Analysis of Essentia models on the MTG-Jamendo dataset

David Bedoya
Universitat Pompeu Fabra
josedavid.bedoya01@estudiant.upf.edu

March 29, 2020

Abstract

This document reports on a comparison of the performance of some deep learning models that run on a large dataset of audio. The models are a subset of the TensorFlow classifiers available in Essentia, and the dataset is a chunk of the MTG-Jamendo dataset.

1 Introduction

Essentia provides a set of auto-tagging models that use the MUSICNN and VGG architectures, each of which is pre-trained on the MSD and MTT research datasets. From these auto-tagging MUSICNN models and a VGGish_AudioSet model, transfer learning classifiers are generated, hereby the former models are referred as source tasks, and the latter as target tasks[1].

The main goal of this project is to compare the performance of some of the Essentia classifier models on a manually annotated chunk of 565 audio samples from the MTG-Jamendo dataset. The code is openly available[2], and has precise explanations for each step of the process.

2 Methodology

2.1 Annotations

The first step is to create the ground truth annotations for each of the 565 audio samples. This is carried out manually using The MTG-Jamendo-Annotator, with which 12 classification tasks are annotated in two groups, namely mood and miscellaneous. (see column **annotation classes** in Table 1). These annotations are stored in json format in the folder named *annotations*.

[1] For a more detailed description of the Essentia models, please refer to:
https://mtg.github.io/essentia-labs//news/2020/01/16/tensorflow-models-released/

[2] https://github.com/jdavibedoya/essentia-models_mtg-jamendo

target task	annotation classes	model classes
mood_acoustic	not_acoustic: 0, acoustic: 1	acoustic: a_0, not_acoustic: a_1
mood_electronic	not_electronic: 0, electronic: 1	electronic: a_0, not_electronic: a_1
mood_aggressive	not_aggressive: 0, aggressive: 1	aggressive: a_0, not_aggressive: a_1
mood_relaxed	not_relaxed: 0, relaxed: 1	not_relaxed: a_0, relaxed: a_1
mood_happy	not_happy: 0, happy: 1	happy: a_0, not_happy: a_1
mood_sad	not_sad: 0, sad: 1	not_sad: a_0, sad: a_1
mood_party	not_party: 0, party: 1	not_party: a_0, party: a_1
timbre	dark: 0, bright: 1	N/A
tonal_atonal	atonal: 0, tonal: 1	tonal: a_0, atonal: a_1
danceability	non_danceable: 0, danceable: 1	danceable: a_0, non_danceable: a_1
voice_instrumental	instrumental: 0, voice: 1	instrumental: a_0, voice: a_1
gender	male: 0, female: 1, instrumental: 2	female: a_0, male: a_1

Table 1: Target tasks and classes.

2.2 Predictions

Predictions for all the classifications tasks[3] are computed using two source tasks, namely MUSICNN_MSD and VGGish_AudioSet. In the code, each audio file is loaded just once per source task, and only its first 3 minutes are processed. The code is also designed to store the predictions before loading the next audio sample and to avoid recomputing them.

These predictions are stored by source task, that is, the folder named *predictions* contains a sub-folder for each source task. The format of the stored predictions is similar to that used of the annotations, however this time two values $[a_0, a_1]$ are stored because the output layer of the architecture of the transfer learning classifiers consists of 2 units with sigmoid activations. It is important to note that the order of some classes of the models are swapped with respect to the order of the classes of the annotations (see column **model classes** in Table 1).

3 Evaluation

Each prediction is classified according to the activation of the highest value, for instance, in agreement with the **model classes** column in Table 1, $"tonal\_atonal" : [0.40, 0.51]$ is classified as atonal. Weighted f1-score[4] is the metric used here for the evaluation of the classification models on the dataset. This metric takes into account the skewness present in the ground truth annotations. To determine which source task provides the best overall results, the following metrics are used here:

- Weighted average of the weighted f1-scores of the target tasks. This average must be weighted since the support of the gender task is different from the rest, this is due to the fact that this task has three annotated classes, but the models only predict two of them, so the data corresponding to the annotated third class are discarded.
- Number of target tasks in which the source task performed best.

[3] Except for the timbre task, since there is no model for it.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

4 Results

Table 2 presents the results of the metrics proposed in Section 3. Fig. 1 shows the performance of each classification model on the dataset, and for the source task identified as the best, the Fig. 2 shows confusion matrices for all 11 target tasks.

source task	weighted average	the best in
MUSICNN_MSD	72.69 %	mood_acoustic mood_aggressive mood_happy mood_sad tonal_atonal danceability gender
VGGish_AudioSet	72.13 %	mood_acoustic mood_electronic mood_relaxed mood_party voice_instrumental

Table 2: Results of the metrics used to identify the best source task.

5 Discussion

From the results, MUSICNN_MSD is identified as the source task that performed best for both metrics. However, it can be seen that there is not a big performance difference for the weighted average metric of both source tasks, it is only 0.6%. In most cases, the difference in the performance of the target tasks is not big either, except for the tonal_atonal task, for which the VGGish_AudioSet weighted f1-score is less than that of MUSICNN_MSD by 13.31%. The lowest performance can be observed for the mood tasks happy and sad, which could indicate that the distribution of the audios on which the models are trained is significantly different from that of the MTG-Jamendo dataset chunk, or that there is a lack of agreement between the criteria with which both distributions were annotated.

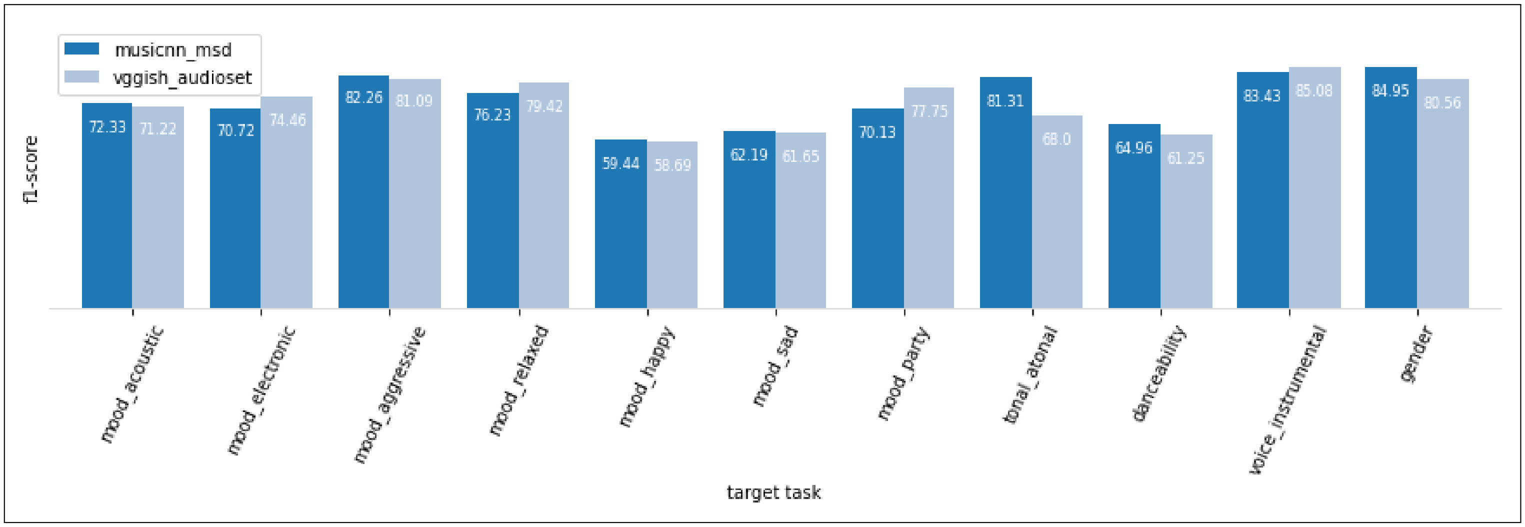

Figure 1: Weighted f1-scores.

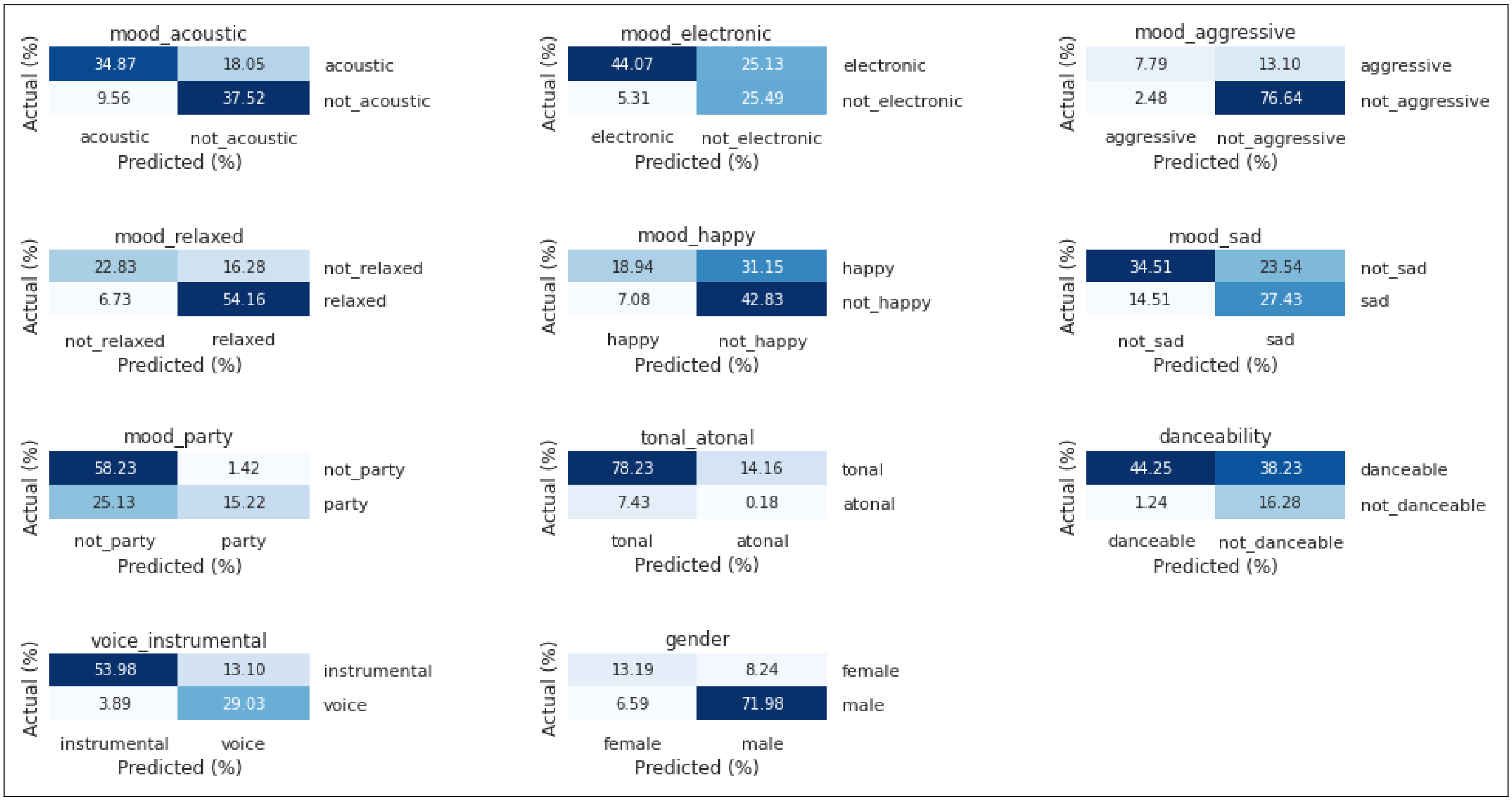

Figure 2: Confusion matrices - musicnn_msd.