Learning to Recognize Musical Genre from Audio

Challenge Overview

Michaël Defferrard
EPFL, Lausanne, Switzerland
michael.defferrard@epfl.ch

Sharada P. Mohanty
EPFL, Lausanne, Switzerland
sharada.mohanty@epfl.ch

Sean Carroll
EPFL, Lausanne, Switzerland
sean.carroll@epfl.ch

Marcel Salathé
EPFL, Lausanne, Switzerland
marcel.salathe@epfl.ch

ABSTRACT

todo

CCS CONCEPTS

• **Information systems** → **Music retrieval**; • **Computing methodologies** → *Supervised learning*;

KEYWORDS

Music Information Retrieval (MIR), Challenge, Open Data

ACM Reference format:
Michaël Defferrard, Sharada P. Mohanty, Sean Carroll, and Marcel Salathé. 2018. Learning to Recognize Musical Genre from Audio. In *Proceedings of The Web Conference, Lyon, France, April 2018 (WWW'18),* 3 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Like never before, the web has become a place for sharing creative work — such as music — among a global community of artists and art lovers. While music and music collections predate the web, the web enabled much larger scale collections. Whereas people used to own a handful of vinyls or CDs, they nowadays have instant access to the whole of published musical content via online platforms such as Spotify, iTunes, Youtube, FMA, Jamendo, Bandcamp, etc. Such dramatic increase in the size of music collections created two challenges: (i) the need to automatically organize a collection (as users and publishers cannot manage them manually anymore), and (ii) the need to automatically recommend new songs to a user knowing his listening habits. An underlying task in both those challenges is to be able to group song in semantic categories.

The task of this challenge, one of the four official challenges of the Web Conference (WWW2018) challenges track, is to recognize the musical genre of a piece of music of which only a recording is available. Genres are broad, e.g. pop or rock, and each song only has one target genre. Other metadata, e.g. the song title or artist name, shall not be used for the prediction. The submitted algorithms shall learn to map an audio signal, i.e. a time series, to one of the 16 target genres.

The data for this challenge comes from the recently published FMA dataset [?], which is a dump of the Free Music Archive[1] (FMA), an interactive library of high-quality and curated audio which is freely and openly available to the public. That dataset is a collection of 917 GiB and 343 days of Creative Commons-licensed audio from 106,574 tracks from 16,341 artists and 14,854 albums, arranged in a hierarchical taxonomy of 161 genres. It provides full-length and high-quality audio, pre-computed features, together with track- and user-level metadata, tags, and free-form text such as biographies.

2 MUSIC GENRE RECOGNITION

Music genres are categories that have arisen through a complex interplay of cultures, artists, and market forces to characterize similarities between compositions and organize music collections. Yet, the boundaries between genres still remain fuzzy, making the problem of music genre recognition (MGR) a nontrivial task [?]. While its utility has been debated, mostly because of its ambiguity and cultural definition, it is widely used and understood by end-users who find it useful to discuss musical categories [?]. As such, it is one of the most researched areas in the Music Information Retrieval (MIR) field.[2]

Given the extent to which MGR is studied, we can expect many groups from the MIR community to be interested. Moreover, given the well-defined scope of the challenge, we can expect members of the wider machine learning community to participate as well. Both communities are well versed in challenges, with e.g. Kaggle and NIPS for the ML community and the MIREX for the MIR community.

3 THE CHALLENGE

To avoid overfitting and cheating, the challenge will happen in two rounds. The final ranking will be based on results from the second round. The training data for both rounds consisted of the FMA medium subset, which is composed of 25,000 clips of 30 seconds, categorized in 16 genres. The categorization is unbalanced with 21 to 7,103 clips per genre, but only one genre per track. As the data is public, we collected new test data for the second round to prevent access to the test set.

In the first round, participants have been provided a test set of 30,000 clips of 30 seconds each and had to submit the predicted

WWW'18, April 2018, Lyon, France
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] https://freemusicarchive.org

[2] See e.g. [?] for an extensive list of ~500 references to works targeting MGR as of 2012.

genre for each of these clips. The platform evaluated the predictions and ranked the participants upon submission. A subset of these clips were sampled from the FMA large dataset, while ensuring that the 30 seconds window present in the final test set do not overlap with any of the 30 seconds clips provided in the training set. The other subset was sampled from songs in the FMA full dataset, which are not present in the medium subset.

For the second round, the participants had to wrap their models in a docker container which encapsulated the prediction code and their trained model. We evaluated those against a new unseen test set. These 30s clips were sampled from new contributions to the Free Music Archive. The participants were also asked to post online the code of their approach as well as an executive summary of the method used to be eligible for the second round.

Both rounds used the same evaluation metric. The primary score was the mean log loss and the secondary score was the mean F_1 score. The mean log loss is defined by

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{nc}\ln(p_{nc}), \tag{1}$$

where $N = 35000$ is the number of examples in the test set, $C = 16$ is the number of class labels, i.e. genres, y_{nc} is a binary value indicating if the n-th instance belongs to the c-th label, p_{nc} is the probability according to your submission that the n-th instance belongs to the c-th label, ln is the natural logarithmic function. The F_1 score for a particular class c is given by

$$F_1^c = 2\frac{p^c r^c}{p^c + r^c}, \quad p^c = \frac{tp^c}{tp^c + fp^c}, \quad r^c = \frac{tp^c}{tp^c + fn^c}, \tag{2}$$

where p^c is the precision for class c, r^c is the recall, tp^c refers to the number of true positives, fp^c refers to the number of false positives, fn^c refers to the number of false negatives. The final mean F_1 score is then defined as

$$F_1 = \frac{1}{C}\sum_{c=1}^{C} F_1^c. \tag{3}$$

The challenge is hosted on crowdAI, a public platform for open challenges. Instructions to participate, training and test data, graded submissions, and the leaderboard are available on the challenge page.[3] Moreover, we developed a starter kit[4] which includes code to handle the data and make a submission. It also features some examples and a baseline.

Participants were encouraged to check out the FMA paper [?] for a detailed description of the data as well as the github repository[5] for Jupyter notebooks showing how to use the data, exploring it, and training baseline models.[6]

4 RESULTS

At the end of the first round of the challenge, we had engaged a total of 246 participants, who either made a submission, downloaded the datasets, or contributed in the discussion forums. From these 246 participants 59 participants made atleast one submission, with some of the top participants making as many as 110 submissions through the period of the 1st Round of the challenge. A total of 671 submissions were made in the first round, and of these, 77 were invalid submissions and 576 submissions were successfully graded. From the 576 successfully graded submissions, a total of 364 submissions had a score better than the baseline submission provided by the organisers, based on their own previous work. Figure **??** shows the distribution of the F_1*Score* and the *LogLoss* for valid submissions in the challenge. The current rank #1 on the leaderboard of the challenge has a F_1*Score* of 0.909 and a *LogLoss* of 0.330, improving the state of art **(TODO: Add @michael's paper's citation)** on this problem by **@michael: can you compute this value** ?. Figure **??** shows the progression of scores of all the participants through the Round-1 of the challenge, with each line representing one active participant in the challenge and their top-score at the point of time referenced by the X-axis.

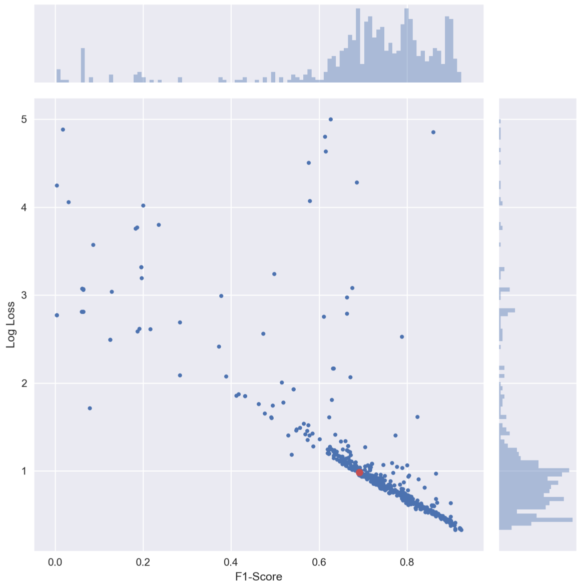

Figure 1: Joint plot of the distribution of the F_1*score* and the *LogLoss* of all valid submissions made by the participants. The point marked in red represents the baseline submission prepared by the organisers at the beginning of the challenge. The plot only considers submissions with *LogLoss* < 5 for interpretability.

5 CONCLUSION

That challenge is part of an effort to promote open evaluation in machine learning for music data, which the release of the open FMA dataset was the first step [?]. The goal of the initiative is to establish a reference benchmark based on open data. MIR research has historically suffered from the lack of publicly available benchmark datasets, which stem from the commercial interest in music by record labels, and therefore imposed rigid copyright. The FMA's solution was to aim for tracks which license permits redistribution. All data and code produced during the project and challenge are released under the CC BY 4.0 and MIT licenses.

[3] https://www.crowdai.org/challenges/www-2018-challenge-learning-to-recognize-musical-genre
[4] https://github.com/crowdAI/crowdai-musical-genre-recognition-starter-kit
[5] https://github.com/mdeff/fma
[6] This challenge used the `rc1` version of the data.

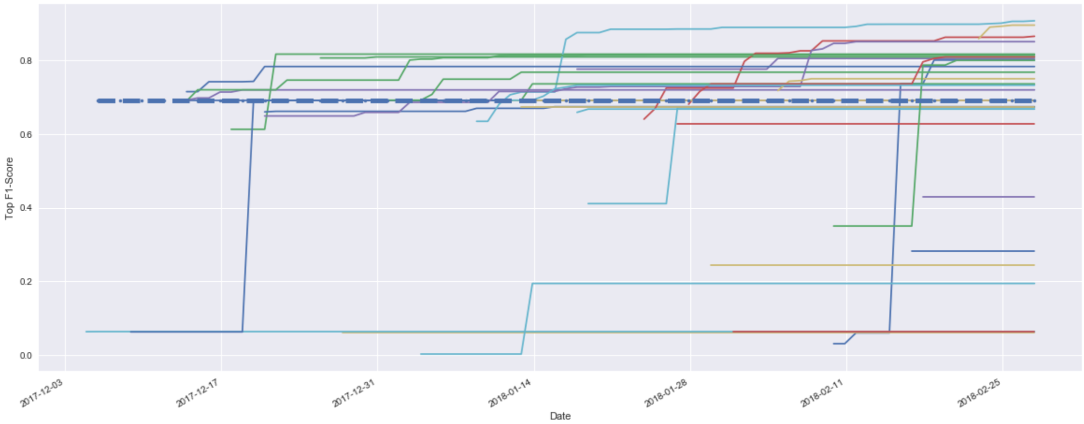

Figure 2: Plot of progression of scores of all the participants through Round-1 of the challenge. Each line represents one active participant in the challenge, and their top-score at the point of time referenced by the X-axis. The dotted blue line represents the baseline score from the baseline submission.