TASK 1B DCASE 2021: AUDIO-VISUAL SCENE CLASSIFICATION WITH SQUEEZE-EXCITATION CONVOLUTIONAL RECURRENT NEURAL NETWORKS

Technical Report

Javier Naranjo-Alcazar[1,2], *Sergi Perez-Castanos*[2], *Maximo Cobos*[2], *Francesc J. Ferri*[2], *Pedro Zuccarello*[1]

[1] Instituto Tecnológico de Informática, València, Spain {jnarnajo, pzuccarello}@iti.es
[2] Universitat de València, Burjassot, Spain, {pecaser@alumni.uv.es, {maximo.cobos, francesc.ferri}@uv.es}

ABSTRACT

Automatic scene classification has always been one of the core tasks in every edition of the DCASE challenge. Until this edition, such classification was performed using only audio data, and so the problematic was defined as Acoustic Scene Classification (ASC). In this 2021 edition, audio data is accompanied with visual data, providing additional information that can be jointly exploited for achieving higher recognition accuracy. The proposed approach makes use of two separate networks which are respectively trained in isolation on audio and visual data, so that each network specializes in a given modality. After training each network, the fusion of information from the audio and visual subnetworks is performed at two different stages. The early fusion stage combines features resulting from the last convolutional block of the respective subnetworks at different time steps to feed a bidirectional recurrent structure. The late fusion stage combines the output of the early fusion stage with the independent predictions provided by the two subnetworks, resulting in the final prediction. For the visual subnetwork, a VGG16 architecture pretrained on the Places365 dataset is used, applying a fine-tuning strategy over the Challenge dataset. On the other hand, the audio subnetwork is trained from scratch and uses squeeze-excitation techniques as in previous contributions from this team. As a result, the final accuracy of the system is 92% on development split, outperforming the baseline by 15 percentage points.

Index Terms— Deep Learning, Convolutional Neural Network, Acoustic Scene Classification, Gammatone, DCASE2021

1. INTRODUCTION

Automated sound analysis can lead to robust solutions in many areas/applications, either using audio alone or by combining it with other sources of information, such as images or other types of sensors. One of these applications is scene classification. This problematic can be understood as a supervised problem in which a scene must be classified into one of the predefined classes (e.g. park, airport, etc.) Until the 2021 edition, the DCASE challenge had always proposed Acoustic Scene Classification (ASC) as a core task to be addressed, i.e. the detection of different scenes, using only audio recordings [1, 2, 3, 4].

The aim of DCASE 2021 Task 1b is to improve a scene classifier by using not only audio but also the corresponding video information. The use of multiple modalities opens new opportunities that, presumably, may lead to higher recognition accuracies than the ones obtained from each modality separately [5]. The dataset provided for the task is known as TAU Urban Audio Visual Scenes 2021 [6]. In turn, this dataset is divided into two splits, the development split and the evaluation split. While the development split contains scenes recorded in 10 cities, the evaluation one contains scenes from 12 cities (there are two cities unseen in the development set). The total number of hours in the development split is 34 hours. One thing to note is that both video and audio data are provided as clips having a duration of 10 seconds. However, it is intended that the system will be able to classify clips having only a duration of one second. Another important aspect is that the use of external data is allowed in this task. In fact, the submission described in this report makes use of the *places365* dataset [7].

The approach proposed in this work consists of two components or modules (an audio module and a visual module) that are further trained together to achieve a more robust solution. The visual model is based on a VGG16 convolutional neural network (CNN) pre-trained with the *places365* dataset. The training procedure of this component is based on a transfer learning strategy with fine tuning that makes use of the dataset provided in the Challenge. On the other hand, the audio ASC system is very similar to those presented by the team in previous years. First, a 3-channel audio input is obtained using a Gammatone filter bank representation with 64 bands. Finally, a CNN incorporating squeeze-excitation (SE) techniques is trained from scratch. The hyperparameters used in this module are very similar to the ones used in previous submissions. Finally, the audio and video modules with frozen weights are combined into a multimodal recurrent structure that performs information fusion both at early and late stages. It should be noted that both the independent components and the final system outperform the baseline. More information on these two modules and their combination is given in Section 2, while the results obtained in the development stage are presented in Section 3.

2. METHOD

2.1. Audio Module

2.1.1. Audio Input Representation

Following the idea of last year submissions [8, 9] a multi-channel 3D audio representation is selected. The chosen option has been the one previously known in our works as LRD (left-right-difference) [8, 9]. In this submission, a Gammatone filter bank-based representation has been chosen for providing a slightly superior performance than other alternatives (e.g. Mel-scale filter banks) in preliminary tests. Note that Gammatone filter banks have also been widely adopted in other machine listening applications as well as in other participations of this team [10, 11, 12, 13].

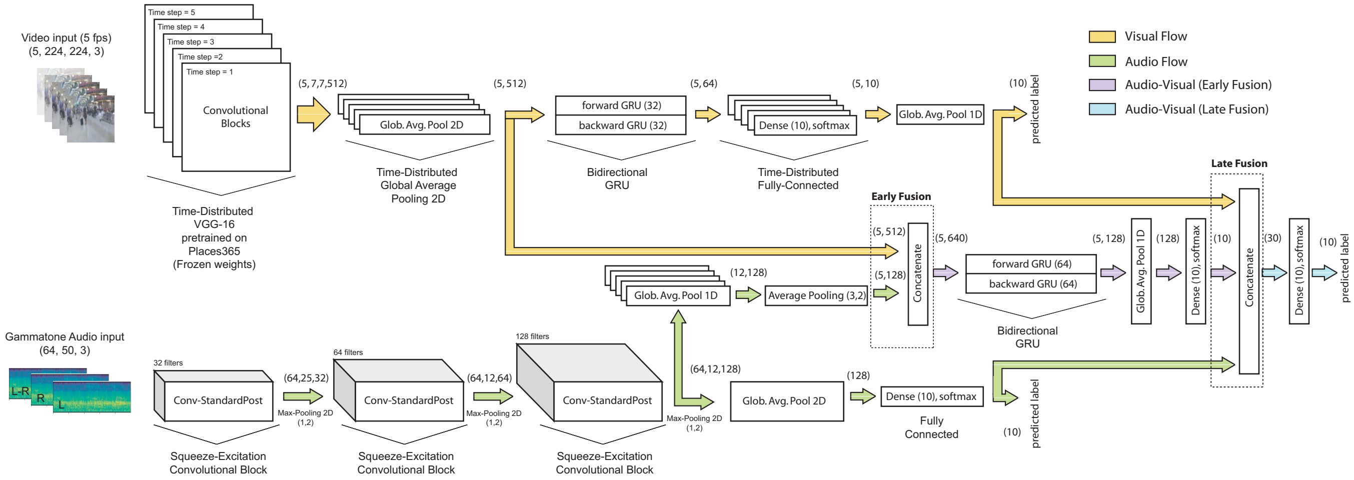

Figure 1: Proposed network achitecture for audiovisual scene classification.

All representations are calculated with a window size of 40 ms with 50% overlapping, using a sampling rate of 44.1 kHz. According to past editions and state-of-the-art research, spectral resolution can be a decisive factor [14]. In our case, it has been decided to use 64 bands. Gammatone representations were computed by using the Auditory Toolbox presented in [15] with Python implementation. Taking the above details into account, one second of audio results in a tensor input of size (64, 50, 3).

2.1.2. Audio Subnetwork

The convolutional network trained with the audio information (see Figure 1) is composed of blocks known as *Conv-StandardPOST*. These blocks were proposed in [16]. The aim of these blocks is to achieve improved accuracy by recalibrating the internal feature maps using residual [17] and squeeze-excitation techniques [18, 19]. For more insight about this choice, please see [16] where *Conv-StandardPOST* is fully explained and compared to other competing blocks. After each Max Pooling layer, a Dropout [20] layer with 0.3 rate is also implemented in order to avoid overfitting. The output feature maps from the convolutional blocks are summarized by global average pooling and fed to a fully-connected layer with softmax activation for classification. This subnetwork is trained from scratch on the whole dataset using only audio data.

2.2. Visual Module

2.2.1. Visual Input Representation

The visual input is adapted to match the pre-trained VGG16 architecture, which accepts color images of size 224×224 pixels. Moreover, as visual scene recognition does not require a very high frame rate (images do not change that much from frame to frame), the videos from the dataset are subsampled for obtaining a frame rate of 5 frames per second (fps). Therefore, a one-second video clip results in a tensor shape of (5, 224, 224, 3).

2.2.2. Visual Subnetwork

The visual module is based on the VGG16 CNN architecture [21] pretrained on the *places365* dataset [7]. With the aim of processing temporal information extracted from multiple frames, a time-distributed structure with frozen weights is considered. The outputs from each time step (5 temporal steps) are globally averaged channel-wise, resulting in a sequence of 512 output features. This sequence is fed to a bidirectional GRU layer and the returned sequences are processed by a time-distributed fully-connected layer with softmax activation, resulting in a predicted label for each time step. The final label is taken as the temporal average of the predictions. This subnetwork is trained on the visual data only, with trainable weights only on the recurrent and final dense layer.

2.3. Full Audio-Visual Network

The complete audio and visual modules described above are then merged into a full audio-visual framework that combines information from both modalities at two different levels. On an early fusion stage, the output of the last convolutional block of the audio and visual modules are concatenated into a sequence of 640 features. To achieve this, the feature maps of the audio module are turned into a temporal sequence matching the temporal resolution of the visual data (i.e. 5 fps) using global and average pooling operators. A bidirectional GRU processes the sequence, and a new prediction is created by stacking a global average pooling and a dense layer. A late fusion stage receives the predictions from the independent modalities as well as the one resulting from their combination and produces the final prediction with a dense layer with softmax activation.

2.4. Experimental details

2.4.1. Training

The whole network is trained in three steps. The first step corresponds to the training of the audio module from scratch using audio data. The second step trains the recurrent and classification parts of the visual module (the convolutional blocks use frozen weights from the pre-trained network). In the last step, the whole audio-visual network is trained using frozen weights from the audio and visual modules. A fine-tuning strategy is finally followed, unfreezing all the weights and using a very small learning rate. The loss function used at each training step was categorical cross-entropy. The optimizer used was Adam [22] with default parameters. The models were trained with a maximum of 200 epochs. Batch size was set to 32 for training the independent subnetworks and 16 for the complete audio-visual network due to memory constraints. The learning rate started with a maximum value of 0.001 decreasing with a factor of 0.5 in case of no improvement in validation accuracy after 20 epochs. In the last fine-tuning with all trainable weights, the starting learning rate was 10^{-5}. The training is considered as early finished in case of no improvement in validation accuracy after 50 epochs. Due to the competition context, mixup [23] with $\alpha = 0.4$ has been implemented. Keras with Tensorflow backend was used to implement the models of this submission.

3. RESULTS

The results obtained by the scene classification system, using each module separately and together, are shown below. As observed in Table 1, the baseline is exceeded in all 3 cases. The audio network improves the baseline by 4 percentage points while the visual network improves the baseline by about 22 points. As can be noticed (and this being the aim of this Challenge) merging both sources of information leads to a more accurate system. The best performing module (visual) improves by almost 4 points when combined with auditory information, leading to a final accuracy of 90.0%.

	Audio	Visual	Audio-Visual
Challenge Baseline	65.1	64.9	77.0
Proposed system	69.0	86.5	90.0

Table 1: Accuracy (%) results obtained compare with the proposed baseline

4. CONCLUSION

Nowadays, there are a multitude of machine learning solutions adapted to data from different domains (image, speech, audio, ...). However, mixed solutions exploiting jointly data from multiple domains are not as well explored. Task 1b of the 2021 edition of the DCASE Challenge proposes a modification of the classic ASC task to turn it into an audio-visual task where, apart from the audio of the scene, the video is also available. Following the line of research of this team during the two previous editions, an ASC system using squeeze-excitation techniques using Gammatone audio spectrograms and merged with a well-known architecture for computer vision (VGG16), has been proposed. The temporal structure of the data is handled by a recurrent architecture, performing data fusion from both modules both at an early and late stages. The results show that merging both sources of information makes the system more accurate, achieving a 15 percentage point accuracy improvement with respect to the Challenge baseline.

5. ACKNOWLEDGEMENTS

The participation of Dr. Cobos and Dr. Ferri is supported by ERDF and the Spanish Ministry of Science, Innovation and Universities under Grant RTI2018-097045-B-C21, as well as grants AICO/2020/154 and AEST/2020/012 from Generalitat Valenciana.

6. REFERENCES

[1] A. Mesaros, T. Heittola, and T. Virtanen, “Acoustic scene classification: an overview of dcase 2017 challenge entries,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2018, pp. 411–415.

[2] ——, “Acoustic scene classification in dcase 2019 challenge: Closed and open set classification and data mismatch setups,” 2019.

[3] M. Valenti, A. Diment, G. Parascandolo, S. Squartini, and T. Virtanen, “Dcase 2016 acoustic scene classification using convolutional neural networks,” in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99.

[4] 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, 2015.

[5] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, “Audio-visual scene classification: analysis of dcase 2021 challenge submissions,” 2021.

[6] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, “A curated dataset of urban scenes for audio-visual scene analysis,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: https://arxiv.org/abs/2011.00030

[7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[8] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, M. Cobos, and F. J. Ferri, “Cnn depth analysis with different channel inputs for acoustic scene classification,” *arXiv preprint arXiv:1906.04591*, 2019.

[9] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, “Task 1 dcase 2020: Asc with mismatch devices and reduced size model using residual squeeze-excitation cnns,” DCASE2020 Challenge, Tech. Rep, Tech. Rep., 2020.

[10] S. Tabibi, A. Kegel, W. K. Lai, and N. Dillier, “Investigating the use of a gammatone filterbank for a cochlear implant coding strategy,” *Journal of neuroscience methods*, vol. 277, pp. 63–74, 2017.

[11] Z. Zhang, S. Xu, S. Cao, and S. Zhang, “Deep convolutional neural network with mixup for environmental sound classification,” in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2018, pp. 356–367.

[12] S. Perez-Castanos, J. Naranjo-Alcazar, P. Zuccarello, and M. Cobos, “Anomalous sound detection using unsupervised and semi-supervised autoencoders and gammatone audio representation,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 145–149.

[13] ——, “Listen carefully and tell: An audio captioning system based on residual learning and gammatone audio representation,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 150–154.

[14] K. J. Piczak, "The details that matter: Frequency resolution of spectrograms in acoustic scene classification," *Detection and Classification of Acoustic Scenes and Events*, pp. 103–107, 2017.

[15] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep*, vol. 10, no. 1998, 1998.

[16] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112 287–112 296, 2020.

[17] 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.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2018.00745

[19] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

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

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

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.