Low-complexity acoustic scene classification for multi-device audio: analysis of DCASE 2021 Challenge systems

Irene Martín-Morató, Toni Heittola, Annamaria Mesaros, Tuomas Virtanen

Computing Sciences, Tampere University, Finland

Task description

The task targeted development of low-complexity solutions with good generalization properties. The provided baseline system is based on a CNN architecture and post-training quantization of parameters. The system is trained using all the available training data, without any specific technique for handling device mismatch, and obtains an accuracy of 47.7%, with a log loss of 1.473.

TAU Urban Acoustic Scenes 2020 Mobile

Recordings from multiple European cities[1]

- Scenes: Airport, indoor shopping mall, metro station, pedestrian street, public square, street with medium level of traffic, travelling by a tram, travelling by a bus, travelling by an underground metro and urban park.
- Recording devices: A, B, C, and D (real devices) and 11 simulated ones (S1-S11 devices).
- 64 hours of audio available in the development set and 22 hours in the evaluation set.

Baseline

- Three CNN layers and one fully connected layer, followed by the softmax output layer.
- Quantization to 16 bits (float16) after training.
- Input shape of 40×500 for each 10 second audio file, log mel-band energies, calculated with an analysis frame of 40 ms and 50% hop size.
- Final model size of the system after quantifying is 90.3 KB.

System complexity requirements

- Model complexity limit of 128 KB for the non-zero parameters.
- This limit allows 32768 in float32 (32-bit float) representation, (32768 parameter values * 32 bits per parameter / 8 bits per byte= 131072 bytes = 128 KB).

Submissions

- **99 submissions** from 30 teams.
- Most of the submitted systems outperformed the baseline.
- The most used techniques among the submissions were residual networks and weight quantization.
- 18 submitted systems had over 70% accuracy and log loss under 0.8

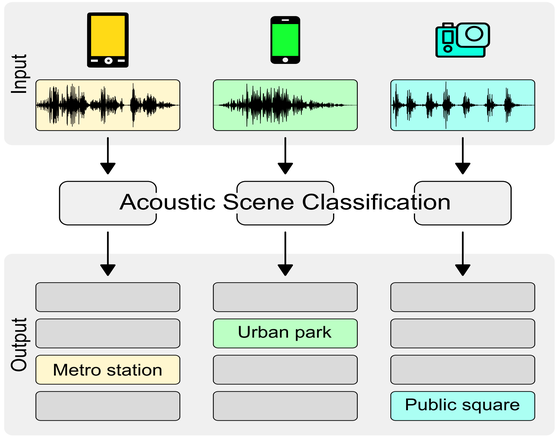

Figure 1: Acoustic scene classification for audio recordings.

RESULTS

Rank	System	Logloss ±95% CI	Acc ±95% CI(%)	Size (KB)	Weights	Sparsity	Learning	Architecture
1	Kim_QTI_2	0.72±0.03	76.1±0.94	121.9	int8	✓	KD	BC-ResNet
3	Yang_GT_3	0.74±0.02	73.4±0.97	125.0	int8	✓	KD	Ensemble
9	Koutini_CPJKU_3	0.83±0.03	72.1±0.99	126.2	float16	✓	grouping CNN	CP_ResNet
12	Heo_Clova_4	0.87±0.02	70.1±1.01	124.1	float16	-	KD	ResNet
13	Liu_UESTC_3	0.88±0.02	69.6±1.01	42.5	1-bit	-	-	ResNet
17	Byttebier_IDLab_4	0.91±0.02	68.8±1.02	121.9	int8	✓	grouping CNN	ResNet
19	Verbitskiy_DS_4	0.92±0.02	68.1±1.03	121.8	float16	-	-	EfficientNet
22	Puy_VAI_3	0.94±0.02	66.2±1.04	122.0	float16	-	focal loss	Separable CNN
25	Jeong_ETRI_2	0.95±0.03	67.0±1.04	113.9	float16	-	-	Trident ResNet
28	Kim_KNU_2	1.01±0.03	63.8±1.06	125.1	int8	-	mean-teacher	Shallow inception
85	Baseline	1.73±0.05	45.6±1.10	90.3	float16	-	-	CNN

Table 1: Performance on the evaluation set and complexity management techniques for selected top systems (best system of each team). "KD" refers to Knowledge Distillation and "BC" stands for Broadcasting.

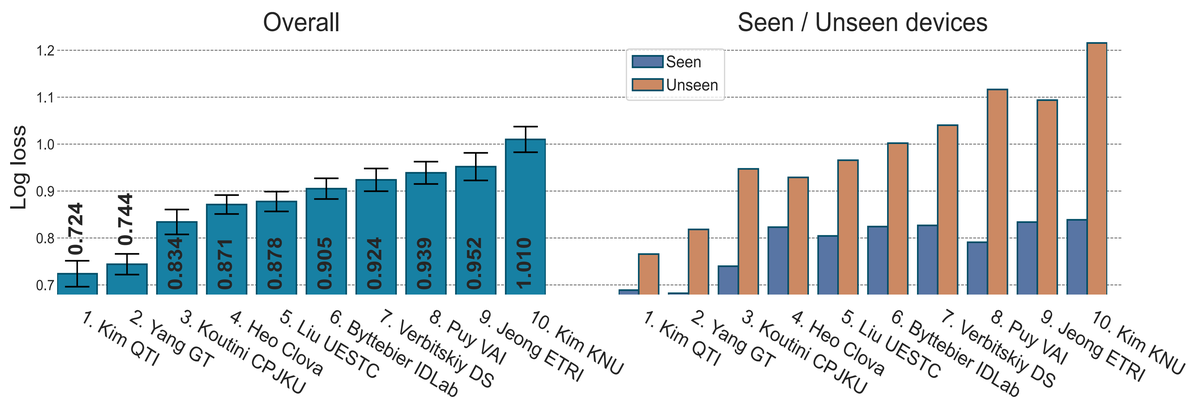

Figure 2: Classification log loss for the 10 top teams (best system per team) on the evaluation dataset.

Features and augmentation techniques

Top 10 teams make use of mel energies as feature representation. Augmentation techniques are also used, with most popular techniques being mixup (used by 20 teams) and specAugment (used by 10 teams).

Architectures

Residual architectures are the most popular, being used by a total of 15 teams. Followed by the use of modified versions of networks based on residual blocks such as MobileNet [2] or EfficientNet [3].

System complexity

A notably small model, with size 42.5KB, ranked 13th, with a 0.878 log loss and 69.60% accuracy. The model compression is performed with 1-bit quantization. The top 10 systems are close to the allowed model size limit, ranging from 110 KB to 126.81 KB, with the system ranked first having a size of 121.9KB.

Device and class-wise performance

All systems have higher performance on the devices seen during training (A, B, C, S1, S2, S3) than on the unseen ones (D, S7, S8, S9, S10), with a difference in accuracy of almost 3% (statistically significant) for the system ranked first. Data mismatch due to the unseen devices is more challenging than the mismatch created by different cities, due to the different properties of the recorded audio, which are related to the device-specific processing

Conclusions

- ◇ The method for calculating the model complexity includes only the parameters of the network, with exceptions in the case of employing embeddings.
- ◇ Multiple techniques to improve robustness, like data augmentation, with methods directed towards obtaining light models, e.g., knowledge distillation, weights quantization, and sparsity.
- ◇ acoustic scene classification is still relevant for the audio community, and in particular, to the development of solutions applicable for real-life devices.

References

[1] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. Acoustic scene classification in DCASE 2019 challenge: closed and open set classification and data mismatch setups. In *Proc. of the DCASE 2019 Workshop*, New York, Nov 2019.

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[3] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *36th Int. Conf. on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.

PAPER CODE TASK

[1] https://doi.org/10.5281/zenodo.3819968,https://doi.org/10.5281/zenodo.3685828