно. Несколько объединяющих слоев также применяется между свёрточными слоями для увеличения площади, покрываемой ядрами следующих слоев.

Выходной тензор финального свёрточного слоя конкатенируется в одномерный вектор и используются в качестве входных данных нескольких полносвязных слоев, которые можно описать следующим образом:

$$T_i = g_L(X_L \mid \Theta_L) = h(WX_L + b), \Theta_L = [W, b], \quad (3)$$

Для многоклассовой классификации количество нейронов выходного слоя равно количеству классов K. После применения функции Softmax, используемой в качестве функции активации к каждому элементу вектора T_L, выходной сигнал нейрона t_i можно интерпретировать как вероятность принадлежности кадра к классу i:

$$Soft\max(t_{L,i}) = \frac{e^{t_{L,i}}}{\sum_{j=1}^{K} e^{t_{L,i}}}, \quad (4)$$

В процессе обучения параметры сети корректируются в соответствии с обратным распространением ошибки для минимизации функции потерь. Интересной особенностью этой модели является то, что при инициализации весов первого слоя может использоваться банк фильтров Gammatone. Gammatone фильтр – это линейный фильтр, имеющий импульсную характеристику, описываемую как синусоидальный сигнал, модулированный по амплитуде гамма-распределением.

Одна из конфигураций сети 1DCNN представлена на рисунке 1. В верхней части схемы описана размерность признаков на каждом слое нейронной сети, в нижней части – применяемые слои, где C – свёрточный слой, P – объединяющий слой, F – полносвязный слой.

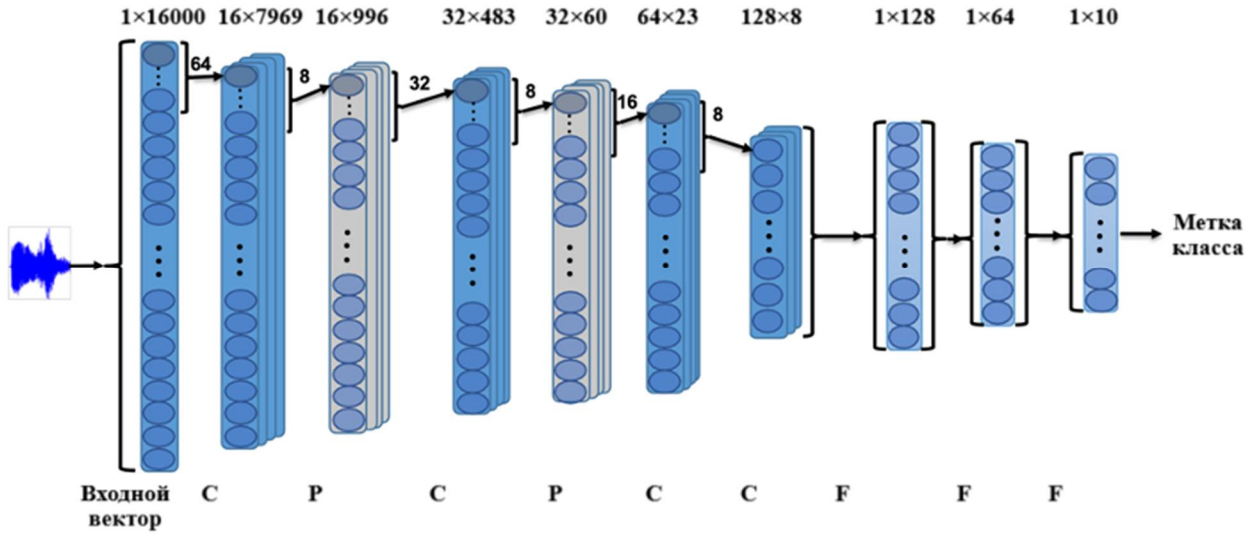

Рисунок 1 – Архитектура 1DCNN [5]
Figure 1 – 1DCNN [5] architecture

Преимущества подобных моделей – малое число параметров и простота реализации. В момент обработки аудиозаписи нет необходимости рассчитывать Мел-спектрограммы или более затратные по количеству вычислений признаки. И, если задача заключается в классификации звуков, которые возможно отличить, используя лишь временные паттерны, подобные модели могут быть применены, однако их точность сильно уступает моделям, использующим другие архитектуры. Также использование идеи объединения результатов классификации сильно зависит от набора данных. Если аудиозаписи не состоят целиком из звуков, которые необходимо классифицировать, деление на кадры может привести к ухудшению качества набора данных – кадры, содержащие тишину или посторонние звуки, будут размечены как принадлежащие к другому классу.

Несмотря на эти ограничения, данный подход применялся во многих исследованиях, например в [6], подобную модель с добавлением остаточных соединений используют для решения задачи классификации пола и регрессионной оценки роста и возраста говорящего по аудиозаписи. В [7] подобную модель используют в системе, определяющей виды птиц, система состоит из постоянно работающей легковесной модели 1DCNN, которая детектирует пение птиц, и другой модели, требующей расчета спектрограмм, классифицирующей уже вид птицы. Такой подход позволяет снизить количество вычислений, а следовательно, и энергозатратность.

ESResNet[8] (Environmental Sound Residual neural Network, остаточная нейронная сеть для звуков окружающей среды) – архитектура остаточной свёрточной нейронной сети для классификации звуков. Предобработка аудиозаписи x для данной сети заключается в:

1) В приведении к фиксированному размеру t первым подходом.

2) В расчете логарифмической спектрограммы S с помощью оконного преобразования Фурье:

$$S = 10Lg\,|\,X(\tau,\omega)\,|^2 \tag{5}$$

$$X(\tau,\omega) = \sum_{n=-\infty}^{+\infty} x(n)\cdot w(n-\tau)\cdot e^{-j\omega n}, \tag{6}$$

где w – оконная функция Блэкмана-Харриса:

$$w(k) = a_0 - a_1\cos(\frac{2\pi k}{N}) + a_2\cos(\frac{4\pi k}{N}) - a_3\cos(\frac{6\pi k}{N}), \tag{7}$$

с размером окна 37,5 мс и сдвигом 12,7 мс. В результате тензор спектрограммы имеет размерность $(\mathrm{F}_{disk} \times \lfloor t / 0{,}0127 \rfloor$ где F_{disc} – частотное разрешение.

В целях использования архитектуры ResNet спектрограмма разделяется на 3 частотные полосы, которые интерпретируются как каналы изображения, после чего входной тензор имеет форму $(3 \times \lfloor F_{disk} / 3 \rfloor \times \lfloor t / 0.0127 \rfloor)$.

Сама ResNet призвана решить проблему затухающих градиентов путем добавления соединений быстрого доступа между свёрточными слоями [9]. Рассмотрим функцию $G(x)$, которая представляет собой последовательность свёрточных слоев, примененных к тензору x. Тогда блок с соединением быстрого доступа определяется как функция $H(x)$:

$$H(x) = h\left(BN_{y_2,\beta_2}\left(G_2\left(BN_{y_1,\beta_1}(\mathrm{G}_1(BN_{y_0,\beta_0}\left(G_0(x)\right)))\right)+x\right)\right), \tag{8}$$

где h – функция активации; $BN_{\gamma,\beta}$ – слой пакетной нормализации:

$$BN_{\gamma,\beta}(x_i) = \gamma \cdot \left(\frac{x_i - \mu_\beta}{\sqrt{\sigma_B^2 + \in}}\right) + \beta, \tag{9}$$

где μ_B – математическое ожидание пакета; σ_B^2 – дисперсия пакета; γ и β – обучаемые параметры для каждой размерности; ϵ – необучаемый параметр, в стандартной реализации принимающий значение $1e^{-5}$.

Параллельно некоторым блокам ResNet добавлены блоки внимания, которые состоят из последовательных слоев объединения (по максимуму), сверток и пакетной нормализации.

$$H'(x) = H_2\left(H_1\left(H_0(x)\right)\right) \odot h\left(BN\left(G_1'\left(G_0'\left(Pool(x)\right)\right)\right)\right), \tag{10}$$

где $\odot$ – поэлементное умножение; $Pool(x)$ – объединяющий слой.

Архитектура сети EsResNet представлена на рисунке 2. Черным обозначен полносвязный слой, жёлтым, голубым и фиолетовым обозначены блоки разных типов, состоящие из слоев свертки (красные), пакетной нормализации (зеленые) и объединяющих слоев (серые). Для блоков указано количество матриц признаков при входе и выходе из него, для блоков свертки указан также размер ядер.

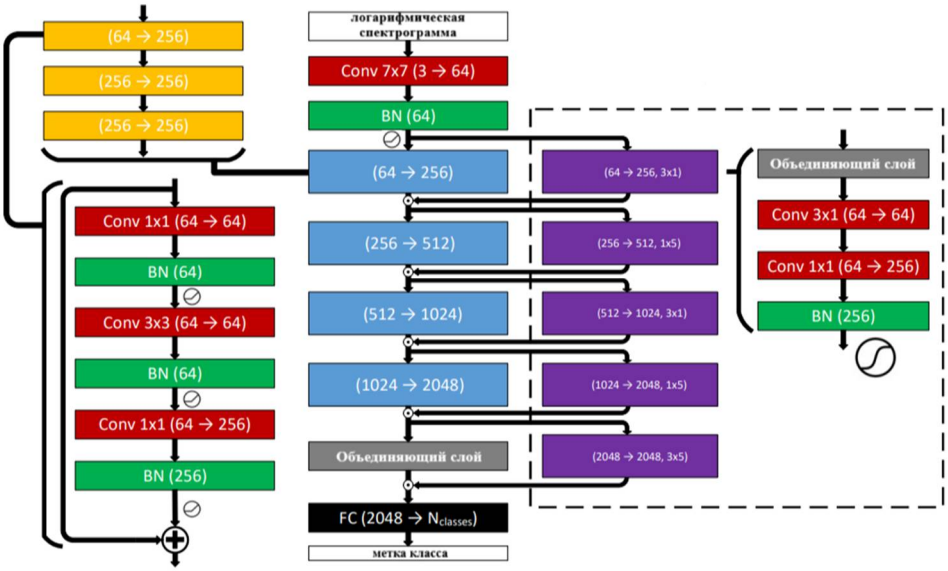

Рисунок 2 – Архитектура EsResNet [8]
Figure 2 – EsResNet architecture[8]

Благодаря сходству большой части архитектуры с оригинальной ResNet, возможно инициализировать веса совпадающих блоков, используя веса предобученной модели ResNet на наборе данных для классификации изображений ImageNet [10]. Этот перенос знаний позволяет повысить точность модели на несколько процентов.

Для обработки стереоаудиозаписей сеть до последнего полносвязного слоя применяется на каждом канале, два вектора размером 2048 суммируются и подаются в полносвязный слой. Такая обработка 2-канальной аудиозаписи вместо 1-канальной улучшает метрику качества.

ESResNet использовали в ряде исследований, в [11], заменив преобразование Фурье (6) на сплайн-вейвлет и заменив базовую архитектуру ResNet на ResNeXt [12] (ResNet neXt, остаточная сеть нового поколения), добились улучшения точности в задаче классификации звуков городской среды. В [13] используют EsResNet как основу для модели, обученной без учителя. В [14] описывают метод улучшения качества модели путем обучения ESResNet с дополнительным параллельным блоком многослойного перцептрона и гибридной функцией потерь.

AST [15] (Audio Spectrogram Transformer, трансформер для спектрограмм аудио данных) – архитектура нейронной сети для классификации звуков, основанная на модели ViT [16] (Vision Transformer, трансформер для визуальных данных). Входным тензором является Мел-спектрограмма, для получения которой используется окно Хемминга размером 25 мс с шагом 10 мс и 128 Мел фильтров. Соответственно при длительности аудиозаписи t секунд форма входного тензора равна ($128 \times 100t$).

Этот входной тензор делится с помощью двумерного скользящего окна на патчи размером (16×16) с шагом 10 (перекрытием 6 в оба измерения). Следовательно, количество патчей $N = 12\lceil (100t - 16)/10 \rceil$.

Каждый патч с помощью линейной проекции преобразуется в одномерный вектор E_d, называемый эмбеддингом патча. После к каждому патчу в соответствии с его позицией добавляется вектор, называемый эмбеддингом позиции [17]. Эмбеддинги позиций в данной архитектуре являются абсолютными и обучаемыми, таким образом существует матрица, P_{N+d} которая корректируется в процессе обучения, и итоговый эмбеддинг патча с позицией i равен

$$E_i' = E_i + P_i. \tag{11}$$

Фактически, это строго ограничивает длительность аудиозаписей, так как в момент обучения сети необходимо знать максимальное количество патчей. Но, применив билинейную интерполяцию по оси времени к матрице P, можно адаптировать уже обученную модель к иному количеству патчей.

Далее, эмбеддинги патчей и эмбеддинг специального токена подаются в последовательные слои энкодеров из архитектуры трансформера [18]. Выход, соответствующий специальному токену, подается в последовательные полносвязные слои, иначе – многослойный перцептрон, MLP, аналогично тому, как это показано в формуле (3).

Архитектура AST представлена на рисунке 3.

Использование стандартной архитектуры энкодеров трансформера, полностью совпадающей с одним из визуальных трансформеров, позволяет воспользоваться методикой передачи знаний и инициализировать веса AST, используя веса обученного на задаче классификации изображений визуального трансформера. Это позволяет уменьшить размер набора данных, необходимого для обучения AST. Преимущественно AST превосходят свёрточные нейронные сети в метриках качества, однако уступают в производительности – сложность вычислений и количество необходимой памяти растут квадратично с увеличением количества патчей. AST является основой других моделей [19, 20]. В [21] заменили блоки внимания в энкодере на блоки Фурье и добились повышения качества и уменьшения размера модели в задаче классификации медицинских звуков (кашель, насморк, ...).

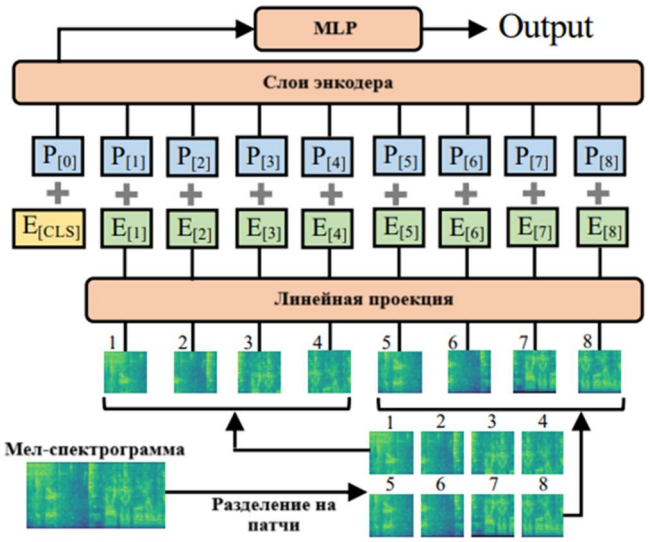

Рисунок 3 – Архитектура AST[15]
Figure 3 – AST architecture[15]

PaSST [19] (Patchout faSt Spectrogram Transformer, быстрый трансформер для спектрограмм с исключением патчей) – архитектура нейронной сети для классификации звуков, основанная на модели AST. Эмбеддинги позиций разделены на эмбеддинги времени и частоты – каждый патч имеет номер по оси времени i от 0 до N^T и номер по оси частоты j от 0 до N^F, итоговый эмбеддинг патча $E^i_{i,j}$ вместо (11) равен:

$$E'_{i,j} = E_{i,j} + T_i + F_j, \tag{12}$$

где T, F – обучаемые матрицы размером $(\mathrm{N}^F \times d)$ и $(\mathrm{N}^T \times d)$ соответственно. При увеличении длительности аудиозаписи меняется только N^T, а следовательно, достаточно интерполировать только матрицу T.

В целях оптимизации количества патчей применяется их структурированное выпадение: случайным образом выбирается несколько строк и столбцов, и все патчи, принадлежащие им, не попадают в трансформер. Исключение половины патчей не только увеличивает скорость работы и уменьшает затраты памяти в 4 раза, но и аналогично механизму исключения нейронов [22] улучшает метрики качества. Архитектура PaSST с выпадением патчей и разделением эмбеддингов изображена на рисунке 4.

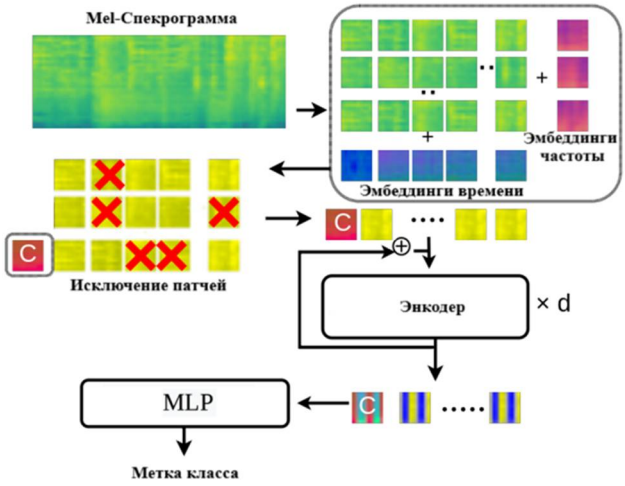

Рисунок 4 – Визуализация выпадения патчей в архитектуре PaSST [19]
Figure 4 – Visualization of patch dropout in the PaSST [19] architecture

Также в процессе обучения применяется множество аугментаций аудиоданных:

– смешивание [23] аудиозаписей и смешивание спектрограмм;

– маскировка [24] частотных и временных интервалов;

– смещение аудиосигнала по оси времени;

– случайное изменение громкости: изменение амплитуды сигнала на -7..+7 дБ.

При использовании архитектуры PaSST создана [25] модель для классификации акустической сцены по звуку. В [26] PaSST изучают как модель выделения эмбеддингов, на основе которой уже строят классификатор стилей музыки. В [27] подробно сравнивают использование эмбеддингов, генерируемых PaSST, на разных уровнях энкодера. С использованием PaSST в качестве учителя созданы модели [28] и [30].

mn_40as [28] (Mobile Network with width 4.0 trained on AudioSet, нейронная сеть для мобильных устройств с увеличенным в 4 раза числом параметров, обученная на наборе данных AudioSet) – пример модели, использующей подход Transformer-to-CNN [29] (дистилляция знаний из трансформера в свёрточную нейронную сеть) в контексте классификации звуков, позволяющий добиться точности моделей с большим числом параметров, основанных на AST, используя модели, основанные на двумерных свёрточных нейронных сетях. Для получения фиксированной длины аудио аудиозапись дополняется нулями, выделение признаков состоит в расчете Мел-спектрограмм аудиозаписи. Дальше эта спектрограмма рассматривается как изображение и подается в модель с архитектурой MobileNetV3-Large [31] (огромная нейронная сеть для мобильных устройств, версия 3).

Для повышения точности малой модели-ученика используется метод дистилляции знаний из модели-учителя архитектуры PaSST. Для этого, помимо обычных потерь при классификации, модель-ученик сводит к минимуму потери при дистилляции, основанные на прогнозах модели-учителя. Итоговую функцию потерь можно сформулировать как взвешенную сумму потерь классификации L_i и потерь дистилляции L_{kd}:

$$Loss = \lambda L_l(\delta(z_S), y) + (1-\lambda)L_{kd}(\delta(z_S), \delta(z_T / \tau)), \quad (13)$$

где z_S, z_T – выходные векторы классификатора модели-ученика и модели-учителя; y – действительные метки; δ – функция активации; λ – весовой коэффициент; τ – температура обучения.

В данном подходе потери классификации и дистилляции рассчитывали как перекрёстную энтропию:

$$L(p, y) = -\sum_{i=1}^{C} y_i \log(p_i). \quad (14)$$

Такой метод позволяет улучшить метрику качества и увеличить производительность, как показано в [28].

Подробно подход дистилляции знаний между моделями разных архитектур и в разных направлениях изучается в [32]. В [33] описывается способ совмещения офлайн (запуск учителя и ученика происходит отдельно) дистилляции знаний и аугментаций на больших наборах данных за счет оптимизации необходимой для хранения информации.

Экспериментальные исследования

Данные для обучения моделей классификации звуков окружающей среды

Существует множество наборов данных, предназначенных для обучения и оценки моделей классификации звуков окружающей среды. В исследовательских целях мы выбрали 3 открытых, доступных набора данных – ESC-50, UrbanSound8K, FSD50K.

ESC-50 [34] (Environmental Sound Classification – 50, классификация звуков окружающей среды на 50 классов) – набор данных, состоящий из аудиозаписей продолжительностью 5 секунд, организованных в 50 классов (по 40 файлов в каждом классе), каждый класс принадлежит одной из 5 категорий.

UrbanSound8K [35] (8 тысяч звуков города) – представляет собой набор аудиоданных, содержащий 8732 помеченных звуковых фрагмента длительностью до 4 секунд звуков городской среды, разделенных на 10 классов.

FSD50K [36] (Freesound Database 50K, база данных бесплатных звуков) – это открытый набор из 51197 звуковых файлов длительностью от 0,3, до 30 секунд, неравномерно распределенных по 200 классам, взятым из набора данных AudioSet [37].

Сравнение наборов данных представлено на таблице 1.

Таблица 1 – Сравнение наборов данных

Table 1 – Dataset comparison

Наборы данных	ES-50	UrbanSound8K	FSD50K
Классы	50	10	200
Формат данных	wav, PCM 16bit, 44.1 kHz mono	wav, PCM 16bit, 16-48 kHz stereo	wav, PCM 16bit, 44.1 kHz mono
Объем (шт.)	2000	8732	51197
Длительность (одного файла)	5c	<= 4c	0,3-30c
Длительность (общая)	2,7 часа	27 часов	108,3 часа
Разделение	5 секций	10 секций	Обучающая и валидационная выборка

Оценки методов классификации звуков

Оценка алгоритмов классификации важна для определения их сильных и слабых сторон, выбора моделей для решения определенной задачи и сравнения результатов разных моделей. Ниже – некоторые из используемых метрик.

Средняя доля правильных ответов при K секционной кросс валидации при классификации на N классов.

$$acc_{K-fold} = \frac{1}{K*N}\sum_{j=1}^{K}\sum_{i=1}^{N}\left[y_i = g_j(x_i)\right].$$

Средняя точность (mAP) – это широко используемый показатель для оценки производительности моделей классификаторов.

$$mAP = \left(\frac{1}{N}\right)\sum\nolimits_{i=1}^{N} AP_i,$$

где N – количество классов объектов на изображении, AP_i:

$$AP_i = \int_0^1 p_i(t)dt,$$

где $p_i(t)$ – точность классификатора для класса i с порогом t (точность – доля истинно положительных классификаций в общем числе положительных классификаций).

Сравнение методов классификации звуков на наборах данных

Приведем таблицу (таблица 2) сравнения основных методов классификации звуков на наборах данных UrbanSound8K, FSD50K и ESC-50. Данные взяты из исследований [5, 8, 15, 19, 28].

Столь низкая доля правильных ответов 1DCNN на наборе данных ESC-50 по сравнению с UrbanSound8K, полученная в результате экспериментальных исследований, может быть объяснена тем, что в наборе ESC-50 аудиозаписи не состоят целиком из звуков событий, в связи с чем подход 1DCNN при делении на кадры получает много кадров, не содержащих звуков событий, но размеченных как таковые. Также из представленного авторами статьи [5] исходного кода видно, что при обучении и проверке этого метода на наборе данных UrbanSound8K секции были сформированы случайным образом, что могло привести к утечке

данных из тестового набора в обучающий. В совокупности с фактором малого числа аудиозаписей на один класс в наборе ESC-50 метод 1DCNN дает сильно отстающие результаты по сравнению с другими подходами.

Таблица 2 – Оценки перечисленных моделей на наборах данных UrbanSound8K, ESC-50, FSD50K

Table 2 – Estimates of the listed models on Urban Sound 8 K, ESC-50, FSD50K datasets

	UrbanSound8K,$acc_{10-fold}$	ESC-50, acc_{5-fold}	FSD50K, mAP
1DCNN	0,89 [5]	0,35	-
EsResNet	0,854 [8]	0,915 [8]	-
AST	-	0,957 [15]	-
PaSST	-	0,968 [19]	0,653 [19]
MN_40s	-	0,9745 [28]	0,656 [28]

Из таблицы сравнения видно, что применение методов, основанных на трансформерах (AST, PaSST), значительно улучшило качество моделей по сравнению с методами, основанными на свёрточных нейронных сетях.

Метод дистилляции знаний позволил обучить масштабированную модель архитектуры MobileNetV3-Large и добиться улучшения точности метода классификации.

Заключение

Классификации звуков окружающей среды – это одна из ключевых задач в области автоматизации обработки аудиосигналов и машинного обучения. В обзорной статье были описаны основные архитектуры нейронных сетей, позволяющие решить эту задачу. Были рассмотрены конкретные модели, такие как 1DCNN, EsResNet, AST, PaSST, принципы их работы, а также их преимущества и недостатки. Указаны необходимые методы предобработки данных. Описаны методы переноса и дистилляции знаний.

Рассмотрены наборы данных для обучения моделей классификации звуков окружающей среды, описаны метрики и оценки, применяемые на этих наборах данных. В целом классификации звуков – это сложная задача, которая требует знания в области компьютерного зрения, машинного обучения и даже в обработке естественного языка. Однако огромное количество новых методов способствует непрерывному улучшению точности и эффективности моделей для классификации звуков.

Библиографический список

1. **Pan D.Y.** Digital audio compression // *Digital Technical Journal*. 1993, vol. 5, no. 2, pp. 28-40.

2. **Козырев М.О., Орлов М.Ю.** Оконные функции и преобразование Фурье // *Инновационные научные исследования: теория, методология, практика,* 2017. С. 21-25.

3. **Нифонтов С.В., Либеров Р.В.** Анализ техник извлечения характеристик речевых сигналов // *Электронный журнал: наука, техника и образование*. 2017. № 2. С. 107-114.

4. **Вашкевич М.И., Лихачёв Д.С., Азаров И.С.** Система анализа и классификации голосового сигнала на основе пертрубационных параметров и кепстрального представления в психоакустических шкалах // *Доклады Белорусского государственного университета информатики и радиоэлектроники*. 2022. Т. 20. № 1. С. 73-82.

5. **Abdoli S., Cardinal P., Koerich A.L.** End-to-end environmental sound classification using a 1D convolutional neural network // *Expert Systems with Applications*. 2019, vol. 136, pp. 252-263.

6. **Jaid U.H., AbdulHassan A.K.** End-to-End Speaker Profiling Using 1D CNN Architectures and Filter Bank Initialization .*International Journal of Online & Biomedical Engineering*. 2023, vol. 19, no. 10.

7. **Mun'ím Ahmad Zabidi M., Idris M.Y.I., Mahmud R.** Fowl Play: Identifying Birds by Bioacoustics and Deep Learning. 2023. Avalible at: https://www.researchgate.net/publication/371563132_Fowl_Play_Identifying_Birds_by_Bioacoustics_and_Deep_Learning

8. **Guzhov A. et al.** Esresnet: Environmental sound classification based on visual domain models . 2020 25th International Conference on Pattern Recognition (ICPR). 2021, pp. 4933-4940.

9. **Павликов А.Е., Городничев М.Г.** Обзор технологий определения положения тела человека // *Модели, системы, сети в экономике, технике, природе и обществе*. 2023, С. 81-97.

10. **Russakovsky O. et al.** Imagenet large scale visual recognition challenge. *International journal of computer vision*. 2015, vol. 115, p. 211-252.

11. **Guzhov A. et al.** Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio // 2021 International Joint Conference on Neural Networks (IJCNN). 2021, p. 1-8.

12. **Xie S. et al.** Aggregated residual transformations for deep neural networks // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, p. 1492-1500.

13. **Hornauer S. et al.** Unsupervised discriminative learning of sounds for audio event classification // *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021. p. 3035-3039.

14. **Nasiri A., Hu J.** SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification // arXiv e-prints. 2021 p. arXiv: 2103.01929.

15. **Gong Y., Chung Y.A., Glass J.** AST: Audio Spectrogram Transformer // arXiv e-prints. 2021, p. arXiv: 2104.01778.

16. **Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.** An image is worth 16x16 words: Transformers for image recognition at scale. arXiv e-prints. 2021, p. arXiv: 2010.11929

17. **Wang Y.A., Chen Y.N.** What Do Position Embeddings Learn? An Empirical Study of Pre-Trained Language Model Positional Encoding . arXiv e-prints. 2020, p. arXiv: 2010.04903.

18. **Vaswani A. et al.** Attention is all you need. Advances in neural information processing systems. 2017, vol. 30.

19. **Koutini K. et al.** Efficient Training of Audio Transformers with Patchout // arXiv e-prints. 2021. pp. arXiv: 2110.05069.

20. **Gazneli A. et al.** End-to-End Audio Strikes Back: Boosting Augmentations Towards An Efficient Audio Classification Network // arXiv e-prints. 2022. p. arXiv: 2204.11479.

21. **Hassan K.M.N., Haque M.A.** ASFNet: Audio Spectrogram Fourier Network for Efficient Medical Sound Event Detection. // TechRxiv. 2023.

22. **Srivastava N. et al.** Dropout: a simple way to prevent neural networks from overfitting // The journal of machine learning research. 2014, vol. 15, no. 1, p. 1929-1958.

23. **Zhang H. et al.** mixup: Beyond Empirical Risk Minimization // arXiv e-prints. 2017. p. arXiv: 1710.09412.

24. **Park D.S. et al.** SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition // arXiv e-prints. 2019, p. arXiv: 1904.08779.

25. **Morocutti T. et al.** Device-Robust Acoustic Scene Classification via Impulse Response Augmentation // arXiv e-prints. 2023. p. arXiv: 2305.07499.

26. **Alonso-Jiménez P., Serra X., Bogdanov D.** Efficient Supervised Training of Audio Transformers for Music Representation Learning // Ismir 2023 Hybrid Conference. 2023.

27. **Koutini K. et al.** Learning General Audio Representations With Large-Scale Training of Patchout Audio Transformers // HEAR: Holistic Evaluation of Audio Representations. 2022, pp. 65-89.

28. **Liu Y. et al.** Hint-dynamic Knowledge Distillation // arXiv e-prints. 2022. pp. arXiv: 2211.17059.

29. **Chia Y.K., Witteveen S., Andrews M.** Transformer to CNN: Label-scarce distillation for efficient text classification // arXiv e-prints. 2019.pp. arXiv: 1909.03508.

30. **Schmid F., Koutini K., Widmer G.** Dynamic Convolutional Neural Networks as Efficient Pretrained Audio Models // arXiv e-prints. 2023. pp. arXiv:2310.15648.

31. **Howard A. et al.** Searching for mobilenetv3 // Proceedings of the IEEE/CVF international conference on computer vision. 2019, pp. 1314-1324.

32. **Gong Y. et al.** CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification // arXiv e-prints. 2022, pp. arXiv: 2203.06760.

33. **Dinkel H. et al.** CED: Consistent ensemble distillation for audio tagging // arXiv e-prints. 2023, pp. arXiv: 2308.11957.

34. **Piczak K.J.** ESC: Dataset for environmental sound classification // Proceedings of the 23rd ACM international conference on Multimedia. 2015, p. 1015-1018.

35. **Salamon J., Jacoby C., Bello J.P.** A dataset and taxonomy for urban sound research // Proceedings of the 22nd ACM international conference on Multimedia. 2014, pp. 1041-1044.

36. **Fonseca E. et al**. Fsd50k: an open dataset of human-labeled sound events // *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2021, vol. 30, pp. 829-852.

37. **Gemmeke J.F. et al.** Audio set: An ontology and human-labeled dataset for audio events // 2017 *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 2017, pp. 776-780.

UDC 004.891.3

OVERVIEW OF METHODS FOR CLASSIFYING SOUNDS OF URBAN ENVIRONMENT

G. M. Mkrtchyan, postgraduate student, Assistant of the Department of MCaIT MTUCI, Moscow, Russia; orcid.org/0000-0002-5802-5513, e-mail: g.m.mkrtchyan@mtuci.ru
N. A. Kravchenko, student of MTUCI, Moscow, Russia;
orcid.org/0009-0006-8897-2331, n.a.kravchenko@mtuci.ru

Classifying the sounds of urban environment is a complex task that has some common facets both with the task of classifying images and the task of processing natural language. The article describes the methods of audio data preparation, and presents some types of deep neural network architectures used to classify sounds of urban environment such as 1DCNN, EsResNet, AST, PaSST. The advantages and disadvantages of such architectures are discussed. The methods of knowledge distillation and transfer used to increase the effectiveness of the methods used are considered. ***The aim of the work*** *is to compare the results of model training on several datasets, including ESC-50, UrbanSound8K and FSD50K, based on mAP and Accuracy metrics.*

Keywords: *convolutional neural network, end-to-end 1DCNN, ESResNet, AST, PaSST, knowledge transfer, knowledge distillation, UrbanSound8k, ESC-50, FSD50K, audio signal classification, feature extraction, spectrogram, datasets, evaluation metrics.*

DOI: 10.21667/1995-4565-2024-88-53-64

References

1. **Pan D.Y.** Digital audio compression. *Digital Technical Journal*. 1993, vol. 5, no. 2, pp. 28-40.

2. **Kozyrev M.O., Orlov M.Ju.** Okonnye funkcii i preobrazovanie Fur'e. *Innovacionnye nauchnye issledovanija: teorija, metodologija, praktika.* 2017, pp. 21-25. (in Russian).

3. **Nifontov S.V., Liberov R.V.** Analiz tehnik izvlechenija harakteristik rechevyh signalov. *Jelektronnyj zhurnal: nauka, tehnika i obrazovanie*. 2017, no 2, pp. 107-114. (in Russian).

4. **Vashkevich M.I., Lihachjov D.S., Azarov I.S.** Sistema analiza i klassifikacii golosovogo signala na osnove pertrubacionnyh parametrov i kepstral'nogo predstavlenija v psihoakusticheskih shkalah. *Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radiojelektroniki*. 2022, vol. 20, no. 1, pp. 73-82. (in Russian).

5. **Abdoli S., Cardinal P., Koerich A.L.** End-to-end environmental sound classification using a 1D convolutional neural network. *Expert Systems with Applications*. 2019, vol. 136, pp. 252-263.

6. **Jaid U.H., AbdulHassan A.K.** End-to-End Speaker Profiling Using 1D CNN Architectures and Filter Bank Initialization. *International Journal of Online & Biomedical Engineering*. 2023, vol. 19, no. 10.

7. **Mun'ím Ahmad Zabidi M., Idris M.Y. I., Mahmud R.** Fowl Play: Identifying Birds by Bioacoustics and Deep Learning. 2023. Avalible at: https://www.researchgate.net/publication/371563132_Fowl_Play_Identifying_Birds_by_Bioacoustics_and_Deep_Learning

8. **Guzhov A. et al.** Esresnet: Environmental sound classification based on visual domain models . *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 4933-4940.

9. **Pavlikov A.E., Gorodnichev M.G.** Overview of technologies for determining the position of the human body. *Models, systems, networks in economics, technology, nature and society*. 2023, pp. 81-97. (in Russian).

10. **Russakovsky O. et al.** Imagenet large scale visual recognition challenge. *International journal of computer vision.* 2015, vol. 115, pp. 211-252.

11. **Guzhov A. et al.** Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1-8.

12. **Xie S. et al.** Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492-1500.

13. **Hornauer S. et al.** Unsupervised discriminative learning of sounds for audio event classification. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, pp. 3035-3039.

14. **Nasiri A., Hu J.** SoundCLR: Contrastive Learning of Representations For Improved Environmental Sound Classification. *arXiv e-prints*. 2021, pp. arXiv: 2103.01929.

15. **Gong Y., Chung Y.A., Glass J.** AST: Audio Spectrogram Transformer. *arXiv e-prints*. 2021, pp. arXiv: 2104.01778.

16. **Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.** An image is worth 16x16 words: Transformers for image recognition at scale. arXiv e-prints. 2021, pp. arXiv: 2010.11929

17. **Wang Y.A., Chen Y.N.** What Do Position Embeddings Learn? An Empirical Study of Pre-Trained Language Model Positional Encoding. *arXiv e-prints*. 2020, pp. arXiv: 2010.04903.

18. **Vaswani A. et al.** Attention is all you need. *Advances in neural information processing systems*. 2017, vol. 30.

19. **Koutini K. et al.** Efficient Training of Audio Transformers with Patchout. *arXiv e-prints*. 2021. pp. arXiv: 2110.05069.

20. **Gazneli A. et al.** End-to-End Audio Strikes Back: Boosting Augmentations Towards An Efficient Audio Classification Network. *arXiv e-prints*. 2022, pp. arXiv: 2204.11479.

21. **Hassan K.M.N., Haque M.A.** ASFNet: Audio Spectrogram Fourier Network for Efficient Medical Sound Event Detection. *TechRxiv*. 2023.

22. **Srivastava N. et al.** Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*. 2014, vol. 15, no. 1, pp. 1929-1958.

23. **Zhang H. et al.** mixup: Beyond Empirical Risk Minimization. *arXiv e-prints*, 2017, pp. arXiv: 1710.09412.

24. **Park D.S. et al.** SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv e-prints*, 2019, pp. arXiv: 1904.08779.

25. **Morocutti T. et al.** Device-Robust Acoustic Scene Classification via Impulse Response Augmentation. *arXiv e-prints*. 2023, pp. arXiv: 2305.07499.

26. **Alonso-Jiménez P., Serra X., Bogdanov D.** Efficient Supervised Training of Audio Transformers for Music Representation Learning. *Ismir 2023 Hybrid Conference*, 2023.

27. **Koutini K. et al.** Learning General Audio Representations With Large-Scale Training of Patchout Audio Transformers. *HEAR: Holistic Evaluation of Audio Representations*. 2022, pp. 65-89.

28. **Liu Y. et al.** Hint-dynamic Knowledge Distillation. *arXiv e-prints*. 2022, pp. arXiv: 2211.17059.

29. **Chia Y.K., Witteveen S., Andrews M.** Transformer to CNN: Label-scarce distillation for efficient text classification. *arXiv e-prints*. 2019, pp. arXiv: 1909.03508.

30. **Schmid F., Koutini K., Widmer G.** Dynamic Convolutional Neural Networks as Efficient Pretrained Audio Models. *arXiv e-prints.* 2023, pp. arXiv:2310.15648.

31. **Howard A. et al.** Searching for mobilenetv3. *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1314-1324.

32. **Gong Y. et al.** CMKD: CNN/Transformer-Based Cross-Model Knowledge Distillation for Audio Classification. *arXiv e-prints*. 2022, pp. arXiv: 2203.06760.

33. **Dinkel H. et al.** CED: Consistent ensemble distillation for audio tagging. *arXiv e-prints*. 2023, pp. arXiv: 2308.11957.

34. **Piczak K.J.** ESC: Dataset for environmental sound classification. *Proceedings of the 23rd ACM international conference on Multimedia.* 2015, pp. 1015-1018.

35. **Salamon J., Jacoby C., Bello J.P.** A dataset and taxonomy for urban sound research. *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 1041-1044.

36. **Fonseca E. et al**. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 2021,vol. 30, pp. 829-852.

37. **Gemmeke J.F. et al.** Audio set: An ontology and human-labeled dataset for audio events. 2017 *IEEE international conference on acoustics, speech and signal processing (ICASSP).* 2017, pp. 776-780.