Taller de Aprendizaje Automático

Segundo Proyecto - Freesound AudioTagging 2019

Instituto de Ingeniería Eléctrica
Facultad de Ingeniería

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

Montevideo, 2025

Motivación

- En la actualidad se generan inmensas cantidades de grabaciones
- Necesidad de estructurar los datos

Desafío

kaggle

Search

Sign In | Register

Create

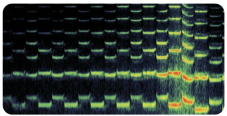
- Home
- Competitions
- Datasets
- Models
- Code
- Discussions
- Learn
- More

FREESOUND · RESEARCH CODE COMPETITION · 6 YEARS AGO

Late Submission

Freesound Audio Tagging 2019

Automatically recognize sounds and apply tags of varying natures



Overview | Data | Code | Models | Discussion | Leaderboard | Rules

Overview

Start
Apr 4, 2019

Close
Jun 18, 2019

Merger & Entry

Competition Host
Freesound

Prizes & Awards
$5,000
Awards Points & Medals

Participation
5,039 Entrants
520 Participants
880 Teams
677 Submissions

Tags
Audio

Description

One year ago, Freesound and Google's Machine Perception hosted an audio tagging competition challenging Kagglers to build a general-purpose auto tagging system. This year they're back and taking the challenge to the next level with multi-label audio tagging, doubled number of audio categories, and a *noisier than ever*

Problema a resolver

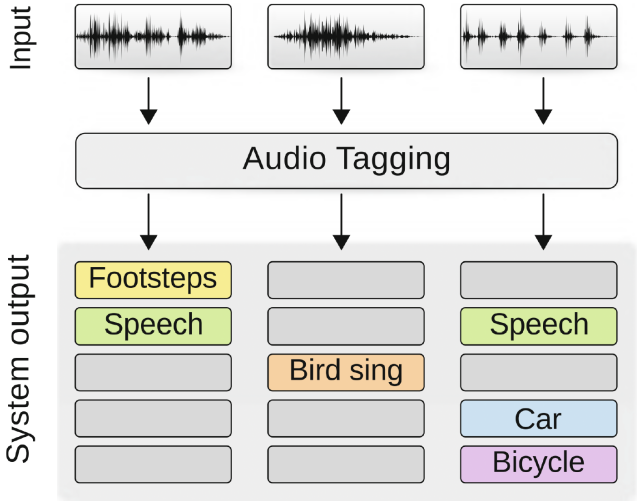


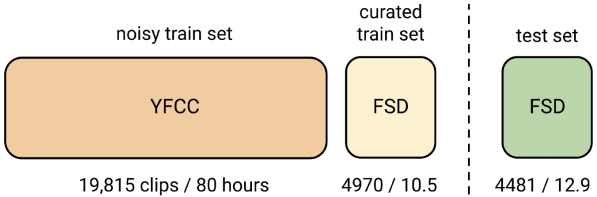
Figure: Esquema de un problema de Audio Tagging [5]

Objetivos del desafío

- Construir un modelo de Audio Tagging capaz de reconocer eventos sonoros de naturaleza diversa.
- Aprovechar subconjuntos de datos de entrenamiento con anotaciones de fiabilidad variable.

Datos del desafío

- Los datos provienen de Freesound y Flickr



Aspect	curated train	noisy train	test
Clips/class	∼75*	300	∼ 50 - 150
Total clips	4970	19,815	4481
Labels/clip	1.2	1.2	1.4
Clip length	∼0.3 - 30s	∼15s	∼0.3 - 30s
Total duration	∼10.5h	∼80h	∼12.9h
Labelling	correct (inexhaustive)	noisy	exhaustive

Figure: Distribución de los datos [2]

- Datos etiquetados con 80 clases posibles.

Clases del desafío

Human sounds
- Human voice
- Whistling
- Respiratory sounds
- Human locomotion
- Digestive
- Hands
- Heart sounds, heartbeat
- Otoacoustic emission
- Human group actions

Source-ambiguous sounds
- Generic impact sounds
- Surface contact
- Deformable shell
- Onomatopoeia
- Silence
- Other sourceless

Animal
- Domestic animals, pets
- Livestock, farm animals, working animals
- Wild animals

Sounds of things
- Vehicle
- Engine
- Domestic sounds, home sounds
- Bell
- Alarm
- Mechanisms
- Tools
- Explosion
- Wood
- Glass
- Liquid
- Miscellaneous sources
- Specific impact sounds

Music
- Musical instrument
- Music genre
- Musical concepts
- Music role
- Music mood

Natural sounds
- Wind
- Thunderstorm
- Water
- Fire

Channel, environment and background
- Acoustic environment
- Noise
- Sound reproduction

Figure: Ontología de AudioSet[3]

Métrica

- Se utiliza la métrica *label-weighted* **label-ranking** average **precision** (lwlrap)

$$\text{Prec}(s, c) = \frac{1}{\text{Rank}(s, c)} \sum_{r=1}^{\text{Rank}(s,c)} \mathbf{1}\left[Lab(s, r) \in C(s)\right]$$

Expresión de *label-ranking precision*[2]

- **s** – Indica la muestra con la que estamos trabajando
- **c** – Indica la clase para la cual calculamos la precisión
- **C(s)** – Lista de todas las etiquetas ground-truth de la muestra **s**
- **Rank(s, c)** – Ranking de la clase **c** en la predicción sobre **s**
- **Lab(s, r)** – Etiqueta predicha en la posición **r** para la muestra **s**

Ejemplo de Cálculo

Clases del problema $\longrightarrow$ [dog barking – siren – engine idling]

Sea una muestra **s_1** con las siguientes etiquetas:

C(s_1) $\longrightarrow$ [siren – engine idling]

Predicciones del modelo $\longrightarrow$ [0.7 – 0.2 – 0.5]

A partir de las predicciones obtenemos el ranking:

Rank(s_1, dog barking) = 1 Rank(s_1, engine idling) = 2 Rank(s_1, siren) = 3

A partir de las predicciones obtenemos las etiquetas predichas en cada ranking:

Lab(s_1, 1) = dog barking Lab(s_1, 2) = engine idling Lab(s_1, 3) = siren

Ejemplo de Cálculo Detallado

Queremos calcular la precisión de la clase `engine idling` para la muestra s_1:

- $\text{Rank}(s_1, \text{engine idling}) = 2$
- $C(s_1) = [\texttt{siren}, \texttt{engine idling}]$
- $\text{Lab}(s_1, 1) = \texttt{dog barking}$
- $\text{Lab}(s_1, 2) = \texttt{engine idling}$

Aplicamos la fórmula:

$$\text{Prec}(s_1, \text{engine idling}) = \frac{1}{2}\left(1[\texttt{dog barking} \in C(s_1)] + 1[\texttt{engine idling} \in C(s_1)]\right)$$

Evaluación:

$$\text{Prec}(s_1, \text{engine idling}) = \frac{1}{2}(0 + 1) = 0.5$$

Ejemplo de Cálculo (cont.)

Sea una segunda muestra **s_2** con la siguiente etiqueta:

C(s_2) $\longrightarrow$ [`dog barking`]

Predicciones del modelo $\longrightarrow$ [0.4 – 0.6 – 0.3]

Ranking de etiquetas:

`Rank(s_2, engine idling)` $= 1$ `Rank(s_2, dog barking)` $= 2$ `Rank(s_2, siren)` $= 3$

$$\text{Prec}(s_2, \text{dog barking}) = \frac{1}{2} \times (1\,[\text{Lab}(s_2, 1) \in C(s_2)] + 1\,[\text{Lab}(s_2, 2) \in C(s_2)])$$

$$\text{Prec}(s_2, \text{dog barking}) = \frac{1}{2} \times 1 = 0.5$$

Métrica

- Se utiliza la métrica **label-weighted** *label-ranking* **average** *precision* (lwlrap)

$$lwlrap = \frac{1}{\sum_s |C(s)|} \sum_s \sum_{c \in C(s)} \text{Prec}(s, c)$$

- $|C(s)|$ – Indica la cantidad de etiquetas de la muestra **s**
- Por la formulación de lwlrap, solo nos interesa calcular esta precisión para las clases que sí tiene la muestra **s**

Cálculo Final de *lwlrap*

$$lwlrap = \frac{1}{\sum_s |C(s)|} \sum_s \sum_{c \in C(s)} \text{Prec}(s, c)$$

Para la muestra s_1 $C(s_1) = [\texttt{siren}, \texttt{engine idling}]$

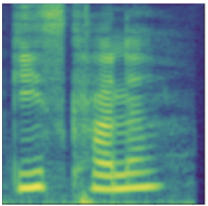
- $\text{Prec}(s_1, \text{engine idling}) = 0.5$
- $\text{Prec}(s_1, \text{siren}) = \frac{1}{3}(0 + 1 + 1) = \frac{2}{3} \approx 0.6667$

Para la muestra s_2 $C(s_2) = [\texttt{dog barking}]$

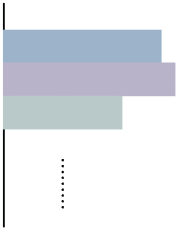
- $\text{Prec}(s_2, \text{dog barking}) = 0.5$

$$\text{lwlrap} = \frac{0.5 + 0.6667 + 0.5}{2 + 1} = \frac{1.6667}{3} \approx \boxed{0.5556}$$

Baseline



MobileNet v1



Audios

Espectrograma
Log-Mel

Modelo

Score

Datos disponibles - InClass

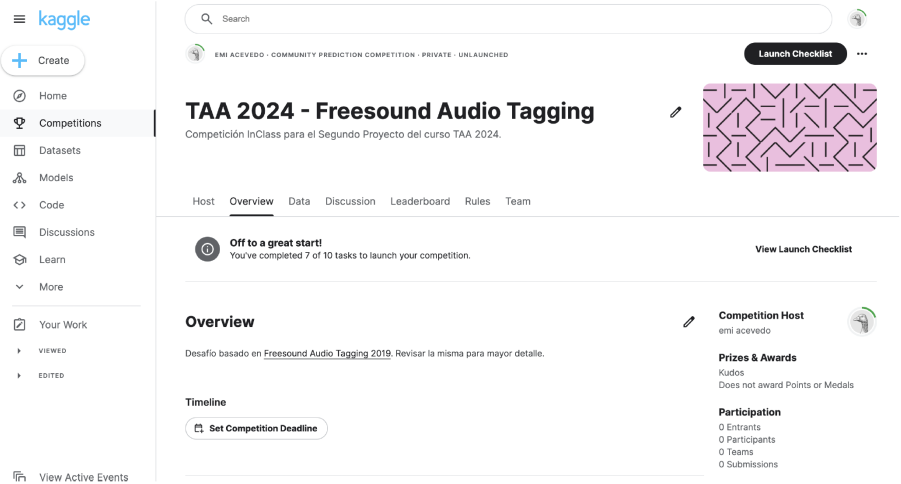


Figure: Link a la competencia

Referencias Relevantes

- Documentación del desafío
- Ejemplo de problema de audio en Tensorflow
- Implementación de la Métrica lwlrap

Referencias I

[1] Steven Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* (1980).

[2] Eduardo Fonseca et al. "Audio Tagging with Noisy Labels and Minimal Supervision". In: *Proceedings of the European Signal Processing Conference (EUSIPCO)*. arXiv:1906.02975. 2020.

[3] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[4] Thomas F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[5] Tuomas Virtanen, Mark D. Plumbley, and Daniel Ellis. *Computational Analysis of Sound Scenes and Events*. Ed. by T. Virtanen, M. D. Plumbley, and D. Ellis. Springer, 2018.