Audio-Visual Scene Classify

Yuan Feng, 521030910369
Shanghai JiaoTong University

1. Introduction

Scene classification is commonly studied in both audio and video domains. For acoustic scene classification the input is typically a short audio recording, while for visual scene classification tasks the input can be an image or a short video clip. However, these two tasks are unimodal, meaning they use either audio information or visual information.

In our daily lives, we perceive the world through multiple senses (sight and hearing). Additionally, the classification methods in individual domains have generally matured, and therefore, multimodal analysis has become a pursued research direction for further improvement.

Recent research has shown that the joint learning of acoustic features and visual features can bring additional benefits in various tasks. Therefore, in this task, we implement the use of both audio information and video information for scene classification, namely **Audio-Visual Scene Classification** [1].

As shown in the Figure 1 below.

2. Experiment Tasks Overview

In this project, I primarily conducted the following experiments:

1. Understanding the different ways of obtaining representations and the ways of multimodal fusion in multimodal learning work.
2. Analyze the conflicting modal results, that is, separately test the results on each unimodal, observe the categories where multimodal fusion is more effective, and analyze the reasons.
3. Replace the feature fusion method with the two types of fusion methods: early feature fusion and late decision fusion, and observe the performance.
4. Try to improve the model performance by replacing features, modifying the model, and tuning hyperparameters.
5. Provide the scripts and configurations for the final optimal results.

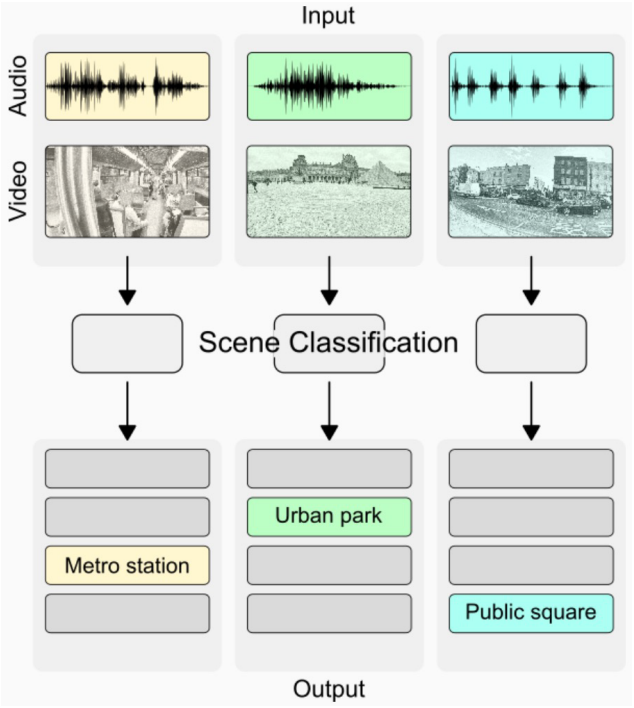

Figure 1. Audio-Visual Scene Classification

3. Experiments

3.1. Multimodal Fusion

In multimodal tasks, an important part is the way of modal fusion, that is, how to combine the features of different modalities. Different fusion methods will have different predictive effects. In previous studies [2], authors have already proposed **early** fusion and **intermediate** fusion. In this project, we take intermediate fusion as the baseline, and introduce a new **late** fusion method. We will try and compare these three methods respectively.

The schematic diagrams of early fusion mode and intermediate fusion mode are shown in Figure 2.

As for the **late fusion** mode, the process is as follows: first, the Audio specific layers and Video specific layers go through their respective Classification layers separately, and then the results are concatenated, finally passing through a

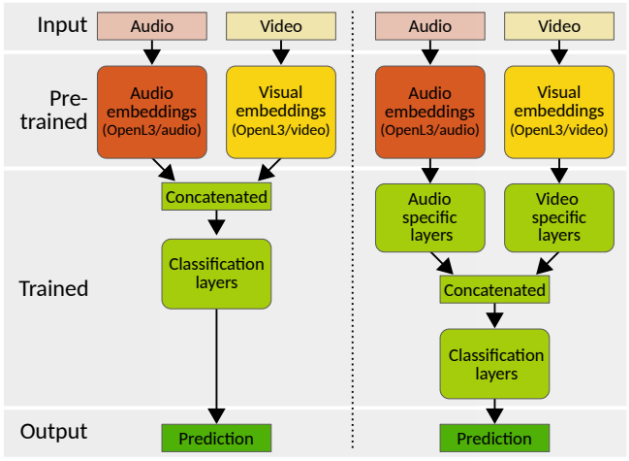

Figure 2. Early & Intermediate Feature Fusion

fully connected layer to obtain the prediction.

As shown in the comparison results in Table 1, it can be found that the accuracy of the intermediate fusion mode is still relatively better.

Table 1. Different Methods of Fusion

	Early	**Intermediate**	**Late**
ACC	0.797	**0.798**	0.792
Loss	**0.528**	0.541	0.540

3.2. Compared to Unimodal

To analyze the categories where multimodal fusion is more effective, we separately tested the classification results using unimodal features. Figure 2 and 3 show the performance comparison between the multimodal model and the models using only audio features or only video features, respectively.

In the tables, the numbers 1 to 10 in the Type column represent 10 scene labels, specifically: airport, bus, metro, metro_station, park, public_square, shopping_mall, street_pedestrian, street_traffic, tram.

The evaluation metrics used are P/R/F, where M represents Multimodal Fusion, A represents Audio-Only, V represents Video-Only, and P, R, F represent Precision, Recall, and F1-score, respectively.

From the table, it can be seen that the multimodal feature fusion method generally outperforms the unimodal methods across most of the metrics for each scene.

Among all the scenes, the ones numbered 2, 4, 8, and 10, namely bus, metro_station, street_pedestrian, and tram, show more significant improvements across the various metrics. In these scenes, the fusion of audio and video features works particularly well, greatly enhancing the performance indicators.

Table 2. Compared to Audio-Only

Type	M-P	M-R	M-F	A-P	A-R	A-F
1	0.79	0.66	0.72	0.75	0.68	0.71
2	0.92	0.86	0.89	0.83	0.69	0.75
3	0.84	0.88	0.86	0.67	0.71	0.69
4	0.85	0.88	0.86	0.66	0.69	0.67
5	0.96	0.85	0.90	0.93	0.75	0.83
6	0.63	0.71	0.67	0.58	0.68	0.63
7	0.74	0.77	0.76	0.71	0.66	0.69
8	0.74	0.67	0.70	0.62	0.62	0.62
9	0.85	0.90	0.87	0.78	0.85	0.81
10	0.71	0.80	0.75	0.63	0.73	0.67

Table 3. Compared to Video-Only

Type	M-P	M-R	M-F	V-P	V-R	V-F
1	0.79	0.66	0.72	0.62	0.37	0.46
2	0.92	0.86	0.89	0.64	0.80	0.71
3	0.84	0.88	0.86	0.80	0.74	0.77
4	0.85	0.88	0.86	0.79	0.87	0.83
5	0.96	0.85	0.90	0.87	0.80	0.83
6	0.63	0.71	0.67	0.55	0.52	0.53
7	0.74	0.77	0.76	0.58	0.74	0.65
8	0.74	0.67	0.70	0.59	0.62	0.60
9	0.85	0.90	0.87	0.77	0.75	0.76
10	0.71	0.80	0.75	0.48	0.42	0.44

Of course, in practice, the joint model does not always have an advantage over the separate acoustic or visual models, especially in more complex and diverse scenarios. Therefore, we need to analyze different scenes and consider whether the modalities conflict, as well as the possibility of cross-modal knowledge transfer when one modality is not available.

Overall accuracy comparison as shown in the Table 4.

Table 4. Accuracy Comparison with Unimodal

	Audio-Only	**Video-Only**	**Multimodal**
ACC	0.706	0.674	**0.798**
Loss	0.808	0.854	**0.541**

3.3. Weighted Method

Based on the above unimodal experiments, we found that the model performance using only video features is slightly worse than using only audio information. This hints that we could use a weighted feature fusion approach, where on top of the late fusion, we directly obtain the predictions from the two modalities, and then multiply one modality with a weight before adding it to the other modality. The weight can be adjusted during the training process.

Compared to the baseline, the performance of the weighted fusion method is shown in Table 5.

Table 5. Comparison with Weighted Method

	Baseline	**Weighted**
ACC	0.798	**0.816**
Loss	0.541	**0.507**

It can be observed that the Weighted Method significantly outperforms the baseline model. This suggests that the features from different modalities cannot be simply averaged for fusion, and instead require different weights to be applied.

4. Best Result

In summary, based on the analysis above, I have adopted the Weighted Method as the model with the best performance. The configuration file can be found at `config/best.yaml`.

And the model file is `models_best.py`.

Results are shown in Figure 3.

	precision	recall	f1-score	support
airport	0.80	0.78	0.79	281
bus	0.87	0.85	0.86	327
metro	0.85	0.89	0.87	360
metro_station	0.86	0.87	0.87	386
park	0.95	0.80	0.87	386
public_square	0.71	0.78	0.74	387
shopping_mall	0.81	0.75	0.78	387
street_pedestrian	0.76	0.73	0.75	421
street_traffic	0.85	0.92	0.88	402
tram	0.73	0.80	0.76	308
accuracy			0.82	3645
macro avg	0.82	0.82	0.82	3645
weighted avg	0.82	0.82	0.82	3645

accuracy: 0.816
overall log loss: 0.507

Figure 3. Best Results

References

[1] Shanshan Wang, Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. Audio-visual scene classification: analysis of dcase 2021 challenge submissions, 2021. 1

[2] Shanshan Wang, Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A curated dataset of urban scenes for audio-visual scene analysis, 2021. 1