DCASE 2021 Task 1B: Technique Report

AIT Group: Lam Pham, Anahid Jalali, Alexander Schindler, Mina Schütz, Jasmin Lampert, Sven Schlarb, Ross King
Competence Unit Data Science & Artificial Intelligence,
Center for Digital Safety & Security,
Austrian Institute of Technology, Austria.

***Abstract*—This report shows a deep learning framework for audio-visual scene classification (SC). Our extensive experiments, which are conducted on DCASE Task 1B development dataset, achieve the best classification accuracy of 82.2%, 91.1%, and 93.9% with audio input only, visual input only, and both audio-visual input, respectively.**

***Index Terms*—Audio-visual scene, pre-trained model, Imagenet, AudioSet, deep learning framework.**

I. Introduction

Analysing both audio and visual (or image) information from videos has opened a variety of real-life applications such as detecting the sources of sound in videos [1], lip-reading by using audio-visual alignment [2], or source separation [3]. Joined audio-visual analysis shows effective compared to the visual only data proven in tasks of video classification [4], multi-view face recognition [5], emotion recognition [6], or video recognition [7]. Although a number of audio-visual datasets exit, they mainly focus on human for specific tasks such as detecting human activity [8], action recognition [9], classifying sport types [10], [11], or emotion detection [6]. DCASE community [12] has released an audio-visual dataset used for DCASE 2021 Task 1B challenge of classifying ten different scene contexts [13]. We therefore evaluate this dataset by leveraging deep learning techniques.

II. Deep learning frameworks proposed

As we aim to evaluate individual roles of audio and visual features within SC task, deep learning frameworks using either audio or visual input are presented in separate sections.

A. Audio-based deep learning frameworks

In audio-based deep learning frameworks, audio recordings are firstly transformed into spectrograms, referred to as front-end low-level feature extraction. As using ensemble of either different spectrogram inputs [14]–[17] or different deep neural networks [18], [19] has been a rule of thumb to improve audio-based SC performance, we therefore uses three spectrogram transformation methods: Mel filter (MEL) [20], Gammatone filter (GAM) [21], and Constant Q Transform (CQT) [20]. These spectrograms are then split into ten 50%-overlapping patches, each which represents for 1-second audio segment. To enforce back-end classifiers, mixup data augmentation [22], [23] is applied on these patches of spectrogram before feeding into a VGGish network for classification as shown in Table I. As description shown in Table I, the VGGish network architecture contains sub-blocks which perform convolution (Conv), batch normalization (BN) [24], rectified linear units (ReLU) [25], average pooling (AV), global average pooling (GAP), dropout (Dr) [26], fully-connected (FC) and Softmax layers. In total, we have 12 convolutional layers and 2 fully-connected layers containing trainable parameters that makes the proposed network architecture like VGG14 [27]. We refer three audio-spectrogram based frameworks proposed to as *audio-CQT-Vgg14, audio-GAM-Vgg14*, and *audio-MEL-Vgg14*, respectively.

TABLE I
The VGG14 network architecture used for audio-based frameworks.

Network architecture	Output
BN - Conv [3×3]@64 - ReLU - BN - Dr (25%)	128×128×64
BN - Conv [3×3]@64 - ReLU - BN - AP - Dr (25%)	64×64×64
BN - Conv [3×3]@128 - ReLU - BN - Dr (30%)	64×64×128
BN - Conv [3×3]@128 - ReLU - BN - AP - Dr (30%)	32×32×128
BN - Conv [3×3]@256 - ReLU - BN - Dr (35%)	32×32×256
BN - Conv [3×3]@256 - ReLU - BN - Dr (35%)	32×32×256
BN - Conv [3×3]@256 - ReLU - BN - Dr (35%)	32×32×256
BN - Conv [3×3]@256 - ReLU - BN - AP - Dr (35%)	16×16×256
BN - Conv [3×3]@512 - ReLU - BN - Dr (35%)	16×16×512
BN - Conv [3×3]@512 - ReLU - BN - Dr (35%)	16×16×512
BN - Conv [3×3]@512 - ReLU - BN - Dr (35%)	16×16×512
BN - Conv [3×3]@512 - ReLU - BN - GAP - Dr (35%)	512
FC - ReLU - Dr (40%)	1024
FC - Softmax	$C = 10$

TABLE II
The network architectures [28] proposed for visual based frameworks

Network architectures	Image inputs
1/ Xception	299×299×3
2/ Vgg19	224×224×3
3/ Resnet50	224×224×3
4/ InceptionV3	299×299×3
5/ MobileNetV2	224×224×3
6/ DenseNet121	299×299×3
7/ NASNetLarge	331×331×3

B. Visual-based deep learning frameworks

We use the network architectures from Keras application library [28], which are considered as benchmarks for evaluating ImageNet dataset [29] as shown in Table II, for visual-based deep learning frameworks. In order to directly train image frame inputs with the network architectures in Table II, we reduce the C dimension of the final fully connected layer ($C = 1000$ that equals to the number of object detection defined in ImageNet dataset) to $C = 10$ that matches the number of scene categories classified. The visual-based deep learning frameworks proposed are referred to as *visual-image-Xception, visual-image-Vgg19, visual-image-Resnet50, visual-image-InceptionV3, visual-image-MobileNetV2, visual-image-DenseNet121*, and *visual-image-NASNetLarge*, respectively. Same as audio-based approaches, the final classification accuracy of visual-based frameworks is obtained by applying late fusion of individual frameworks.

III. Evaluation Setting

A. TAU Urban Audio-Visual Scenes 2021 dataset [13]

This dataset is referred to as DCASE Task 1B Development, which was proposed for DCASE 2021 challenge [12]. The dataset in slightly unbalanced and contains both acoustic and visual information, recorded from 12 large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna. It consists of 10 scene classes: airport, shopping mall (indoor), metro station (underground), pedestrian street, public square, street (traffic), traveling by tram, bus and metro (underground), and urban park, which can be categorised into three meta-class of indoor, outdoor, and transportation. To evaluate, we follow the DCASE 2021 Task 1B challenge [12], separate this dataset into training (Train.) and evaluation (Eval.) subsets. Then, Train. subset is used for training frameworks proposed and Eval. subset is used for evaluating.

B. Deep learning framework implementation

We use Tensorflow framework to build all classification models in this report. As we apply mixup data augmentation [22], [23] to enforce back-end classifiers, the labels of the mixup data input are no longer one-hot. We therefore train back-end classifiers with Kullback-Leibler (KL) divergence loss [30] rather than the standard cross-entropy loss over all N mixup training samples: The training is carried out for 100 epochs using Adam [31] for optimization.

C. Late fusion strategy for multiple predicted probabilities

As back-end classifiers work on patches of spectrograms or image frames, the predicted probability of an entire spectrogram or all image frames of a video recording is computed by averaging of all images or patches' predicted probabilities. Let us consider $\mathbf{P^n} = (\mathbf{p_1^n}, \mathbf{p_2^n}, ..., \mathbf{p_C^n})$, with C being the category number and the n^{th} out of N image frames or patches of spectrogram fed into a learning model, as the probability of a test instance, then the average classification probability is denoted as $\mathbf{\bar{p}} = (\bar{p}_1, \bar{p}_2, ..., \bar{p}_C)$ where,

$$\bar{p_c} = \frac{1}{N}\sum_{n=1}^{N} p_c^n \quad for \quad 1 \leq n \leq N \tag{1}$$

and the predicted label $\hat{y}$ for an entire spectrogram or all image frames evaluated is determined as:

$$\hat{y} = argmax(\bar{p}_1, \bar{p}_2, ..., \bar{p}_C) \tag{2}$$

TABLE III
Performance comparison of audio-based frameworks

Audio Based Frameworks	Acc.
audio-CQT-Vgg14	68.3
audio-GAM-Vgg14	69.6
audio-MEL-Vgg14	72.2
MAX Fusion	78.0
MEAN Fusion	79.7
PROD Fusion	**80.4**

TABLE IV
Performance comparison of visual-based frameworks

Visual Based Frameworks	Acc.
visual-image-Xception	85.9
visual-image-Vgg19	83.8
visual-image-Resnet50	86.3
visual-image-InceptionV3	88.9
visual-image-MobileNetV2	84.4
visual-image-DenseNet121	87.8
visual-image-NASNetLarge	86.9
MAX Fusion	90.2
MEAN Fusion	90.5
PROD Fusion	**91.1**

To evaluate ensembles of multiple predicted probabilities obtained from different frameworks, we proposed three late fusion schemes, namely MEAN, PROD, and MAX fusions. In particular, we conduct experiments over individual frameworks, thus obtain predicted probability of each framework as $\mathbf{\bar{p_s}} = (\bar{p}_{s1}, \bar{p}_{s2}, ..., \bar{p}_{sC})$ where C is the category number and the s^{th} out of S framework evaluated. Next, the predicted probability after late MEAN fusion $\mathbf{p_{f-mean}} = (\bar{p}_1, \bar{p}_2, ..., \bar{p}_C)$ is obtained by:

$$\bar{p_c} = \frac{1}{S}\sum_{s=1}^{S} \bar{p}_{sc} \quad for \quad 1 \leq s \leq S \tag{3}$$

The PROD strategy $\mathbf{p_{f-prod}} = (\bar{p}_1, \bar{p}_2, ..., \bar{p}_C)$ is obtained by,

$$\bar{p_c} = \frac{1}{S}\prod_{s=1}^{S} \bar{p}_{sc} \quad for \quad 1 \leq s \leq S \tag{4}$$

and the MAX strategy $\mathbf{p_{f-max}} = (\bar{p}_1, \bar{p}_2, ..., \bar{p}_C)$ is obtained by,

$$\bar{p_c} = max(\bar{p}_{1c}, \bar{p}_{2c}, ..., \bar{p}_{Sc}) \tag{5}$$

Finally, the predicted label $\hat{y}$ is determined by (2):

IV. Experimental Results And Discussion

A. Analysis of audio-based deep learning frameworks for scene classification

As Table III shows accuracy results obtained from audio-based deep learning frameworks, we can see that all late fusion methods help to improve the performance significantly, achieve the highest score of 80.4% from PROD fusion.(Note that these frameworks and DCASE baseline only use audio data input).

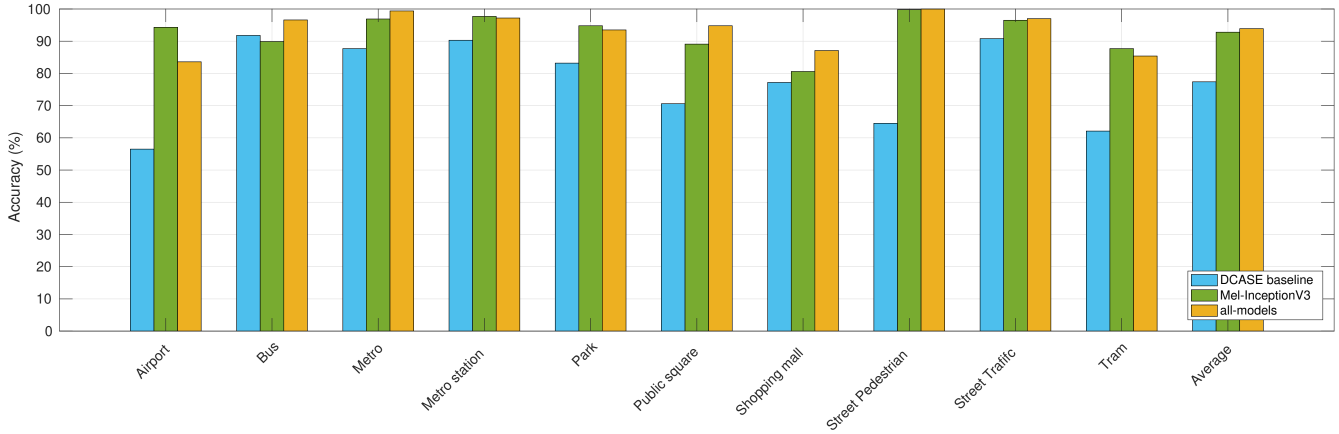

Fig. 1. Performance comparison (Acc.%) of DCASE baseline, *MEL-InceptionV3* and *all-models* across all scene categories

B. Analysis of visual-based deep learning frameworks for scene classification

As obtained results are shown in Table IV, we can see that the PROD fusion of seven visual-image based frameworks achieves the best accuracy of 91.1%, improving DCASE baseline by 13.7% (Note that these frameworks and DCASE baseline only use visual data input).

Compare performance between audio-based and visual-based approaches, the PROD fusion of seven visual based frameworks (91.1%) outperforms the best result of 80.4% from PROD fusion of *audio-CQT-Vgg14, audio-GAM-Vgg14, audio-MEL-Vgg14* mentioned in Section IV-A.

C. Combine both visual and audio features for scene classification

We then evaluate a combination of audio and visual features by proposing two PROD fusions: (1) three audio based frameworks (*audio-CQT-Vgg14, audio-GAM-Vgg14, audio-MEL-Vgg14*) and top-3 visual based frameworks (*visual-image-DenseNet121, visual-image-InceptionV3, visual-image-NASNetLarge*) referred to as *all-models*, and (2) one audio based framework (*audio-MEL-Vgg14*) and one visual based framework (*visual-image-InceptionV3*) referred to as *MEL-InceptionV3*. As results shown in Fig. 1, *all-models* helps to achieve the highest accuracy classification score of 93.9%, improving DCASE baseline by 16.5% and showing improvement on all scene categories. Although *MEL-InceptionV3* only fuses two frameworks, it achieves 92.8%, showing competitive to *all-models* fusing 6 frameworks.

D. Early detecting scene context

We further evaluate whether deep learning frameworks proposed can help to detect scene context early. To this end, we evaluate 10 different frameworks: (1-2-3) 3 individual audio based frameworks (*audio-CQT-Vgg14, audio-GAM-Vgg14, audio-MEL-Vgg14*), (4) PROD fusion of these three audio based frameworks referred to as *all-audio-models*, (5-6-7) 3 visual based frameworks (*visiual-image-NASNetLarge, visual-image-Densnet121, visual-image-InceptionV3*), (8) PROD fusion of these three visual-image based frameworks referred to as *all-visual-models*, (9) *MEL-InceptionV3*, and (10) *all-models*. As the results shown in Fig. 2, while performance of audio-based frameworks is improved by time, visual-based frameworks show stable. As a result, when we combine audio and visual features, which are evaluated in *MEL-InceptionV3* and *all-models*, the performance is improved by time and stable after 6 seconds. Notably, accuracy scores of both *MEL-InceptionV3* and *all-models* are larger than 90.0% at the first second, which is potentially for real-life applications integrating the function of early detecting scene context.

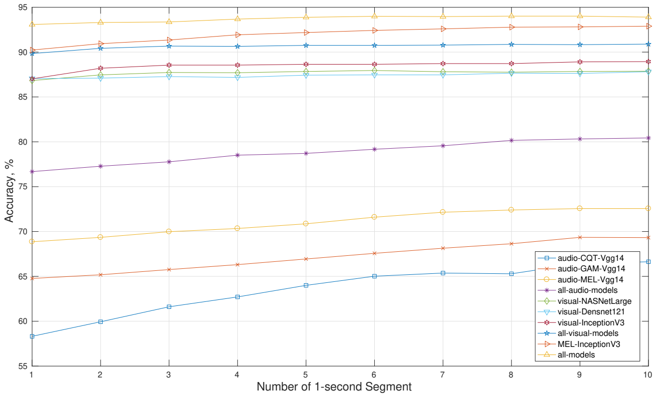

Fig. 2. Performance of individual audio based frameworks (*audio-CQT-Vgg14, audio-GAM-Vgg14, audio-MEL-Vgg14*), PROD fusion of three audio based frameworks (*all-audio-models*), individual visual based frameworks (*visual-image-NASNetLarnge, visual-image-Densnet121, visual-image-InceptionV3*), PROD fusion of three visual based frameworks (*all-visual-model*), PROD fusion of audio-based and visual-based frameworks (*MEL-InceptionV3, all-models*) with the increasing number of 1-second input segments

V. Conclusion

We conducted extensive experiments and explored various deep learning based frameworks for classifying 10 categories of urban scene. Our method, which uses an ensemble of audio-based and visual-based frameworks, achieves the best classification accuracy of 93.9% on DCASE Task 1B development set. The obtained results outperform DCASE baseline, improving by 17.1% with only audio data input, 26.2% with only visual data input, and 16.5% with both audio-visual data.

References

[1] R. Arandjelović and A. Zisserman, "Objects that sound," in *ECCV*, 2018.

[2] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444–3453.

[3] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. H. McDermott, and A. Torralba, "The sound of pixels," *ArXiv*, vol. abs/1804.03160, 2018.

[4] N. Takahashi, M. Gygli, and L. V. Gool, "Aenet: Learning deep audio features for video analysis," *IEEE Transactions on Multimedia*, vol. 20, pp. 513–524, 2018.

[5] C. Sanderson and B. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Proc. ICB*, 2009.

[6] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, pp. 3030–3043, 2018.

[7] R. Gao, T.-H. Oh, K. Grauman, and L. Torresani, "Listen to look: Action recognition by previewing audio," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 454–10 464.

[8] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 961–970.

[9] K. Soomro, A. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *ArXiv*, vol. abs/1212.0402, 2012.

[10] R. Gade, M. Abou-Zleikha, M. G. Christensen, and T. Moeslund, "Audio-visual classification of sports types," in *IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 768–773.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[12] Detection and Classification of Acoustic Scenes and Events Community, *DCASE 2021 challenges*, http://dcase.community/challenge2021.

[13] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," *arXiv preprint arXiv:2011.00030*, 2020.

[14] L. Pham, I. Mcloughlin, H. Phan, R. Palaniappan, and A. Mertins, "Deep feature embedding and hierarchical classification for audio scene classification," in *International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

[15] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. Mcloughlin, "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework," *Digital Signal Processing*, vol. 110, p. 102943, 2021.

[16] L. Pham, I. McLoughlin, H. Phan, R. Palaniappan, and Y. Lang, "Bag-of-features models based on C-DNN network for acoustic scene classification," in *Proc. AES*, 2019.

[17] L. Pham, I. Mcloughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification," in *Proc. INTERSPEECH*, 2019, pp. 3634–3638.

[18] D. Ngo, H. Hoang, A. Nguyen, T. Ly, and L. Pham, "Sound context classification basing on join learning model and multi-spectrogram features," *ArXiv*, vol. abs/2005.12779, 2020.

[19] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Proc. DCASE*, 2018, pp. 34–38.

[20] B. McFee, R. Colin, L. Dawen, D. Ellis, M. Matt, B. Eric, and N. Oriol, "librosa: Audio and music signal analysis in python," in *Proceedings of The 14th Python in Science Conference*, 2015, pp. 18–25.

[21] D. P. W. . Ellis, "Gammatone-like spectrogram," 2009. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/ gammatonegram

[22] K. Xu, D. Feng, H. Mi, B. Zhu, D. Wang, L. Zhang, H. Cai, and S. Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim Conference on Multimedia*, 2018, pp. 14–23.

[23] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *ICLR*, 2018.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[25] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[26] 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.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[28] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[30] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.