Final Project: Audio Event Recognition

December 2017

1 Introduction

Audio event recognition, the human-like ability to identify and relate sounds from audio, is a nascent problem in machine perception. One of the comparable problems is object detection in image processing.

2 Database

Audioset is used in this project. The raw audios are not released and hard to download through YouTube, thus the features provided by Google will be used in this project. Feature files can be downloaded via on the FTP site [1]. The basic information about the database is as follows, for more information, please contact TA.

- Each utterance is about 10 seconds, features are extracted at 1Hz
- Features are 128 dimensional vectors, extracted from a bottleneck layer of ResNet. Details of the model can be found in the paper [1]. Google also released this model (TensorFlow format) [2]
- Each utterance may have several labels and there are 527 kinds of labels in total.

Features are stored in the pickle format, and we have made a script to show how you can read them.

3 Evaluation Metrics

In the evaluation part, you need to use the test partition of Audioset by applying your classifier to 1 sec frames taken from each segment, averaging the scores, then for each category ranking all segments by their scores. You need to evaluate the performance of your model with two metrics:

[1] http://202.120.38.125:9999/
[2] https://github.com/tensorflow/models/tree/master/research/audioset

- AP, also known as average precision. It computes the area under precision-recall curve. In the evaluation section, record the set segments classified for each category as A_i and denote the set of segments in each category as B_i. Precision p_i is defined as $p_i = \frac{|A_i \bigcap B_i|}{|A_i|}$ and recall is defined as $r_i = \frac{|A_i \bigcap B_i|}{|B_i|}$. Compute the area of precision-recall curve for recall from 0 to 1 gets the average precision for each category.
- AUC, computes the area under the ROC curve, which is a TPR-FPR curve. Denote the set of all segments as C. TPR(true positive rate) is defined as $\frac{|A_i \bigcap B_i|}{|B_i|}$ and FPR is defined as $\frac{|A_i - B_i|}{|C - B_i|}$. Compute the area of ROC for FPR from 0 to 1 gets the AUC score for each category.

Final scores will be given according to the performances on the evaluation set (refers to the folder ./eval). However, the evaluation set shouldn't be used before the completion of training. That is to say, it is cheating if you use the evaluation set to train your model or choose the optimal model according to the performances on the evaluation set. We will give serious punishment if such behaviors are found.

4 Requirements

The results based on the balanced training set must be provided, whereas results obtained based on the unbalanced training set are optional.

References

[1] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE ICASSP*, 2017.