SOUNDSCAPE EMOTION RECOGNITION VIA DEEP LEARNING

Jianyu Fan
Simon Fraser University
jianyuf@sfu.ca

Fred Tung
Simon Fraser University
ftung@sfu.ca

William Li
Simon Fraser University
dla135@sfu.ca

Philippe Pasquier
Simon Fraser University
pasquier@sfu.ca

ABSTRACT

Deep learning has proven very effective in image and audio classification tasks. Is it possible to improve the performance of emotion recognition tasks based on deep learning approaches? We introduce the strength of deep learning in the context of soundscape emotion recognition (SER). To the best of our knowledge, this is the first study to use deep learning for SER. The main aims are to evaluate the performance of the Convolutional Neural Network (CNN) trained from scratch, the Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) trained from scratch, the CNN trained through supervised fine-tuning, the Support Vector Machines for Regression (SVR), and the combination of CNN and SVR (Transfer Learning) for predicting the perceived emotion of soundscape recordings. The results show that deep learning is a promising approach for improving the performance for SER. Moreover, the fine-tuned VGGish outperforms the other deep-learning frameworks regarding predicting valence. The best performance of predicting arousal is obtained by the CNN trained from scratch. Finally, we analyze the performance of predicting perceived emotion for soundscape recordings in each of Schafer's soundscape categories.

1. INTRODUCTION

A soundscape recording is "a recording of sounds at a given locale at a given time, obtained with one or more fixed or moving microphones" [1]. The research in soundscape emotion recognition (SER) investigates computational systems to recognize the perceived emotion of soundscape recordings. One application of such research is to build automatic sound design systems that help sound designers to create sound effects that evoke target emotions. It can also be an effective tool for engineers to design emotion-based recommendation systems for retrieval of soundscape recordings.

In the last few years, the development of deep learning techniques has greatly improved the performance of audio and image classification tasks. These breakthroughs are caused by the powerful hardware, larger datasets and the designs of neural network architectures [2]. Now, there are publicly available annotated soundscape recordings datasets that can be used in SER studies [3]. Although they are far from being as large as datasets such as AudioSet [4], it is possible to apply the deep learning approaches to improve the performance of SER.

In this work, our goal is to maximize performance and compare five state-of-the-art architectures for the prediction of perceived valence and arousal of soundscape recordings. The five frameworks are fine-tuned Convolutional Neural Network (CNN), CNN trained from scratch, Long Short-Term Memory Recurrent Neural Networks, (LSTM-RNN) trained from scratch, Support Vector Machines for Regression (SVR), and transfer learning. To the best of our knowledge, this is the first study using deep learning approaches to perform SER.

The paper is organized as follows. Section 2 provides background material on SER works, as well as deep neural networks and kernel methods. In Section 3, our dataset and data augmentation are described. The five machine learning frameworks are presented in Section 4. Their performance is discussed in Section 5, while the paper ends in Section 6 with conclusions.

2. BACKGROUND

2.1 Soundscape Emotion Recognition

A great deal of the literature has discussed emotion induction as it relates to music and movies, but not much has been written about the perceived emotion of soundscapes. To obtain the important emotional attributes in soundscapes recordings, Berglund et al. [5] conducted a survey where 100 listeners were asked to evaluate 30 outdoor soundscapes recordings based on 116 perceptual-emotional attributes. Then, the authors used principal component analysis on the survey data to select two critical dimensions: pleasantness and eventfulness. The author indicated that these two dimensions are corresponding to the two dimensions (valence and arousal) [6] in the circumplex model of emotion developed by Russell. Valence represents the pleasantness of a stimulus. Arousal indicates the level of eventfulness [6].

Later, Brocolini et al. [7] investigated the relationship between perceived pleasantness of soundscapes and other subjective variables. The authors asked 120 people to rate the pleasantness of soundscapes, incorporating "global" perceptions such as visual and air quality pleasantness, on a continuous scale from 0 to 10. They found that the acoustic scene has a strong impact on the evaluation of pleasantness.

Thorogood and Pasquier [8] designed the Impress system for predicting perceived pleasantness and eventfulness for soundscapes recordings. The authors selected audio excerpts from the Freesound database and used a segmentation algorithm [9] to search for regions with a consistent soundscape characteristic greater or equal to 4 seconds. The segmentation algorithm was designed based on perceptual categories including background, foreground, and background with foreground sound. Next, the authors extracted low-level audio features and applied the bag-of-frames approach (BOF) [10] to represent audio signals. Multiple linear regression models were designed for the mapping between features and ratings of the perceived emotion of soundscape recordings provided by one expert user. Based on this work, Fan et al. [11] curated a corpus of audio files extracted from the Sound Ideas sound effects library[1] and the World Soundscape Project library[2]. The authors collected annotated soundscape recordings from an online survey where 20 participants annotated 120 soundscape excerpts. Then, they analyzed the level of agreement between annotators and built a gold standard model to predict perceived emotion of soundscape recordings. Their evaluation showed that the models provide strong prediction for both arousal (R^2: 0.816) and valence (R^2: 0.567).

Lundén et al. [12] investigated yet another method mapping audio features of soundscape recordings onto the 2-D emotional space. The authors extracted 93 excerpts from 77 soundscape recordings and invited 33 participants to rate the soundscapes recordings on 2-D emotional space. A Gaussian mixture model is used to cluster audio features. The authors did outlier detection and used the resulting dissimilarity matrix to train two support vector regression models. Evaluation of the model showed a good fit of the Mel-frequency cepstral coefficients (MFCCs) to responses of models of predicting both eventfulness (R^2: 0.83) and pleasantness (R^2: 0.74).

Later, Fan et al. designed a crowdsourcing experiment to collect annotations of perceived valence and arousal of 1,213 soundscape excerpts [3] and published the dataset: Emo-Soundscapes, which is described in section 3. The authors used a ranking-based annotation method instead of rating-based methods. The authors also defined protocols to assess performance of SVR. The results are human competitive (arousal, R^2: 0.853; valence, R^2: 0.622).

2.2 Kernel Methods, Deep Neural Networks and Affective Computing

SVR is one of the most common kernel methods in machine learning [13]. The model maps the data into a high-dimensional feature space based on a non-linear function induced by the selected kernel. SVRs have been used extensively in the affective computing field for music emotion recognition [14], SER [3], and affective video content analysis [17].

A CNN consists of stacked convolutional layers followed by one or more fully connected layers [16]. In affective computing, CNNs have been mostly used for facial expression recognition [15]. Researchers have also done affective video content analysis using CNNs [17]. Regarding audio, Liu et al. presented a CNN framework to classify music emotion based on spectrograms [18]. Their method outperforms traditional methods.

Long Short-Term Memory networks (LSTMs) are a special kind of recurrent neural network (RNN) [19]. The output of an RNN depends not just on the network input but also on a hidden state, which is updated with each new input. Unlike a standard RNN, an LSTM-RNN network can learn long-term dependencies. It contains memory blocks that are composed of a memory cell, an input gate, an output gate and a forget gate. These learnable gates accumulate new information to the cell and control the state of the cell.

In affective computing, Weninger et al. utilized LSTM-RNN to predict perceived valence and arousal of songs using psychoacoustic features [20]. Recently, Malik et al. proposed stacked convolutional and recurrent neural networks for music emotion recognition [16]. Their model has fewer parameters compared with the state-of-the-art methods for the same task and yet achieves the best result reported on the MediaEval2015 Music dataset[3].

3. EMO-SOUNDSCAPES DATASET

We use the Emo-Soundscapes dataset curated by Fan et al. [3]. Emo-Soundscapes is a database for soundscape emotion recognition composed of 1213 6-seconds long monophonic soundscape excerpts. Emo-Soundscapes also contains rankings of the perceived emotion of 1213 soundscape recordings in the 2D valence-arousal space. To collect affective annotations, Fan et al. conducted a crowdsourcing study where 1182 trusted annotators from 74 different countries did pairwise comparisons of all soundscape experts regarding perceived valence and perceived arousal. Each pair has been annotated by three annotators. Based on the pairwise comparisons, the database is sorted along the valence and arousal axis. In this study, we convert the rankings to ratings by mapping the range of ranking values, 1 to 1213, to a range of rating values, 1.0 to −1.0, so that the highest ranked excerpt has the highest rating.

There are two sets of soundscape recordings in the Emo-soundscape dataset. The first set has 600 excerpts that are extracted from soundscape recordings downloaded from Freesound.org[4] shared under Creative Commons licenses[5]. Fan et al. retrieved these soundscape excerpts based on the audio quality and the keywords that are selected following Schafer's soundscape taxonomy [21]. Table 1 shows Schafer's taxonomy. There are six categories. In the first set, there are 100 excerpts per category. The second set contains 613 excerpts that are mixed using the selected excerpts from the first set.

[1] https://www.sound-ideas.com/

[2] https://www.sfu.ca/~truax/wsp.html

[3] http://www.multimediaeval.org/mediaeval2015

[4] http://freesound.org/

[5] https://creativecommons.org/licenses/

Categories	Examples
Natural sounds	Bird, thunder, rain, wind
Human sounds	Laugh, whisper, shouts
Sounds and society	Party, concert, store
Mechanical sounds	Engine, factory
Quiet and silence	Quiet part, silent forest
Sounds as indicators	Clock, church bells

Table 1. Murray Schafer's Taxonomy [3, 21].

We adopt a windowing method to perform data augmentation to artificially enlarge the training set. We chose the window size of 4096 and a step size of 2048. The sample rate of each excerpt is 44100 Hz. First, we cropped the beginning of each excerpt to make the duration of the remaining part of the original excerpt equal to 100 step size, which is 4.644 seconds. This is because there are usually differences regarding the timbre between the beginning of each excerpt and the remaining parts. Second, we segmented the remaining part of the original soundscape excerpt to generate more augmented excerpts. For one soundscape excerpt, we kept selecting 30 consecutive windows as one augmented excerpt and moving one step ahead until we reach the end of a soundscape excerpt. Here, the step size is 20480, which is 10 times of the original step size. One augmented excerpt is 1.393 seconds long. After the data augmentation, we end up having 8491 excerpts. The annotations of each augmented excerpt are the same as the annotations of the original soundscape excerpt.

We used two different sets of audio feature extraction methods. The first method is to use a pre-trained deep neural network for audio classification [22] to extract latent features. The second set of audio features used in this study are 54 dimensions of handcrafted features, which include loudness, energy, perceptual spread, perceptual sharpness, spectral flatness, spectral rolloff, spectral flux, spectral slop, spectral variation, spectral shape, temporal shape, zero cross rate, and 13 MFCCs. Regarding features extraction, we applied the window size of 4096 and the step size of 2048. Both YAAFE [23] and MIRToolbox [24] are used for the feature extraction. Since we extract features from 30 consecutive windows in a soundscape excerpt, and we have 54 dimensions of features for one window, we end up with having a 54×30 feature vector for each augmented excerpt.

4. FRAMEWORKS FOR EMOTION RECOGNITION

In this section, we describe the five frameworks. We train and test all the models twice: once for predicting perceived arousal and again for predicting perceived valence.

4.1 Deep Learning

4.1.1 Fine-tuning

This first framework is based on the fine-tuning strategy. The concept of fine-tuning is to use a model pre-trained on a large dataset, replace its last layers by new layers dedicated to the new task, and fine-tune the weights of the pre-trained network by continuing the backpropagation. The main motivation is that the most generic features of a deep neural network are contained in the earlier layers and should be useful for solving many different tasks. However, later layers of a deep neural network become more and more specific to the task for which the network has been originally trained.

In this work, we fine-tune the VGG-like audio classification model[6] (VGGish) trained on a large YouTube dataset proposed by Hershey et al. [22]. The authors exploited ideas from the image classification task and compared several CNN architectures for the audio classification task. They introduced the YouTube-100M dataset that contains 100 million YouTube videos [22]. Each video is labeled with one or more tags. The audio was then divided into non-overlapping 960 ms frames. Next, the authors computed the log-Mel spectrograms of multiple frames to create 2D image-like patches as the input to the CNNs. These experiments show that the "analogs of the CNNs do well on the audio classification task, and a model using embedding from these classifiers does much better than raw features on the AudioSet" [22].

To adapt VGGish to our task, the last layer is replaced by a fully connected layer composed of 64 neurons. Then the output later contains one neuron to produce the prediction score for valence/arousal. The loss associated with the output of the model is the mean square error. Thus, the model minimizes the sum of squares of differences between the ground truth and the predicted score across training examples. All the layers of the pre-trained model are fine-tuned. We trained the fine-tuned models using the Adam optimizer with a batch size of 32 examples, learning rate of 1×10^{-4}, and epsilon of 1×10^{-8}.

4.1.2 CNN (Trained from Scratch)

We built and trained a CNN from scratch. We used a grid search method to find the number of kernels in each layer, kernel size, learning rate and decay. The model is composed of two convolutional layers and one fully connected layer. The first convolutional layer filters the 54×30×1 input features with 8 kernels of size 5×5×1 with a stride of 1. The second convolutional layer, connected to the first one, uses 8 kernels of size 3×3×8. We used maxpooling (2×2) for the outputs of both convolutional layers. The dropout rate is 0.15. The fully connected layer, connected to the second convolutional layer, is composed of 256 neurons. The ReLU non-linearity is applied to all the convolutional layers and the fully connected layer. The output layer is composed 1 of neuron. We use the linear activation for the output layer to obtain the predicted score. All the weights are initialized based on a Xavier uniform, which draws samples from a uniform distribution within a range. The range is determined by the number of input units and the number of output units. We trained the CNN using the RMSProp optimizer

[6] https://github.com/tensorflow/models/tree/master/research/audioset

with a batch size of 32 examples, learning rate of 1×10^{-3} and decay of 1×10^{-6}.

4.1.3 LSTM-RNN (Trained from Scratch)

We built and trained an LSTM network from scratch. We used a grid search method to find the number of neurons in each layer and parameters, learning rate and decay. Our model is composed of two stacked LSTM units. The network for predicting arousal has 128 neurons in each LSTM unit, while the network for predicting valence has 64 neurons in each LSTM unit. In both cases, the LSTM units use the tanh non-linearity. The output layer is composed of 1 neuron. We use the linear activation for the output layer to obtain the predicted score. Similar to CNN, all the weights are also initialized based on a Xavier uniform. The dimension of the input is 54×30, while 54 is the dimension of the feature vector extracted from one window and 30 is the number of consecutive windows. We trained the LSTM-RNN using the RMSProp optimizer with a batch size of 32 examples, learning rate of 1×10^{-3} and decay of 1×10^{-6}.

4.2 Standard SVR

This model is similar to the baseline framework presented by Fan et al. [3]: two independent SVRs are trained to predict arousal and valence scores separately. The SVR is fed with the handcrafted features detailed in Section 3. All features are normalized using the standard score. We used the BOF approach proposed by Aucouturier and Defreville [10], which represents signals as the long-term statistical distribution of local spectral features. Next, all features are normalized between [0, 1.0]. We eliminated features whose variance is lower than a threshold (0.02). We choose threshold as a heuristic value [3]. We selected the Radial Basis Function (RBF) kernel and used a grid search method to find the parameters C and gamma.

4.3 Transfer learning

Because the VGGish embedding is more semantically compact than raw audio features, we used the VGGish model as a feature extractor to convert the audio input into a semantically meaningful, high-level 128-D embedding that is then fed as input to the SVR outlined in the previous section. Regarding SVR, we selected the RBF kernel and used a grid search method to find the parameters C and gamma. The VGGish is used to improve the performance of the SVR.

5. PERFORMANCE ANALYSIS

To learn and evaluate the various frameworks, the augmented dataset composed of 8491 1.393-seconds segments is shuffled 10 times. Each time, 10% of the dataset is randomly selected for testing, and the remaining 90% is used for training the model.

5.1 General Performance

Table 2 presents the results of using fine-tuned VGGish, CNN and LSTM-RNN trained from scratch, transfer learning, standard SVR and the combination of transfer learning and standard SVR. We use R^2 and MSE to evaluate the performance of the prediction.

Table 2 shows that the fine-tuned VGGish outperforms the other deep-learning frameworks in terms of predicting valence. This is because pre-training VGGish on the YouTube-100M dataset captures timbre features in its early layers and high-level semantic information in the mid layers, which are useful for predicting perceived emotion for soundscapes. Table 2 shows that the highest R^2 of predicting arousal is obtained by the CNN trained from scratch.

An LSTM-RNN can remember things and find patterns across time to make predictions. However, in our case, the performance of the CNN trained from scratch is better than the LSTM-RNN trained from scratch. We think this is because LSTM-RNN is much more complex than CNN and LSTM-RNN need more data for training. Since the CNN trained from scratch sees the entire 54x30 feature vector as input, it is able to learn filters that capture temporal patterns directly [25, 26].

Regarding standard SVR, the performance for predicting arousal (R^2: 0.850) is almost the same as the previous study (R^2: 0.853) [3]. The performance for predicting valence (R^2: 0. 656) is slightly better than the result in the previous study (R^2: 0.622) [3]. Since we provide the same handcrafted feature set and we use the same method, the improvement is solely caused by data augmentation.

Although the performance of transfer learning is not as good as other frameworks in this study, it still reaches those of previous studies regarding valence [12]. Moreover, we combine the transfer learning and the standard SVR by concatenating embedding extracted by VGGish and the handcrafted features together. When we use the concatenated features as the input for a SVR model, the results are significantly better than either that of standard SVR or of transfer learning. This result reveals that VGGish provides generic mid-level audio representations that can be transferred to the task of predicting the perceived valence and arousal.

Framework	Arousal		Valence	
	R^2	MSE	R^2	MSE
VGGish (Fine-tuned)	0.873	0.040	**0.759**	**0.078**
CNN (Trained from Scratch)	**0.892**	**0.035**	0.712	0.096
LSTM-RNN (Trained from Scratch)	0.873	0.042	0.654	0.115
SVR (Standard)	0.850	0.049	0.656	0.114
SVR (Transfer learning)	0.747	0.083	0.665	0.111
SVR (Standard + Transfer learning)	0.864	0.045	0.717	0.094

Table 2. Prediction results for valence and arousal dimensions (R^2: Coefficient of determination, MSE: Mean Square Error)

5.2 Comparisons between Schafer's Categories

We investigate the SER for each soundscape category. For arousal, we use the best model, CNN trained from scratch. Regarding valence, we use the best model, fine-tuned VGGish. We train the models as we described in Section 4. During the test stage, we select the prediction results of test samples that belong to each category and analyze their performance. It is worth pointing out that there are test samples that are mixed soundscape excerpts, which we do not categorize as any specific category and are not included in this analysis. We only analyze the first set, which is composed of 600 excerpts following Schafer's categories as described in Section 3. Table 3 shows the R^2 of predicting arousal and valence for each category.

Categories	Arousal (R^2)	Valence (R^2)
Mechanical sounds	0.903	0.794
Natural sounds	0.884	0.843
Human sounds	0.907	0.808
Sounds and society	0.865	0.547
Quiet and silence	0.646	0.810
Sounds as indicators	0.848	0.900

Table 3. Results of predicting perceived valence and arousal of soundscape recordings that belong to each category.

The previous study indicates that "sounds as indicators carries strong semantic information, which plays an important role in evoking valence to listeners" [27]. It is difficult to model valence with only timbre features [28]. However, our Fine-tuned VGGish model performs very well in predicting valence for "sounds as indicators." Again, we think this is because the model learned high-level semantic information in the mid layers. When predicting the valence of "sounds and society," the R^2 is low. We find that most annotations of valence for "sounds and society" are neutral, and its distribution is close to the uniform distribution. Therefore, the values of valence of soundscape excerpts belonging to "sounds and society" are difficult to differentiate, and it is difficult for machine learning models to learn.

6. CONCLUSIONS

This work presents the performance of deep-learning approaches for soundscape emotion recognition. We have found that the fine-tuned CNN framework is a promising solution for predicting valence and CNN trained from scratch is good at predicting arousal. Intermediate layers, originally trained to perform audio classification tasks, are generic enough to provide mid-level audio representations that can greatly improve soundscape emotion recognition. However, the limited size of the training set (8491 samples) prevents the LSTM-RNN framework from obtaining good performances in terms of R^2.

In future work, we plan to further explore the deep learning design space, and in particular whether residual connections or dense connectivity [29, 30] can improve SER. We also plan to investigate whether such architectures can be learned automatically from data [31].

Acknowledgments

We would like to acknowledge the Social Sciences and Humanities Research Council of Canada and Natural Sciences and Engineering Research Council.

7. REFERENCES

[1] M. Thorogood, J Fan, and P Pasquier. "Soundscape Audio Signal Classification and Segmentation Using Listeners Perception of Background and Foreground Sound," in J. of the Audio Engineering Society, 2016, vol. 64, no.7/8, pp. 484-492.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Proc. Advances in Neural Information Processing Systems (NIPS2012), 2012, pp. 1097–1105.

[3] J. Fan, M. Thorogood, and P. Pasquier, "Emo-Soundscapes: A Dataset for Soundscape Emotion Recognition," in Proc. Int. Conf. on Affective Computing and Intelligent Interaction (ACII2017), 2017.

[4] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dartaset for audio events," in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP2017), 2017.

[5] B. Berglund, M. Nilsson, and O. Axelsson, "Soundscape Psychophysics in Place," in Proc. Int. Congress and Exhibition on Noise Control Engineering, 2007, pp. 3704–3712.

[6] J. A. Russell, "A Circumplex Model of Affect," in J. Personality and Social Psych., 1980, vol. 39, pp. 1161–1178.

[7] L. Brocolini, C. Lavandier, C. Marquis- Favre, M. Quoy, and M. Lavandier, "Prediction and explanation of sound quality indicators by multiple linear regressions and artificial neural networks," in Acoustics, 2012, pp. 2121- 2126.

[8] M. Thorogood and P. Pasquier, "Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment," in Proc. Int. Conf. on New Interfaces for Musical Expression, (NIME2013), 2013, pp. 256–260.

[9] M. Thorogood, J. Fan and, P. Pasquier, "BF-Classifier: Background/Foreground Classification and Segmentation of Soundscape Recordings," in Audio Mostly, 2015.

[10] J. J. Aucouturier and B. Defreville, "Sounds Like a Park: A Computational Technique to Recognize Soundscapes Holistically, Without Source Identifi-

cation," in Proc. Int. Congress on Acoustics, Spain, 2007.

[11] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Soundscape Affect Recognition Using A Dimensional Approach," in J. of the Audio Engineering Society, 2016, vol. 64, no. 9, pp. 646-653.

[12] P. Lundén, O. Axelsson, M. Hurtig, "On Urban Soundscape Mapping: A Computer can Predict the Outcome of Soundscape Assessments," in Proc. Int. Congress and Exposition on Noise Control Engineering: Towards a Quieter Future, 2016, pp. 4725-4732.

[13] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik et al., "Support vector regression machines," A in Proc. Advances in Neural Information Processing Systems (NIPS1997), 1997, pp. 155–161.

[14] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," in Frontiers in psychology, 2013, pp. 1664–1078.

[15] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," in IEEE Transactions on Affective Computing, 2011, vol. 2, no. 2, pp. 92–105.

[16] M. Malik, S. Adavanne, K. Drossos, T. Virtanen, D. Ticha, and R. Jarina, "Stacked convolutional and recurrent neural networks for music emotion recognition." arXiv preprint arXiv:1706.02292, 2017.

[17] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in Proc. Int. Conf. on MultiMedia Modeling (MMM2014), 2014.

[18] X. Liu, Q. Chen, X. Wu, Y. Liu, and Y. Liu, "CNN based music emotion classification" arXiv preprint arXiv:1704.05665v1, 2017.

[19] S. Hochreiter and J. Schmidhuber, Long short-term memory. Neural Computation 9, 1997，pp. 1735–1780.

[20] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP2014), 2014.

[21] R. M. Schafer, The Soundscape: Our Sonic Environment and the Tuning of the World, Rochester, VT: Destiny Books, 1993.

[22] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP2017), 2017.

[23] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an Easy to Use and Efficient Audio Feature Extraction Software," in Proc. Int. Symposium on Music Information Retrieval (ISMIR2010), 2010, pp. 441–446.

[24] O. Lartillot, P. Toiviainen, and T. Eerola, "A Matlab Toolbox for Music Information Retrieval," in: Data Analysis, Machine Learning and Applications. Springer, Berlin, Heidelberg, 2008.

[25] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in Proc. IEEE International Conference on Computer Vision (ICCV2015), 2015.

[26] N. Mehrasa, Y. Zhong, F. Tung, L. Bornn, and G. Mori, "Learning person trajectory representations for team activity analysis," arXiv preprint arXiv:1706.00893, 2017. [Online]. Available:https://arxiv.org/abs/1706.00893

[27] G. H. Orians, M. Dethier, C. Hirshman, A. Kohn, D. Patten, and T. Young, "Sound Indicators: A Review for the Puget Sound Partnership," Washington Academy of Sciences, 2012.

[28] J. Fan, M. Thorogood, and P. Pasquier, "Automatic Recognition of Eventfulness and Pleasantness of Soundscape," in Proc. Audio Mostly, 2015.

[29] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition," In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[30] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. "Densely connected convolutional networks," In Proc. IEEE Conference on Computer Vision and Pattern Recognition (ICCV2017), 2017.

[31] B. Zoph and Q. V. Le. "Neural architecture search with reinforcement learning," in Proc. International Conference on Learning Representations (ICLR2017), 2017.