HOME ONTOLOGY DATASET
DOWNLOAD ABOUT

Download

We offer the AudioSet dataset for download in two formats:

1. Text (csv) files describing, for each segment, the YouTube video ID, start time, end time, and one or more labels.
2. 128-dimensional audio features extracted at 1Hz. The audio features were extracted using a VGG-inspired acoustic model described in Hershey et. al., trained on a preliminary version of YouTube-8M. The features were PCA-ed and quantized to be compatible with the audio features provided with YouTube-8M. They are stored as TensorFlow Record files. The model used to generate the features is available in the TensorFlow models GitHub repository (see below for more details).

The labels are taken from the AudioSet ontology which can be downloaded from our AudioSet GitHub repository.

The dataset is made available by Google Inc. under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, while the ontology is available under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

Dataset split

The dataset is divided in three disjoint sets: a balanced evaluation set, a balanced training set, and an unbalanced training set. In the balanced evaluation and training sets, we strived for each class to have the same number of examples. The unbalanced training set contains the remainder of annotated segments.

Evaluation - eval_segments.csv

20,383 segments from distinct videos, providing at least 59 examples for each of the

527 sound classes that are used. Because of label co-occurrence, many classes have more examples.

Balanced train - balanced_train_segments.csv

22,176 segments from distinct videos chosen with the same criteria: providing at least 59 examples per class with the fewest number of total segments.

Unbalanced train - unbalanced_train_segments.csv

2,042,985 segments from distinct videos, representing the remainder of the dataset.

CSV file format

Each csv file has a three-line header with each line starting with “#”, and with the first two lines indicating the creation time and general statistics:

```
# Segments csv created Sun Mar  5 10:54:25 2017
# num_ytids=20371, num_segs=20371, num_unique_labels=527, num_positive_labels
```

Each subsequent line has columns defined by the third header line:

```
# YTID, start_seconds, end_seconds, positive_labels
```

for example:

```
-0RWZT-miFs, 420.000, 430.000, "/m/03v3yw,/m/0k4j"
```

means that for the YouTube video -0RWZT-miFs, for the 10 second chunk from t=420 sec to t=430 sec, annotators confirmed the presence of sound classes /m/03v3yw ("Keys jangling") and /m/0k4j ("Car").

Features dataset

Frame-level features are stored as tensorflow.SequenceExample protocol buffers. A tensorflow.SequenceExample proto is reproduced here in text format:

```
context: {
  feature: {
    key  : "video_id"
    value: {
      bytes_list: {
        value: [YouTube video id string]
      }
    }
  }
  feature: {
    key  : "start_time_seconds"
    value: {
      float_list: {
        value: 6.0
      }
    }
  }
  feature: {
    key  : "end_time_seconds"
    value: {
      float_list: {
        value: 16.0
      }
    }
  }
  feature: {
    key  : "labels"
      value: {
        int64_list: {
          value: [1, 522, 11, 172] # The meaning of the labels can be found here.
        }
      }
    }
}
feature_lists: {
  feature_list: {
    key  : "audio_embedding"
    value: {
      feature: {
        bytes_list: {
          value: [128 8bit quantized features]
        }
      }
      feature: {
        bytes_list: {
          value: [128 8bit quantized features]
        }
      }
    }
    ... # Repeated for every second of the segment
  }

}
```

The total size of the features is 2.4 gigabytes. They are stored in 12,228 TensorFlow record files, sharded by the first two characters of the YouTube video ID, and packaged as a tar.gz file.

The labels are stored as integer indices. They are mapped to sound classes via class_labels_indices.csv. The first line defines the column names:

```
index,mid,display_name
```

Subsequent lines describe the mapping for each class. For example:

```
0,/m/09x0r,"Speech"
```

which means that “labels” with value 0 indicate segments labeled with “Speech”.

To download the features, you have the following options:

- Manually download the tar.gz file from one of (depending on region):
    - storage.googleapis.com/us_audioset/youtube_corpus/v1/features/features.tar.gz
    - storage.googleapis.com/eu_audioset/youtube_corpus/v1/features/features.tar.gz
    - storage.googleapis.com/asia_audioset/youtube_corpus/v1/features/features.tar.gz

- Use gsutil rsync, with the command:

  ```
  gsutil rsync -d -r features gs://{region}_audioset/youtube_corpus/v1/features
  ```

  Where {region} is one of “eu”, “us” or “asia”. For example:

  ```
  gsutil rsync -d -r features gs://us_audioset/youtube_corpus/v1/features
  ```

SHA-256 checksum: cd95d500ab2422d4233cb822e25cf73033633e2068eab64d39024e85125cb760

Models and Supporting Code

The VGG-like model, which was used to generate the 128-dimensional features and which we call *VGGish*, is available in the TensorFlow models Github repository, along with supporting code for audio feature generation, embedding postprocessing, and demonstrations of the model in inference and training modes.

You can use the YouTube-8M starter code to train models on the released features from both AudioSet as well as YouTube-8M. The code can be found in the YouTube-8M GitHub repository.

Quality Assessment and rerating

We conducted an internal Quality Assessment task where experts checked 10 random segments for most of the classes. Due to a variety of reasons such as misinterpretation, confusability, and difficulty, a substantial number of sound classes had poor accuracy. We engaged in a rerating process to improve the quality for lower-quality classes by providing better instructions and by labeling segments in clusters. This rerating is about 50% complete at this point. The “v1” release includes the rerating done thus far. For rerated classes/segments, we have re-run the quality assessment to give an updated estimate of the label quality.

Due to the size of the dataset, we have been rerating only up to 1,000 segments for each class (sampled independently per label). This means that for the majority of the classes all segments of eval and balanced_train are, or will, get rerated. At the same time, for classes with substantially more than 1,000 segments in total, the quality in unbalanced_train dataset can be substantially different from the balanced evaluation and train datasets.

We offer two files to trace the quality assessment for each class and specify which segments got rerated:

qa_true_counts.csv

A csv file with the first line defining column names:

```
label_id,num_rated,num_true
```

Subsequent lines contain the quality assessment for each class. For example:

```
/m/05zppz,10,9
```

Indicating that 9 out of 10 example segments for the sound class /m/05zppz (“Male speech, man speaking”) indeed contained this sound.

rerated_video_ids.txt

A text file containing videos that have been labeled in the rerating task. This file consists of one YouTube video ID per line. Any segment in the dataset with these YouTube IDs will only contain rerated labels.

RESEARCH AT GOOGLE

Home
Publications
People
Teams
Outreach
Blog
Work at Google

MORE

Google Scholar
YouTube Tech Talks

FOLLOW US

Google+
Twitter

Google · About Google · Privacy · Terms