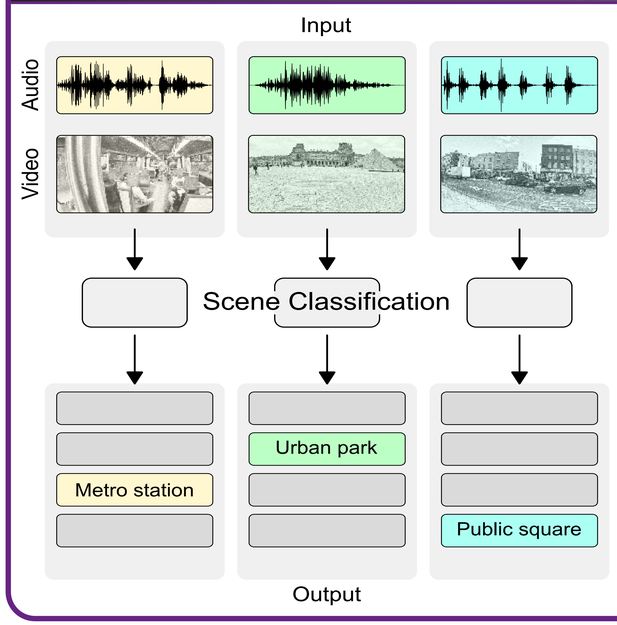
AUDIO-VISUAL SCENE CLASSIFICATION: ANALYSIS OF DCASE 2021 CHALLENGE SUBMISSIONS

Tampere University

Shanshan Wang, Toni Heittola, Annamaria Mesaros, Tuomas Virtanen

DCASE2021 CHALLENGE

T1B: Audio-Visual Scene Classification



- **Scene Classification based on audio and video input**
- Motivation: Humans perceive the world through multiple senses (seeing and hearing)
- Examples: Passing the traffic road by seeing and hearing the surroundings

DataSet: TAU Audio-Visual Urban Scenes 2021

1. Recorded in **12** large European cities: Amsterdam, Barcelona, Helsinki, Lisbon, London, Lyon, Madrid, Milan, Prague, Paris, Stockholm, and Vienna.
2. 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.
3. **Development set** (34 hours): 10s for each audio and video clip
4. **Evaluation set** (20 hours): 1s for each audio and video clip

Evaluation metrics & Submissions

Evaluation metrics

1. Multiclass cross-entropy (log-loss) (used for system ranking)
2. Accuracy (used for comparison with the ASC evaluations from the challenge of previous editions)

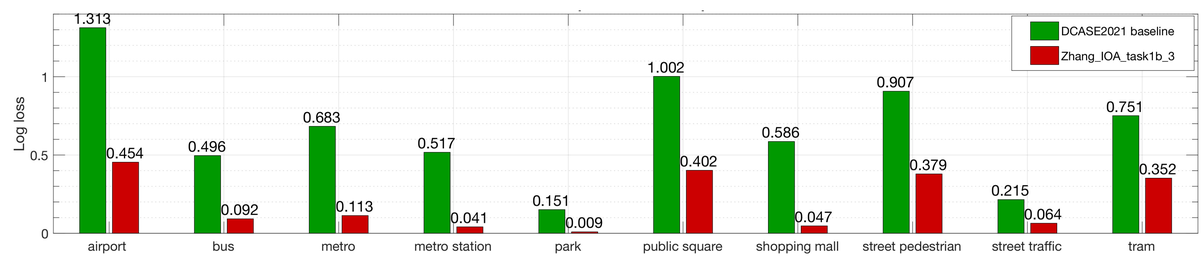
Submissions

1. We received 43 submissions from 13 teams.
2. 15 out of 43 systems have logloss less than 0.34 and accuracy more than 90%.
3. 27 out of 43 systems adopt multimodalities approach.

RESULTS

Rank	Team	Logloss	Accuracy (95% CI)	Fusion Methods	Model Complexity
1	Zhang_IOA_3	0.195	93.8% (93.6 - 93.9)	early fusion	110M
5	Du_USTC_4	0.221	93.2% (93.0 - 93.4)	early fusion	373M
9	Okazaki_LDSLVision	0.257	93.5% (93.3 - 93.7)	audio-visual	636M
10	Yang_THU_3	0.279	92.1% (91.9 - 92.3)	early fusion	121M
16	Hou_UGent_4	0.416	85.6% (85.3 - 85.8)	late fusion	28M
24	DCASE2021 baseline	0.662	77.1% (76.8 - 77.5)	early fusion	711k

Performance and general characteristics of top 5 teams



Class-wise performance comparison between the top 1 system and the baseline system

General Trends

- **Audio Features**: Log-mel Spectrogram
- **Multimodality**: Audio + video, audio +video + text
- **Large pretrained model**: ResNet, VGG, EfficientNet, PANN network
- **Data Augmentation**: Mixup, Frequency masking, Pitch shifting, Color jitter
- **Transfer Learning**: VGG, PANN trained AudioSet, ResNet trained on ImageNet, places365

Discussion and conclusion

Discussion

1. Video-based models shows higher performance than audio-based models. However, the joint model performs the best.
2. Complex models with larger trainable parameters tend to lead higher performance.

Conclusion

1. The choice of evaluation metrics (log loss or accuracy) does not affect the ranking drastically.
2. Performance on development dataset goes in line with the evaluation set, which proves the consistency of the dataset.