Audio-based Eating Analysis and Tracking Utilising Deep Spectrum Features

Shahin Amiriparian[1], Sandra Ottl[1], Maurice Gerczuk[1], Sergey Pugachevskiy[1], and Björn Schuller[1,2]
[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
{shahin.amiriparian, sandra.ottl, maurice.gerczuk, sergey.pugachevskiy, schuller}@informatik.uni-augsburg.de
[2] GLAM – Group on Language, Audio & Music, Imperial College London, U. K.

Abstract—**This This paper proposes a deep learning system for audio-based eating analysis on the ICMI 2018 Eating Analysis and Tracking (EAT) challenge corpus. We utilise Deep Spectrum features which are image classification convolutional neural network (CNN) descriptors. We extract the Deep Spectrum features by forwarding Mel-spectrograms from input audio through deep task-independent pre-trained CNNs, including AlexNet and VGG16. We then use the activations of first (*fc6*), second (*fc7*), and third (*fc8*) fully connected layers from these networks as feature vectors. We obtain the best classification result by using the first fully connected layer (*fc6*) of AlexNet for extracting the features from Mel-spectrograms with a window size of 160 ms and a hop size of 80 ms and a *viridis* colour map. Finally, we build Bag-of-Deep-Features (BoDF) which is the quantisation of the Deep Spectrum features. In comparison to the best baseline results on the test partitions of the Food Type and the Likability sub-challenges, unweighted average recall is increased from 67.2 percent to 79.9 percent and from 54.2 percent to 56.1 percent, respectively. For the test partition of the Difficulty sub-challenge the concordance correlation coefficient is increased from .506 to .509.**

Keywords—Deep Spectrum features; pre-trained convolutional neural networks; audio processing; eating analysis.

I. Introduction

Performing automated eating condition recognition from multimodal data is a new research field which has some promising avenues. In digital health care, research has already been done on using self-monitoring apps to help with the treatment of eating disorders [1]–[4]. So far, these apps require their users to manually and consistently track their eating behaviour. Using audio-visual data that is easily collected by smartphones or other wearable devices and automatically analysing it could make it easier and less intrusive for patients to effectively self-monitor their eating habits. In this respect, the ICMI 2018 Eating Analysis and Tracking (EAT) challenge [5] provides data and tasks that deal with the audiovisual analysis of eating behaviours. The likeability subtask of the challenge is also closely related to the field of sentiment analysis: Analysing audio-visual signals could, for example, reveal if a customer of a restaurant is satisfied with their meal. The method we use for the challenge tasks harnesses the spectral information produced by consuming foods of different texture (e.g. crispy or crunchy) [6]. Our approach combines the strengths of the state-of-the-art Deep Spectrum system[1] with the unsupervised Bag-of-Audio-Words model to form a salient feature representation for the challenge tasks. The Deep Spectrum feature system utilises Convolutional Neural Network (CNN) descriptors extracted from Mel-spectrograms using task independent pre-trained image CNNs. It has been successfully applied to a variety of audio analysis tasks [7]–[9]. Its combination with the Bag-of-Audio-Words model, resulting in so-called Bag-of-Deep-Features (BoDF), has been shown to improve noise robustness over the original Deep Spectrum feature space when applied to audio recorded in adverse, in-the-wild conditions [10].

The rest of the paper is structured in the following way: First, we very briefly outline the challenge dataset and tasks in Section II. In Section III, we describe the process we employed to reach our final BoDF feature representation. We present the results achieved in each sub-challenge with our method in Section IV-A, and conclude our work in Section V also by giving an outlook into our plans for future research in this area.

II. Dataset

The challenge dataset [5] contains audio and video recordings of subjects speaking while consuming different types of foods. In total, the training set consists of 945 clips recorded from both read and spontaneous speech by 20 different subjects eating six types of foods (or none at all). The subjects were further asked to rate their likeability of the consumed food items on a continuous scale between 0 and 1 and also specify on a 5 point Likert scale how difficult they found speaking while eating a certain type of food [11]. For our submission, we only considered the audio modality contained in the challenge dataset. Though our approach only improved on the results for the food type sub-challenge, we also include our results for the other two sub-challenges.

[1] https://github:com/DeepSpectrum/DeepSpectrum

III. Deep Feature Representations

Our feature representations that are later used to perform the food-type classification are derived in the following way: We first create Mel-spectrograms from consecutive overlapping segments of the audio instances in the dataset (cf. Section III-A). We then use the plots of these spectrograms as inputs for a pre-trained image classification CNN (cf. Section III-B) to extract Deep Spectrum features (cf. Section III-C). Finally, we quantise the Deep Spectrum features and form a sparse Bag-of-Deep-Features (BoDF) representation for each instance in the dataset (cf. Section III-D).

A. Mel-Spectromgrams

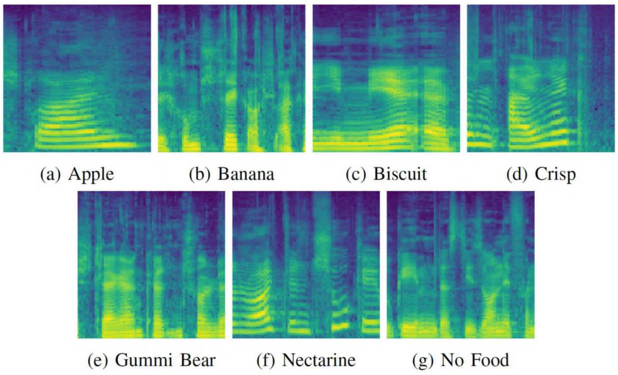
We compute Mel-spectrograms from overlapping segments of the audio instances with a width of 160 ms spaced with a hop size of 80 ms. Mel-spectrograms are derived from the log-magnitude spectrum by dimensionality reduction with a Mel-filter. In our experiments, we use 128 filter banks equally spaced on the Mel-scale:

$$2595 \cdot \log_{10}\left(1+\frac{f}{700}\right)$$

Compared to other scales used for displaying frequencies, this scale is based on how humans perceive frequencies: Lower frequencies can be distinguished with a higher resolution by the human ear. This scale is also used to display the Mel-spectrograms in our experiments. We plot the Mel-spectrograms with the python library librosa [12] to convert them to a format compatible with image-classification CNNs. We further use two different colour mappings for the log-amplitudes in the spectrograms, *viridis* and *magma*. In Figure 1, example Mel-spectrograms from each of the target classes are displayed.

B. Deep Feature Extractors

We use two popular CNN architectures to extract deep representations from the Mel-spectrograms described in Section III-A: The classic AlexNet architecture [14] and the 16-layer variant of VGG architecture introduced by Oxford's Visual Geometry Group [15]. Both networks have been trained on the ImageNet [16] corpus for the task of object categorisation. An overview of their architectural structure is depicted in Table I. AlexNet consists of 5 convolutional layers followed by 3 fully connected layers [14]. Maxpooling is used after the first, second, and fifth convolutional layer and rectified linear units are chosen to provide non-linearity, improving generalisation capabilities. VGG16 distinguishes itself from AlexNet by the use of smaller, 3 x 3 receptive fields in all convolutional layers and also stacks multiple of those layers on top of each other before applying max pooling, resulting in a deeper network architecture.


(a) Apple (b) Banana (c) Biscuit (d) Crisp

(e) Gummi Bear (f) Nectarine (g) No Food

Fig. 1. Example Mel-spectrograms (a – f) extracted from classes of the foodtype sub-challenge. The range of the vertical (frequency) and horizontal (time) axes are [0 – 4,096] Hz and [0 – 0.45] s, respectively. We observe narrow band spectrograms for the classes *Apple*, *Biscuit*, *Crisp*, and *Nectarine* while the classes *Banana*, *Gummi Bear*, and *No Food* have relatively high f0. For all classes we see that the lower frequencies are more dense, i.e. they have a higher amplitude.

We evaluate using activations of each of the fully connected layers (denoted as *fc6*, *fc7* and *fc8* in Table 1) of both networks as intermediate feature representations that are later quantised and bagged to form BoDF. For both networks the last fully connected layers have 1,000 neurons, while the *fc6* and *fc7* layers consist of 4,096 neurons each.

C. Deep Spectrum Features

We use a state-of-the-art system which features an extraction pipeline for our Deep Spectrum feature representations. The components of this pipeline are shown in the left part of Figure 2. As described in Section III-A, we first

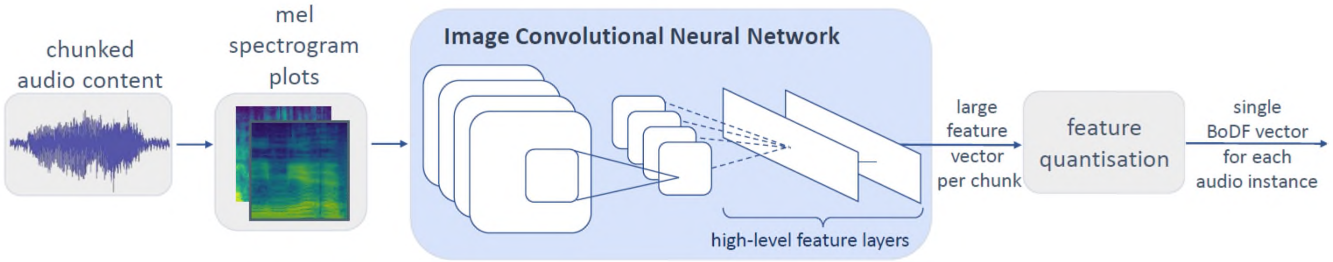
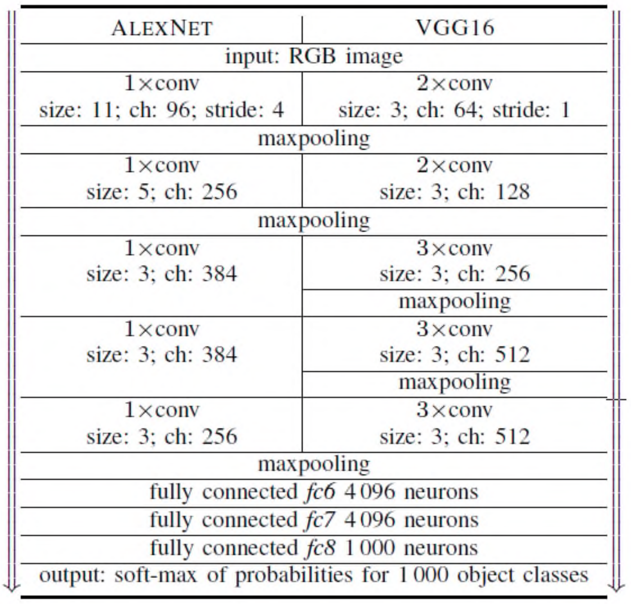


Fig. 2. Plots of Mel-spectrograms are generated from overlapping segments of the audio instances. Afterwards, they are fed to an image Convolutional Neural Network (CNN) and the activations of a specific layers form Deep Spectrum features. At this point, each segment is represented by a large feature vector. Finally, these extracted vectors are bagged to form a single sparse Bag-of-Deep-Features (BoDF) vector for each instance. This last step is performed using openXBOW, our open-source toolkit for the generation of bag-of-words representations [13].

create plots of Mel-spectrograms from overlapping segments (windows of 160 ms with a hop size of 80 ms) of the audio instances. For this, we utilise the audio and music analysis library librosa [12]. Our choice of Mel-spectrograms is based on their successful application to a wide range of audio analysis tasks [10], [17]–[19]. We then feed the plots to an image classification CNN as input and extract the activations of a late layer as feature vectors.

TABLE I. COMPARISON OF ARCHITECTURAL FEATURES OF THE TWO CNNS USED AS DEEP SPECTRUM EXTRACTORS, ALEXNET AND VGG16. CONV IS USED TO ABBREVIATE CONVOLUTIONAL LAYER AND CH DENOTES CHANNELS. THE TABLE IS ADAPTED FROM [7].



ALEXNET	VGG16
input: RGB image	
1×conv size: 11; ch: 96; stride: 4	2×conv size: 3; ch: 64; stride: 1
maxpooling	
1×conv size: 5; ch: 256	2×conv size: 3; ch: 128
maxpooling	
1×conv size: 3; ch: 384	3×conv size: 3; ch: 256
	maxpooling
1×conv size: 3; ch: 384	3×conv size: 3; ch: 512
	maxpooling
1×conv size: 3; ch: 256	3×conv size: 3; ch: 512
maxpooling	
fully connected *fc6* 4 096 neurons	
fully connected *fc7* 4 096 neurons	
fully connected *fc8* 1 000 neurons	
output: soft-max of probabilities for 1 000 object classes	

D. Bag-of-Deep-Features

The last step in creating our feature representation is to quantise and bag the extracted Deep Spectrum features. For this, we first create a codebook by identifying a certain number of 'deep audio words' from the Deep Spectrum features of the training data. A fixed length histogram representation of each audio instance is then formed by quantising the original feature space according to this codebook. The frequency of each deep audio word in a given instance is shown in the histogram [13], [20], [21]. Specifically, we first standardise our Deep Spectrum features and then random sample a codebook with fixed size from the training partition. Histograms for each audio instance are then created by applying each input feature vector (from all partitions) to a fixed number of its closest codebook vectors. We finally apply logarithmic term weighting to these histograms.

TABLE II. UAR ACHIEVED BY DIFFERENT CONFIGURATIONS OF BODF REPRESENTATIONS ON THE TRAINING PARTITION USING LOSO-CV FOR ALL THREE SUB-CHALLENGES. WE EVALUATE TWO COLOUR MAPS FOR PLOTTING THE MEL-SPECTROGRAMS AND EXTRACT ACTIVATIONS OF ALL THREE FULLY CONNECTED LAYERS OF BOTH ALEXNET AND VGG16.

	colour map	ALEXNET			VGG16		
		fc6	fc7	fc8	fc6	fc7	fc8
Food Type	viridis	**77.0**	74.2	72.1	73.1	73.8	72.0
	magma	75.8	76.0	73.3	75.1	74.9	73.6
Likability	viridis	**59.1**	59.0	58.1	58.6	58.0	57.7
	magma	58.7	58.5	57.0	58.4	58.1	57.0
Difficulty	viridis	**.317**	.309	.301	.316	.310	.300
	magma	.315	.311	.300	.312	.303	.301

We tried different combinations of values for the size of the random sampled codebooks and the number of codebook words applied to each input vector by evaluating the performance of the resulting feature representation in Leave-One-Speaker-Out Cross-Validation (LOSO-CV) using a linear Support Vector Machine (SVM) classifier. We achieved the best results with 11,000 random sampled codebook words and assigning the 60 nearest of those words to each input vector.

IV. CLASSIFIER AND EVALUATION METRICS

For the classification tasks, we use the baseline linear SVM codes with our BoDF features. The code optimises the classifier's complexity parameter using LOSO-CV on the training partition on a logarithmic scale between 10^{-4} and 10^{0} with a step size of 10^{1}. As described in the challenge baseline paper, unweighted average recall (UAR) is used as evaluation metric for the subtasks of food type and likeability classification, whereas the concordance correlation coefficient (ccc) is used for the Difficulty sub-challenge.

A. Results

We first run experiments on the training partition to find the best configuration for the BoDF features, evaluating different combinations of network architecture, extraction layer and colour map for the Mel-spectrogram plots. Table II shows the results achieved by each of the tested configurations on all three sub-challenges. Based on these results, we choose to feed AlexNet Mel-spectrograms plotted with the *viridis* colour map and extract the activations of its *fc6* layer as Deep Spectrum features which we then quantise and bag to form our BoDF representation.

Table III compares the challenge baseline with the results achieved by our BoDF system. For the Food Type sub-challenge, our approach improves the UAR from 64.3% to 77.0% on the training partition (with LOSO-CV) and from 67.2% to 79.9% on the test partition. Even though we mainly focused on this sub-challenge our results show slight improvement upon the test set results of the Likability and Difficulty baselines provided by the challenge authors. In the Likability challenge, we achieve a lower UAR on the training partition with LOSO-CV, but slightly improve upon the test set. For the Difficulty sub-challenge, we receive considerably worse result using LOSO-CV on the training partition but beat the challenge baseline on the test partition. A confusion matrix from the Food Type labels of the training set during LOSO-CV is also shown in Figure 3, from which we can observe almost no confusion for the class *No Food* and relatively high confusion between the classes *Apple* and *Nectarine*. The classes *No Food* and *Nectarine* achieve the highest (97.9 %) UAR and the lowest (60.2 %) UAR, respectively.

TABLE III. COMPARISON OF THE BEST RESULTS ACHIEVED BY OUR BODF SYSTEM WITH THE CHALLENGE BASELINES. THE APPLIED FEATURE VECTOR FOR ALL SUB-CHALLENGES IS OBTAINED FROM THE *FC6* LAYER OF ALEXNET WITH *VIRIDIS* COLOUR MAP.

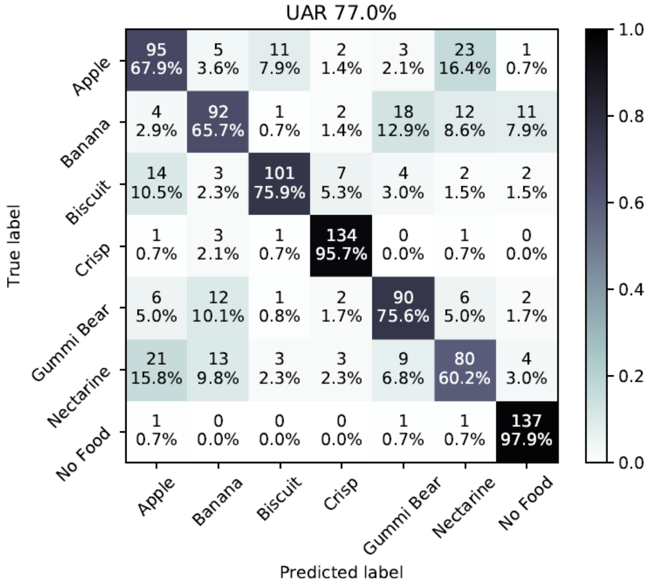
	Baseline		BoDF	
	LOSO-CV	test	LOSO-CV	test
FOOD TYPE	64.3	67.2	77.0	**79.9**
LIKEABILITY	66.5	54.2	59.1	**56.1**
DIFFICULTY	.470	.506	.317	**.509**

V. CONCLUSIONS

We have shown that a system using quantised, deep feature representations can achieve strong performance for automatic audio-based analysis of eating behaviours on the EAT Challenge. Considering the fact that the applied CNNs in our BoDF system are primarily designed for image classification tasks, we demonstrated that using BoDF, it is possible to extract very strong representations from the EAT audio recordings and improve upon all test partitions of the challenge baselines. With regard to our approach and the best baseline results, which are both extracted from quantised (and bagged) features, we see that unsupervised representation learning directly from raw audio files lead to superior performance over hand-crafted, expert-designed features.

We also observed that changing the colour maps of the input Mel-spectrograms leads to different classification results. We assume this is highly related to the distribution of the colours of the images in the ImageNet database [16], which were used for pre-training the CNNs.

In future work, we plan to fine-tune both AlexNet and VGG16 on much larger real-world EAT datasets from social multimedia, utilising our toolkit for efficient large-scale big data collection [22]. It is also interesting to test a wider range of colour maps for the audio plots and other pre-trained CNN architectures, including ResNets [23] or InceptionV4 [24], and analyse their effect on the recognition rate for each sub-challenge. We want to fuse the features and the models obtained from the *fc6*, *fc7*, and *fc8* layers of the used CNNs in order to analyse their complementarity. Finally, as the consumption of food while speaking negatively affects automatic speech recognition systems trained on clear speech [5], audio-based analysis could be a possible solution for adapting those systems to different eating conditions.



	Apple	Banana	Biscuit	Crisp	Gummi Bear	Nectarine	No Food
Apple	95 67.9%	5 3.6%	11 7.9%	2 1.4%	3 2.1%	23 16.4%	1 0.7%
Banana	4 2.9%	92 65.7%	1 0.7%	2 1.4%	18 12.9%	12 8.6%	11 7.9%
Biscuit	14 10.5%	3 2.3%	101 75.9%	7 5.3%	4 3.0%	2 1.5%	2 1.5%
Crisp	1 0.7%	3 2.1%	1 0.7%	134 95.7%	0 0.0%	1 0.7%	0 0.0%
Gummi Bear	6 5.0%	12 10.1%	1 0.8%	2 1.7%	90 75.6%	6 5.0%	2 1.7%
Nectarine	21 15.8%	13 9.8%	3 2.3%	3 2.3%	9 6.8%	80 60.2%	4 3.0%
No Food	1 0.7%	0 0.0%	0 0.0%	0 0.0%	1 0.7%	1 0.7%	137 97.9%

Fig. 3: Confusion Matrix from the best classification labels for the audio files of the test set of the Food Type sub-challenge.

REFERENCES

[1] J. J. Aardoom, A. E. Dingemans, and E. F. Van Furth, “E-health interventions for eating disorders: Emerging findings, issues, and opportunities,” Current psychiatry reports, vol. 18, no. 4, p. 42, 2016.

[2] A. S. Juarascio, S. P. Goldstein, S. M. Manasse, E. M. Forman, and M. L. Butryn, “Perceptions of the feasibility and acceptability of a smartphone application for the treatment of binge eating disorders: Qualitative feedback from a user population and clinicians,” International journal of medical informatics, vol. 84, no. 10, pp. 808–816, 2015.

[3] A. S. Juarascio, S. M. Manasse, S. P. Goldstein, E. M. Forman, and M. L. Butryn, “Review of smartphone applications for the treatment of eating disorders,” European Eating Disorders Review, vol. 23, no. 1, pp. 1–11, 2015.

[4] C. G. Fairburn and E. R. Rothwell, “Apps and eating disorders: A systematic clinical appraisal,” International Journal of Eating Disorders, vol. 48, no. 7, pp. 1038–1046, 2015.

[5] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. E.-D. Mousa, and B. Schuller, “I hear you eat and speak: automatic recognition of eating condition and food type, use-cases, and impact on asr performance,” PloS one, vol. 11, no. 5, p. e0154486, 2016.

[6] C. Dacremont, “Spectral composition of eating sounds generated by crispy, crunchy and crackly foods,” Journal of texture studies, vol. 26, no. 1, pp. 27–43, 1995.

[7] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, “Snore sound classification using imagebased 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, (acceptance rate: 50 %).

[8] 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.

[9] 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.

[10] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, “Bag-of-Deep-Features: Noise-Robust Deep Feature Representations for Audio Analysis,” in Proceedings 31st International Joint Conference on Neural Networks (IJCNN), IEEE. Rio de Janeiro, Brazil: IEEE, July 2018, 8 pages, to appear.

[11] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller, “Eat-: The icmi 2018 eating analysis and tracking challenge,” in Proceedings of the 2018 on International Conference on Multimodal Interaction. ACM, 2018, pp. 559–563.

[12] B. McFee, M. McVicar, O. Nieto, S. Balke, C. Thome, D. Liang, E. Battenberg, J. Moore, R. Bittner, R. Yamamoto, D. Ellis, F.-R. Stoter, D. Repetto, S. Waloschek, C. Carr, S. Kranzler, K. Choi, P. Viktorin, J. F. Santos, A. Holovaty, W. Pimenta, and H. Lee, “librosa 0.5.0,” Feb. 2017. [Online]. Available: https://doi:org/10:5281/zenodo:293021

[13] M. Schmitt and B. Schuller, “openxbow – introducing the passau opensource crossmodal bag-of-words toolkit,” Journal of Machine Learning Research, vol. 18, 2017, 5 pages.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems. Curran Associates, Inc., 2012, vol. 25, pp. 1097–1105. [Online]. Available: http://papers:nips:cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks:pdf

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[17] S. Panwar, A. Das, M. Roopaei, and P. Rad, “A deep learning approach for mapping music genres,” in System of Systems Engineering Conference (SoSE), 2017 12th. IEEE, 2017, pp. 1–5.

[18] J. Salamon and J. P. Bello, “Unsupervised feature learning for urban sound classification,” in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 171–175.

[19] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, “Dcase 2016 acoustic scene classification using convolutional neural networks,” in Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), 2016, pp. 95–99.

[20] S. Pancoast and M. Akbacak, “Bag-of-audio-words approach for multimedia event classification.” in Proceedings of INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association. Portland, OR, USA: ISCA, 2012, pp. 2105–2108.

[21] “Softening quantization in bag-of-audio-words,” in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 1370–1374.

[22] S. Amiriparian, S. Pugachevskiy, N. Cummins, S. Hantke, J. Pohjalainen, G. Keren, and B. Schuller, “CAST a database: Rapid targeted largescale 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 Antonio, TX: IEEE, October 2017, pp. 340–345.

[23] 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.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning.” In AAAI, vol. 4, 2017, p. 12.