НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2024 Том 24 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2024 Vol. 24 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2024-24-5-758-769
УДК 004.93

Совместное распознавание акустических сцен и аудиособытий с помощью многозадачного обучения компактных моделей

Максим Константинович Сурков✉

Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
surkovmax007@mail.ru✉, https://orcid.org/0000-0002-3929-7484

Аннотация

Введение. Задача распознавания метаинформации заключается в выявлении и извлечении данных различной природы (речь, шумы, акустическая сцена, акустические события, аномальные звуки) из входного аудиосигнала. Существуют подходы, способные обеспечить высокую точность распознавания метаинформации различной природы в аудиозаписях. Данные модели часто опираются на глубокие нейронные сети с числом обучаемых параметров более сотни миллионов. Как следствие, такие модели невозможно использовать в реальных коммерческих системах, так как они ограничены в вычислительных ресурсах. Это влияет на работу умных устройств, таких как мобильные телефоны, умные часы, колонки, системы «умный дом». Обычно к умным устройствам предъявляются серьезные требования по энергоэффективности, что влияет на применение тех или иных компонентов в составе таких продуктов. Тактовые частоты процессоров, объемы оперативной и дисковой памяти в таких устройствах сильно ограничены и не способны работать с нейросетевыми моделями с большим числом обучаемых параметров. Подобные ограничения требуют поиска возможных решений, которые бы позволили применять технологии распознавания метаинформации в коммерческих устройствах. Возможным решением могут стать так называемые компактные нейросетевые модели, которые за счет архитектуры и многозадачных алгоритмов обучения способны распознавать метаинформацию в аудиозаписях и используют ограниченное число обучаемых параметров. Коммерческий интерес к данной задаче согласуется и с заинтересованностью научного сообщества. Так, в рамках международного конкурса под названием «Detection and Classification of Acoustic Scenes and Events» организаторами были сформулированы специальные подзадачи — распознавание акустической сцены при использовании низкоресурсных систем («Low-Complexity Acoustic Scene Classification») и детекции аудиособытий («Sound Event Detection with Weak Labels and Synthetic Soundscapes»). Важными исследовательскими вопросами являются как создание оптимальной архитектуры компактной нейронной сети, так и алгоритмов их обучения для получения низкоресурсной высокоточной системы распознавания акустических сцен и аудиособытий. **Метод.** Исследование выполнено на основе корпуса данных задач Challenge «Low-Complexity Acoustic Scene Classification» и «Sound Event Detection with Weak Labels and Synthetic Soundscapes». Предложена архитектура многозадачной нейронной сети, состоящая из общего кодировщика и двух независимых декодировщиков для каждой из двух задач. Рассмотрены классические алгоритмы многозадачного обучения SoftMTL и HardMTL, а также разработаны их модификации CrossMTL, которые опираются на идею переиспользования данных от одной задачи при обучении декодировщика решать вторую задачу, и FreezeMTL, в процессе которого обученные веса общего кодировщика замораживаются после обучения на первой задаче и используются для оптимизации второго декодировщика. **Основные результаты.** Показано, что применение модификации CrossMTL дает возможность существенно увеличить точность классификации акустических сцен и детекции аудиособытий по сравнению с классическими подходами SoftMTL и HardMTL. Алгоритм FreezeMTL позволяет получить модель, демонстрирующую точность классификации сцен в 42,44 % и детекции событий в 45,86 %, что сравнимо с показателями базовых решений задач 2023 года. **Обсуждение.** Предложена компактная нейронная сеть, состоящая из 633,5 тыс. обучаемых параметров, требующая 43,2 млн арифметических операций для обработки аудио длиной в одну секунду. Модель использует на 7,8 % меньше обучаемых параметров и на 40 % меньше арифметических операций по сравнению с наивным применением двух независимых моделей. Разработанную модель можно применить в умных устройствах за счет уменьшения числа обучаемых параметров и арифметических операций, необходимых для ее применения.

Ключевые слова
распознавание акустической сцены, детекция аудиособытий, компактные модели, многозадачные нейронные сети, многозадачное обучение, распознавание метаинформации, умные устройства, нейронные сети

Ссылка для цитирования: Сурков М.К. Совместное распознавание акустических сцен и аудиособытий с помощью многозадачного обучения компактных моделей // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 5. С. 758–769. doi: 10.17586/2226-1494-2024-24-5-758-769

Low-complexity multi task learning for joint acoustic scenes classification and sound events detection

Maxim K. Surkov✉

ITMO University, Saint Petersburg, 197101, Russian Federation

surkovmax007@mail.ru✉, https://orcid.org/0000-0002-3929-7484

Abstract
The task of automatic metainformation recognition from audio sources is to detect and extract data of various natures (speech, noises, acoustic scenes, acoustic events, anomalies) from a given audio input signal. This area is well developed and known to the scientific community and has various approaches with high quality. But, the vast majority of such methods are based on large neural networks with a huge number of weights to be trained. Subsequently, it is impractical to use them in environments with severely limited computing resources. The smart device industry is currently growing rapidly: smartphones, smart watches, voice assistants, TV, smart home. Such products have limitations in both processor and memory. At that moment, the State-of-the-Art way to cope with these conditions is to use so-called low-complexity models. Moreover, in recent years, the interest of the scientific community in the above-mentioned problem has been growing (DCASE Workshop). One of the most crucial subtasks in the global meta information recognition problem is the task of Automatic Scene Classification and the task of Sound Event Detection. The most important scientific questions are the development of both the optimal low-complexity neural network architecture and learning algorithms to obtain a low-resource, high-quality system for classifying acoustic scenes and detecting sound events. In this paper the datasets from DCASE Challenge "Low-Complexity Acoustic Scene Classification" and "Sound Event Detection with Weak Labels and Synthetic Soundscapes" were used. A multitask neural network architecture was proposed consisting of a common encoder and two independent decoders for each of the two tasks. The classical algorithms of multitask learning SoftMTL and HardMTL were considered, and their modifications were developed: CrossMTL, which is based on the idea of reusing data from one task when training the decoder to solve the second task, and FreezeMTL, in which the trained weights of the common encoder are frozen after training on the first task and used to optimize the second decoder. As a result of the experiments, it was shown that the use of the CrossMTL modification can significantly increase the accuracy of the classification of acoustic scenes and event detection in compare with classical approaches SoftMTL and HardMTL. The FreezeMTL algorithm made it possible to obtain a model that provides 42.44 % accuracy in scene classification and 45.86 % accuracy in event detection, which is comparable to the results of the baseline solutions of 2023. In this paper, a low-complexity neural network consisting of 633.5 K trainable parameters was proposed, requiring 43.2 M MACs to process one second audio. This approach uses 7.8 % fewer trainable parameters and 40 % fewer MACs compared to the naive application of two independent models. The developed model can be used in smart devices due to a small number of trainable parameters, as well as a small number of MACs required for its application.

Keywords
acoustic scene classification, sound event detection, compact models, multitask neural networks, multitask learning, meta-information recognition, smart devices, neural networks

For citation: Surkov M.K. Low-complexity multi task learning for joint acoustic scenes classification and sound events detection. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2024, vol. 24, no. 5, pp. 758–769 (in Russian). doi: 10.17586/2226-1494-2024-24-5-758-769

Введение

На сегодняшний день задача распознавания метаинформации в аудиосигнале является актуальной и вызывает большой интерес со стороны научного сообщества и коммерческих компаний. Она заключается в выявлении и извлечении информации различной природы (речь, шумы, акустическая сцена, акустические события, аномальные звуки) из поступающего входного аудиосигнала. На данный момент существуют подходы, способные обеспечить высокую точность распознавания метаинформации различной природы в аудиозаписях [1–6]. В работе [3] представлена модель Conformer, которая объединяет в себе две ключевые идеи в области глубоких нейронных сетей: сверточные слои и механизм внимания. Рассмотрено несколько конфигураций полученной архитектуры с 10, 30 и 118 млн обучаемых параметров. Ученые смогли разработать модель, способную распознавать человеческую речь с наилучшей точностью, по сравнению с другими моделями при использовании тестового корпуса данных Librispeech [7]. В [6] описано несколько ключевых идей, с помощью которых была разработана модель Whisper-AT. Было замечено, что большая нейросетевая модель Whisper [5], состоящая из более 70 млн параметров, обладает свойством инвариантности распознавания речи относительно фонового музыкального шума, а также показано, как можно применить данную модель для задачи одновременного распознавания речи и выявления акустических событий и получить при этом

высокую точность распознавания итоговой системы. Одной из важнейших работ в области распознавания метаинформации является статья под названием All-in-One Transformer [8], где исследователи обратили внимание научного сообщества на универсальность слуховой системы человека. В [8] показано, что современные модели могут обладать тем же свойством, а именно в случае обучения системы решать сразу несколько задач распознавания информации в звуке одновременно, качество на каждой из задач будет выше по сравнению с результатами моделей, которые были обучены решать одну конкретную задачу. Их подход основан на использовании общего кодировщика для задач распознавания речи и акустических событий. Также был предложен способ многозадачного обучения модели, что в итоге привело к увеличению точности всей системы. В работе [9] в качестве базовой модели использован подход на основе коннекционистской временной классификации. В [10] была представлена модель BEATs, которая смогла превзойти лучшие решения в задаче распознавания акустических событий на крупнейшем корпусе данных Audioset [11]. Исследователи предложили обучать нейросетевую модель на основе Visual Transformer [12] вместе с аудиотокенизатором в режиме самообучения с последующим дообучением на задаче распознавания метаинформации. Полученная система, состоящая из более 90 млн параметров, способна распознавать аудиособытия из более 500 классов.

Помимо стандартных задач распознавания метаинформации из звука, существует ряд альтернативных содержательных задач: автоматическая аннотация аудио [13], распознавание эмоций в человеческой речи [14], ответы на вопросы по аудиосигналу [15], анализ музыкальных нот [16] и другие задачи распознавания и анализа метаинформации в аудиосигнале. Для каждой из перечисленных задач существует свое специализированное решение. Отметим, что в 2023 г. группой ученых из Alibaba Group был предложен подход Qwen-Audio [17], позволяющий решить все вышеупомянутые задачи при помощи одной общей модели, которая состоит из кодировщика Whisper-Large-V2 [5] из 640 млн обучаемых параметров и декодировщика QwenLM [18] из 7 млрд обучаемых параметров. В [17] предложен способ многозадачного обучения, который позволил получить систему, обеспечивающую более высокую точность по сравнению с лучшими аналогами, специализированными под каждую конкретную задачу.

Заметим, что каждый из описанных способов основан на использовании моделей, которые содержат десятки или даже сотни миллионов обучаемых параметров. Подобные модели неприменимы в активно развивающейся индустрии умных устройств (телефоны, часы, колонки), так как интеллектуальные устройства ограничены в вычислительных ресурсах, объеме оперативной и дисковой памяти. Данная проблема не оставила без внимания и научное сообщество, которое начало свое активное изучение данной задачи на ежегодной конференции Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, в рамках которой несколько лет назад целенаправленно были поставлены задачи распознавания акустической сцены ASC («Low-Complexity Acoustic Scene Classification») и аудиособытий SED («Sound Event Detection with Weak Labels and Synthetic Soundscapes») при наличии ограничений на число обучаемых параметров и число операций умножения и сложения.

Первая задача заключается в том, чтобы по входной аудиозаписи длиной в одну секунду определить, в какой из 10 акустических сцен была сделана запись: в аэропорту, торговом центре, на станции метро, во время прогулки, на городской площади, на оживленной улице, в трамвае, автобусе, вагоне метро или в парке. В задаче существует два ограничения: размер модели не должен превышать 128 Кбайт, число операций сложения и умножения (MACs) — 30 млн. На данный момент самую высокую точность классификации показывает модель CP-Mobile [19]. Конфигурация модели, состоящая из 61 148 обучаемых параметров в 16-битном формате и требующая 29 419 156 операций, демонстрирует точность в 57 % по метрике accuracy, в то время как базовая модель 2023 г., представляющая собой многослойную сверточную нейронную сеть, достигала 42,9 %. Нейросеть CP-Mobile является 6-слойной сверточной нейронной сетью. Каждый слой внутри сети использует так называемые переходные, стандартные и пространственно-понижающие блоки (рис. 1), которые были разработаны специально для решения задачи классификации акустических сцен Acoustic Scene Classification (ASC).

Вторая задача заключается в том, чтобы во входной аудиозаписи длиной 10 с и по заранее зафиксированному списку аудиособытий определить, в каких частях аудио произошло каждое из событий. Отметим, что возникают ситуации, когда несколько событий происходят одновременно. Также существуют случаи, когда одно и то же событие повторяется несколько раз в разных частях записи, например лай собаки. Кроме того, во входной аудиозаписи каких-то событий может и не быть (рис. 2). Авторы задачи зафиксировали следующий список аудиособытий: звонок будильника, звуки блендера, мяуканье кошки, звук тарелок, лай собаки, звук электрической зубной щетки, звук жарки еды, звук смесителя, человеческая речь, звук пылесоса. Как и в первой задаче, в данной постановке есть несколько ограничений. В качестве обучающих данных предоставлено четыре набора данных с разметкой разной гранулярности: полностью размеченный корпус данных из 3,5 тыс. примеров, где для каждого аудиособытия известно, когда оно произошло; слабо размеченный корпус данных из 1,5 тыс. примеров, где для каждого аудиособытия известно лишь то, присутствует оно на записи или нет; 14 тыс. аудиозаписей без разметки, а также корпус данных из 10 тыс. синтетических примеров, сгенерированных с помощью инструмента Scaper [20].

Кроме ограниченного набора размеченных данных, учтено число операций сложения и умножения, необходимых для использования моделей. На данный момент большинство подходов опирается на архитектуру Convolutional Recurrent Neural Network (CRNN). Данная нейронная сеть состоит из нескольких слоев, каждый из которых представляет специально определенный

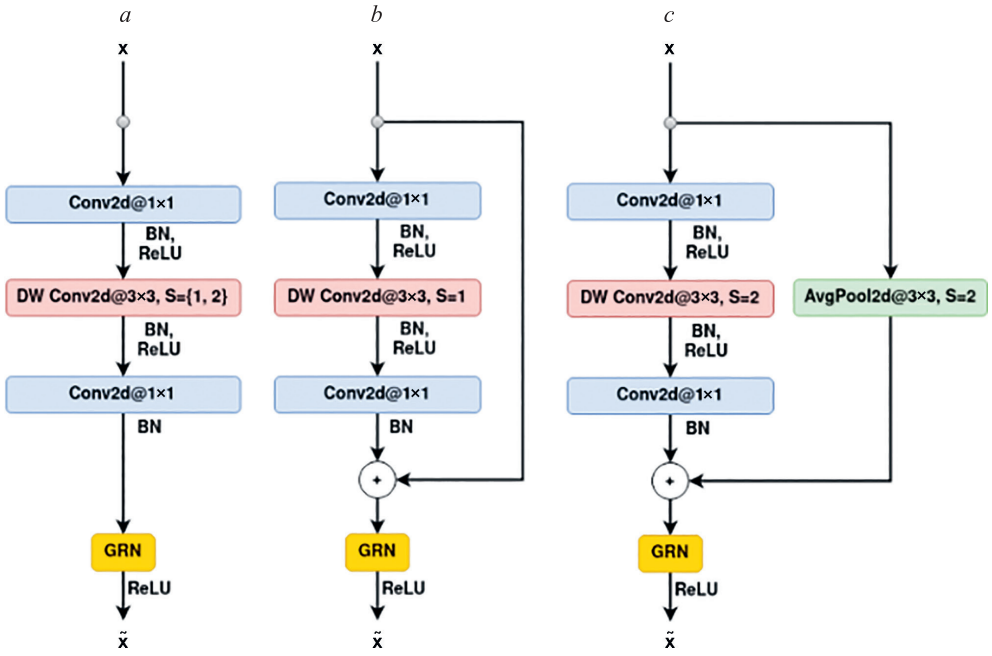

Рис. 1. Блоки одного слоя модели CP-Mobile [19] для классификации акустических сцен: переходный (*a*); стандартный (*b*); пространственно-понижающий (*c*).

BN — батч нормализация; DW — свертка по глубине; GRN — глобальная ответная нормализация; ReLU — выпрямленный линейный блок; **x** — входной сигнал; $\tilde{\mathbf{x}}$ — выходной сигнал

Fig. 1. Block types of one of the CP-Mobile [19] layers for acoustic scene classification problem: Transition Block (*a*); Standard Block (*b*); Spatial Downsampling Block (*c*).

BN — Batch Normalization; DW — Depthwise convolution; GRN — Global Response Normalization; ReLU — Rectified Linear Unit; **x** — input signal; $\tilde{\mathbf{x}}$ – output signal

блок сверток. По окончании применения всех сверток в сети, используется рекуррентная нейронная сеть, которая совершает предсказания для каждого момента времени каждого аудиособытия. Большое количество решений отличаются друг от друга построением сверточных блоков, а также особенностями обучения

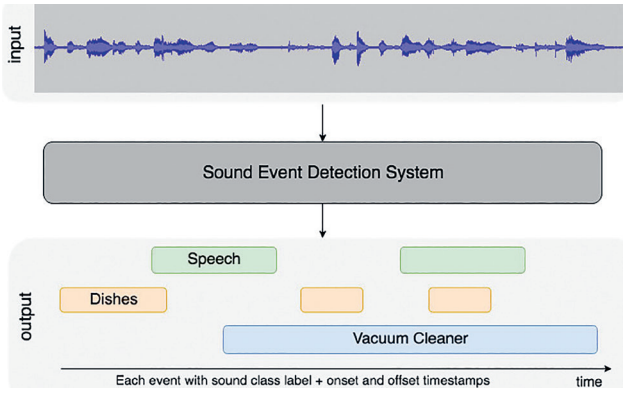

Рис. 2. Описание задачи распознавания аудиособытий.

input — входной аудиосигнал; output — предсказания модели; Sound Event Detection System — система детекции аудиособытий; Speech — речь; Dishes — звуки посуды; Vaccuum Cleaner — звуки пылесоса; time — ось времени, «Each event with sound class label + onset and offset timestamps» — каждое событие с меткой соответствующего класса + время начала и конца события

Fig. 2. Definition of the Sound Event Detection problem

моделей. Самым популярным алгоритмом, активно использующимся для решения задачи распознавания аудиособытий (Sound Event Detection, SED) с использованием данных с разметкой разной гранулярности, является так называемый mean-teacher [21]. В процессе обучения модели помимо основной обучаемой нейросети вычисляется так называемый учитель в виде экспоненциально взвешенного среднего за последние несколько шагов обучения. Итоговая функция потерь определяется как сумма бинарной кросс-энтропии и среднеквадратичной ошибки между предсказаниями обычной сети и ее экспоненциального среднего. Такой алгоритм позволяет эффективно использовать слаборазмеченные и неразмеченные данные. Авторский подход 2023 года основан на использовании 7-слойной сверточной нейронной сети в комбинации с двунаправленной рекуррентной нейросетью. Модель состоит из 1 млн параметров, требует 93 млн операций для обработки 1 с аудио и демонстрирует точность в 43,3 % по метрике Event-Based F1-score.

Таким образом, существует две важнейшие задачи в области распознавания метаинформации в звуке: ASC и SED. Для каждой из них существует ряд высокоточных решений, но такие модели требуют большого количества ресурсов или были разработаны специально для решения одной конкретной задачи: CP-Mobile (для решения задачи ASC), CRNN (для задачи SED). Вероятно, отсутствие универсальных подходов для одновременного решения вышеописанных задач при использовании ограниченного числа обучаемых пара-

Таблица 1. Различия между задачами классификации акустических сцен (ASC) и распознавания аудиособытий (SED)
Table 1. Differences between Acoustic Scene Classification (ASC) and Sound Event Detection (SED) problems

Характеристики задачи	Задачи распознавания	
	ASC	SED
Длительность аудиозаписи, с	1	10
Максимальное число обучаемых параметров	$128 \cdot 10^3$	$1 \cdot 10^6$
Размер размеченного тренировочного корпуса данных, ч	38,8	9,6
Алгоритм обучения	supervised	self supervised (mean-teacher)
Число примеров в батче	256	48

метров и арифметических операций связано с рядом существенных различий в конфигурациях поставленных задач и в оптимальных гиперпараметрах моделей, сильно влияющих на точность предсказаний обученных нейросетей (табл. 1).

В машинном обучении применяется метод обучения нейронных сетей — «многозадачное обучение» (Multi Task Learning, MTL), который представляет собой алгоритм обучения одной нейронной сети, при котором модель учится решать несколько задач одновременно. Одно из первых своих успешных применений MTL нашел в компьютерном зрении, где исследователи предлагали различные архитектуры [22–28] для одновременного решения задач компьютерного зрения: классификации, семантической сегментации, детекции объектов. В задаче распознавания метаинформации из аудиосигнала многозадачное обучение было успешно использовано для увеличения точности предсказаний для одной задачи за счет добавления в исходную нейронную сеть ответвлений для вспомогательных задач [29, 30], а также для решения большого числа задач распознавания метаинформации за счет использования больших языковых моделей совместно с тщательной организацией обучения модели [17].

Таким образом, целью исследователей было одновременное решение как можно большего числа задач с использованием моделей, состоящих из сотен миллионов обучаемых параметров, или улучшение одной конкретной задачи при помощи вспомогательных. Однако не было найдено подходов, которые бы позволяли получить компактную модель, которая может решать одновременно задачи ASC и SED с точностью, близкой к моделям, решающим каждую из задач отдельно. В настоящей работе предложена компактная архитектура многозадачной нейронной сети с общим кодировщиком в виде многослойной сверточной нейронной сети с добавлением двух ответвлений для обеих задач. Для ее обучения применены несколько классических алгоритмов многозадачного обучения, предложен ряд модификаций данных алгоритмов, произведено сравнение методов между собой и с результатами обучения специализированных моделей для каждой из задач распознавания. Предложенная модель требует на 7,8 % меньше обучаемых параметров и на 40 % меньше арифметических операций по сравнению с наивным использованием двух независимых специализированных моделей. Модель способна одновременно решать задачу ASC с точностью 42,43 % по метрике accuracy и задачу SED с точностью 45,89 % по метрике Event-Based F1-score, что сравнимо с точностью работы базовых подходов, предложенных авторами задач в 2023 г.

Архитектура компактной многозадачной нейронной сети

Ключевую роль в алгоритме многозадачного обучения играет архитектура нейронной сети. Необходимо подобрать такую модель, которая бы подходила для решения задач ASC и SED. Дополнительно основной целью модели является экономия ресурсов будущей системы, поэтому необходимо, чтобы архитектура сети содержала меньшее число обучаемых параметров и требовала меньшее число арифметических операций по сравнению с двумя независимыми моделями, решающими каждую из задач отдельно.

Очевидной идеей для построения архитектуры многозадачной нейронной сети является адаптация модели, предназначенной для решения одной из задач, для ее использования при решении второй задачи. В настоящей работе предлагается адаптировать модель CP-Mobile [19] для решения задачи SED. Для этого добавим в сеть еще два блока сверток (CPM BLOCK), а также рекуррентную нейронную сеть (RNN) аналогично структуре большинства высокоточных детекторов событий (рис. 3).

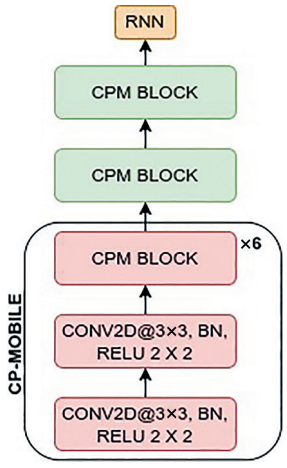

Рис. 3. Архитектура модели CP-Mobile-RNN, где CPM Block — специализированный блок модели CP-Mobile
Fig. 3. CP-Mobile-RNN model architecture, where CPM Block is a specialized unit of the CP-Mobile model

Далее обучим всю систему с помощью алгоритма mean-teacher [21] распознавать аудиособытия. Эксперименты показали, что предложенная модель обеспечивает точность SED в 35,54 %, что на 7,76 % меньше по сравнению с базовым подходом. С другой стороны, адаптация лучших детекторов событий для ASC невозможна из-за того, что они состоят из большого числа обучаемых параметров и требуют большого числа арифметических операций. Однако можно заметить, что, как и высокоточные классификаторы акустических сцен, так и лучшие детекторы аудиособытий имеют схожую структуру. Сначала для входной аудиозаписи строится мел-спектрограмма, после чего к ней применяется несколько сверточных слоев, в результате получается набор эмбеддингов для каждого момента времени. Затем классификатор акустической сцены применяет механизм внимания по оси времени в комбинации с линейным слоем, а детектор событий применяет к полученным эмбеддингам рекуррентную нейронную сеть (рис. 4). Таким образом, в обоих сценариях одинаковым образом строятся эмбеддинги, содержащие полезную информацию из аудио. Однако специализированные модели состоят из специально разработанных слоев с модифицированными свертками, предназначенными для решения конкретных задач. Вероятно, слои модели для решения одной задачи не подходят для решения другой задачи. Потому гипотетически разумным путем является использование обычных сверточных слоев для решения обеих задач. Эксперимент показал, что классическая сверточная нейросеть способна решать задачу ASC с точностью 53,31 %, а задачу SED с точностью 46,64 %, что на 11,1 % больше по сравнению с CP-Mobile-RNN. Видно, что данный подход менее точный по сравнению с самыми точными решениями задачи ASC в виде модели CP-Mobile, но он лучше адаптируется под задачу SED по сравнению с моделью CP-Mobile-RNN. Таким образом, итоговая структура многозадачной модели (MT model или Multi Task model) представляет из себя общий кодировщик, который по входному аудио строит последовательность эмбеддингов, которые затем обрабатываются декодировщиками для каждой из двух задач — механизмом внимания для задачи ASC, рекуррентной нейронной сетью для задачи SED (рис. 4). Стоит отметить, что ключевой особенностью данной модели является то, что в ней используются одни и те же эмбеддинги для двух задач. Другими словами, нет необходимости два раза строить две последовательности эмбеддингов для их последующей обработки, что приводит к существенному сокращению числа арифметических операций при использовании нейросети. В этом можно убедиться, вычислив размеры моделей, а также число арифметических операций, необходимых для их использования (табл. 2).

Алгоритмы обучения многозадачной нейронной сети

Классический подход обучения многозадачных моделей — алгоритм SoftMTL — состоит из нескольких шагов. Нейронная сеть принимает на вход данные от двух задач и применяет к ним общий кодировщик, который вычисляет эмбеддинги для каждого момента времени. После этого полученные вектора обрабатываются декодировщиками для соответствующих задач. Далее вычисляются функции потерь (L_{total}) для каждой задачи и складываются с некоторыми, заранее определенными коэффициентами (рис. 5). Данный алгоритм прост в реализации и позволяет обучать модель решать несколько задач одновременно, однако существуют пары задач,

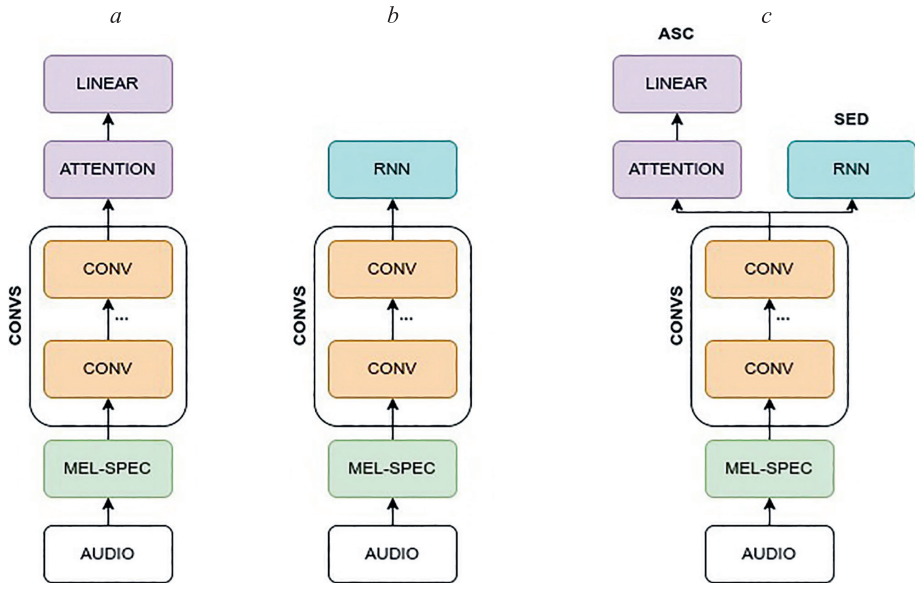

Рис. 4. Модели архитектуры: классификатора акустических сцен (ASC) (*a*); детектора аудиособытий (SED) (*b*); многозадачной модели (MTL) (*c*).

LINEAR — линейный слой; ATTENTION — механизм внимания; CONVS — сверточные слои; MEL-SPEP — мел-спектрограмма

Fig. 4. Model architectures of: acoustic scene classifier (*a*); sound event detector (*b*); multi task model (*c*)

Таблица 2. Число обучаемых параметров и арифметических операций, необходимых для использования моделей, где Baseline-2023 — базовое решение 2023 года

Table 2. The number of trainable model weights and multiply-accumulate operations needed for inference

Модель	Задача	Число обучаемых параметров	MACs для обработки аудио длительностью 1 с
Baseline-2023 ASC	ASC	**$65 \cdot 10^3$**	**$29 \cdot 10^6$**
CP-Mobile	ASC	$61 \cdot 10^3$	$29 \cdot 10^3$
Baseline-2023 SED	SED	**$1 \cdot 10^6$**	**$93 \cdot 10^6$**
CP-Mobile-RNN	SED	$966 \cdot 10^3$	$34 \cdot 10^3$
CNN	ASC	$60 \cdot 10^3$	$29 \cdot 10^3$
CRNN	SED	$628 \cdot 10^3$	$43 \cdot 10^3$
MT	**ASC+SED**	**$634 \cdot 10^3$**	**$43 \cdot 10^3$**

которые гипотетически могут быть не похожими друг на друга, как следствие при обучении нейронной сети линейная комбинация градиентов может приводить к сходимости модели в локальные минимумы, в которых нейросеть будет показывать низкую точность на обеих задачах.

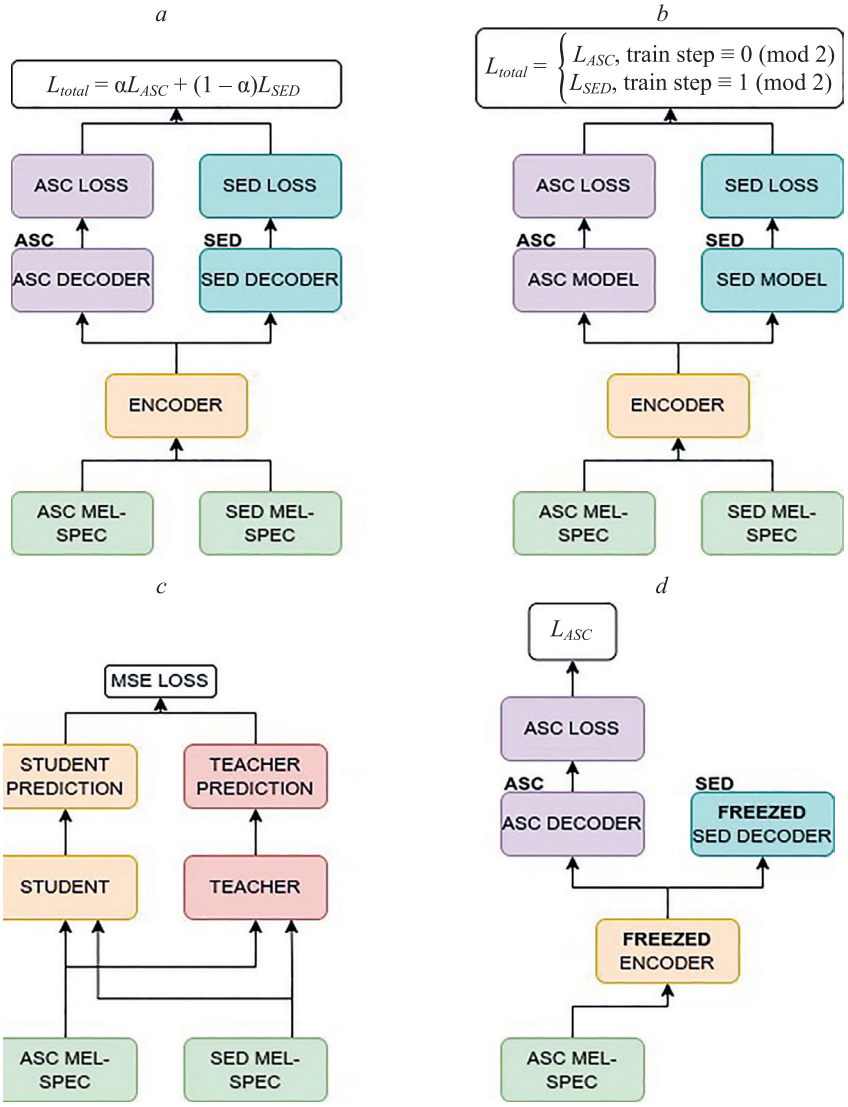

Рис. 5. Алгоритмы обучения многозадачных нейронных сетей: SoftMTL (*a*); HardMTL (*b*); CrossMTL (*c*); FreezeMTL (*d*), где α — коэффициент при функции ошибки для задачи ASC, LOSS — функция ошибки, DECODER — декодировщик, MODEL — модель, ENCODER — кодировщик, STUDENT PREDICTION — предсказание модели-студента, TEACHER PREDICTION — предсказание модели-учителя

Fig. 5. Training algorithms for multi task neural networks: SoftMTL (*a*); HardMTL (*b*); CrossMTL (*c*); FreezeMTL (*d*)

Альтернативный способ обучения — алгоритм HardMTL, когда модель поочередно учится на данных первой и второй задач. Иными словами, модель обновляет веса подсети, отвечающей за решение первой задачи на четных шагах, а веса подсети, отвечающей за решение второй подзадачи — на нечетных шагах обучения (рис. 5).

Изучение алгоритма mean-teacher [21] для обучения модели SED на данных с разметкой разной гранулярности приводит к применению похожего подхода для многозадачного обучения. Назовем обучаемую многозадачную модель студентом. Предлагается, аналогично алгоритму mean-teacher, вычислять экспоненциальное взвешенное среднее модели-студента. Назовем это среднее — моделью-учителем. Заметим, что в классических алгоритмах SoftMTL и HardMTL данные от одной задачи никак не используются при вычислении градиентов и функции ошибки для подсети, решающей вторую задачу, так как для них нет соответствующей разметки. Однако можно применить так называемую дистилляцию знаний из модели-учителя в модель-студента. Предлагается использовать подсеть модели-учителя, решающую первую задачу к данным от второй задачи. Аналогичным образом применяется подсеть модели-студента и вычисляется среднеквадратичная ошибка между предсказаниями моделей студента и учителя. Наконец, данное среднеквадратичное отклонение суммируется со стандартной функцией потерь, которая вычисляется в алгоритмах SoftMTL и HardMTL (рис. 5). Гипотетически описанный подход CrossMTL должен увеличить точность обученных многозадачных моделей за счет использования большего количества данных.

Большой интерес представляет крайний случай алгоритма HardMTL, а именно, когда модель сначала учится решать одну задачу, а затем вторую задачу. Другими словами, вместо поочередного использования батчей данных от двух задач сначала обрабатывается большое число батчей от первой задачи, затем от второй. Однако такой подход в вышеописанном виде нежизнеспособен из-за так называемого явления под названием «Catastrophic Forgetting» [31–33]. Для того чтобы избежать данную проблему, предлагается использовать метод заморозки весов обучаемой модели — применять замороженные параметры в качестве констант. Сначала обучим нейросеть решать одну задачу, затем заморозим веса общего кодировщика, а также декодировшика для данной задачи. После чего дообучим декодировщик второй задачи. Разработанный алгоритм назовем FreezeMTL. Заметим, что при таком подходе декодировщик второй задачи будет использовать эмбеддинги, полученные с помощью кодировщика от первой задачи (рис. 5). В результате возникает несколько вопросов для изучения: каким образом эмбеддинги кодировщика первой задачи влияют на точность предсказаний декодировщика второй задачи, а также какую из двух рассматриваемых задач выбрать для обучения общего кодировщика. Задачи ASC и SED существенно отличаются друг от друга. В результате возможна ситуация, когда эмбеддинги, полученные с помощью кодировщика от первой задачи, могут содержать недостаточно полезной информации для решения второй задачи и точность предсказаний на второй задаче может снизиться. На основании определения задач можно предположить, что SED является более сложной задачей по сравнению с ASC. Тогда эмбеддинги кодировшика для решения данной задачи должны содержать больше полезной информации, поэтому применение описанных эмбеддингов должно привести к получению более точной модели как с точки зрения классификации сцен, так и с точки зрения SED.

Экспериментальное определение наиболее эффективного алгоритма обучения многозадачной модели распознавания метаинформации в звуке

В настоящей работе эксперименты проводились с использованием данных, предоставленных авторами задач DCASE Task 1 и DCASE Task 4. Для задачи ASC использован корпус данных, состоящий из 38,8 ч тренировочных и 8,2 ч валидационных аудиособытий с равномерным распределением акустических сцен в них. Для задачи SED применен датасет, состоящий из 9,6 ч полностью размеченных, 4,4 ч слабо-размеченных, 40 ч неразмеченных, 27,8 ч синтетических и 3,3 ч валидационных данных. Для каждого аудио была построена мел-спектрограмма со стандартными гиперпараметрами: частота дискретизации составляет 16 кГц, длина скользящего окна равна 128 мс, длина шага окна — 10 мс, число мел-фильтрбанков — 128. Во всех экспериментах использованы стандартные аугментации, маскирующие мел-спектрограммы, обнуляя 15 % значений по каждой из осей времени и мел-фильтрбанков.

Для обучения моделей применен оптимизатор AdamW со стандартными гиперпараметрами: скорость обучения — 0,001, бета-коэффициенты для вычисления бегущих средних градиентов равны 0,9 и 0,999. В качестве планировщика скорости обучения выбран стандартный планировщик с линейным увеличением скорости от 0 до 0,001 за первые 10 % шагов обучения с последующим уменьшением до нуля по косинусному закону.

В качестве первого подхода к обучению многозадачной нейросети взят алгоритм SoftMTL, в процессе которого функция потерь вычисляется как среднее значение между функциями потерями для каждой из задач. В результате обучения многозадачной сети с помощью данного алгоритма была получена модель, выдающая точности ASC в 37,59 % и SED — 39,08 % (табл. 3).

Из табл. 3 видно, что точность предсказания многозадачной модели ниже по сравнению с базовыми подходами. Гипотетически, это может быть связано с тем, что во время обучения две задачи конфликтуют между собой и линейная комбинация их градиентов может приводить к тому, что веса модели подбираются не оптимальным образом. Для того чтобы частично избавиться от подобного рода конфликтов градиентов, предлагается использовать алгоритм HardMTL, в процессе которого веса модели поочередно обновляются сначала для первой задачи, потом для второй. В результате применения данного алгоритма была получена модель,

Таблица 3. Результаты обучения моделей с использованием различных алгоритмов обучения многозадачной нейронной сети
Table 3. Experimental results of applying various training algorithms for training of multi task neural network

Алгоритм обучения	ASC accuracy, %	SED Event-based F1-score, %	Число обучаемых параметров	MACs
ASC baseline 2023	**42,90**	—	$65{,}0 \cdot 10^3$	$29{,}0 \cdot 10^6$
ASC SOTA (CP-Mobile [19])	57,00	—	$61{,}0 \cdot 10^3$	$29{,}0 \cdot 10^6$
SED baseline 2023	—	**43,30**	$1{,}0 \cdot 10^6$	$93{,}0 \cdot 10^6$
SED SOTA (FDY-LKA [34])	—	58,30	$9{,}0 \cdot 10^6$	$7{,}0 \cdot 10^9$
SoftMTL	37,59	39,08	$633{,}5 \cdot 10^3$	$43{,}2 \cdot 10^6$
HardMTL	40,92	39,05		
CrossSoftMTL	36,64	39,03		
CrossHardMTL	42,30	39,96		
FreezeMTL (SED, ASC)	**42,44**	**45,86**		
FreezeMTL (ASC, SED)	51,29	30,74		

имеющая точности ASC — 40,92 % и SED – 39,05 %. Видно, что данный подход детектирует события с такой же точностью, как и алгоритм SoftMTL, но существенно точнее классифицирует акустические сцены.

Заметим, что в алгоритмах SoftMTL и HardMTL данные одной задачи никак не используются при обучении декодировщика второй задачи. Гипотетически, большее количество обучающих данных должно привести к более высокоточной модели. Комбинация алгоритмов SoftMTL и CrossMTL под названием CrossSoftMTL позволяет получить модель, обладающей точностями ASC — 36,64 % и SED — 39,03 %, в то время как добавление CrossMTL в HardMTL приводит к обучению модели, которая демонстрирует точности ASC — 42,3 % и SED — 39,96 %. Видно, что алгоритм CrossMTL позволяет улучшить точность модели при его совместном применении с алгоритмом HardMTL, однако его комбинация с SoftMTL приводит к небольшому ухудшению точности ASC.

В заключение рассмотрим предельный случай алгоритма HardMTL, а именно ситуацию, когда модель сначала полностью учится только на данных первой задачи, затем веса общего кодировщика замораживаются, после чего декодировшик оставшейся задачи доучивается на данных своей задачи. Результаты применения данного алгоритма можно найти в табл. 3 в строках с алгоритмами FreezeMTL. Версия (SED, ASC) означает, что сначала модель была обучена решать задачу SED, а затем доучена решать задачу ASC. Исходя из результатов, указанных в табл. 3 можно сделать вывод о том, что точность решения задачи, которую модель была обучена решать первой, существенно выше по сравнению с точностью предсказаний для второй задачи. Это связано с тем, что фактически при обучении модели решать вторую задачу используются эмбеддинги, полученные с помощью кодировщика, который учился решать первую задачу. Заметим, что порядок, в котором модель учится решать задачу существенен. Если сначала модель обучить классифицировать ASC, то точность SED уменьшается до 30,74 %, что ниже результатов всех вышеупомянутых алгоритмов, в то время как при первоначальном обучении модели решать задачу SED точность ASC составляет 42,44 %, что выше, чем у остальных подходов при одновременной точности SED в 45,86 %. Это говорит о том, что эмбеддинги, полученные с помощью кодировщика модели SED содержат больше полезной информации об аудиособытии по сравнению с векторами, сгенерированными моделью ASC.

В результате проведенных экспериментов показано положительное влияние алгоритма CrossMTL на точность предсказаний финальных моделей. Получена многозадачная модель, обученная с помощью алгоритма FreezeMTL (SED, ASC), которая классифицирует ASC с точностью 42,44 %, что всего на 0,46 % меньше по сравнению с базовым подходом, а также позволяет детектировать аудиособытия с точностью 45,86 %, что больше базового подхода на 2,56 %. Более того, полученная модель содержит на 7,8 % меньше обучаемых параметров, а также требует на 40 % меньше арифметических операций по сравнению с наивным использованием двух независимых моделей ASC и SED (табл. 2).

Заключение

Задача распознавания метаинформации в звуке представляет большой интерес как с точки зрения науки, так и с точки зрения индустрии. На данный момент существует множество подходов, позволяющих решать большинство из них при использовании больших нейронных сетей, состоящих из сотен миллионов параметров. Такие высокие требования к ресурсам не позволяют использовать данные методы в умных устройствах. Описанная проблема активно изучается на ежегодной конференции DCASE Workshop, где было представлено множество задач распознавания метаинформации. Для исследования были выбраны две задачи: классификации акустических сцен и детекции аудиособытий. В настоящей работе предложена компактная нейронная сеть, состоящая из 633,5 тыс. обучаемых параметров, требующая 43,2 млн арифметических операций для обработки аудио длиной в 1 с. Данный подход использует на 7,8 % меньше обучаемых параметров и на 40 % меньше арифметических операций по сравнению с наивным применением двух независимых моделей. Рассмотрены классические алгоритмы многозадачного обучения SoftMTL и HardMTL и предложены их

модификации CrossSoftMTL и CrossHardMTL. Изучен предельный случай алгоритма HardMTL, когда модель сначала учится решать одну задачу, затем веса общего кодировщика замораживаются и, наконец, модель доучивается решать вторую задачу. Данный алгоритм позволяет обучить многозадачную модель, одновременно классифицировать акустические события с точностью 42,44 % (меньше базового решения на 0,46 %) и детектировать аудиособытия с точностью 45,86 % (что больше базового решения на 2,56 %).

Литература

1. Kriman S., Beliaev S., Ginsburg B., Huang J., Kuchaiev O., Lavrukhin V., Leary R., Li J., Zhang Y. Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions // Proc. of the ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 6124–6128. https://doi.org/10.1109/icassp40776.2020.9053889
2. Lakhotia K., Kharitonov E., Hsu W.-N., Adi Y., Polyak A., Bolte B., Nguyen T.-A., Copet J., Baevski A., Mohamed A., Dupoux E. On generative spoken language modeling from raw audio // Transactions of the Association for Computational Linguistics. 2021. V. 9. P. 1336–1354.
3. Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented transformer for speech recognition // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2020. P. 5036–5040. https://doi.org/10.21437/interspeech.2020-3015
4. Hsu W.N., Tsai B., Bolte Y.-H.H., Salakhutdinov R., Mohamed A. HuBERT: How much can a bad teacher benefit ASR pre-training? // Proc. of the ICASSP 2021 — 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. P. 6533–6537. https://doi.org/10.1109/icassp39728.2021.9414460
5. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // Proceedings of Machine Learning Research, PMLR. 2023. V. 202. P. 28492–28518.
6. Gong Y., Khurana S., Karlinsky L., Glass J. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2023. P. 2798–2802. https://doi.org/10.21437/interspeech.2023-2193
7. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: an asr corpus based on public domain audio books // Proc. of the IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP). 2015. P. 5206–5210. https://doi.org/10.1109/icassp.2015.7178964
8. Moritz N., Wichern G., Hori T., Le Roux J. All-in-One transformer: Unifying speech recognition, audio tagging, and event detection // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2020. P. 3112–3116. https://doi.org/10.21437/interspeech.2020-2757
9. Karita S., Soplin N.E.Y., Watanabe S., Delcroix M., Ogawa A., Nakatani T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2019. P. 1408–1412. https://doi.org/10.21437/interspeech.2019-1938
10. Chen S., Wu Y., Wang C., Liu S., Tompkins D., Chen Z., Che W., Yu X., Wei F. Beats: Audio pre-training with acoustic tokenizers // Proceedings of Machine Learning Research. 2023. V. 202, P. 4672–4712.
11. Gemmeke J.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C., Plakal M., Ritter M. Audio set: An ontology and human-labeled dataset for audio events // Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017. P. 776–780. https://doi.org/10.1109/icassp.2017.7952261
12. 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 // Proc. of the ICLR 2021 — 9th International Conference on Learning Representations. 2021.
13. Drossos K., Lipping S., Virtanen T. Clotho: An audio captioning dataset // Proc. of the ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 736–740. https://doi.org/10.1109/icassp40776.2020.9052990

References

1. Kriman S., Beliaev S., Ginsburg B., Huang J., Kuchaiev O., Lavrukhin V., Leary R., Li J., Zhang Y. Quartznet: Deep automatic speech recognition with 1D time-channel separable convolutions. *Proc. of the ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6124–6128. https://doi.org/10.1109/icassp40776.2020.9053889
2. Lakhotia K., Kharitonov E., Hsu W.-N., Adi Y., Polyak A., Bolte B., Nguyen T.-A., Copet J., Baevski A., Mohamed A., Dupoux E. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 2021, vol. 9, pp. 1336–1354.
3. Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. Conformer: Convolution-augmented transformer for speech recognition. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 5036–5040. https://doi.org/10.21437/interspeech.2020-3015
4. Hsu W.N., Tsai B., Bolte Y.-H.H., Salakhutdinov R., Mohamed A. HuBERT: How much can a bad teacher benefit ASR pre-training?. *Proc. of the ICASSP 2021 — 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6533–6537. https://doi.org/10.1109/icassp39728.2021.9414460
5. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. *Proceedings of Machine Learning Research*, 2023, vol. 202, pp. 28492–28518.
6. Gong Y., Khurana S., Karlinsky L., Glass J. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2023, pp. 2798–2802. https://doi.org/10.21437/interspeech.2023-2193
7. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: an asr corpus based on public domain audio books. *Proc. of the IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210. https://doi.org/10.1109/icassp.2015.7178964
8. Moritz N., Wichern G., Hori T., Le Roux L. All-in-One transformer: Unifying speech recognition, audio tagging, and event detection. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 3112–3116. https://doi.org/10.21437/interspeech.2020-2757
9. Karita S., Soplin N.E.Y., Watanabe S., Delcroix M., Ogawa A., Nakatani T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 1408–1412. https://doi.org/10.21437/interspeech.2019-1938
10. Chen S., Wu Y., Wang C., Liu S., Tompkins D., Chen Z., Che W., Yu X., Wei F. Beats: Audio pre-training with acoustic tokenizers. *Proceedings of Machine Learning Research*, 2023, vol. 202, pp. 4672–4712.
11. Gemmeke J.F., Ellis D.P.W., Freedman D., Jansen A., Lawrence W., Moore R.C., Plakal M., Ritter M. Audio set: An ontology and human-labeled dataset for audio events. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. https://doi.org/10.1109/icassp.2017.7952261
12. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. et al. An image is worth 16×16 words: Transformers for image recognition at scale. *Proc. of the ICLR 2021 — 9th International Conference on Learning Representations*, 2021.
13. Drossos K., Lipping S., Virtanen T. Clotho: An audio captioning dataset. *Proc. of the ICASSP 2020 — 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,

14. Poria S., Hazarika D., Majumder N., Naik G., Cambria E., Mihalcea R. MELD: A multimodal multi-party dataset for emotion recognition in conversations // Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 527–536. https://doi.org/10.18653/v1/p19-1050
15. Lipping S., Sudarsanam P., Drossos K., Virtanen T. Clotho-AQA: A crowdsourced dataset for audio question answering // Proc. of the 30th European Signal Processing Conference (EUSIPCO). 2022. P. 1140–1144. https://doi.org/10.23919/eusipco55093.2022.9909680
16. Engel J., Resnick C., Roberts A., Dieleman S., Norouzi M., Eck D., Simonyan K. Neural audio synthesis of musical notes with wavenet autoencoders // Proceedings Conference on Machine Learning, PMLR. 2017. V. 70. P. 1068–1077.
17. Chu Y., Xu J., Zhou X., Yang Q., Zhang S., Yan Z., Zhou C., Zhou J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models // arXiv. 2023. arXiv:2311.07919. https://doi.org/10.48550/arXiv.2311.07919
18. Bai J., Bai S., Chu Y., Cui Z. Qwen technical report // arXiv. 2023. arXiv:2309.16609. https://doi.org/10.48550/arXiv.2309.16609
19. Schmid F., Morocutti T., Masoudian S., Koutini K., Widmer G. CP-JKU submission to dcase23: Efficient acoustic scene classification with cp-mobile: Technical Report / Detection and Classification of Acoustic Scenes and Events (DCASE). 2023. 5 p.
20. Salamon J., MacConnell D., Cartwright M., Li P., Bello J.P. Scaper: A library for soundscape synthesis and augmentation // Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2017. P. 344–348. https://doi.org/10.1109/waspaa.2017.8170052
21. Tarvainen A., Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results // Advances in Neural Information Processing Systems. 2017. V. 30. P. 1196–1205.
22. Zhang Z., Luo P., Loy C.C., Tang X. Facial landmark detection by deep multi-task learning // Lecture Notes in Computer Science. 2014. V. 8694. P. 94–108. https://doi.org/10.1007/978-3-319-10599-4_7
23. Dai J., He K., Sun J. Instance-aware semantic segmentation via multi-task network cascades // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 3150–3158. https://doi.org/10.1109/cvpr.2016.343
24. Zhao X., Li H., Shen X., Liang X., Wu Y. A modulation module for multi-task learning with applications in image retrieval // Lecture Notes in Computer Science. 2018. V. 11205. P. 415–432. https://doi.org/10.1007/978-3-030-01246-5_25
25. Liu S., Johns E., Davison A.J. End-to-end multi-task learning with attention // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. P. 1871–1880. https://doi.org/10.1109/cvpr.2019.00197
26. Ma J., Zhao Z., Yi X., Chen J., Hong L., Chi E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts // Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018. P. 1930–1939. https://doi.org/10.1145/3219819.3220007
27. Misra I., Shrivastava A., Gupta A., Hebert M. Cross-stitch networks for multi-task learning // Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 3994–4003. https://doi.org/10.1109/cvpr.2016.433
28. Ruder S., Bingel J., Augenstein I., Søgaard A. Latent multi-task architecture learning // Proceedings of the AAAI Conference on Artificial Intelligence. 2019. V. 33. N 01. P. 4822–4829. https://doi.org/10.1609/aaai.v33i01.33014822
29. Krause D.A., Mesaros A. Binaural signal representations for joint sound event detection and acoustic scene classification // Proc. of the 30th European Signal Processing Conference (EUSIPCO). 2022. P. 399–403. https://doi.org/10.23919/eusipco55093.2022.9909581
30. Khandelwal T., Das R.K. A multi-task learning framework for sound event detection using high-level acoustic characteristics of sounds // Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2023. P. 1214–1218. https://doi.org/10.21437/interspeech.2023-909
31. French R.M. Catastrophic forgetting in connectionist networks // Trends in Cognitive Sciences. 1999. V. 3. N 4. P. 128–135. https://doi.org/10.1016/s1364-6613(99)01294-2
32. McCloskey M., Cohen N.J. Catastrophic interference in connectionist networks: The sequential learning problem // Psychology of Learning and Motivation. 1989. V. 24. P. 109–165. https://doi.org/10.1016/s0079-7421(08)60536-8

2020, pp. 736–740. https://doi.org/10.1109/icassp40776.2020.9052990
14. Poria S., Hazarika D., Majumder N., Naik G., Cambria E., Mihalcea R. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536. https://doi.org/10.18653/v1/p19-1050
15. Lipping S., Sudarsanam P., Drossos K., Virtanen T. Clotho-AQA: A crowdsourced dataset for audio question answering. *Proc. of the 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1140–1144. https://doi.org/10.23919/eusipco55093.2022.9909680
16. Engel J., Resnick C., Roberts A., Dieleman S., Norouzi M., Eck D., Simonyan K. Neural audio synthesis of musical notes with wavenet autoencoders. *International Conference on Machine Learning*. PMLR, 2017, vol. 70, pp. 1068–1077.
17. Chu Y., Xu J., Zhou X., Yang Q., Zhang S., Yan Z., Zhou C., Zhou J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv*, 2023, arXiv:2311.07919. https://doi.org/10.48550/arXiv.2311.07919
18. Bai J., Bai S., Chu Y., Cui Z. Qwen technical report. *arXiv*, 2023, arXiv:2309.16609. https://doi.org/10.48550/arXiv.2309.16609
19. Schmid F., Morocutti T., Masoudian S., Koutini K., Widmer G. *CP-JKU submission to dcase23: Efficient acoustic scene classification with cp-mobile: Technical Report*. Detection and Classification of Acoustic Scenes and Events (DCASE), 2023. 5 p.
20. Salamon J., MacConnell D., Cartwright M., Li P., Bello J.P. Scaper: A library for soundscape synthesis and augmentation. *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 344–348. https://doi.org/10.1109/waspaa.2017.8170052
21. Tarvainen A., Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 2017, vol. 30, pp. 1196–1205.
22. Zhang Z., Luo P., Loy C.C., Tang X. Facial landmark detection by deep multi-task learning. *Lecture Notes in Computer Science*, 2014, vol. 8694, pp. 94–108. https://doi.org/10.1007/978-3-319-10599-4_7
23. Dai J., He K., Sun J. Instance-aware semantic segmentation via multi-task network cascades. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3150–3158. https://doi.org/10.1109/cvpr.2016.343
24. Zhao X., Li H., Shen X., Liang X., Wu Y. A modulation module for multi-task learning with applications in image retrieval. *Lecture Notes in Computer Science*, 2018, vol. 11205, pp. 415–432. https://doi.org/10.1007/978-3-030-01246-5_25
25. Liu S., Johns E., Davison A.J. End-to-end multi-task learning with attention. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1871–1880. https://doi.org/10.1109/cvpr.2019.00197
26. Ma J., Zhao Z., Yi X., Chen J., Hong L., Chi E.H. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1930–1939. https://doi.org/10.1145/3219819.3220007
27. Misra I., Shrivastava A., Gupta A., Hebert M. Cross-stitch networks for multi-task learning. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003. https://doi.org/10.1109/cvpr.2016.433
28. Ruder S., Bingel J., Augenstein I., Søgaard A. Latent multi-task architecture learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 01, pp. 4822–4829. https://doi.org/10.1609/aaai.v33i01.33014822
29. Krause D.A., Mesaros A. Binaural signal representations for joint sound event detection and acoustic scene classification. *Proc. of the 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 399–403. https://doi.org/10.23919/eusipco55093.2022.9909581
30. Khandelwal T., Das R.K. A multi-task learning framework for sound event detection using high-level acoustic characteristics of sounds. *Proc. of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2023, pp. 1214–1218. https://doi.org/10.21437/interspeech.2023-909
31. French R.M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 1999, vol. 3, no. 4, pp. 128–135. https://doi.org/10.1016/s1364-6613(99)01294-2
32. McCloskey M., Cohen N.J. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning*

33. Kirkpatrick J., Pascanu R., Rabinowitz N., Veness J., Desjardins G., Rusu A.A., Milan K., Quan J., Ramalho T., Grabska-Barwinska A., Hassabis D., Clopath C., Kumaran D., Hadsell R. Overcoming catastrophic forgetting in neural networks // Proceedings of the national academy of sciences. 2017. V. 114. N 13. P. 3521–3526. https://doi.org/10.1073/pnas.1611835114
34. Kim J.W., Lee G.W., Kim H.K., Seo Y.S., Song I.H. Semi-supervised learning-based sound event detection using frequency-channel-wise selective kernel for DCASE challenge 2022 Task 4: Technical Report / Detection and Classification of Acoustic Scenes and Events (DCASE), 2022. 4 p.

and Motivation, 1989, vol. 24, pp. 109–165. https://doi.org/10.1016/s0079-7421(08)60536-8
33. Kirkpatrick J., Pascanu R., Rabinowitz N., Veness J., Desjardins G., Rusu A.A., Milan K., Quan J., Ramalho T., Grabska-Barwinska A., Hassabis D., Clopath C., Kumaran D., Hadsell R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 2017, vol. 114, no. 13, pp. 3521–3526. https://doi.org/10.1073/pnas.1611835114
34. Kim J.W., Lee G.W., Kim H.K., Seo Y.S., Song I.H. *Semi-supervised learning-based sound event detection using frequency-channel-wise selective kernel for DCASE challenge 2022 Task 4: Technical Report*. Detection and Classification of Acoustic Scenes and Events (DCASE), 2022, 4 p.

Автор

Сурков Максим Константинович — аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, https://orcid.org/0000-0002-3929-7484, surkovmax007@mail.ru

Author

Maxim K. Surkov — PhD Student, ITMO University, Saint Petersburg, 197101, Russian Federation, https://orcid.org/0000-0002-3929-7484, surkovmax007@mail.ru

Статья поступила в редакцию 20.06.2024
Одобрена после рецензирования 07.07.2024
Принята к печати 27.09.2024

Received 20.06.2024
Approved after reviewing 07.07.2024
Accepted 27.09.2024