applied sciences

MDPI

Article

A Review of Deep Learning Based Methods for Acoustic Scene Classification

Jakob Abeßer

Semantic Music Technologies, Fraunhofer IDMT, Ehrenbergstraße 31, 98693 Ilmenau, Germany; jakob.abesser@idmt.fraunhofer.de

Received: 18 February 2020; Accepted: 9 March 2020; Published: 16 March 2020

Abstract: The number of publications on acoustic scene classification (ASC) in environmental audio recordings has constantly increased over the last few years. This was mainly stimulated by the annual Detection and Classification of Acoustic Scenes and Events (DCASE) competition with its first edition in 2013. All competitions so far involved one or multiple ASC tasks. With a focus on deep learning based ASC algorithms, this article summarizes and groups existing approaches for data preparation, i.e., feature representations, feature pre-processing, and data augmentation, and for data modeling, i.e., neural network architectures and learning paradigms. Finally, the paper discusses current algorithmic limitations and open challenges in order to preview possible future developments towards the real-life application of ASC systems.

Keywords: acoustic scene classification; machine listening; deep neural networks

1. Introduction

Recognizing different indoor and outdoor acoustic environments from recorded acoustic signals is an active research field that has received much attention in the last few years. The task is an essential part of auditory scene analysis and involves summarizing an entire recorded acoustic signal using a pre-defined semantic description like "office room" or "public place". Those semantic entities are denoted as acoustic scenes and the task of recognizing them as acoustic scene classification (ASC) [1].

A particularly challenging task related to ASC is the detection of audio events that are temporarily present in an acoustic scene. Examples of such audio events include vehicles, car horns, and footsteps, among others. This task is referred to as acoustic event detection (AED), and it substantially differs from ASC as it focuses on the precise temporal detection of particular sound events.

State-of-the-art ASC systems have been shown to outperform humans on this task [2]. Therefore, they are applied in numerous application scenarios such as context-aware wearables and hearables, hearing aids, health care, security surveillance, wild-life monitoring in nature habitats, smart cities, IoT, and autonomous navigation.

This article summarizes and categorizes deep learning based algorithms for ASC in a systematic fashion based on the typical processing steps illustrated in Figure 1. Section 2, Section 3, and Section 4 discuss techniques to represent, pre-process, and augment audio signals for ASC. Commonly used neural network architectures and learning paradigms are detailed in Section 5 and Section 6. Finally, Section 7 discusses the open challenges and limitations of current ASC algorithms before Section 8 concludes this article. Each section first provides an overview of previously used approaches. Then, based on the published results of the the Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 and 2019 challenges, the most promising methods are highlighted. It must be noted that evaluating and comparing the effectiveness of different methods is often complicated by the use of different evaluation datasets.

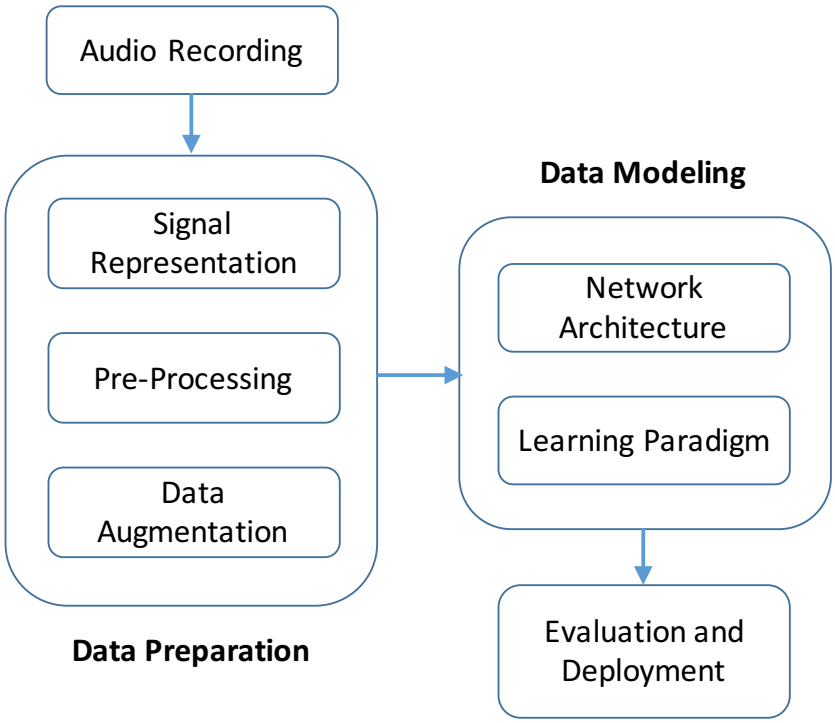

Figure 1. The flowchart summarizes the article's structure and lists the typical processing flow of an acoustic scene classification (ASC) algorithm.

As a complementary read to this article, Barchiesi et al. published an in-depth overview of ASC methods using "traditional" feature extraction and classification techniques prior to the general transition to deep learning based methods in [3]. Other related survey articles focus on deep learning methods for AED [4,5] or summarize algorithms submitted for various machine listening tasks including ASC for a particular year of the DCASE challenge such as [6]. Methodologies and common datasets for evaluating ASC algorithms are not further addressed in this article.

2. Signal Representations

Datasets for the tasks of ASC or AED contain digitized audio recordings. The resulting acoustic signals are commonly represented as waveforms that denote the amplitude of the recorded signal over discrete time samples. In most cases, ASC or AED systems perform the tasks of interest on derived signal representations, which will be introduced in the following section.

2.1. Monaural vs. Multi-Channel Signals

ASC algorithms commonly process monaural audio signals. Sound sources in acoustic scenes are spatially distributed by nature. If multi-channel audio recordings are available, the inherent spatial information can be exploited to better localize sound sources. The joint localization and detection of sound events was first addressed in Task 3 of the DCASE 2019 challenge (http://dcase.community/challenge2019/task-sound-event-localization-and-detection).

In addition to the left/right channels, a mid/side channel representation can be used as additional signal representation [7,8]. As an example for using a larger number of audio channels, Green and Murphey [9] classified acoustic scene recordings of fourth-order Ambisonics by combining spatial features describing the direction of sound arrival with band-wise spectral diffuseness measures. Similarly, Zirliński and Lee combined spatial features from binaural recordings with spectro-temporal features to characterize the foreground/background sound distribution in acoustic scenes [10]. Current state-of-the-art ASC algorithms process either mono or stereo signals without a clear trend.

2.2. Fixed Signal Transformations

Most neural network architectures applied for ASC require multi-dimensional input data (compare Section 5). The most commonly used time-frequency transformations are the short-time Fourier transform (STFT), the Mel spectrogram, and the wavelet spectrogram. The Mel spectrogram is based on a non-linear frequency scale motivated by human auditory perception and provides a more compact spectral representation of sounds compared to the STFT. ASC algorithms process only the magnitude of the Fourier transform while the phase is discarded.

Wavelets can be computed in a one-step [11,12] or cascaded fashion [13] to decompose time-domain signals into a set of basis function coefficients. The deep scattering spectrum [13] decomposes a signal using a sequential cascade of wavelet decompositions and modulation operations. The scalogram [14,15] uses multiple parallel wavelet filters with logarithmically spaced support and bandwidth to provide invariance to both time-warping and local signal translations. Time-averaged statistics based on computer vision algorithms like local binary patterns (LPB) or histogram of oriented gradients (HOG) can be used to summarize such two-dimensional signal transformations [16]. In addition to basic time-frequency transformations, perceptually-motivated signal representations are used as input to deep neural networks. Such representations for instance characterize the distribution (e.g., Mel-frequency cepstral coefficients (MFCC) [17], sub-band power distribution [18], and gammatone frequency cepstral coefficients [19]) and modulation of the spectral energy (e.g., amplitude modulation bank features [20] and temporal energy variation [21]). Feature learning techniques based on hand-crafted audio features and traditional classification algorithms such as support vector machines (SVM) have been shown to underperform deep learning based ASC algorithms [22,23].

High-dimensional feature representations are often redundant and can lead to model overfitting. Therefore, before being processed by the neural network, the feature space dimensionality can be further reduced: One approach is to aggregate sub-bands of spectrograms using local binary pattern (LBP) histograms [24] or sub-band power distribution (SBD) features [18]. A second approach is to map features to a randomized low-dimensional feature space as proposed by Jimenez et al. [25].

The best performing ASC algorithms from the recent DCASE challenges used almost exclusively spectrogram representations based on logarithmic frequency spacing and logarithmic magnitude scaling such as log Mel spectrograms as the network input. Occasionally, end-to-end learning based on raw waveforms was used.

2.3. Learnable Signal Transformations

Three different approaches have been used to the best of our knowledge in ASC systems to avoid fixed pre-defined signal transformations. The first approach is to apply end-to-end learning where neural networks directly process raw audio samples. Examples of such network architectures are AclNet and AclSincNet[26], as well as SoundNet [27]. As a potential advantage against spectrogram based methods, the signal phase is not discarded.

The second approach is to interpret the signal transformation step as a learnable function, commonly denoted as “front-end”, which can be jointly trained with the classification back-end [28]. The third approach is to use unsupervised learning to derive semantically meaningful signal representations. Amiriparian et al. combined representations learned using a deep convolutional generative adversarial network (DCGAN) and using a recurrent sequence to sequence autoencoder (S2SAE) [29]. Similarly, environmental audio recordings can be decomposed into suitable basis functions using well-established matrix factorization techniques such non-negative matrix factorization (NMF) [30] and shift-invariant probabilistic latent component analysis (SIPLCA) [31]. Up to this point, the best-performing ASC algorithms of the recent DCASE challenges still focus on fixed spectrogram based signal representations instead of learnable signal transformations.

3. Pre-Processing

Feature standardization is commonly used to speed up the convergence of gradient descent based algorithms [8]. This process changes the feature distribution to have zero mean and unit variance. In order to compensate for the large dynamic range in environmental sound recordings, logarithmic scaling is commonly applied to spectrogram based features. Other low-level audio signal pre-processing methods include dereverberation and low-pass filtering [32].

Both ASC and AED face the challenge that foreground sound events in acoustic scenes are often overshadowed by background noises. Lostanlen et al. used per-channel energy normalization (PCEN) [33] to reduce stationary noise and to enhance transient sound events in environmental audio recordings [34]. This algorithm performs an adaptive, band-wise normalization and decorrelates the frequency bands. Wu et al. enhanced edge-like structures in Mel spectrograms using two edge detection methods from image processing based on the difference of Gaussians (DoG) and Sobel filtering [35]. The background drift of the Mel spectrogram is removed using median filtering. Similarly, Han et al. used background subtraction and applied median filtering over time [7] to remove irrelevant noise components from the acoustic scene background and the recording device.

Several filtering approaches are used as pre-processing for ASC algorithms. For example, Nguyen et al. applied a nearest neighbor filter based on the repeating pattern extraction technique (REPET) algorithm [36] and replaced the most similar spectrogram frames by their median prior to the classification [37]. This allowed emphasizing repetitive sound events in acoustic scenes such as from sirens or horns. As another commonly used filtering approach, harmonic-percussive source separation (HPSS) decomposes the spectrogram into horizontal and vertical components and provides additional feature representations for ASC [7,32,38]. While most of the discussed pre-processing techniques have been proposed just very recently, logarithmic magnitude scaling is the only well-established method, which is consistently used among the best performing ASC algorithms.

4. Data Augmentation Techniques

Training deep learning models usually requires large amounts of training data to capture the natural variability in the data to be modeled. The size of machine listening datasets increased over the last few years, but lagged behind computer vision datasets such as the ImageNet dataset with over 14 million images and over 21 thousand object classes [39]. The only exception to this day is the AudioSet dataset [40] with currently over 2.1 million audio excerpts and 527 sound event classes. This section summarizes techniques for data augmentation to address this lack of data.

The first group of data augmentation algorithms generates new training data instances from existing ones by applying various signal transformations. Basic audio signal transformation includes time stretching, pitch shifting, dynamic range compression, as well as adding random noise [41–43]. Koutini et al. applied spectral rolling by randomly shifting spectrogram excerpts over time [44].

Several data augmentation methods allow simulating overlap between multiple sound events and the resulting occlusion effects in the spectrogram. Mixup data augmentation creates new training instances by mixing pairs of features and their corresponding targets based on a given mixing ratio [45]. Another approach adopted from the computer vision field is SpecAugment, where features are temporally warped and blocks of the features are randomly masked [46]. Similarly, random erasing involves replacing random boxes in feature representations by random numbers [47]. In the related research task of bird audio detection, Lasseck combined several data augmentation techniques in the time domain (e.g., mosaicking random segments, time stretching, time interval dropout) and time-frequency domain (e.g., piece-wise time/frequency stretching and shifting) [48].

A second group of data augmentation techniques synthesizes novel data instances from scratch. The most common synthesis approaches are based on generative adversarial networks (GAN) [49], where class-conditioned synthesis models are trained using an adversarial training strategy by imitating existing data samples. While data synthesis is usually performed in the audio signal domain [15,50], Mun et al. instead synthesized intermediate embedding vectors [51]. Kong et al.

generated acoustic scenes using the SampleRNN model architecture [52]. Recently proposed ASC algorithms use either mixup data augmentation or GAN based methods to augment the available amount of training data.

5. Network Architectures

ASC algorithms mostly use CNN based network architectures since they usually provide a summarizing classification of longer acoustic scene excerpts. In contrast, AED algorithms commonly use convolutional recurrent neural networks (CRNN) as they focus on a precise detection of sound events [4]. This architecture combines convolutional neural networks (CNN) as the front-end for representation learning and a recurrent layer for temporal modeling. State-of-the-art ASC algorithms almost exclusively use CNN architectures. Hence, the main focus is on CNN based ASC methods in Section 5.1. Other methods using feedforward neural networks (FNN) and CRNN are briefly discussed in Section 5.2 and Section 5.3, respectively. Network architectures and the corresponding hyper-parameters are usually optimized manually. As an exception, Roletscheck et al. automated this process and compared various architectures, which were automatically generated using a genetic algorithm [53].

5.1. Convolutional Neural Networks

Traditional CNN architectures use multiple blocks of successive convolution and pooling operations for feature learning and down-sampling along the time and feature dimensions, respectively. As an alternative, Ren et al. used atrous CNNs, which are based on dilated convolutional kernels [54]. Such kernels allow achieving a comparable receptive field size without intermediate pooling operation. Koutine et al. showed that ASC systems can be improved if the receptive field is regularized by restricting its size [55].

In most CNN based architectures, only the activations of the last convolutional layer are connected to the final classification layers. As an alternative, Yang et al. followed a multi-scale feature approach and further processed the activations from all intermediate feature maps [56]. Additionally, the authors used the Xception network architecture, where the convolution operation is split into a depthwise (spatial) convolution and a pointwise (channel) convolution to reduce the number of trainable parameters. A related approach is to factorize two-dimensional convolutions into two one-dimensional kernels to model the transient and long-term characteristics of sounds separately [19,57]. The influence of different symmetric and asymmetric kernel shapes were systematically evaluated by Wang et al. [58].

Several extensions to the common CNN architecture were proposed to improve the feature learning. Basbug and Sert adapted the spatial pyramid pooling strategy from computer vision, where feature maps are pooled and combined on different spatial resolutions [59]. In order to learn frequency-aware filters in the convolutional layers, Koutini et al. proposed to encode the frequency position of each input feature bin within an additional channel dimension (frequency-aware CNNs) [44]. Similarly, Marchi et al. added the first and second order time derivative of spectrogram based features as additional input channels in order to facilitate detecting transient short-term events that have a rapid increase in magnitude [60].

5.2. Feedforward Neural Networks

Feedforward neural networks (FNN) are used in several ASC algorithms. Bisot et al. used an FNN architecture to concatenate features from an NMF decomposition and a constant-Q transform of the audio signal [61]. Takahashi et al. combined an FNN with multiple Gaussian mixture model (GMM) classifiers to model the individual acoustic scenes [62].

5.3. Convolutional Recurrent Neural Networks

The third category of ASC algorithms is based on convolutional recurrent neural networks (CRNN). Li et al. combined in two separate input branches CNN based front-ends for feature learning with bidirectional gated recurrent units (BiGRU) for temporal feature modeling [13]. In contrast to

a sequential ordering of convolutional and recurrent layers, parallel processing pipelines using long short-term memory (LSTM) layers were used in [50,63]. Two recurrent network types used in ASC systems require fewer parameters and less training data compared to LSTM layers: gated recurrent neural networks (GRNN) [11,12,64] and time-delay neural networks (TDNN) [20,65].

6. Learning Paradigms

Building on the basic neural network architectures introduced in Section 5, approaches to further improve ASC systems are summarized in this section. After discussing methods for closed/open set classification in Section 6.1, extensions to neural networks such as multiple input networks (Section 6.2) and attention mechanisms (Section 6.3) are presented. Finally, both multitask learning (Section 6.4) and transfer learning (Section 6.5) will be discussed as two promising training strategies to improve ASC systems.

6.1. Closed/Open Set Classification

Most ASC tasks in public evaluation campaigns such as the DCASE challenge assume a closed-set classification scenario with a fixed predefined set of acoustic scenes to distinguish. In real-world applications however, the underlying data distributions of acoustic scenes is often unknown and can furthermore change over time with new classes becoming relevant. This motivates the use of open-set classification approaches, where an algorithm can also classify a given audio recording as an "unknown" class. This scenario was first addressed as part of the DCASE 2019 challenge in Task 1C "Open-set Acoustic Scene Classification" [66].

Saki et al. proposed the multi-class open-set evolving recognition (MCOSR) algorithm to tackle open-set ASC [67]. Unknown samples are first rejected by a recognition model before the algorithm tries to identify underlying (hidden) classes in these samples in an unsupervised manner. Finally, the recognition model can be updated using the novel classes. Wilkinghoff and Kurth combined a closed-set classification algorithm and an outlier detection algorithm based on deep convolutional autoencoders (DCAE) to recognize unknown samples in an open-set ASC scenario [68]. Lehner et al. evaluated the model's classification confidence to identify unknown samples [69]. Therefore, a threshold was applied on the highest logit value at the input of the final neural network layer.

6.2. Multiple Input Networks

As discussed before, most ASC algorithms use a convolutional front-end to learn characteristic patterns in multi-dimensional feature representations. As a general difference from image processing, the time and frequency axes in spectrogram based feature representations do not carry the same semantic meaning. In order to train networks to detect spectral patterns, which are characteristic for certain frequency regions, several authors split a spectrogram into two [70] or multiple [71] sub-bands and used networks with multiple input branches. Using the same idea of distributed feature learning, other ASC systems individually process the left/right or mid/side channels [8] or filtered signal variants, which are obtained using harmonic/percussive separation (HPSS) [38] or nearest neighbor filtering (NNF) [37]. Instead of feeding multiple signal representations to the network as individual input branches, Dang et al. proposed to concatenate both MFCC and log Mel spectrogram features along the frequency axis as input features [72]. The majority of state-of-the-art ASC algorithms process spectrogram based input data using network architectures with a single input and output branch.

6.3. Attention

The temporal segments of an environmental audio recording contribute differently to the classification of its acoustic scene. Neural attention mechanisms allow neural networks to focus on a specific subset of its input features. Attention mechanisms can be incorporated at different positions within neural network based ASC algorithms. Li et al. incorporated gated linear units (GLU) in several steps of the feature learning part of the network ("multi-level attention") [13]. GLUs

implement pairs of mutually gating convolutional layers to control the information flow in the network. Attention mechanisms can also be applied in the pooling of feature maps [73]. Wang et al. used self-determination CNNs (SD-CNNs) to identify frames with higher uncertainty due to overlapping sound events. A neural network can learn to focus on local patches within the receptive field if a network-in-network architecture is used [74]. Here, individual convolutional layers are extended by micro neural networks, which allow for more powerful approximations by additional non-linearities. Up to now, attention mechanisms have been rarely used in ASC algorithms, but often applied in AED algorithms, where the exact localization of sound events is crucial.

6.4. Multitask Learning

Multitask learning involves learning to solve multiple related classification tasks jointly with one network [75]. By learning shared feature representations, the performance on the individual tasks can be improved and a better generalization can be achieved.

A natural approach is to train one model to perform ASC and AED in a joint manner [76] as acoustic events are the building blocks of acoustic scenes. Sound events and acoustic scenes naturally follow a hierarchical relationship. While most publications perform a "flat" classification, Xu et al. exploited a hierarchical acoustic scene taxonomy and group acoustic scenes to the three high-level scene classes "vehicle", "indoor", and "outdoor" [77]. The authors used a hierarchical pre-training approach, where the network learned to predict the high-level scene class as the main task and the low-level scene class such as car or tram as the auxiliary task. Using a similar scene grouping approach, Nwe et al. trained a CNN with several shared convolutional layers and three three branches of task-specific convolutional layers to predict the most likely acoustic scene within each scene group [78]. So far, multitask learning is not common in state-of-the-art ASC algorithms.

6.5. Transfer Learning

Many ASC algorithms rely on well-proven neural network architectures from the computer vision domain such as AlexNet [73,79], VGG16 [12], Xception [56], DenseNet [44], GoogLeNet [79], and Resnet [38,69]. Transfer learning allows the fine-tuning of models that are pretrained on related audio classification tasks. For instance, Huang et al. used the AudioSet dataset to pretrain four different neural network architectures and fine-tune them using a task-specific development set [26]. Similarly, Singh et al. took a pretrained SoundNet [80] network as basis for their experiments [27,81]. Ren et al. used the VGG16 model as the seed model, which was pre-trained for object recognition in images [12]. Kumar et al. pre-trained a CNN in a supervised fashion using weak label annotation of the AudioSet dataset. The authors compared three transfer learning strategies to adapt the model to novel AED and ASC target tasks [82]. In the ASC tasks of the DCASE challenge, the use of publicly-available external ASC datasets is explicitly allowed. However, most recently winning algorithms did not use transfer learning from these datasets, but focused instead on the provided training datasets combined with data augmentation techniques (see Section 4).

6.6. Result Fusion

Many ASC algorithms include result fusion steps where intermediate results from different time frames or classifiers are merged. Similar to computer vision, features learned in different layers of the network capture different levels of abstraction of the audio signal. Therefore, some systems apply early fusion and combine intermediate feature representations from different layers of the network as multiscale features [27,56,81]. Ensemble learning is a common late fusion technique where the prediction results of multiple classifiers are combined [8,12,22,26,37]. The predicted class scores can be averaged [83] or used as features for an additional classifier [84]. Averaging the classification results over multiple segments, as well as over multiple classifiers has become an essential part of state-of-the-art ASC algorithms. However, in real-life applications, using multiple models for

an averaged decision is often not feasible due to the available computational resources and processing time constraints.

7. Open Challenges

This section discusses several open challenges that arise from deploying ASC algorithms to real-world application scenarios.

7.1. Domain Adaptation

The performance of sound event classification algorithms often suffers from covariate shift, i.e., a distribution mismatch between training and test datasets. When being deployed in real-world application scenarios, ASC systems usually face novel acoustic conditions that are caused by different recording devices or environmental influences. Domain adaption methods aim to increase the robustness of classification algorithms in such scenarios by adapting them to data from a novel target domain [85]. Depending on whether labels exist for the target domain data, supervised and unsupervised methods are distinguished. Supervised domain adaptation usually involves fine-tuning a model on new target domain data after it was pre-trained on the annotated source domain data.

One unsupervised domain adaptation strategy is to alter the target domain data such that their distribution becomes closer to that of the source domain data. As an example, Kosmider used "spectral correction" to compensate for different frequency responses of the recording equipment. He estimated a set of frequency-dependent magnitude coefficients from the source domain data and used them for spectrogram equalization of the target domain data [86]. Despite its simplicity, this approach led to the best classification results in Task 1B "Acoustic Scene Classification with mismatched recording devices" of the 2019 DCASE Challenge. However, it required time-aligned audio recordings from several microphones during the training phase. Mun and Shon performed an independent domain adaptation of both the source and target domain to an additional domain using a factorized hierarchical variational autoencoder [87].

As a second unsupervised strategy, Gharib et al. used an adversarial training approach such that the intermediate feature mappings of an ASC model followed a similar distribution for both the source and target domain data [85]. This approach was further improved using the Wasserstein generative adversarial networks (WGAN) formulation [88]. When analyzing the best-performing systems in the microphone mismatch ASC Task (1B) of the 2019 DCASE challenge, it becomes apparent that combining approaches for domain adaptation and data augmentation jointly improved the robustness of ASC algorithms against changes of the acoustic conditions.

7.2. Ambiguous Allocation between Sound Events and Scenes

Acoustic scenes often comprise multiple sound events, which are not class-specific, but instead appear in a similar way in various scene classes [2,21]. As an example, sound recordings that are recorded in different vehicle types such as car, tram, or train often exhibit prominent speech from human conversations or automatic voice announcements. At the same time, class-specific sound components like engine noises, road surface sounds, or door opening and closing sounds appear at a lower level in the background. Wu and Lee used the gradient-weighted class activation mappings (GradCAM) to show that CNN based ASC models in general have the capability to ignore high-energy sound events and focus on quieter background sounds instead [35].

7.3. Model Interpretability

Despite their superior performance, deep learning based ASC models are often considered as "black boxes" due to their high complexity and large number of parameters. One main challenge is to develop methods that allow better interpreting the model predictions and internal feature representations. As discussed in Section 6.3, attention mechanisms allow neural networks to

focus on relevant subsets of the input data. Wang et al. investigated an attention based ASC model and demonstrated that only fractions of long-term scene recordings were relevant for its classification [74]. Similarly, Ren et al. visualized internal attention matrices obtained for different acoustic scenes [54]. The results confirmed that either stationary and short-term signal components were most relevant for particular acoustic scenes.

Another common strategy to investigate the class separability in intermediate feature representations are dimension reduction techniques such as t-SNE [27]. Techniques such as layer-wise relevance propagation (LRP) [89] allow interpreting neural networks by investigation the pixel-wise contributions of input features to classification decisions.

7.4. Real-World Deployment

Many challenges arise when ASC models are deployed in smart city [90,91] or industrial sound analysis [92] scenarios. The first challenge is the model complexity, which is limited if data privacy concerns require the classification to be performed directly on mobile sensor devices. Real-time processing requirements often demand for fast model prediction with low latency. In a related study, Sigtia et al. contrasted the performance of different audio event detection methods with the respective computational costs [93]. In comparison to traditional methods such as support vector machines (SVM) and Gaussian mixture models (GMM), fully-connected neural networks achieve the best performance while requiring the lowest number of operations.

This often requires a model compression step, where trained classification models are reduced in size and redundant components need to be identified. Several attempts were made to make networks more compact by decreasing the number of operations and increasing the memory efficiency. For instance, the MobileNetV2 architecture is based on a novel layer module, which mainly uses convolutional operations to avoid large intermediate tensors [94]. A similar approach was followed by Drossos et al. for AED [95]. The authors replaced the common CNN based front-end by depthwise separable convolutions and the RNN backend with dilated convolutions to reduce the number of model parameters and required training time. In the MorphNet approach proposed by Gordon et al., network layers were shrunk and expanded in an iterative procedure to optimize network architectures in order to match given resource constraints [96]. Finally, Tan and Lee showed in the EfficientNet approach that a uniform scaling of the network dimensions depth, width, and resolution of a convolutional neural network led to highly effective networks [97].

A second challenge arises from the audio recording devices in mobile sensor units. Due to space constraints, microelectro-mechanical system (MEMS) microphones are often used. However, scientific datasets used for training ASC models are usually recorded with high-quality electret microphones [98]. As discussed in Section 7.1, changed recording conditions affect the input data distribution. Achieving robust classification systems in such a scenario requires the application of domain adaptation strategies.

8. Conclusions and Future Directions

In the research field of acoustic scene classification, a rapid increase of scientific publications has been observed in the last decade. This progress was mainly stimulated by recent advances in the field of deep learning such as transfer learning, attention mechanisms, and multitask learning, as well as the release of various public datasets. The DCASE community plays a major role in this development by organizing annual evaluation campaigns on several machine listening tasks.

State-of-the-art ASC algorithms have matured and can be applied in context-aware devices such as hearables and wearables. In such real-world application scenarios, novel challenges need to be faced such as microphone mismatch and domain adaptation, open-set classification, as well as model complexity and real-time processing constraints. As a consequence, one future challenge is to extend current evaluation campaign tasks to evaluate the classification performance, as well as the computational requirements of submitted ASC algorithms jointly. This issue will be first addressed in this year's DCASE challenge task "Low-Complexity Acoustic Scene Classification ".

The general demand for deep learning based classification algorithms for larger training corpora can be faced with novel techniques from unsupervised and self-supervised learning, as is shown in natural language processing, speech processing, and image processing. Another interesting future direction is the application of lifelong learning capabilities to ASC algorithms [99]. In many real-life scenarios, autonomous agents continuously process the sound of their environment and need to be adaptable to classify novel sounds while maintaining knowledge about previously learned acoustic scenes and events.

Funding: This work has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement. No 786993 and was supported by the German Research Foundation (AB 675/2-1).

Acknowledgments: The author would like to thank Hanna Lukashevich, Stylianos Mimilakis, David S. Johnson, and Sascha Grollmisch for valuable discussions and proof-reading, as well as the anonymous reviewers whose comments greatly improved this manuscript.

Conflicts of Interest: The author declares no conflict of interest.

References

1. *Computational Analysis of Sound Scenes and Events*; Virtanen, T., Plumbley, M.D., Ellis, D., Eds.; Springer International Publishing: Berlin, Germany, 2018; doi:10.1007/978-3-319-63450-0. [CrossRef]
2. Mesaros, A.; Heittola, T.; Virtanen, T. Assessment of Human and Machine Performance in Acoustic Scene Classification: DCASE 2016 Case Study. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017; pp. 319–323.
3. Barchiesi, D.; Giannoulis, D.D.; Stowell, D.; Plumbley, M.D. Acoustic Scene Classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34, doi:10.1109/MSP.2014.2326181. [CrossRef]
4. Xia, X.; Togneri, R.; Sohel, F.; Zhao, Y.; Huang, D. A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection. In *Circuits, Systems, and Signal Processing*; Springer: Berlin, Germany, 2019; pp. 3433–3453, doi:10.1007/s00034-019-01094-1. [CrossRef]
5. Dang, A.; Vu, T.H.; Wang, J.C. A survey of Deep Learning for Polyphonic Sound Event Detection. In Proceedings of the International Conference on Orange Technologies (ICOT), Singapore, 8–10 December 2017; pp. 75–78, doi:10.1109/ICOT.2017.8336092. [CrossRef]
6. Mesaros, A.; Heittola, T.; Benetos, E.; Foster, P.; Lagrange, M.; Virtanen, T.; Plumbley, M.D. Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 379–393, doi:10.1109/TASLP.2017.2778423. [CrossRef]
7. Han, Y.; Park, J.; Lee, K. Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
8. Mars, R.; Pratik, P.; Nagisetty, S.; Lim, C. Acoustic Scene Classification from Binaural Signals using Convolutional Neural Networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 149–153, doi:10.33682/6c9z-gd15. [CrossRef]
9. Green, M.C.; Murphy, D. Acoustic Scene Classification using Spatial Features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
10. Zieliński, S.K.; Lee, H. Feature Extraction of Binaural Recordings for Acoustic Scene Classification. In Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS). Poznań, Poland, 9–12 September 2018; pp. 585–588, doi:10.15439/2018F182. [CrossRef]
11. Qian, K.; Ren, Z.; Pandit, V.; Yang, Z.; Zhang, Z.; Schuller, B. Wavelets Revisited for the Classification of Acoustic Scenes. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
12. Ren, Z.; Pandit, V.; Qian, K.; Yang, Z.; Zhang, Z.; Schuller, B. Deep Sequential Image Features for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.

13. Li, Z.; Hou, Y.; Xie, X.; Li, S.; Zhang, L.; Du, S.; Liu, W. Multi-Level Attention Model with Deep Scattering Spectrum for Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Shanghai, China, 8–12 July 2019; pp. 396–401, doi:10.1109/ICMEW.2019.00074. [CrossRef]
14. Chen, H.; Zhang, P.; Bai, H.; Yuan, Q.; Bao, X.; Yan, Y. Deep convolutional neural network with scalogram for audio scene modeling. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Hyderabad, India, 2–6 September 2018; pp. 3304–3308, doi:10.21437/Interspeech.2018-1524. [CrossRef]
15. Chen, H.; Liu, Z.; Liu, Z.; Zhang, P.; Yan, Y. Integrating the Data Augmentation Scheme with Various Classifiers for Acoustic Scene Modeling. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
16. Ye, J.; Kobayashi, T.; Toyama, N.; Tsuda, H.; Murakawa, M. Acoustic scene classification using efficient summary statistics and multiple spectro-temporal descriptor fusion. *Appl. Sci.* **2018**, *8*, 1–12, doi:10.3390/app8081363. [CrossRef]
17. Li, Y.; Li, X.; Zhang, Y.; Wang, W.; Liu, M.; Feng, X. Acoustic Scene Classification Using Deep Audio Feature and BLSTM Network. In Proceedings of the 6th International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, China, 16–17 July 2018; pp. 371–374, doi:10.1109/ICALIP.2018.8455765. [CrossRef]
18. Bisot, V.; Essid, S.; Richard, G. HOG and Subband Power Distribution Image Features for Acoustic Scene Classification. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 719–723, doi:10.1109/EUSIPCO.2015.7362477. [CrossRef]
19. Sharma, J.; Granmo, O.C.; Goodwin, M. Environment Sound Classification using Multiple Feature Channels and Deep Convolutional Neural Networks. *arXiv* **2019**, *14*, 1–11.
20. Moritz, N.; Schröder, J.; Goetze, S.; Anemüller, J.; Kollmeier, B. Acoustic Scene Classification using Time-Delay Neural Networks and Amplitude Modulation Filter Bank Features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Budapest, Hungary, 3 September 2016.
21. Park, S.; Mun, S.; Lee, Y.; Ko, H. Acoustic Scene Classification Based on Convolutional Neural Network using Double Image Features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
22. Fonseca, E.; Gong, R.; Bogdanov, D.; Slizovskaia, O.; Gomez, E.; Serra, X. Acoustic Scene Classification by Ensembling Gradient Boosting Machine and Convolutional Neural Networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
23. Maka, T. Audio Feature Space Analysis for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
24. Abidin, S.; Togneri, R.; Sohel, F. Enhanced LBP Texture Features from Time Frequency Representations for Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 626–630.
25. Jiménez, A.; Elizalde, B.; Raj, B. DCASE 2017 Task 1: Acoustic Scene Classification using Shift-Invariant Kernels and Random Features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
26. Huang, J.; Lu, H.; Lopez-Meyer, P.; Maruri, H.A.C.; Ontiveros, J.A.d.H. Acoustic Scene Classification using Deep Learning-Based Ensemble Averaging. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 94–98.
27. Singh, A.; Rajan, P.; Bhavsar, A. Deep Multi-View Features from Raw Audio for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 229–233.
28. Chen, H.; Zhang, P.; Yan, Y. An Audio Scene Classification Framework with Embedded Filters and a DCT-Based Temporal Module. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 835–839.

29. Amiriparian, S.; Freitag, M.; Cummins, N.; Gerczuk, M.; Pugachevskiy, S.; Schuller, B. A Fusion of Deep Convolutional Generative Adversarial Networks and Sequence to Sequence Autoencoders for Acoustic Scene Classification. In Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 977–981, doi:10.23919/EUSIPCO.2018.8553225. [CrossRef]
30. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Feature Learning with Matrix Factorization Applied to Acoustic Scene Classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1216–1229, doi:10.1109/TASLP.2017.2690570. [CrossRef]
31. Benetos, E.; Lagrange, M.; Dixon, S. Characterisation of Acoustic Scenes using a Temporally-Constrained Shift-Invariant Model. In Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK, 17–21 September 2012; pp. 1–7.
32. Seo, H.; Park, J.; Park, Y. Acoustic Scene Classification using Various Pre-Processed Features and Convolutional Neural Networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 3–6.
33. Wang, Y.; Getreuer, P.; Hughes, T.; Lyon, R.F.; Saurous, R.A. Trainable Frontend for Robust and Far-Field Keyword Spotting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5670–5674, doi:10.1109/ICASSP.2017.7953242. [CrossRef]
34. Lostanlen, V.; Salamon, J.; Cartwright, M.; McFee, B.; Farnsworth, A.; Kelling, S.; Bello, J.P. Per-channel energy normalization: Why and how. *IEEE Signal Process. Lett.* **2019**, *26*, 39–43, doi:10.1109/LSP.2018.2878620. [CrossRef]
35. Wu, Y.; Lee, T. Enhancing Sound Texture in CNN based Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 815–819, doi:10.1109/ICASSP.2019.8683490. [CrossRef]
36. Rafii, Z.; Pardo, B. Music/Voice Separation using the Similarity Matrix. In Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, 8–12 October 2012; pp. 583–588.
37. Nguyen, T.; Pernkopf, F. Acoustic Scene Classification using a Convolutional Neural Network Ensemble and Nearest Neighbor Filters. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
38. Mariotti, O.; Cord, M.; Schwander, O. Exploring Deep Vision Models for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
39. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252, doi:10.1007/s11263-015-0816-y. [CrossRef]
40. 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. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
41. Abeßer, J.; Mimilakis, S.I.; Gräfe, R.; Lukashevich, H. Acoustic Scene Classification By Combining Autoencoder-Based Dimensionality Reduction and Convolutional Neural Networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
42. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283, doi:10.1109/LSP.2017.2657381. [CrossRef]
43. Xu, J.X.; Lin, T.C.; Yu, T.C.; Tai, T.C.; Chang, P.C. Acoustic Scene Classification Using Reduced MobileNet Architecture. In Proceedings of the IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 10–12 December 2018; pp. 267–270, doi:10.1109/ISM.2018.00038. [CrossRef]
44. Koutini, K.; Eghbal-zadeh, H.; Widmer, G. Receptive-Field-Regularized CNN Variants for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 124–128.

45. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
46. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 2–15 November 2019; Volume 2019, pp. 2613–2617, doi:10.21437/Interspeech.2019-2680. [CrossRef]
47. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *arXiv* **2017**, arXiv:1708.04896.
48. Lasseck, M. Acoustic bird detection with deep convolutional neural networks. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE), Surrey, UK, 19–20 November 2018; pp. 143–147.
49. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; pp. 2672–2680.
50. Mun, S.; Shon, S.; Kim, W.; Han, D.K.; Ko, H. Deep Neural Network Based Learning and Transferring Mid-Level Audio Features for Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 796–800, doi:10.1097/IOP.0000000000000348. [CrossRef]
51. Mun, S.; Park, S.; Han, D.K.; Ko, H. Generative Adversarial Networks based Acoustic Scene Training Set Augmentation and Selection using SVM Hyperplane. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
52. Kong, Q.; Xu, Y.; Iqbal, T.; Cao, Y.; Wang, W.; Plumbley, M.D. Acoustic Scene Generation with Conditional SampleRNN. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 925–929.
53. Roletscheck, C.; Watzka, T.; Seiderer, A.; Schiller, D.; André, E. Using an Evolutionary Approach To Explore Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
54. Ren, Z.; Kong, Q.; Han, J.; Plumbley, M.D.; Schuller, B.W. Attention based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 56–60, doi:10.1109/ICASSP.2019.8683434. [CrossRef]
55. Koutini, K.; Eghbal-zadeh, H.; Widmer, G.; Kepler, J. CP-JKU Submissions to DCASE'19: Acoustic Scene Classification and Audio Tagging with REceptive-Field-Regularized CNNs. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 1–5.
56. Yang, L.; Chen, X.; Tao, L. Acoustic Scene Classification using Multi-Scale Features. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
57. Cho, J.; Yun, S.; Park, H.; Eum, J.; Hwang, K. Acoustic Scene Classification Based on a Large-Margin Factorized CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 45–49, doi:10.33682/8xh4-jm46. [CrossRef]
58. Wang, C.Y.; Wang, J.C.; Wu, Y.C.; Chang, P.C. Asymmetric Kernel Convolution Neural Networks for Acoustic Scenes Classification. In Proceedings of the IEEE International Symposium on Consumer Electronics (ISCE), Kuala Lumpur, Malaysia, 14–15 November 2017; pp. 11–12.
59. Basbug, A.M.; Sert, M. Acoustic Scene Classification Using Spatial Pyramid Pooling with Convolutional Neural Networks. In Proceedings of the 13th IEEE International Conference on Semantic Computing (ICSC), Newport, CA, USA, 30 January–1 February 2019; pp. 128–131, doi:10.1109/ICOSC.2019.8665547. [CrossRef]
60. Marchi, E.; Tonelli, D.; Xu, X.; Ringeval, F.; Deng, J.; Squartini, S.; Schuller, B. Pairwise Decomposition with Deep Neural Networks and Multiscale Kernel Subspace Learning for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Budapest, Hungary, 3 September 2016.

61. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Nonnegative Feature Learning Methods for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
62. Takahashi, G.; Yamada, T.; Ono, N.; Makino, S. Performance Evaluation of Acoustic Scene Classification using DNN-GMM and Frame-Concatenated Acoustic Features. In Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Honolulu, HI, USA, 2–15 November 2018; pp. 1739–1743, doi:10.1109/APSIPA.2017.8282314. [CrossRef]
63. Bae, S.H.; Choi, I.; Kim, N.S. Acoustic Scene Classification using Parallel Combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Budapest, Hungary, 3 September 2016.
64. Zöhrer, M.; Pernkopf, F. Gated Recurrent Networks Applied to Acoustic Scene Classification and Acoustic Event Detection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Budapest, Hungary, 3 September 2016.
65. Jati, A.; Nadarajan, A.; Mundnich, K.; Narayanan, S. Characterizing dynamically varying acoustic scenes from egocentric audio recordings in workplace setting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
66. Mesaros, A.; Heittola, T.; Virtanen, T. Acoustic Scene Classification in DCASE 2019 Challenge:Closed and Open Set Classification and Data Mismatch Setups. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 164–168, doi:10.33682/m5kp-fa97. [CrossRef]
67. Saki, F.; Guo, Y.; Hung, C.Y. Open-Set Evolving Acoustic Scene Classification System. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 219–223.
68. Wilkinghoff, K.; Frank Kurth. Open-Set Acoustic Scene Classification with Deep Convolutional Autoencoders. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 258–262.
69. Lehner, B.; Koutini, K.; Schwarzlmüller, C.; Gallien, T.; Widmer, G. Acoustic Scene Classification with Reject Option based on Resnets. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
70. Mcdonnell, M.D.; Gao, W. Acoustic Scene Classification Using Deep Residual Networks With Late Fusion of Separated High and Low Frequency Paths. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
71. Phaye, S.S.R.; Benetos, E.; Wang, Y. Subspectralnet—Using Sub-Spectrogram based Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 825–829.
72. Dang, A.; Vu, T.H.; Wang, J.C. Acoustic Scene Classification using Convolutional Neural Networks and Multi-Scale Multi-Feature Extraction. In Proceedings of the IEEE International Conference on Consumer Electronics (ICCE), Hue City, Vietnam, 18–20 July 2018, doi:10.1109/ICCE.2018.8326315. [CrossRef]
73. Ren, Z.; Kong, Q.; Qian, K.; Plumbley, M.D.; Schuller, B.W. Attention based Convolutional Neural Networks for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
74. Wang, C.Y.; Santoso, A.; Wang, J.C. Acoustic Scene Classification using Self-Determination Convolutional Neural Network. In Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Honolulu, HI, USA, 2–15 November 2018; pp. 19–22, doi:10.1109/APSIPA.2017.8281995. [CrossRef]
75. Ruder, S. An Overview of Multi-Task Learning in Deep Neural Networks. *arXiv* **2017**, arXiv:1706.05098.
76. Bear, H.L.; Nolasco, I.; Benetos, E. Towards joint sound scene and polyphonic sound event recognition. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), Graz, Austria, 2–15 November 2019; Volume 2019, pp. 4594–4598, doi:10.21437/Interspeech.2019-2169. [CrossRef]
77. Xu, Y.; Huang, Q.; Wang, W.; Plumbley, M.D. Hierarchical Learning for DNN-Based Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Budapest, Hungary, 3 September 2016.

78. Nwe, T.L.; Dat, T.H.; Ma, B. Convolutional Neural Network with Multi-Task Learning Scheme for Acoustic Scene Classification. In Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Honolulu, HI, USA, 2–15 November 2018; pp. 1347–1350, doi:10.1109/APSIPA.2017.8282241. [CrossRef]
79. Boddapati, V.; Petef, A.; Rasmusson, J.; Lundberg, L. Classifying environmental sounds using image recognition networks. *Proc. Comput. Sci.* **2017**, *112*, 2048–2056, doi:10.1016/j.procs.2017.08.250. [CrossRef]
80. Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems (NIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2016; pp. 892–900.
81. Singh, A.; Thakur, A.; Rajan, P.; Bhavsar, A. A Layer-Wise Score Level Ensemble Framework for Acoustic Scene Detection. In Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018; pp. 837–841, doi:10.23919/EUSIPCO.2018.8553052. [CrossRef]
82. Kumar, A.; Khadkevich, M.; Fugen, C. Knowledge Transfer from Weakly Labeled Audio Using Convolutional Neural Network for Sound Events and Scenes. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Alberta, AB, Canada, 15–20 April 2018; pp. 326–330, doi:10.1109/ICASSP.2018.8462200. [CrossRef]
83. Zeinali, H.; Burget, L.; Cernocky, J. Convolutional Neural Networks and X-Vector Embeddings for DCASE2018 Acoustic Scene Classification Challenge. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
84. Weiping, Z.; Jiantao, Y.; Xiaotao, X.; Xiangtao, L.; Shaohu, P. Acoustic Scene Classification using Deep Convolutional Neural Networks and Multiple Spectrogram Fusions. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Munich, Germany, 16–17 November 2017.
85. Gharib, S.; Drossos, K.; Emre, C.; Serdyuk, D.; Virtanen, T. Unsupervised Adversarial Domain Adaptation for Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
86. Kosmider, M. Calibrating Neural Networks for Secondary Recording Devices. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
87. Mun, S.; Shon, S. Domain Mismatch Robust Acoustic Scene Classification Using Channel Information Conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 845–849, doi:10.1109/ICASSP.2019.8683514. [CrossRef]
88. Drossos, K.; Magron, P.; Virtanen, T. Unsupervised Adversarial Domain Adaptation based on the Wasserstein Distance for Acoustic Scene Classification. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 259–263.
89. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **2015**, *10*, 1–46, doi:10.1371/journal.pone.0130140. [CrossRef] [PubMed]
90. Bello, J.P.; Silva, C.; Nov, O.; DuBois, R.L.; Arora, A.; Salamon, J.; Mydlarz, C.; Doraiswamy, H. SONYC: A System for the Monitoring, Analysis and Mitigation of Urban Noise Pollution. *Commun. ACM (CACM)* **2019**, *62*, *68–77*. [CrossRef]
91. Abeßer, J.; Götze, M.; Clauß, T.; Zapf, D.; Kühn, C.; Lukashevich, H.; Kühnlenz, S.; Mimilakis, S. Urban Noise Monitoring in the Stadtlärm Project—A Field Report. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019.
92. Grollmisch, S.; Abeßer, J.; Liebetrau, J.; Lukashevich, H. Sounding Industry: Challenges and Datasets for Industrial Sound Analysis (ISA). In Proceedings of the 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019; pp. 1–5.
93. Sigtia, S.; Stark, A.M.; Krstulović, S.; Plumbley, M.D. Automatic Environmental Sound Recognition: Performance Versus Computational Cost. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 2096–2107, doi:10.1109/TASLP.2016.2592698. [CrossRef]
94. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520, doi:10.1109/CVPR.2018.00474. [CrossRef]

95. Drossos, K.; Mimilakis, S.I.; Gharib, S.; Li, Y.; Virtanen, T. Sound Event Detection with Depthwise Separable and Dilated Convolutions. *arXiv* **2020**, arXiv:2002.00476.
96. Gordon, A.; Eban, E.; Nachum, O.; Chen, B.; Wu, H.; Yang, T.J.; Choi, E. MorphNet: Fast & Simple Resource-Constrained Structure Learning of Deep Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018 ; pp. 1586–1595, doi:10.1109/CVPR.2018.00171. [CrossRef]
97. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, 9–15 June 2019.
98. Mesaros, A.; Heittola, T.; Tuomas Virtanen. A Multi-Device Dataset for Urban Acoustic Scene Classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), Surrey, UK, 19–20 November 2018.
99. Parisi, G.I.; Kemker, R.; Part, J.L.; Kanan, C.; Wermter, S. Continual Lifelong Learning with Neural Networks: A Review. *Neural Netw.* **2019**, *113*, 54–71. [CrossRef] [PubMed]