Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks

Shahin Amiriparian, Maurice Gerczuk, Sandra Ottl, Lukas Stappen, Alice Baird, Lukas Koebe, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Amiriparian, Shahin, Maurice Gerczuk, Sandra Ottl, Lukas Stappen, Alice Baird, Lukas Koebe, and Björn Schuller. 2020. "Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks." *EURASIP Journal on Audio, Speech, and Music Processing* 2020 (1): 19. https://doi.org/10.1186/s13636-020-00186-0.

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks

Shahin Amiriparian[1*], Maurice Gerczuk[1], Sandra Ottl[1], Lukas Stappen[1], Alice Baird[1], Lukas Koebe[1] and Björn Schuller[1,2]

Abstract

In this paper, we investigate the performance of two deep learning paradigms for the audio-based tasks of acoustic scene, environmental sound and domestic activity classification. In particular, a convolutional recurrent neural network (CRNN) and pre-trained convolutional neural networks (CNNs) are utilised. The CRNN is directly trained on Mel-spectrograms of the audio samples. For the pre-trained CNNs, the activations of one of the top layers of various architectures are extracted as feature vectors and used for training a linear support vector machine (SVM). Moreover, the predictions of the two models—the class probabilities predicted by the CRNN and the decision function of the SVM—are combined in a decision-level fusion to achieve the final prediction. For the pre-trained CNN networks we use as feature extractors, we further evaluate the effects of a range of configuration options, including the choice of the pre-training corpus. The system is evaluated on the acoustic scene classification task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017) workshop, ESC-50 and the multi-channel acoustic recordings from DCASE 2018, task 5. We have refrained from additional data augmentation as our primary goal is to analyse the general performance of the proposed system on different datasets. We show that using our system, it is possible to achieve competitive performance on all datasets and demonstrate the complementarity of CRNNs and ImageNet pre-trained CNNs for acoustic classification tasks. We further find that in some cases, CNNs pre-trained on ImageNet can serve as more powerful feature extractors than AudioSet models. Finally, ImageNet pre-training is complimentary to more domain-specific knowledge, either in the form of the convolutional recurrent neural network (CRNN) trained directly on the target data or the AudioSet pre-trained models. In this regard, our findings indicate possible benefits of applying cross-modal pre-training of large CNNs to acoustic analysis tasks.

Keywords: Domestic activity classification, Deep learning, Convolutional recurrent neural networks, Deep spectrum, Decision-level fusion

*Correspondence: shahin.amiriparian
[1] Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany
Full list of author information is available at the end of the article

1 Introduction

We are regularly surrounded by dynamic audio events, from which some are quite pleasant, such as singing birds or nice music tracks, other less so, like the sound of a chainsaw or a siren. Even at a young age, humans have the ability to analyse and understand a large number of audio activities and the interconnections between them, whilst filtering out a wide range of distractions [1]. In the era of machine learning, computer audition systems for intelligent housing systems [2, 3], recognition of acoustic scenes [4, 5] and sound event detection [4, 6, 7] are being developed. Therefore, it is essential for such systems to perform with high accuracy in real-world conditions. Despite recent developments in the field of audio analysis, contemporary machine learning systems are still facing a major challenge to perform the mentioned tasks with human-like precision. Moreover, deep learning-based technologies lack a mechanism to generalise well when faced with data scarcity problems. In this regard, we follow a threefold strategy by (i) proposing a cross-modal transfer learning strategy in the form of ImageNet pre-trained convolutional neural networks (CNNs) to cope with the limited data challenges, (ii) utilising a CRNN for learning tempo-spatial characteristics of audio signals, and (iii) fusing various neural network strategies to check for further performance improvements.

In particular, we investigate the performance of our methodologies to solve a 9-class audio-based classification problem of daily activities performed in a domestic environment [8], and further evaluate the system for acoustic scene and environmental sound classification.

Recently, Vecchiotti et al. [9] demonstrated the efficacy of CNNs for the task of voice activity detection in a multipurpose domestic environment, and Versperini et al. [10] showed that CNNs can achieve great performance when applied to the detection of rare audio events. At the same time, recurrent neural networks (RNNs) have been widely utilised in order to model the sequential nature of audio data and capture their long-term temporal dependencies [11–15]. With respect to the above-mentioned literature, we propose our hybrid CRNN approach to obtain representations from both CNNs and RNNs. It is worth mentioning that CRNNs, which have been first proposed for document classification [16], are considered as state-of-the-art in various audio recognition tasks, including music classification [17], acoustic event detection (AED) [18] and recognition of specific acoustic vocalisation [19]. Furthermore, they have been successfully applied for speech enhancement [20] and detection of rare audio events, for example, in smart home systems [7].

In addition to our proposed CRNN system, we investigate the efficacy of a transfer learning approach by utilising VGG16 and VGG19 [21], ResNet [22] and DenseNet [23] models for the aforementioned audio classification problem [8, 24]. These models are popular CNN architectures pre-trained on the ImageNet corpus [25]. The main reason behind using pre-trained CNNs is the robust performance that such systems have found across various audio classification and recognition tasks [26, 27]. We further want to investigate if the features learnt for the task of visual object recognition can provide additional information for acoustic scene classification complimentary to training a deep CRNN model on the audio data from scratch. For this, we implemented a late fusion strategy based on support vector machine (SVM) classifiers which are trained on the predictions obtained from our two systems. Finally, we compare ImageNet pre-training to random weight initialisation and models trained on large-scale audio classification tasks in the form of openl3 models [28, 29] and PANNs [30].

The remainder of this paper is organised as follows. In the proceeding section, the datasets used in our experiments are presented. Then, the structure of our proposed framework is introduced in Section 3. Afterwards, the experimental results are discussed and analysed in Section 4. Finally, conclusions and future work plans are given in Section 5.

2 Datasets

We evaluate our proposed systems on three datasets. The first set originates from the “Monitoring of domestic activities based on multi-channel acoustics” task of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2018)[1] [8, 24]. It contains audio data labelled with the particular domestic activity occurring in the recording. The data has been recorded with 7 microphone arrays, each consisting of four linearly arranged microphones. Those microphone arrays were placed in a studio sized holiday home and the person living there was continuously recorded for the period of 1 week. The continuous recordings were then split into 72,984 single audio segments of 10-s length and labelled with 9 different activities (absence, cooking, dish washing, eating, other, social activity, vacuum cleaning, watching TV and working). Segments containing more than one household activity were discarded. The development data of the challenge consists of audio samples recorded by four microphone arrays at different locations. For the evaluation, partition data of seven microphone arrays is used, consisting of the four microphone arrays available in the development partition, and three unknown microphone arrays [8]. We use the exact setup as provided by the challenge organisers. For detailed information about this dataset, the interested reader is referred to [8, 24].

[1] http://dcase.community/challenge2018/task-monitoring-domestic-activities

Further, we show the efficacy of the proposed fusion approach on two additional datasets: the acoustic scene classification challenge (task 1) of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017) workshop [31] and the environmental sound classification dataset ESC-50 [32]. DCASE 2017 contains 4680 10-s audio samples of 15 distinct acoustic scenes in the development partition and another 1620 samples for model evaluation. Furthermore, a cross-validation setup is provided for the development partition which we also use for our experiments. ESC-50's 2000 samples of environmental sounds are spread evenly across 50 categories. As for DCASE, a cross-validation setup is also given. In order to have a similar setup for our experiments, we use four of the five folds during training and development while setting the fifth fold aside for evaluation. This allows us to optimise model parameters using 4-fold cross-validation and afterwards test the best configurations on unseen data.

3 Methods and experimental settings

An overview of our deep learning framework is given in Fig. 1. First, Mel-spectrograms are extracted from the audio data (cf. Section 3.1). After this, the extracted spectrograms are forwarded through the CRNN (cf. Section 3.2) and Deep Spectrum (cf. Section 3.3) systems. Subsequently, our CRNN is trained on these Mel-spectrograms, and deep feature representations are extracted by a range of CNN networks which serve as input for SVM classification. Finally, in a decision-level fusion, the results achieved by different configurations are fused (cf. Section 3.4). We have decided to choose SVM classifiers for our experiments as they have consistently performed well on Deep Spectrum features [1, 26, 27] and are very efficient in high-dimensional feature space [33].

3.1 Spectrogram extraction

To create the Mel-spectrograms from the audio data, we apply periodic Hann windows with length 0.32 s and overlap 0.16 s. From these, we then compute 128 of log-scaled Mel-frequency bands. Mel-spectra features have been shown to be useful for audio tasks, such as speech processing and acoustic scene classification [14, 19, 27, 34]. The Mel-spectra are then normalised, so that the maximum amplitude is at 0 dB. In our initial experiments on DCASE 2018, we also clip the spectrograms at different amplitudes—− 30 dB, − 45 dB and − 60 dB—to minimise the effect of background noise and eliminate higher amplitude signals that are not correlated with the class of the audio recordings.

3.2 CRNN framework

As indicated in Section 1, deep models trained by CNNs and RNNs are suitable for AED and an array of other audio classification tasks. CNNs are trained by learning filters that are shifted in time and frequency. This automatically enables them to extract high-level features that are shift-invariant in both the frequency and time axes [35, 36]. This also means that those features will mostly contain short-term temporal context. Due to the inherent nature of CNNs, the ability to extract long-term temporal context is limited. In contrast, an RNN can extract long-term temporal features and struggles to capture short-term and shift-invariant information [37].

The advantages of CNNs and RNNs can be leveraged by combining them into a CRNN, replacing a specified amount of the final layers of the CNN with recurrent layers.

3.2.1 *DCASE 2018*

Our CRNN for DCASE 2018, task 5 consists of 3 convolutional blocks where each block contains one convolutional

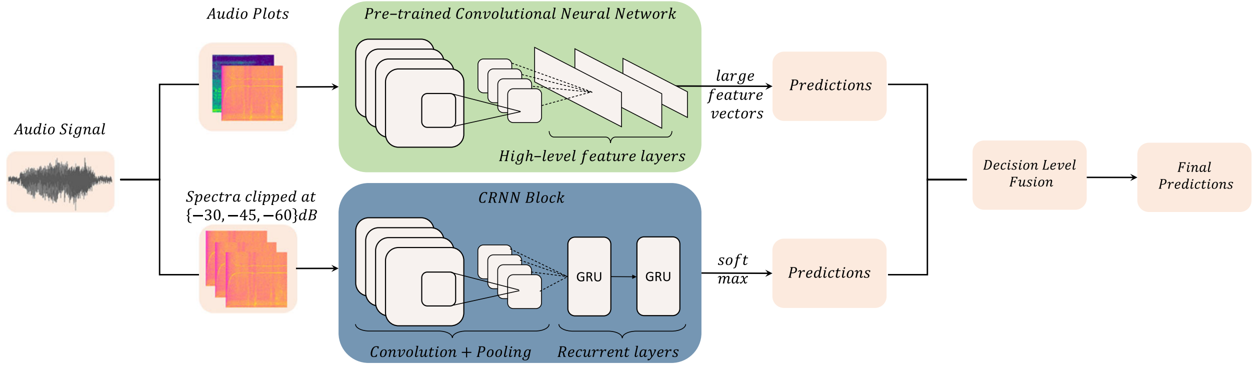


Fig. 1 An overview of our deep learning framework composed of a pre-trained CNN (here exemplified with VGG16) used as feature extractor and a CRNN block. First, the spectrograms are created from the audio recordings. Afterwards, using the pre-trained CNN and our CRNN, block predictions are obtained. In the last step, a decision-level fusion is conducted to get the final predictions. For a detailed account on the framework, refer to Section 3

layer, batch normalisation along the channel axis [38], exponential linear units (ELUs) as activation function [39], two-dimensional max-pooling and a dropout layer with 30 % dropout [40]. The convolutional layers use a 5 × 5, 4 × 4 and 3 × 3 convolutional kernel. We have used a max-pooling with a size of 2 for the time dimension and a size of 32 for the frequency dimension in the third convolutional layer. The three convolutional blocks are followed by two gated recurrent units [41] each with 256 hidden units. We then apply a final dropout with 30% to minimise the possible overfitting effects [40]. The probabilities for each class are computed by a softmax layer with 9 logits.

The loss of the CRNN is calculated with the cross entropy on the logits and the network is trained with the ADAM optimiser with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a combination of two learning rates $lr \in [\,0.01, 0.001]$, and batch size $\in [\,64, 128]$ were evaluated. A learning rate decay of 0.002 was adjusted and the network was trained for 30 epochs.

3.2.2 DCASE 2017 and ESC-50

On DCASE 2017 and ESC-50, the CRNN is slightly adapted to use 4 convolutional blocks and smaller 3 × 3 convolutional kernels throughout based on initial experiments. We use the same optimiser with a learning rate of 0.001 and train for 50 epochs on DCASE 2017 and 100 epochs on ESC-50 due to the smaller dataset sizes. Furthermore, we refrained from using amplitude clipping at different rates; instead, we clip every spectrogram below − 80 dB.

3.3 Pre-trained CNNs as feature extractors

In addition to CRNNs, we also employ the DEEP SPECTRUM toolkit[2] [42] to extract deep features from the audio samples with VGG16, VGG19 [21], 50-layer ResNet [22] and DenseNet121 [23] networks that have been pre-trained on ImageNet. In combination with differing machine learning algorithms, these features have performed well for various audio-based recognition tasks [1, 26, 27, 43].

For the extraction of these features, Mel-spectrograms (with 128 Mel-frequency bands) are first plotted from the audio clips with the *matplotlib* library and the resulting images are then forwarded through the networks. For VGG16 and VGG19, we use the neuron activations of the second to last fully connected layer as representations, while for the ResNet and DenseNet networks, global average pooling is applied to the convolutional base to form the audio features. For the work presented herein, we also evaluate the ImageNet pre-training against random initialisation of weights and using features extracted from models trained on audio data in with the open source toolkits openl3[3] [28, 29] and PANNs[4] [30]. While openl3 uses mel-spectrograms as input just as DEEP SPECTRUM, PANNs employs a hybrid *wavegram* feature, combining a small 1D CNN trained directly on the raw audio waveform with mel-spectrograms by concatenation along the channel axis. Both approaches make use of CNNs as feature extractor. For openl3, we further use the network trained on environmental sounds instead of the one for music recognition as the former fits better with our target tasks. As classifier, we use a linear SVM to which we feed the DEEP SPECTRUM features after applying input standardisation. We optimise the classifier's complexity parameter on a logarithmic scale from 10^{-9} to 1 to achieve the best macro averaged F1 score on the suggested 4-fold cross-validation (CV) setup. The same procedure is also applied on DCASE 2017 and ESC-50.

3.4 Decision-level fusion

In order to assess whether the different systems trained in our experiments are complimentary to each other, we apply a decision-level fusion approach to the predictions of the CRNN models and the SVM classifiers trained on the deep features extracted by the various CNNs. On all three datasets, we perform classifier stacking by utilising the predictions generated by all of the individual trained models, i. e. we concatenate the class probabilities or decision function values generated by CRNN and SVM models to form a new set of features. These features are then used to train another linear SVM as a meta-model to predict the correct class labels. We use feature standardisation and optimise the SVM's complexity parameter on a logarithmic scale from 10^{-9} to 1 with the official 4-fold cross-validation schemes on the development partition for both DCASE tasks, and the first four folds of ESC-50. The best performing configuration is then trained on the whole development partition and used to predict the class labels on the test set.

4 Experimental settings and results

For training and evaluating our deep learning models, we utilise 4-fold CV plus a held-out test set partitioning described in the previous sections. For the Domestic Activity Classification task of DCASE 2018 [8], we use macro average F1 as evaluation measure, as was done in the challenge. In the case of DCASE 2017 and ESC-50 which have balanced class distribution, we use accuracy as measure to make comparisons to the state-of-the-art easier. The results and experimental settings for the three tasks will be presented in their own respective sections, starting with the domestic activity classification task of

[2] https://github.com/DeepSpectrum/DeepSpectrum

[3] https://github.com/marl/openl3

[4] https://github.com/qiuqiangkong/panns_inference

DCASE 2018. The final results for all three tasks are aggregated in Table 3.

4.1 DCASE 2018, task 5

In each fold, we train our CRNNs for each the Mel-spectrogram data and evaluate it on the test partition. We perform mean fusion of the class probabilities generated by each of the fourfold models to arrive at the final predictions on the test set. We further experiment with different learning rate and batch size combinations, specifically $lr \in [0.001, 0.01]$ and $batchsize \in [64, 128]$. Additionally, here, we investigate the impact on performance of clipping the spectrograms at − 30 dB, − 45 dB and − 60 dB. Optimal results are achieved with a batch size of 64 and learning rate set at 0.01. Predictions from these three CRNN configurations are then fused among themselves and with those of the other systems.

The results provided in Table 1 demonstrate that the CRNN systems perform best when trained with a batch size of 64 and a learning rate of 0.01. We choose one model for each of the clipping values for evaluation on the test set and for decision-level fusion. On the development partition, a lone CRNN model performs best when clipping noise is below − 60 dB, achieving an F1 score of 78.8 %. Clipping more noise (at − 45 dB and − 30 dB) results in worse performance on the development partition, indicating a loss of useful information found in the input signal. When looking at the results on the evaluation partition, clipping noise below − 45 dB leads to the strongest result of 79.3% F1. This behaviour might be caused by the introduction of recordings from microphones which are not present in the development partition. Therefore, clipping further might counteract the influence of the unfamiliar sound characteristics of these microphones. Furthermore, noise clipping has a regulating effect on CRNN training, acting against overfitting on the recording setting of the development partition. While clipping less of the input signal allows the model to perform better on the development set, it in turn loses some of its generalisation capabilities.

The training procedure of the SVM models utilising various CNN networks as feature extractors is as described in Section 3.3. For DCASE 2018, we also evaluated the impact on classifier performance resulting from choosing different colour maps for the plots of the mel-spectrograms used in the DEEP SPECTRUM system. In Table 2, results with five different colour mappings for an ImageNet pre-trained 50-layer ResNet are presented. From these results, it can be seen that choosing different colour mappings only has a marginal effect on classification accuracy. Based on these findings, we do not use multiple colour maps for the remaining databases.

Of larger interest are the results achieved with different configurations of model architecture and pre-training, as can be seen in Table 3. Notably, ImageNet pre-trained DenseNet121 and ResNet50 achieve the highest performance on the test partition measured by macro average F1, with 81.1% and 80.3%, respectively. For all network architectures, pre-training on ImageNet improves the saliency of the extracted features when applied to domestic activity classification when compared to using randomly initialised weights. These performance deltas are in the range of 5 to 10 percentage points. Compared to the two evaluated audio pre-trained CNNs, ImageNet pre-trained CNNs further are very favourable. While PANN achieves a higher F1 score of 84.6% than any of the DEEP SPECTRUM systems, openl3 features perform worse than every other feature extractor, even when taking the randomly initialised image CNNs into account. When late fusion is applied to the different system configurations,

Table 1 Performance of CRNNs. All results are given in macro average F1. T_{amp} amplitude threshold, *lr* learning rate. All results are measured in macro average F1

		CRNN		
lr	Batch size	T_{amp} [dB]	Devel	Test
0.001	128	− 30	66.5	–
0.001	128	− 45	57.8	–
0.001	64	− 30	55.4	–
0.01	128	− 30	66.2	–
0.01	128	− 60	69.2	–
0.01	64	− 30	70.7	72.3
0.01	64	− 45	73.7	79.3
0.01	64	− 60	78.8	74.2
Fusion of best 3 CRNNs			81.4	82.2
		DCASE 2018, task 5 baseline [8]		
0.0001	256	–	84.5	83.1

Table 2 Evaluation of the impact different colour maps have on the efficiency of an ImageNet pre-trained ResNet as an audio feature extractor on DCASE 2018's domestic activity classification task. All results are given in macro average F1

Network	Pre-training	Colour map	Devel	Test
ResNet	ImageNet	Cividis	**82.6**	80.2
ResNet	ImageNet	Gray	82.2	79.4
ResNet	ImageNet	Hot	82.0	79.8
ResNet	ImageNet	Magma	81.9	**80.3**
ResNet	ImageNet	Viridis	81.2	79.9

several observations can be made. First of all, fusing the different Deep Spectrum configurations makes the resulting classification system more robust and improves performance over the best individual system to 84.3 on the test partition. However, adding the Deep Spectrum systems with random weights into the fusion does not improve over just fusing all ImageNet pre-trained models. Fusing Deep Spectrum with the CRNN trained only on the target domain data leads to a slightly improved F1 of 85.5. However, combining audio and image pre-training in

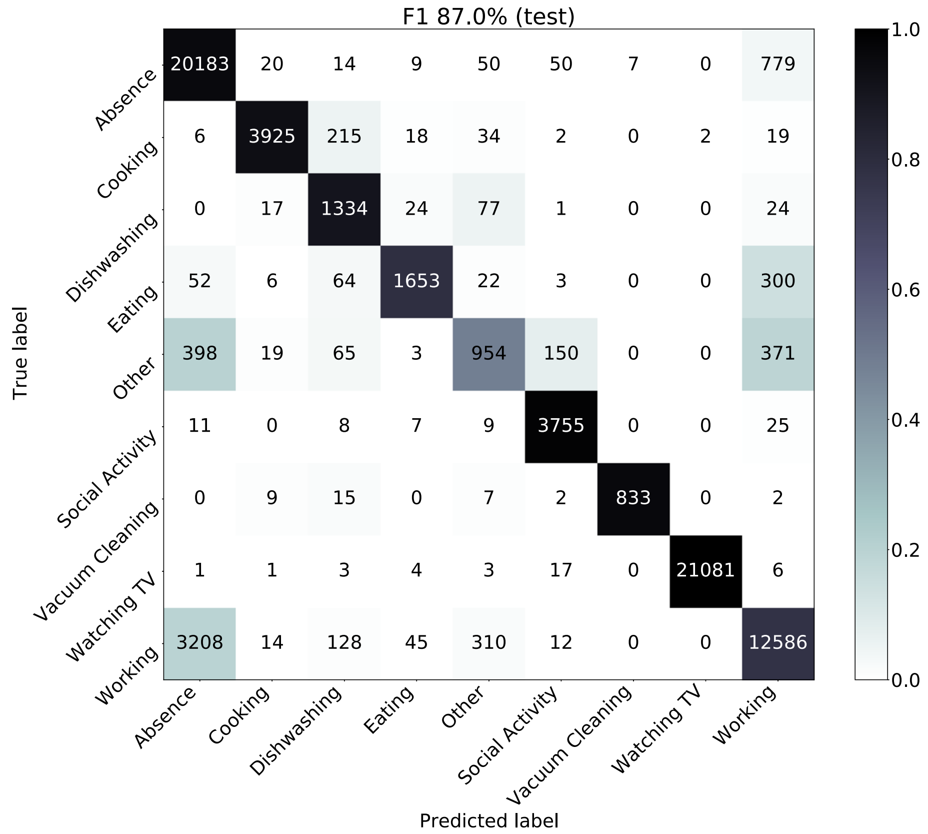
Table 3 Results of Deep Spectrum, pre-trained audio models, CRNN and their fusion on DCASE 2018, task 5, DCASE 2017, task 1 [31] and ESC-50

		DCASE18 (F1 [%])		DCASE17 (Acc [%])		ESC-50 (Acc [%])	
		Devel	Test	Devel	Test	Devel	Test
Proposed Deep Spectrum [42]							
Network	Pre-training						
Densenet121	ImageNet	**82.8**	**81.1**	**78.9**	**64.4**	**73.6**	**75.0**
Densenet121	None	77.7	75.7	71.2	59.0	45.3	44.0
ResNet50	ImageNet	81.9	80.3	76.5	55.9	70.3	72.0
ResNet50	None	70.1	69.9	72.7	61.0	44.8	44.8
VGG16	ImageNet	79.4	77.0	70.1	54.1	63.0	64.8
VGG16	None	73.3	71.6	72.2	57.8	44.6	45.3
VGG19	ImageNet	78.6	77.9	71.8	57.1	62.9	62.5
VGG19	None	74.6	73.2	72.0	61.0	42.9	46.0
Pre-trained audio models							
openl3 [28]	AudioSet	73.3	68.4	**79.3**	**67.7**	69.8	70.8
PANN [30]	AudioSet	**84.6**	**84.6**	69.3	65.7	**91.0**	**89.3**
Proposed fusion							
Proposed CRNN		81.4	82.2	68.9	59.2	62.3	68.8
ImageNet pre-trained Deep Spectrum		84.4	84.0	77.7	63.5	70.7	73.5
All untrained Deep Spectrum		77.9	78.1	74.5	63.3	44.9	46.8
All Deep Spectrum		84.6	84.3	78.7	67.3	69.8	75.8
CRNN + Deep Spectrum		85.0	85.5	80.6	70.0	73.5	78.8
Deep Spectrum + AudioSet nets		**87.0**	**87.0**	**82.7**	71.2	**90.9**	**92.3**
CRNN + AudioSet nets + Deep Spectrum		86.8	86.8	82.5	**71.7**	89.6	90.8
Baselines and SOTA							
Challenge baselines with CNNs [8, 31, 32]		84.5	85.0	74.8	61.0	72.4*	72.4*
CNN + Data augmentation [2]		90.0	88.4	–	–	–	–
Data augmentation with GANs [44]		–	–	87.1	83.3	–	–
Fine-tuned PANNs [30]		–	–	–	–	94.7*	94.7*

*ESC-50 baseline given for 5-fold CV which is different from the evaluated 4-fold plus test setup. The best results of every type of evaluated system are marked in bold a cross-modal fashion by fusing DEEP SPECTRUM, openl3 and PANN shows a larger performance improvement to the highest F1 of 87.0%. This perceived complementarity of features indicates the viability of transfer learning across modalities. The confusion matrix of this best result is also displayed in Fig. 2. While this result falls shortly behind the top performing submission of the challenge which utilises data augmentation with generative adversarial networks (GANs) at 88.4 %, it improves on the strong baseline of 85.0 %.

4.2 DCASE 2017, task 1

In the case of DCASE 2017's acoustic scene classification task, the CRNN trained only on the corpus performs slightly below the challenge's baseline system, achieving an accuracy of 59.2% on the test set. Using deep CNNs as feature extractors leads to better results. With an ImageNet pre-trained DenseNet121, an accuracy of 64.4% on the test set can be achieved. This compares very favourably to features from the audio pre-trained models from openl3 and PANN which reach 67.7% and 65.7% test set accuracy but should intuitively be far better suited to audio analysis tasks than image CNN descriptors. However, on this database, an interesting observation regarding the pre-training of the CNNs can be made: For all DEEP SPECTRUM systems apart from the one based on DenseNet121, pre-training on ImageNet leads to less salient features than randomly initialising the network weights. This disparity is most pronounced with the 50-layer ResNet where random weights lead to an accuracy increase of 5.1 percentage points. By applying the proposed late fusion on all ImageNet pre-trained and all randomly initialised DEEP SPECTRUM systems separately, it can be seen that test set performance is on the same level, while ImageNet pre-training only makes a positive impact during validation. On the other hand, the fusion of both sets of features indicates that the features are complimentary here, increasing test set performance to 67.3 % and thus matching the performance of the openl3 network pre-trained on environmental sounds. Finally, fusing the DEEP SPECTRUM systems with the audio pre-trained models and the CRNN trained directly on the target data leads to the best results, at over 70.0% accuracy—a strong increase over the individual systems. These results further indicate the suitability of adding ImageNet pre-training to audio classification, but additionally shows that randomly initialised CNNs should be considered as well. Finally, the results of the audio pre-trained models on their own are



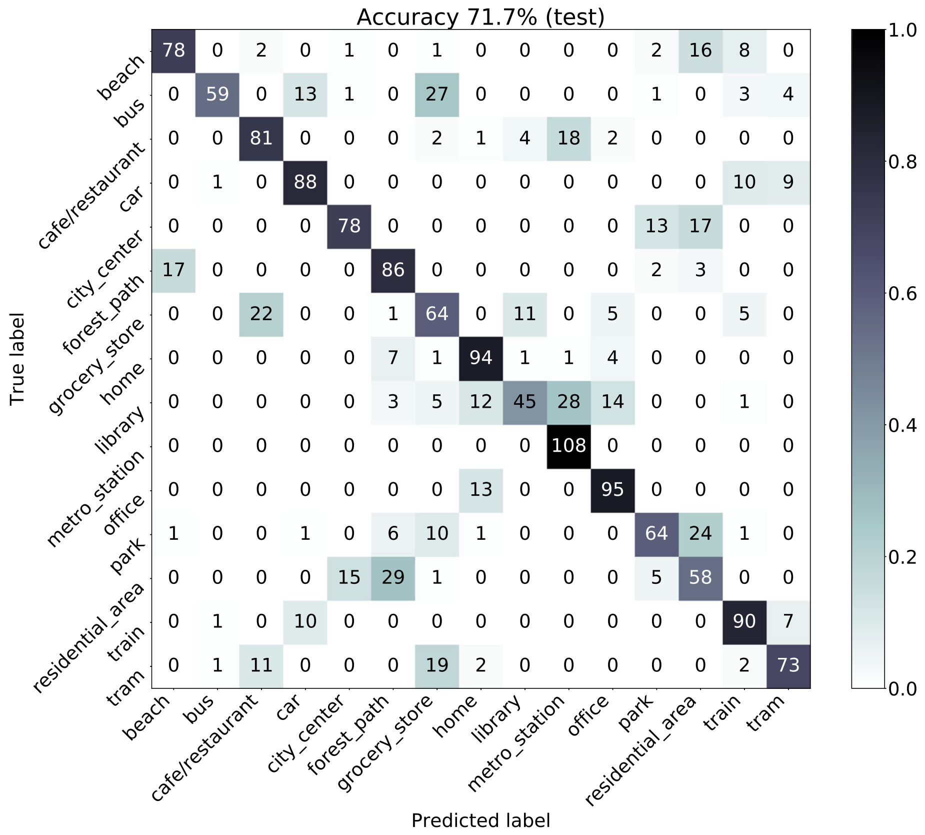
	Absence	Cooking	Dishwashing	Eating	Other	Social Activity	Vacuum Cleaning	Watching TV	Working
Absence	20183	20	14	9	50	50	7	0	779
Cooking	6	3925	215	18	34	2	0	2	19
Dishwashing	0	17	1334	24	77	1	0	0	24
Eating	52	6	64	1653	22	3	0	0	300
Other	398	19	65	3	954	150	0	0	371
Social Activity	11	0	8	7	9	3755	0	0	25
Vacuum Cleaning	0	9	15	0	7	2	833	0	2
Watching TV	1	1	3	4	3	17	0	21081	6
Working	3208	14	128	45	310	12	0	0	12586

Fig. 2 The confusion matrix (CM) of the best prediction on the test set of the DCASE 2018, task 5 dataset

also the worst among the three tasks, indicating that for DCASE 2017, pre-training is not as efficient as for the other databases, regardless of source domain. A confusion matrix for the best fusion configuration can be found in Fig. 3.

4.3 ESC-50

For ESC-50, the CRNN trained only on target data achieves a test set accuracy of 68.8% which is worse than the dataset's official baseline at 72.4%. However, it has to be noted here that the dataset uses a 5-fold cross-validation setup whereas in this paper, we transformed this to 4-fold (fold 1 to 4) cross-validation and a held-out test set (the 5th fold) in order to apply the late fusion scheme like in the rest of the experiments. This circumstance leads to each CRNN model being trained on 20.0% less data. The systems utilising CNNs as feature extractors perform better on this database with the results being relatively in line with those on DCASE18 task 5. Despite their unrelated ImageNet pre-training, Deep Spectrum features are effective for environmental sound classification, especially when extracted from DenseNet121 or a 50-layer ResNet, the former reaching 75.0% accuracy on the test set. Both VGG networks, on the other hand, perform substantially worse, only reaching an accuracy of 64.8% on the test set. For all Deep Spectrum systems however, ImageNet pre-training outperforms random weight initialisation by a wide margin—non pre-trained nets only reach around 45.0% accuracy. For this task, the best Deep Spectrum features are also better than openl3 which reaches 70.8% accuracy. PANN features, on the other hand, are substantially better than both openl3 and Deep Spectrum at 89.3% accuracy. This result matches the findings that the author's presented in [30] where a fine-tuned PANN stands as the current state-of-the-art on ESC-50. Unlike for the DCASE tasks, fusing the Deep Spectrum systems amongst themselves does not lead to better accuracy over using just the best performing DenseNet121 pre-trained on ImageNet. By combining Deep Spectrum with the CRNN—which alone has quite low performance—results are improved by about 3 percentage points to 78.8%. Audio pre-trained models and Deep Spectrum also seem to be complimentary here, with their fusion reaching 92.3%. Adding the CRNN into



	beach	bus	cafe/restaurant	car	city_center	forest_path	grocery_store	home	library	metro_station	office	park	residential_area	train	tram
beach	78	0	2	0	1	0	1	0	0	0	0	2	16	8	0
bus	0	59	0	13	1	0	27	0	0	0	0	1	0	3	4
cafe/restaurant	0	0	81	0	0	0	2	1	4	18	2	0	0	0	0
car	0	1	0	88	0	0	0	0	0	0	0	0	0	10	9
city_center	0	0	0	0	78	0	0	0	0	0	0	13	17	0	0
forest_path	17	0	0	0	0	86	0	0	0	0	0	2	3	0	0
grocery_store	0	0	22	0	0	1	64	0	11	0	5	0	0	5	0
home	0	0	0	0	0	7	1	94	1	1	4	0	0	0	0
library	0	0	0	0	0	3	5	12	45	28	14	0	0	1	0
metro_station	0	0	0	0	0	0	0	0	0	108	0	0	0	0	0
office	0	0	0	0	0	0	0	13	0	0	95	0	0	0	0
park	1	0	0	1	0	6	10	1	0	0	0	64	24	1	0
residential_area	0	0	0	0	15	29	1	0	0	0	0	5	58	0	0
train	0	1	0	10	0	0	0	0	0	0	0	0	0	90	7
tram	0	1	11	0	0	0	19	2	0	0	0	0	0	2	73

Fig. 3 The confusion matrix (CM) of the best prediction on the test set of the DCASE 2017, task 1 database. Confusion is high for the acoustic scene "residential area" which is often mistaken for "city" or "forest_path"

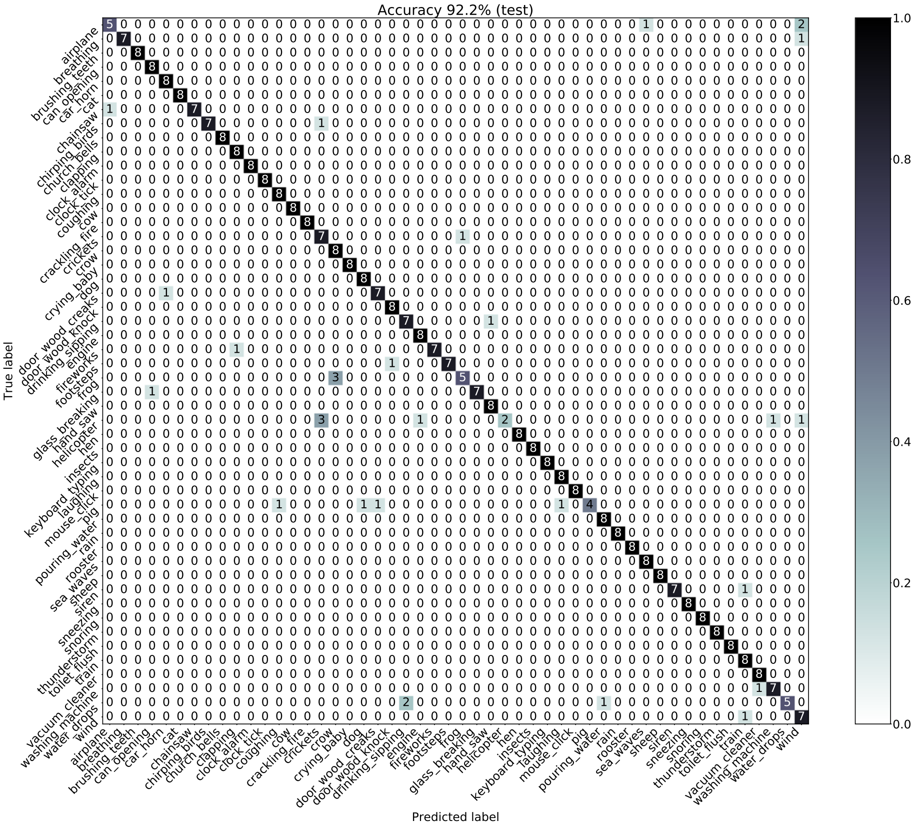


Fig. 4 The confusion matrix (CM) of the best prediction on the fifth fold of the ESC-50 database. The highest confusion can be observed for the classes "frog" and "crow"

this mix, however, has a performance degrading effect. The best result on ESC-50 is also visualised via a confusion matrix in Fig. 4.

5 Conclusions and future work

We have proposed a deep learning framework composed of an image-to-audio transfer learning system, audio pre-trained CNNs and a CRNN. Furthermore, we performed various decision-level fusion strategies between the applied neural networks. We have tested our methodologies for audio-based classification of 15 acoustic scenes (DCASE 2017, task 1 [31]), 50 environmental sounds (ESC-50 [32]) and 9 domestic activities (DCASE 2018, task 5 [8]). We have demonstrated the suitability of our approaches for all of the mentioned tasks. In particular, we have shown that even though the domain gap between audio and images is considerably larger than what is usually found in the field of transfer learning, ImageNet pre-trained CNNs are powerful feature extractors when applied directly to spectrograms, oftentimes matching or outperforming specialised audio feature extraction networks. We further evaluated the ImageNet pre-training against random weight initialisation and found it to be more effective in general. Moreover, various late fusion configurations indicated a complementarity between Deep Spectrum features and more domain-specific knowledge, either in the form of our proposed CRNN or audio pre-trained networks. Whilst our systems did not outperform the current state-of-the-art on the included databases, the findings presented herein motivate further exploration of cross-modal pre-training for audio classification tasks.

In future work, we want to evaluate the impact of ImageNet pre-training against AudioSet pre-training as well as training from scratch in low-data settings. Furthermore, we want to investigate traditional fine-tuning and more involved domain transfer methods, such as domain adversarial neural networks (DANNs) [45] with our Deep Spectrum system.

Abbreviations

AED: Acoustic event detection; BLSTM: Bidirectional long short-term memory; CNN: Convolutional neural network; CRNN: Convolutional recurrent neural network; DANN: Domain adversarial neural network; CV: Cross-validation; ELU: Exponential linear unit; GAN: Generative adversarial network; GRU: Gated recurrent unit; IoT: Internet of Things; LSTM: Long short-term memory; RNN: Recurrent neural network; SVM: Support vector machine

Acknowledgements
This research was conducted as part of the EU's Horizon 2020 project.

Authors' contributions
S.A., M.G., A.B., and L.K. conceptualised the study and run the machine learning experiments. All authors did literature analysis, manuscript preparation, and editing, as well as revised, developed, read, and approved the final manuscript.

Funding
This work is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 688835 (RIA DE-ENIGMA) and the BMW Group research. Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
The first data originates from the "Monitoring of domestic activities based on multi-channel acoustics" task Task 5 of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2018) workshop and can be found under the following link: http://dcase.community/challenge2018/task-monitoring-domestic-activities. The second dataset is from the acoustic scene classification challenge (Task 1) of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017) workshop [31] and can be found under the following link: http://dcase.community/challenge2017/task-acoustic-scene-classification. The third dataset is the environmental sound classification dataset ESC-50 [32] and can be found under the following link: https://github.com/karolpiczak/ESC-50.

Competing interests
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author details
[1]Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany. [2]GLAM – the Group on Language, Audio & Music, Imperial College London, London, UK.

Received: 10 July 2019 Accepted: 27 October 2020
Published online: 02 December 2020

References

1. S. Amiriparian, Deep representation learning techniques for audio signal processing. PhD thesis, Technische Universität München (2019)
2. T. Inoue, P. Vinayavekhin, S. Wang, D. Wood, N. Greco, R. Tachibana, Domestic activities classification based on CNN using shuffling and mixing data augmentation. Technical report, DCASE2018 Challenge (2018)
3. K. Nakadai, D. R. Onishi, Partially-shared convolutional neural network for classification of multi-channel recorded audio signals. Technical report, DCASE2018 Challenge (2018)
4. A. Mesaros, T. Heittola, T. Virtanen, in *24th European Signal Processing Conference (EUSIPCO), 2016*. Tut database for acoustic scene classification and sound event detection (IEEE, Budapest, 2016), pp. 1128–1132
5. D. de Benito-Gorron, A. Lozano-Diez, D. T. Toledano, J. Gonzalez-Rodriguez, Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP J. Audio Speech Music. Process. **2019**(1), 9 (2019)
6. E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1291–1303 (2017)
7. S. Amiriparian, S. Julka, N. Cummins, B. Schuller, in *Proceedings of 44. Jahrestagung Für Akustik (DAGA)*. Deep convolutional recurrent neural networks for rare sound event detection (DAGA, Munich, 2018), pp. 1522–1525
8. G. Dekkers, L. Vuegen, T. van Waterschoot, B. Vanrumste, P. Karsmakers, DCASE 2018 Challenge-Task 5: monitoring of domestic activities based on multi-channel acoustics. Technical report, KU Leuven (2018). https://arxiv.org/abs/1807.11246
9. P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, F. Piazza, in *Multidisciplinary Approaches to Neural Computing*. Convolutional neural networks with 3-d kernels for voice activity detection in a multiroom environment (Springer, Cham, 2018), pp. 161–170
10. F. Vesperini, D. Droghini, E. Principi, L. Gabrielli, S. Squartini, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Hierarchic conv nets framework for rare sound event detection (IEEE, Rome, 2018), pp. 1497–1501
11. A. Graves, N. Jaitly, A.-r. Mohamed, in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*. Hybrid speech recognition with Deep Bidirectional LSTM (IEEE, Olomouc, 2013), pp. 273–278
12. A. Graves, A.-r. Mohamed, G. Hinton, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Speech recognition with deep recurrent neural networks (IEEE, Vancouver, 2013), pp. 6645–6649
13. A. Graves, N. Jaitly, in *Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copy-right 2014 by the authors*. Towards end-to-end speech recognition with recurrent neural networks, (2014), pp. 1764–1772
14. S. Amiriparian, M. Freitag, N. Cummins, B. Schuller, in *Proceedings of the 2nd Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*. Sequence to sequence autoencoders for unsupervised representation learning from audio (IEEE, Munich, Germany, 2017), pp. 17–21
15. S. Amiriparian, M. Freitag, N. Cummins, M. Gerzcuk, S. Pugachevskiy, B. W. Schuller, in *Proceedings of 26th European Signal Processing Conference (EUSIPCO)*. A fusion of deep convolutional generative adversarial networks and sequence to sequence autoencoders for acoustic scene classification (IEEE, Rome, Italy, 2018), pp. 982–986. EURASIP
16. D. Tang, B. Qin, T. Liu, in *Proceedings of the 2015 conference on empirical methods in natural language processing*. Document modeling with gated recurrent neural network for sentiment classification, (Lisbon, 2015), pp. 1422–1432
17. K. Choi, G. Fazekas, M. Sandler, K. Cho, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Convolutional recurrent neural networks for music classification (IEEE, New Orleans, 2017), pp. 2392–2396
18. G. Parascandolo, H. Huttunen, T. Virtanen, in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference On*. Recurrent neural networks for polyphonic sound event detection in real life recordings (IEEE, Shanghai, 2016), pp. 6440–6444
19. S. Amiriparian, A. Baird, S. Julka, A. Alcorn, S. Ottl, S. Petrović, E. Ainger, N. Cummins, B. Schuller, in *Proceedings of INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*. Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks (ISCA, Hyderabad, India, 2018), pp. 2334–2338
20. H. Zhao, S. Zarar, I. Tashev, C.-H. Lee, Convolutional-recurrent neural networks for speech enhancement. arXiv preprint arXiv:1805.00579, 2401–2405 (2018)
21. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. CoRR. **abs/1409.1556** (2014)
22. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Deep residual learning for image recognition, (Las Vegas, 2016), pp. 770–778
23. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Densely connected convolutional networks, (Honolulu, 2017), pp. 4700–4708
24. G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, T. van Waterschoot, B. Vanrumste, M. Verhelst, P. Karsmakers, in *IEEE Conference on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017*. The SINS database for detection of daily activities in a home environment using an acoustic sensor network, (Munich, 2017), pp. 32–36
25. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al*, Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
26. S. Amiriparian, N. Cummins, M. Gerczuk, S. Pugachevskiy, S. Ottl, B. Schuller, "are you playing a shooter again?!" deep representation learning for audio-based video game genre recognition. IEEE Trans. Game. **12**(2), 145–154 (2020)
27. S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, B. Schuller, in *Proceedings of the 31st International Joint Conference on Neural Networks (IJCNN)*. Bag-of-deep-features: noise-robust deep feature representations for audio analysis (IEEE, Rio de Janeiro, Brazil, 2018), pp. 2419–2425
28. R. Arandjelovic, A. Zisserman, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Look, listen and learn, (Venice, 2017), pp. 609–617

29. J. Cramer, H.-H. Wu, J. Salamon, J. P. Bello, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Look, listen, and learn more: design choices for deep audio embeddings (IEEE, Brighton, 2019), pp. 3852–3856
30. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, Panns: large-scale pretrained audio neural networks for audio pattern recognition. arXiv preprint arXiv:1912.10211. **28**, 2880–2894 (2019)
31. A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 - Workshop on Detection and Classification of Acoustic Scenes and Events*. Dcase 2017 challenge setup: tasks, datasets and baseline system, (Munich, 2017)
32. K. J. Piczak, in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ESC: dataset for environmental sound classification (ACM Press, New York, pp. 1015–1018. https://doi.org/10.1145/2733373.2806390
33. A. C. Braun, U. Weidner, S. Hinz, Classification in high-dimensional feature spaces–assessment using SVM, IVM and RVM with focus on simulated EnMAP data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(2), 436–443 (2012)
34. K. Choi, G. Fazekas, M. Sandler, K. Cho, in *2018 26th European Signal Processing Conference (EUSIPCO)*. A comparison of audio signal preprocessing methods for deep neural networks on music tagging (IEEE, Rome, 2018), pp. 1870–1874
35. M. Norouzi, M. Ranjbar, G. Mori, in *2009 Computer Vision and Pattern Recognition (CVPR)*. Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning (IEEE, Miami, 2009)
36. O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(10), 1533–1545 (2014)
37. E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection. CoRR. **25**(6), 1291–1303 (2017). http://arxiv.org/abs/1702.06286
38. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 448–456 (2015)
39. D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
40. N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
41. J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR. **abs/1412.3555** (2014). http://arxiv.org/abs/1412.3555
42. S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, B. Schuller, in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Snore sound classification using image-based deep spectrum features (ISCA, Stockholm, Sweden, 2017), pp. 3512–3516
43. S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, B. Schuller, in *Proceedings of the 7th Biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)*. Sentiment analysis using image-based deep spectrum features (IEEE, San Antonio, TX, 2017), pp. 26–29
44. S. Mun, S. Park, D. Han, H. Ko, Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. Technical report, DCASE2017 Challenge (2017)
45. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2016)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.