DCASE2021 WORKSHOP

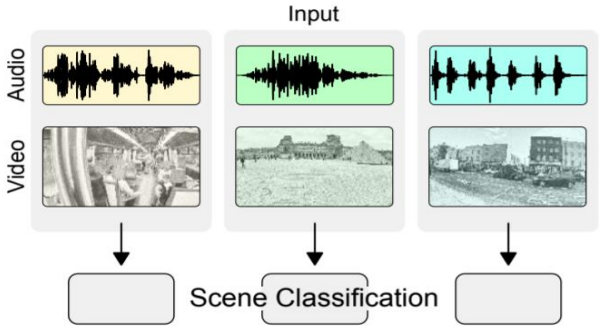
A Multi-Modal Fusion Approach for Audio-Visual Scene Classification Enhanced by CLIP Variants

{Soichiro Okazaki, Quan Kong, Tomoaki Yoshinaga}*1
*1 Lumada Data Science Lab., Hitachi, Ltd., Toyko, Japan

HITACHI
Inspire the Next

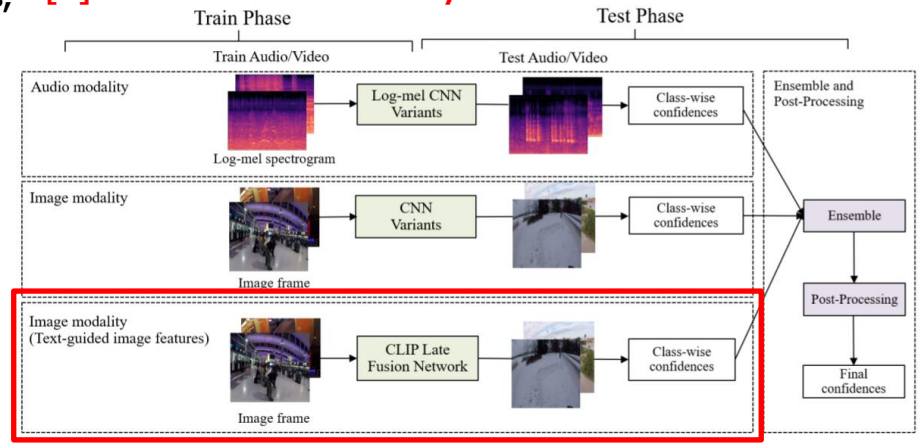
1. Introduction

- Task: Audio-Visual Scene Classification
- Data: Synchronous 1 sec. audio-video files, train: 86460 files, val: 36450 files, 10 scene classes (e.g. bus, park, tram)



2. Approach Overview

- Ensemble of three domain models:
  [1] Audio models by CNNs, [2] Video models by CNNs,
  [3] Another Video model by CLIP Late Fusion Network



3. The Effect of CLIP models

- OpenAI CLIP models are used for boosting accuracy
- Extracted video features from CLIP variants are concatenated and trained in CLIP late fusion network
- The characteristics of CLIP late fusion network:
  - ✓ Fast training
  - ✓ Lightweight model
  - ✓ Solid performance in DCASE Scene Classification

Table 1: The architecture of CLIP Late Fusion Network.

RN50x4 (dim:640)	RN101 (dim:512)	ViT-B/32 (dim:512)
Linear(640, 512)	Linear(512, 512)	Linear(512, 512)
BatchNorm1d(512)	BatchNorm1d(512)	BatchNorm1d(512)
ReLU()	ReLU()	ReLU()
Dropout(p=0.2)	Dropout(p=0.2)	Dropout(p=0.2)
Linear(512, 256)	Linear(512, 256)	Linear(512, 256)
concatenation of 256*3 dimension		
Linear(256*3, 128)		
Linear(128, 10)		

Table 4: The effect of CLIP Late Fusion Network (C04). With adding CLIP Late Fusion Network, the recognition performance are boosted in both logloss and accuracy metric for validation dataset.

Description	CLIP	Logloss	Accuracy
A04/V04 Fusion	no	0.293	92.4
A04/V04/C04 Fusion	**yes**	**0.238**	**95.8**
A04/V04 Fusion with p.p.	no	0.205	93.0
A04/V04/C04 Fusion with p.p.	**yes**	**0.149**	**96.1**

4. Results (of Validation Set)

- Video models have good score than audio models due to short time dataset
- Off-the-shelf CLIPs have good score with raw classnames as prompts (C01-03)

Index	Architecture	Audio	Video	Notes	Logloss	Accuracy
B01	OpenL3's model	log-mel CNN	-	Baseline model of Audio-only	1.048	65.1
A01	RegNet-6.4F	log-mel CNN	-	Training with 1 sec. audio files	0.711	76.6
A02	ResNeSt-50d	log-mel CNN	-	Training with 1 sec. audio files	0.732	76.9
A03	TF-Efficientnet-B1-NS	log-mel CNN	-	Training with 1 sec. audio files	0.821	77.2
A04	A01-A03's models	log-mel CNN	-	Ensemble of A01-A03	0.721	78.1
B02	OpenL3's model	-	CNN	Baseline model of Visual-only	1.648	64.9
V01	RegNet-6.4F	-	CNN	-	0.328	90.0
V02	ResNeSt-50d	-	CNN	-	0.367	91.7
V03	HRNet-W18	-	CNN	-	0.336	90.9
V04	V01-V03's models	-	CNN	Ensemble of V01-V03	0.316	92.4
C01	ResNet-101	-	CLIP CNN	No Training	0.671	76.7
C02	ResNet-50x4	-	CLIP CNN	No Training	0.668	74.5
C03	ViT-B/32	-	CLIP ViT	No Training	0.725	72.5
C04	C01-C03's models	-	CLIP CNN&ViT	Late Fusion of C01-C03	0.273	90.9
B03	OpenL3's model	log-mel CNN	CNN	Baseline model of Audio-Visual	0.658	77.0
E01	A04/V04/C04's models	log-mel CNN	CNN / CLIP CNN&ViT	Ensemble of A04/V04/C04	0.238	95.8
E02	**A04/V04/C04's models**	**log-mel CNN**	**CNN / CLIP CNN&ViT**	**E01 with Post-Processing**	**0.149**	**96.1**

5. Conclusion

- The performance of Audio-Visual Scene Classification can be boosted by CLIPs.
- Off-the-shelf CLIP models have good recognition performance (※75% accuracy) for 10 defined scene classes in DCASE 2021 Task1B dataset.
- CLIP late fusion network is lightweight and can be trained fast.
- CLIP late fusion network have different characteristics from Video models even though they are trained on the same dataset. (※Details can be seen in our paper)
- Our approach with CLIPs achieved 3rd place in DCASE 2021 Task1B Challenge.