Conjunto de datos GTZAN - Clasificación de géneros musicales

Acerca del conjunto de datos

Contexto

Música. Los expertos han estado tratando durante mucho tiempo de comprender el sonido y lo que diferencia una canción de otra. Cómo visualizar el sonido. Lo que hace que un tono sea diferente de otro.
Con suerte, estos datos pueden brindar la oportunidad de hacer precisamente eso.

Contenido

- **géneros originales**: una colección de 10 géneros musicales (blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, rock) con 100 archivos de audio cada uno, todos con una duración de 30 segundos (el famoso conjunto de datos GTZAN, el MNIST de sonidos)
- **imágenes originales**: una representación visual de cada archivo de audio. Una forma de clasificar los datos es a través de redes neuronales. Debido a que los NN (como CNN, lo que usaremos hoy) generalmente toman algún tipo de representación de imagen, los archivos de audio se convirtieron a Mel Spectrograms para que esto sea posible.
- **2 archivos CSV**: contienen características de los archivos de audio. Un archivo tiene para cada canción (30 segundos de duración) una media y una varianza calculadas sobre múltiples funciones que se pueden extraer de un archivo de audio. El otro archivo tiene la misma estructura, pero las canciones se dividieron antes en archivos de audio de 3 segundos (aumentando así 10 veces la cantidad de datos que alimentamos en nuestros modelos de clasificación). *Con datos, más es siempre mejor*.

Version 1 (1.41 GB)

Data

- (Carpeta) genres_original
  - (Carpeta)blues
  - (Carpeta)classical
  - (Carpeta)country
  - (Carpeta)disco
  - (Carpeta)hiphop
  - (Carpeta)jazz
  - (Carpeta)metal
  - (Carpeta)pop
  - (Carpeta)reggae
  - (Carpeta)rock
- (Carpeta) images_original
  - (Carpeta)blues
  - Carpeta)classical
  - (Carpeta)country
  - (Carpeta)disco
  - (Carpeta)hiphop
  - (Carpeta)jazz
  - (Carpeta)metal
  - (Carpeta)pop
  - (Carpeta)reggae
  - (Carpeta)rock
- features_30_sec.csv
- features_3_sec.csv

Resumen

- 2001 files
- 120 columns

Agradecimientos

- El conjunto de datos GTZAN es el conjunto de datos público más utilizado para la evaluación en la investigación de escucha de máquinas para el reconocimiento de géneros musicales (MGR). Los archivos se recolectaron en 2000-2001 de una variedad de fuentes, incluidos CD personales, radio, grabaciones de micrófonos, para representar una variedad de condiciones de grabación (http://marsyas.info/downloads/datasets.html) .
- Este fue un proyecto de equipo para la universidad, por lo que el esfuerzo de crear las imágenes y las funciones no fue solo mío. Entonces, quiero agradecer a **James Wiltshire, Lauren O'Hare y Minyu Lei** por ser los mejores compañeros de equipo y por divertirse tanto y aprender tanto durante los 3 días que trabajamos en esto.

UrbanSound8K - Clasificación

Datos

UrbanSound8K

8732 extractos de sonido etiquetados

Última actualización: hace 3 años (Versión 1)

INPUT (7.09 GB)

Data Sources

- (Carpeta) UrbanSound8K
  - Fold0(aire acondicionado)
  - fold1(bocina de auto)
  - fold2(niños jugando)
  - fold3(ladrido de perro)
  - fold4(perforacion)
  - fold5(motor ralenti)
  - fold6(disparo)
  - fold7(martillo neumatico)
  - fold8(sirena)
  - fold9(musica callejera)

  UrbanSound8K.csv

Output

UrbanSound8kResults.csv

Acerca de este conjunto de datos

Este conjunto de datos contiene 8732 extractos de sonidos etiquetados (<=4 s) de sonidos urbanos de 10 clases: `air_conditioner`, `car_horn`, `children_playing`, `dog_bark`, `drilling`, `enginge_idling`, `gun_shot`, `jackhammer`, `sireny` `street_music`. Las clases se extraen de la taxonomía de sonidos urbanos. Para obtener una descripción detallada del conjunto de datos y cómo se compiló, consulte nuestro artículo. Todos los extractos están tomados de grabaciones de campo cargadas en www.freesound.org . Los archivos se clasifican previamente en diez pliegues (carpetas denominadas pliegue1-pliegue10) para ayudar en la reproducción y comparación con los resultados de la clasificación automática informados en el artículo anterior.

Además de los extractos de sonido, también se proporciona un archivo CSV que contiene metadatos sobre cada extracto.

ARCHIVOS DE AUDIO INCLUIDOS

8732 archivos de audio de sonidos urbanos (ver descripción arriba) en formato WAV. La frecuencia de muestreo, la profundidad de bits y la cantidad de canales son los mismos que los del archivo original cargado en Freesound (y, por lo tanto, pueden variar de un archivo a otro).

ARCHIVOS DE METADATOS INCLUIDOS

```
UrbanSound8k.csv
```

Este archivo contiene información de metadatos sobre cada archivo de audio en el conjunto de datos. Esto incluye:

- slice_file_name: El nombre del archivo de audio.
- fsID: el ID de Freesound de la grabación de la que se toma este extracto (fragmento)
- inicio: La hora de inicio del segmento en la grabación original de Freesound
- end: la hora de finalización del segmento en la grabación original de Freesound
- prominencia: una calificación (subjetiva) de prominencia del sonido. 1 = primer plano, 2 = fondo.
- fold: el número de pliegue (1-10) al que se ha asignado este archivo.
- classID:
  un identificador numérico de la clase de sonido:
  0 = acondicionador de aire
  1 = bocina de auto
  2 = niños jugando
  3 = ladrido de perro
  4 = perforando
  5 = motor en ralentí
  6 = disparo
  7 = martillo neumático
  8 = sirena
  9 = música callejera
- clase: el nombre de la clase: aire acondicionado, bocina de automóvil, niños jugando, ladrido de perro, perforación, motor en ralentí, disparo, martillo neumático, sirena, música callejera.

Agradecimientos

Solicitamos amablemente que los artículos y otros trabajos en los que se utilice este conjunto de datos citen el siguiente documento:

J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research", 22nd ACM International Conference on Multimedia, Orlando USA, Nov. 2014.

Más información en https://urbansounddataset.weebly.com/urbansound8k.html

Etiquetado de audio Freesound 2019

Reconocer automáticamente los sonidos y aplica etiquetas de diferentes naturalezas.

Data Explorer

26.15 GB

- sample_submission.csv
- test.zip
- train_curated.csv
- train_curated.zip
- train_noisy.csv
- train_noisy.zip

Descripción del conjunto de datos

Si está interesado en el conjunto de datos **FSDKaggle2019** utilizado para esta competencia, descárguelo de **Zenodo**. La versión de Zenodo contiene todos los archivos del conjunto de datos, **incluido el conjunto de prueba completo y las etiquetas**, así como archivos csv actualizados con metadatos y licencias.

Descargue FSDKaggle2019 en https://doi.org/10.5281/zenodo.3612637

Citación

Si usa el conjunto de datos FSDKaggle2019 o el código de referencia, cite nuestro documento DCASE 2019:

Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. *Free sound datasets: a platform for the creation of open audio datasets*. In Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), pp 486-493. Suzhou, China, 2017.

Archivos dañados detectados en el conjunto de trenes seleccionados

Los siguientes 5 archivos de audio en el conjunto de trenes seleccionado tienen una etiqueta incorrecta debido a un error en el proceso de cambio de nombre de archivo:
`f76181c4.wav`, `77b925c2.wav`, `6a1f682a.wav`, `c7db12aa.wav`, `7752cc8a.wav`
Se descubrió que el archivo de audio `1d44b0bd.wav`en el conjunto de trenes seleccionados estaba dañado (no contiene señal) debido a un error en la conversión de formato.

Motivación de tareas

Las técnicas actuales de aprendizaje automático requieren conjuntos de datos grandes y variados para proporcionar un buen rendimiento y generalización. Sin embargo, etiquetar manualmente un conjunto de datos lleva mucho tiempo, lo que limita su tamaño. Los sitios web como Freesound o Flickr alojan grandes volúmenes de audio y metadatos aportados por los usuarios, y las etiquetas se pueden inferir automáticamente a partir de los metadatos y/o hacer predicciones con modelos previamente entrenados. Sin embargo, estas etiquetas inferidas automáticamente pueden incluir un nivel sustancial de ruido de etiquetas.

La principal pregunta de investigación abordada en esta competencia es cómo explotar adecuadamente una pequeña cantidad de datos confiables etiquetados manualmente y una mayor cantidad de datos de audio web ruidosos en una tarea de etiquetado de audio de etiquetas múltiples con una configuración de vocabulario grande. Además, dado que los datos provienen de

diferentes fuentes, la tarea fomenta los enfoques de adaptación de dominios para lidiar con un posible desajuste de dominios.

Conjunto de datos de audio

El conjunto de datos utilizado en este desafío se llama **FSDKaggle2019** y emplea clips de audio de las siguientes fuentes:

- Freesound Dataset (FSD): conjunto de datos que se está recopilando en el MTG-UPF a partir de contenidos Freesound organizados con AudioSet Ontology
- Las bandas sonoras de un grupo de videos de Flickr tomados del conjunto de datos Yahoo Flickr Creative Commons 100M (YFCC)

Los datos de audio se etiquetan utilizando un vocabulario de 80 etiquetas de la ontología AudioSet de Google [1], que cubre diversos temas: guitarra y otros instrumentos musicales, percusión, agua, digestivo, sonidos respiratorios, voz humana, locomoción humana, manos, acciones de grupos humanos, Insectos, animales domésticos, vidrio, líquido, vehículos de motor (carretera), mecanismos, puertas y una variedad de sonidos domésticos. La lista completa de categorías se puede consultar en `sample_submission.csv`la parte inferior de esta página.

Etiquetas de verdad de tierra

Las etiquetas de verdad básica se proporcionan a nivel de clip y expresan la presencia de una categoría de sonido en el clip de audio, por lo que pueden considerarse etiquetas o etiquetas *débiles* . Los clips de audio tienen una duración variable (aproximadamente de 0,3 a 30 s; consulte más detalles a continuación).

El contenido de audio de FSD ha sido etiquetado manualmente por humanos siguiendo un proceso de etiquetado de datos utilizando la plataforma Freesound Annotator . La mayoría de las etiquetas tienen acuerdo entre anotadores, pero no todas. Se pueden encontrar más detalles sobre el proceso de etiquetado de datos y Freesound Annotator en [2].

Las bandas sonoras de YFCC se etiquetaron utilizando heurísticas automatizadas aplicadas al contenido de audio y los metadatos de los clips originales de Flickr. Por lo tanto, se puede esperar una cantidad sustancial de ruido de etiqueta. El ruido de la etiqueta puede variar ampliamente en cantidad y tipo según la categoría, incluidos los ruidos dentro y fuera del vocabulario. Más información sobre algunos de los tipos de ruido de etiqueta que se pueden encontrar está disponible en [3].

Formato y Licencia

Todos los clips se proporcionan como archivos de audio mono PCM de 16 bits y 44,1 kHz sin comprimir. Todos los clips utilizados en este concurso se publican bajo licencias Creative Commons (CC), algunos de los cuales requieren la atribución a sus autores originales y otros prohíben su reutilización comercial. Para poder cumplir con los términos de las licencias CC, se publicará una lista completa de clips de audio con sus licencias asociadas y una referencia al contenido original (en Freesound o Flickr) al final de la competencia. Hasta entonces, los archivos de audio proporcionados solo se pueden utilizar con el único fin de participar en el concurso.

Conjunto de entrenamiento

El **conjunto de entrenamiento** está destinado a ser para el desarrollo del sistema. La idea es limitar la supervisión proporcionada (es decir, los datos etiquetados manualmente), promoviendo así

enfoques para lidiar con el ruido de las etiquetas. El conjunto de trenes se compone de **dos subconjuntos** de la siguiente manera:

Subconjunto seleccionado

El subconjunto curado es un pequeño conjunto de datos etiquetados manualmente de FSD .

- Número de clips/clase: 75 excepto en algunos casos (donde hay menos)
- Número total de clips: 4970
- Número medio de etiquetas/clip: 1,2
- Duración total: 10,5 horas

La duración de los clips de audio oscila entre 0,3 y 30 s debido a la diversidad de categorías de sonido y las preferencias de los usuarios de Freesound a la hora de grabar/cargar sonidos. Puede suceder que algunos de estos clips de audio presenten material acústico adicional más allá de las etiquetas de verdad proporcionadas.

subconjunto ruidoso

El subconjunto ruidoso es un conjunto más grande de datos de audio web ruidosos de videos de Flickr tomados del conjunto de datos YFCC [5].

- Número de clips/clase: 300
- Número total de clips: 19815
- Número medio de etiquetas/clip: 1,2
- Duración total: ~80 horas

La duración de los clips de audio oscila entre 1 y 15 segundos, y la gran mayoría dura 15 segundos.

Teniendo en cuenta los números anteriores, la distribución de datos por clase disponible para el entrenamiento es, para la mayoría de las clases, 300 clips del subconjunto ruidoso y 75 clips del subconjunto curado, lo que significa 80 % ruidoso - 20 % curado a nivel de clip (no a nivel de duración del audio, considerando los clips de duración variable).

Conjunto de prueba

El **conjunto de prueba** se utiliza para la evaluación del sistema y consta de datos etiquetados manualmente de FSD . Dado que la mayoría de los datos del tren provienen de YFCC , se puede esperar cierta discrepancia en el dominio acústico entre el tren y el equipo de prueba. Todo el material acústico presente en el set de prueba está rotulado, salvo error humano, considerando el vocabulario de 80 clases utilizado en la competencia.

El conjunto de prueba se divide en dos subconjuntos, para las tablas de clasificación **públicas** y **privadas .** En esta competencia, la presentación se realizará a través de Kaggle Kernels. Solo se proporciona el subconjunto de prueba correspondiente a la tabla de clasificación *pública* (sin datos reales).

Las presentaciones deben realizarse con modelos de inferencia que se ejecutan en Kaggle Kernels. Sin embargo, los participantes pueden decidir entrenar también en Kaggle Kernels o fuera de línea (ver Requisitos de Kernels para más detalles).

Esta es una competencia solo de kernels con dos etapas. La primera etapa comprende el plazo de presentación hasta la fecha límite del 10 de junio. Después de la fecha límite, en la segunda etapa, Kaggle volverá a ejecutar los núcleos seleccionados en un conjunto de prueba invisible. **El conjunto de prueba de la segunda etapa es aproximadamente tres veces más grande que el primero.** Debe planificar la memoria, el disco y el tiempo de ejecución de su núcleo en consecuencia.

archivos

- **train_curated.csv**: etiquetas de verdad del terreno para el subconjunto **curado** de los archivos de audio de entrenamiento (ver Campos de datos a continuación)
- **train_noisy.csv**: etiquetas de verdad del terreno para el subconjunto **ruidoso** de los archivos de audio de entrenamiento (ver Campos de datos a continuación)
- **sample_submission.csv**: un archivo de envío de muestra en el formato correcto, incluida la clasificación correcta de las categorías de sonido; contiene la lista de archivos de audio que se encuentran en la carpeta **test.zip** (correspondiente a la tabla de clasificación pública)
- **train_curated.zip**: una carpeta que contiene los archivos de entrenamiento de audio (.wav) del **subconjunto** seleccionado
- **train_noisy.zip**: una carpeta que contiene los archivos de entrenamiento de audio (.wav) del subconjunto **ruidoso**
- **test.zip**: una carpeta que contiene los archivos de prueba de audio (.wav) para la tabla de clasificación pública

columnas

Cada fila de los archivos **train_curated.csv** y **train_noisy.csv** contiene la siguiente información:

- **fname** : el nombre del archivo de audio, por ejemplo,`0006ae4e.wav`
- **etiquetas** : la(s) etiqueta(s) de clasificación de audio (verdad básica). Tenga en cuenta que el número de etiquetas por clip puede ser uno, por ejemplo, `Bark`o más, por ejemplo, `"Walk_and_footsteps,Slam"`.

Referencias

[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 Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2017. [PDF]

[2] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. "Freesound Datasets: A Platform for the Creation of Open Audio Datasets." In Proceedings of the International Conference on Music Information Retrieval, 2017. [PDF]

[3] Eduardo Fonseca, Manoj Plakal, Daniel P. W. Ellis, Frederic Font, Xavier Favory, and Xavier Serra. "Learning Sound Event Classifiers from Web Audio with Noisy Labels." In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2019. [PDF]

[4] Frederic Font, Gerard Roma, and Xavier Serra. "Freesound technical demo." Proceedings of the 21st ACM international conference on Multimedia, 2013. https://freesound.org

[5] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li, YFCC100M: The New Data in Multimedia Research, Commun. ACM, 59(2):64–73, January 2016

Giannakopoulos (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis

Kaur (2022). Fall Detection from Audios with Audio Transformers

Dastres (2021). A Review in Advanced Digital Signal Processing Systems

McFee (2015). A SOFTWARE FRAMEWORK FOR MUSICAL DATA AUGMENTATION

Schluter (2015). EXPLORING DATA AUGMENTATION FOR IMPROVED SINGING VOICE DETECTION WITH NEURAL NETWORKS