A Fusion of Deep Convolutional Generative Adversarial Networks and Sequence to Sequence Autoencoders for Acoustic Scene Classification

Shahin Amiriparian[1,2], Michael Freitag[3], Nicholas Cummins[1], Maurice Gerczuk[1], Sergey Pugachevskiy[1], Björn Schuller[1,4]
shahin.amiriparian@tum.de

***Abstract*— Unsupervised representation learning shows high promise for generating robust features for acoustic scene analysis. In this regard, we propose and investigate a novel combination of features learnt using both a deep convolutional generative adversarial network (DCGAN) and a recurrent sequence to sequence autoencoder (S2SAE). Each of the representation learning algorithms is trained individually on spectral features extracted from audio instances. The learnt representations are: (i) the activations of the discriminator in case of the DCGAN, and (ii) the activations of a fully connected layer between the decoder and encoder units in case of the S2SAE. We then train two multilayer perceptron neural networks on the DCGAN and S2SAE feature vectors to predict the class labels. The individual predicted labels are combined in a weighted decision-level fusion to achieve the final prediction. The system is evaluated on the development partition of the acoustic scene classification data set of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2017). In comparison to the baseline, the accuracy is increased from 74.8 % to 86.4 % using only the DCGAN, to 88.5 % on the development set using only the S2SAE, and to 91.1 % after fusion of the individual predictions.**

***Index Terms*— unsupervised feature learning, generative adversarial networks, sequence to sequence autoencoders, acoustic scene classification**

I. Introduction

The choice of data representation heavily influences the performance of machine learning algorithms, hence determining what constitutes an adequate representation is a key research topic [1]. To date, audio processing tasks, including acoustic scene classification have been dominated by 'hand-crafted' features, for instance Mel-Frequency Cepstral Coefficients [2], [3]. However, recently, unsupervised representation learning, in particular deep representation learning, has started to attract increasing research attention [4]–[8]. Such approaches are desirable, as they are theoretically capable of learning more abstract, thus robust representations than shallow approaches [1]. Further, they enable machines to automatically learn the discriminative characteristics, without label information, thus eliminating the associated manual efforts with these tasks.

[1]Shahin Amiriparian, Nicholas Cummins, Maurice Gerczuk, Sergey Pugachevskiy, and Björn Schuller are with the ZD.B Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany {shahin.amiriparian, sergey.pugachevskiy, nicholas.cummins, bjoern.schuller}@informatik.uni-augsburg.de

[2]Shahin Amiriparian is also with the Machine Intelligence & Signal Processing Group, Technische Universität München, Germany

[3]Michael Freitag is with the University of Passau, Germany

[4]Björn Schuller is also with GLAM – Group on Language, Audio and Music, Imperial College London, U. K.

Unsupervised representation learning for computer audition tasks is not a novel concept; approaches such as sparse feature learning [9] and bag-of-audio-words [10], have been proposed in the literature and developed. The increasing presence of deep unsupervised representation learning approaches for computer audition tasks can be highlighted by the recent Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [11], [12]. While most entrants to the 2016 and 2017 challenges were based on conventional audio feature representations, there has been an increase in the presence of entries using deep unsupervised methods, especially when compared to the original 2013 challenge [3].

In particular, approaches based on Convolutional Neural Networks (CNNs) [13], Restricted Boltzmann Machines (RBMs) [14], Deep Non-negative Matrix Factorisation (DNMF) [15], and Recurrent Neural Networks (RNNs) [8] have shown promise for the task of acoustic scene classification. However, while the aforementioned systems were highly competitive in their respective challenges, the reality that such systems did not consistently outperform approaches based on conventional audio features, shows there are still improvements to be made in terms of generating more meaningful and task salient feature representations.

Deep unsupervised representation learning paradigms are arguably more established within the field of image processing [1]. The use of CNN architectures such as *ResNet* [16], *AlexNet* [17], and *VGG19* [18] is now considered as an established feature extractor for image tasks, including object or scene recognition [19], [20]. Moreover, these pre-trained CNNs have also been successfully applied in audio recognition tasks [21]–[24]. A more recent development in image processing, yet to be fully utilised in computer audition, has been the advent of Generative Adversarial Networks (GANs) [25]. GANs, in particular, Deep Convolutional GANs (DCGANs) have shown state-of-the-art performance in image classification tasks [26].

GANs have also shown promising results in a small number of computer audition tasks, including automatic speech recognition [27] and computational paralinguistics [28]. However, to the best of the authors' knowledge, the work presented herein is the first time this technique has been proposed for the task of acoustic scene classification. We present an approach which fuses representations learnt from Mel-spectrograms of audio files using both a DCGAN and a recurrent sequence to sequence autoencoder (S2SAE), extracted with the state-of-the-art auDeep toolkit [5].

ISBN 978-90-827970-1-5

The rest of this contribution is organised as follows. Section II introduces the acoustic scene database. Section III outlines our deep learning methods for unsupervised representation learning from the audio files. The experimental settings and results are outlined in Section IV, before concluding the paper in Section V.

II. Acoustic Scene Data Set

The DCASE 2017 acoustic scene classification challenge was carried out on the TUT Acoustic Scenes 2017 data set [12]. This data set contains binaural audio samples of 15 acoustic scenes recorded at distinct geographic locations. For each location, between 3 and 5 minutes of audio were initially recorded and then split into 10 second chunks. The development set for the challenge contains 4 680 instances, with 312 instances per class, and the evaluation set contains 1 620 instances. A four-fold cross-validation setup is provided by the challenge organisers for the development set. In each fold, about 75 % of the samples are used as the training split, and the remaining samples are used as the evaluation split. Samples from the same original recording are always included in the same split. Our study is conducted on the development set only, there is a notable mismatch in recording conditions between the partitions which causes observable confounding effects. This is evidenced by the lack of relationship between the development and evaluation scores in the 2017 challenge[1]. For further details on the challenge data and the cross fold validation setup, the interested reader is referred to [12].

III. System Architecture

Our approach is composed of two components for unsupervised feature learning: 1) a deep convolutional generative adversarial networks (DCGAN), and 2) a sequence to sequence autoencoder (S2SAE). First, we extract the activations of the discriminator for the DCGAN and the activations of a fully connected layer between the decoder and encoder units of the S2SAE. As depicted in Figure 1, separate classifiers are trained on the individual feature sets (8 in total), and the resulting prediction probabilities are fused. Fusion is done in two stages: first, the results on DCGAN-features and S2SAE-features are fused with separately optimised weights, and subsequently, the resulting prediction probabilities (one for DCGAN and one for S2SAE) are fused with optimised weights (cf. Section IV-C).

A. Spectrogram Generation

To create the power spectra of the acoustic data, we apply periodic Hann windows with length l and overlap $0.5l$. From these, we then compute a given number N_{mel} of log-scaled Mel-frequency bands; Mel-spectra features have previously been shown to be useful for the task of acoustic scene classification [8]. We also normalise the Mel-spectra values in $[-1;1]$, as the outputs of the S2SAE are constrained to this interval. The acoustic scene corpus contains audio samples which have been recorded in stereo [12]. Following

[1] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results

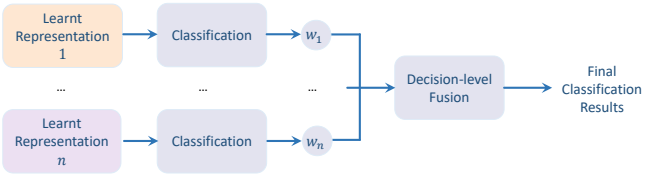

Fig. 1: Illustration of the decision-level fusion system of multiple learnt representation vectors for the audio recordings. We first compute the predictions and confidence scores individually on each representation. Subsequently, the decisions are weighted regarding to predetermined weights and fused to obtain the final classification results.

the winners of the DCASE 2016 acoustic scene classification challenge [29], we thus extract Mel-spectrograms from each individual channel, as well as from the mean and difference of the two channels. Features are then learnt independently on the Mel-spectrograms, and we investigate decision-level fusion of these representations.

B. Deep Convolutional Generative Adversarial Network

In a GAN, the generative model, or generator, is set to compete against a discriminative model, or discriminator, in an adversarial setting. The discriminator is trained to distinguish accurately whether a given sample has been produced by the generator or drawn from the data distribution. At the same time, the objective of the generator is to fool the discriminator into misclassifying the generated samples as real samples. These competing training objectives force both models to continuously improve their methods until the distribution learnt by the generator closely matches the real data distribution [25].

DCGANs are GANs that use CNNs in the generator and discriminator. They have been shown to learn strong representations, which achieve state-of-the-art classification accuracy on image classification tasks [26]. DCGANs can be applied for representation learning from acoustic data by training them on the visual representation of the input audio data, e. g. (Mel-)spectrograms.

Based on the results reported by Radford et al. [26], the following DCGAN architecture is selected for the purposes of this paper (cf. Figure 2). Both the generator and the discriminator contain the same number N_{layer}^{DCGAN} of convolutional layers, with a fixed stride of two. The output layer of the generator and the input layer of the discriminator have the spatial dimensions of the spectrograms that should be processed. The convolutional layer connected to the output layer of the generator and the input layer of the discriminator contains N_{maps}^{DCGAN} feature maps. In each additional layer on top of this layer in the discriminator, or below this layer in the generator, the number of feature maps is doubled and the spatial dimensions are reduced by half. Finally, 100-dimensional Gaussian noise is used as input to the generator, where it is projected and reshaped to the dimensionality required by the first convolutional layer. As suggested by Radford et al., the feature maps have kernels with size 5×5 [26]. An example spectrogram generated by the DCGAN is shown in Figure 3.

ISBN 978-90-827970-1-5 © EURASIP 2018

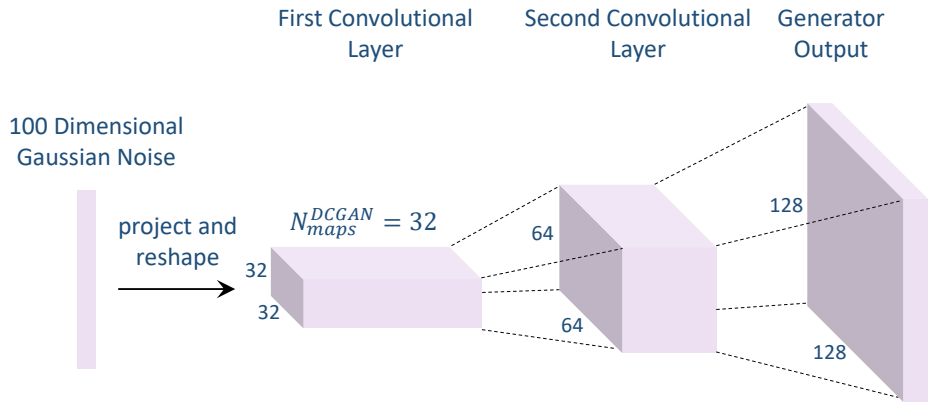

Fig. 2: Illustration of the generator in a DCGAN with $N_{layer}^{DCGAN} = 2$ and $N_{maps}^{DCGAN} = 32$, used to generate spectrograms with hypothetical dimensions 128×128. In each convolutional layer below the output layer, the spatial dimensions are halved. The convolutional layer immediately below the output layer contains N_{maps}^{DCGAN} feature maps, and this number is doubled in each further layer. The discriminator mirrors the architecture of the generator, and information would flow right-to-left in the illustration. N_{layer}^{DCGAN}: Number of convolutional layers in the generator and discriminator CNNs; N_{maps}^{DCGAN}: Number of feature maps in the output layer of the generator, and the input layer of the discriminator.

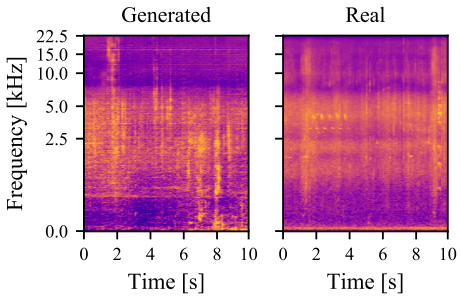

Fig. 3: An example of the spectrogram generated by a DCGAN on the acoustic scene corpus. The DCGAN has learnt to generate spectrograms that are qualitatively very similar to real training examples. The power spectral density in the generated spectrogram is within the range $[-1; 1]$, and no mapping onto the Decibel scale is given. For this reason, no such scale is displayed in the plot.

C. Recurrent Sequence to Sequence Autoencoders

We use a similar implementation for the S2SAE as given in [5], [8]. First, Mel-spectrograms are extracted from the raw acoustic data. Subsequently, an autoencoder is trained on these spectra, which are viewed as time-dependent sequences of frequency vectors. After autoencoder training, the learnt representations of the Mel-spectrograms are then generated for use as feature vectors for the corresponding instances.

Mel-spectra are considered as time-dependent sequences of frequency vectors in $[-1; 1]^{N_{mel}}$, each of which describes the amplitudes of the N_{mel} Mel-frequency bands within one audio segment. This sequence is fed to a multilayered *encoder* RNN that updates its hidden state in each time step based on the input frequency vector. Therefore, the final hidden state of the encoder RNN contains information about the whole input sequence. This final hidden state is transformed using a fully connected layer, and another multilayered *decoder* RNN is used to reconstruct the original input sequence from the transformed representation. For full details, the interested reader is referred to [8].

The encoder RNN consists of N_{layer} layers, each containing N_{unit} Gated Recurrent Units (GRUs). During training, we use the root mean square error (RMSE) between the decoder output and the target sequence as the objective function. To cope with overfitting, i. e. fitting the training model to the noise instead of the underlying signal, we apply dropout [30] to the inputs and outputs of the recurrent layers, but not to the hidden states. After the training process, the activations of the fully connected layer are extracted as the learnt representations of spectrograms.

D. Classifier

A multilayer perceptron (MLP), similar to the one applied in the baseline system [12], is employed for classification. Our MLP has two hidden fully connected layers with rectified linear activation, and a softmax output layer. Each hidden layer contains 150 units, and the output layer has one unit for each class label (15 units total). Training is performed using cross entropy between the ground truth and the network output as the objective function, with dropout applied to all layers except the output layer.

IV. Experimental Settings and Results

A. Common Experimental Settings

We have implemented both representation learning approaches [8], [26] outlined above as part of the auDeep toolkit[2] for deep representation learning from audio [5]. auDeep is implemented in Python, and relies on TensorFlow[3] for the core DCGAN and S2SAE implementations. All neural networks, i. e. the S2SAEs, DCGANs, and MLPs, are trained using the Adam optimiser. Autoencoders are

[2] https://github.com/auDeep/auDeep/
[3] https://www.tensorflow.org/

ISBN 978-90-827970-1-5 © EURASIP 2018

trained for 50 epochs in batches of 64 samples with a fixed learning rate of 0.001, and we apply 20 % dropout to the outputs of each recurrent layer. Furthermore, we clip gradients with absolute value above 2 [31]. The DCGANs are trained for 10 epochs in batches of 32 examples, and a fixed learning rate of 0.0002 and momentum $\beta_1 = 0.5$ is used. The MLPs used for classification are trained for 400 epochs without batching or gradient clipping, and 40 % dropout is applied to the hidden layers.

B. Hyperparameter Selection

Our deep feature learning systems contain a wide range of adjustable hyperparameters that prohibits an exhaustive analysis of the parameter space. For our experiments, we use similar DCGAN and S2SAE hyperparameters as those applied in [26] and [8] (cf. Section IV-B.1 and Section IV-B.2).

1) DCGAN: For DCGAN, previous work is consulted extensively to guide parameter selection [26]. The DCGAN architecture is selected based on the results reported by Radford et al. [26], who use $N_{layer}^{DCGAN} = 4$ and $N_{maps}^{DCGAN} = 64$. A slightly less complex DCGAN architecture with $N_{layer}^{DCGAN} = 3$ and $N_{maps}^{DCGAN} = 32$ is used in this paper.

2) S2SAE: For S2SAE, we select the following parameters according to [8]: LSTM cells with $N_{layer} = 2$ layers and $N_{unit} = 256$ units; a unidirectional encoder RNN, and a bidirectional decoder RNN; Mel-spectrograms extracted with the window length $l = 0.20$ seconds, the window overlap $0.5l = 0.10$ seconds, and $N_{mel} = 320$ Mel-frequency bands.

C. Fusion Experiments

We extract four sets of spectrograms from the mean and difference of channels, and from the left and right channels individually (cf. Section III-A). On each set of spectrograms, a DCGAN and a S2SAE are trained, and the learnt representations are extracted as features for the audio instances. This results in four feature sets for each approach herein identified by the spectrogram type from which they have been extracted (i. e. 'mean', 'difference', 'left', and 'right'). The 'right' feature set (84.5 %) for the DCGAN and the 'mean' feature set (86.0 %) for the S2SAE achieved the highest individual classification accuracy (cf. Table I).

For each of the 8 individual feature sets (4 for DCGAN and 4 for S2SAE) we train a classifier and fuse the resulting prediction probabilities in two steps. First, we fuse the results on DCGAN-features and S2SAE-features and with separately optimised weights. We then fuse the resulting prediction probabilities (one for DCGAN and one for S2SAE) with optimised weights. In order to determine optimal weights for n representations, all combinations of weights $w_1, \ldots, w_n \in [0, 1]$ with $\sum_{i=1}^{n} w_i = 1$ are sampled in steps of 0.1, and the weights that yield the highest classification accuracy are selected.

In the first fusion step, we achieve 86.4 % accuracy on the fused DCGAN predictions, and 88.5 % on the fused S2SAE results (cf. Table I). In the final step, we obtain the highest classification accuracy of 91.1 % on the fused prediction probabilities of DCGAN and S2SAE.

TABLE I: Comparison of the classification accuracies of our proposed systems with the challenge baseline. We extract four different feature sets of spectrograms from the mean (M) and difference (D) of channels, and from the left (L) and right (R) channels separately. We obtain the highest accuracy after fusing the prediction probabilities of S2SAE and DCGAN. *CV*: Cross Validation.

System	Features	CV Accuracy [%]
Baseline	200 (per frame)	74.8
Proposed: DCGAN		
Mean (M)	3 072	84.1
Left (L)	3 072	83.4
Right (R)	3 072	84.5
Difference (D)	3 072	83.5
Fused (M + L + R + D)		86.4
Proposed: S2SAE		
Mean (M)	1 024	86.0
Left (L)	1 024	84.9
Right (R)	1 024	84.0
Difference (D)	1 024	82.0
Fused (M + L + R + D)		88.5
Proposed: DCGAN + S2SAE		**91.1**

V. Conclusions and Future Work

This work analysed the effectiveness of applying deep unsupervised representation learning algorithms for the task of acoustic scene classification. In this regard, we proposed a novel combination of features generated using a DCGAN and a S2SAE. Results presented indicate that fusing the prediction probabilities of each classifier trained on each representation, it is possible to improve upon the challenge baseline from 74.8 % to 91.1 %, representing a relative percentage increase of 21.8 %. This result indicate the two techniques complement each other in the task of acoustic scene recognition.

Despite CNNs which typically require inputs of fixed dimensionality the proposed S2SAE is able to learn a fixed length representation from variable length audio signals while considering their time-dependent nature. Further, we gave evidence that adversarial networks learn strong representations from spectral features. The applied DCGAN is able to generate spectral images which are highly similar to real training examples. This finding can help in extending the DCGAN framework for other audio related tasks, such as speech synthesis. Both S2SAE and DCGAN are unsupervised representation learning approaches and can be applied for big data, and are less susceptible to overfitting. In future work, we will be testing our system over a wide range of different acoustic classification tasks. We also want to explore the benefits of collecting further data from social multimedia using our purpose built software [32] to train the DCGAN and S2SAE with more real-world audio recordings.

VI. Acknowledgements

This research has received funding from the European Unions's Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

ISBN 978-90-827970-1-5 © EURASIP 2018

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[2] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, Feb. 2018.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[4] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain, 2016, pp. 892–900.

[5] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *Journal of Machine Learning Research*, vol. 19, 2018, 5 pages, to appear.

[6] N. D. Lane, P. Georgiev, and L. Qendro, "DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '15. ACM, 2015, pp. 283–294.

[7] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley, "Unsupervised feature learning based on deep models for environmental audio tagging," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1230–1241, June 2017.

[8] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to sequence autoencoders for unsupervised representation learning from audio," in *DCASE2017*. Munich, Germany: IEEE, Nov. 2017, pp. 17–21.

[9] K. Lee, Z. Hyung, and J. Nam, "Acoustic scene classification using sparse feature learning and event-based pooling," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2013, pp. 1–4.

[10] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 16–34, May 2015.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO 2016)*. Budapest, Hungary: IEEE, Aug. 2016, pp. 1128–1132.

[12] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, Nov. 2017, pp. 85–92.

[13] I. Bae, S. H.and Choi and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proceedings of DCASE'16, satellite to EUSIPCO'16*. IEEE, 2016, pp. 11–15.

[14] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, Sept. 2014, pp. 506–510.

[15] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Nonnegative feature learning methods for acoustic scene classification," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (2015)*. San Diego, CA, USA: arXiv, 2015, https://arxiv.org/abs/1409.1556, 14 pages.

[19] Y. Donahue, J.and Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proceedings of ICML'14*, vol. 32. Beijing, P. R. China: JMLR, 2014, pp. 647–655.

[20] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of CVPR'14*. Columbus, OH, USA: IEEE, 2014, pp. 806–813.

[21] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 3512–3516.

[22] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," in *Proceedings of INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 849–853.

[23] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proceedings of the 2nd International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2017) held in conjunction with the 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, AAAC. San Antonio, TX: IEEE, October 2017, pp. 26–29.

[24] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 25th ACM International Conference on Multimedia, MM 2017*. Mountain View, CA: ACM, Oct. 2017, pp. 478–484.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[26] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[27] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, "Invariant representations for noisy speech recognition," *CoRR*, vol. abs/1612.01928, 2016. [Online]. Available: http://arxiv.org/abs/1612.01928

[28] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proceedings of the 2017 International Conference on Digital Health*. London, UK: ACM, 2017, pp. 53–57.

[29] H. Eghbal-Zadeh, B. Lehner, M. Dorfer, and G. Widmer, "CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks," *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2016.

[30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[31] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112.

[32] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, "CAST a database: Rapid targeted large-scale big data acquisition via small-world modelling of social media platforms," in *Proc. 7th biannual Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, AAAC. San Antionio, TX: IEEE, Oct. 2017, pp. 340–345.

ISBN 978-90-827970-1-5