Analysis of Tensorflow models for Essentia on MTG-jamendo dataset

Francesca Ronchini

March 2020

1 Introduction

This work aims to analyse and evaluate transfer learning using Tensorflow models for Essentia [1], concentrating the research on two already pre-trained models: *MusicCNN* and *VGG-like (VGGish).* The models have been pre-trained as part of Jordi Pons' Ph.D. Thesis [6] on two research datasets, respectively*Million Song Dataset (MSD)* [2] and *AudioSet* [3] for different source tasks categories [1]. This study is based on *Mood* and *Miscellaneous* source task on a particular chunk (9th) of MTG-jamendo dataset [4], with the goal of evaluating how well each model performs when transfer learning is applied, comparing the results with manually annotated ground truth.
More information about Essentia models can be found at [1] whereas for more information about MTG-jamendo-dataset the reader is referred to [4].

The project has been divided in three main steps:

1. Annotation of all the audio files inside the selected chunk (9th) for both source tasks categories, *mood* and *miscellaneous.*
2. Classification of all the audio files using the Tensorflow models for Essentia (*MusicCNN_MSD* and *VGGish_AudioSet* for both source tasks categories).
3. Evaluation of Essentia models for both source tasks in terms of overall source task and single classifier *accuracy* and *error rate.*

In the following sections, each step will be detailed. Code repositories are openly available for reproducibility. [1]

2 Annotation

The first step of the study is the manual annotation of the 9th chunk of the MTG-jamendo dataset [4], which contain 581 tracks. Each track has been an-

[1] https://github.com/RonFrancesca/AnalysisEssentiamodels-MTGJamendodataset

notated for both *mood* and *miscellaneous* task. Table 1 and Table 2 present the class according to which a song should have been annotated for the two source tasks.

Class	Values
mood_acoustic	acoustic, not acoustic
mood_aggressive	aggressive, not aggressive
mood_electronic	electronic, not electronic
mood_happy	happy, not happy
mood_party	party, not party
mood_relaxed	relaxed, not relaxed
mood_sad	sad, not sad

Table 1: Mood category classes

Class	Values
tonal_atonal	tonal, atonal
voice_instrumental	voice, instrumental
danceability	danceable, not danceable
gender	instrumental, female, male
timbre	bright, dark

Table 2: Miscellaneous category classes

The class *timbre* has not been considered in this work.

The audio file have been manually annotated using the MTG-jamendo annotator tool [5], according to the following criteria.

2.1 Mood category

- *Acoustic*: a track has been annotated as *acoustic* if not played with instrument which require electricity or considered as electronic (such as keyboards, electric guitar, etc.). Otherwise, it has have been annotated as *not acoustic.*

- *Aggressive*: a track has been annotated as *aggressive* if conveying sense of anxiety and somehow fear, usually accompanied by aggressive lyric of the song. Otherwise, it has have been annotated as *not aggressive.*

- *Electronic*: a track has been annotated as *electronic* if played with at least one instrument which require electricity or considered as electronic (such as keyboards, electric guitar, etc.). Otherwise, it has have been annotated as *not electronic.*

- *Happy*: a track has been annotated as *happy* if conveying positiveness and happiness, considering also the text of the lyric if any. Otherwise, it has have been annotated as *not happy*.
- *Party*: a track has been annotated as *party* according to if it would have been expected to be listened or played at parties. Otherwise, it has have been annotated as *not party*.
- *Relaxed*: a track has been annotated as *relaxed* if conveying relaxing and calm feeling. Otherwise, it has been annotated as *not relaxed*.
- *Sad*: a track has been annotated as *sad* if conveying sadness, also considering eventual lyrics of the song. Otherwise, it has have been annotated as *not sad*.

2.2 Miscellaneous

- *Danceability*: a track has been annotated as *danceable* if conveying desire to dance, if it would have been expected to be listened or played in a club or if belonging to some category of dance (such as salsa, etc.). Otherwise, it has have been annotated as *not danceable*.
- *Gender*: a track has been annotated as *female* or *male* according to which voice was more prominent in the song. If no voice was listened, it has have been annotated as *instrumental*.
- *Tonal_atonal*: a track has been annotated as *tonal* if a tonal scale, harmony or hierarchy of relation between notes could have been recognized in the audio file. Otherwise, it has have been annotated as *atonal*.
- *Voice_instrumental*: a track has been annotated as *voice* if at least one singing voice would have been heard during the song. Otherwise, it has have been annotated as *instrumental*.

3 Classification

After all the tracks have been annotated, an algorithm have been developed in order to make predictions for both *mood* and *miscellaneous* source task using the pre-trained TensorFlow models for Essentia [1]. For each audio file in the chuck 9th of the MTG-jamendo dataset [4], each classifier for each source task have been executed in order to make prediction and classify the track. In particular, 14 classifiers have been run for the *mood* category (7 for the *MusicCNN_MSD* model and 7 for the *VGGish_AudioSet* model) and 8 classifiers have been run for the *miscellaneous* category (4 for the *MusicCNN_MSD* model and 4 for the *VGGish_AudioSet* model).

The predictions, properly processed, have been saved in json files to make the comparison between the annotation considered as ground truth and the prediction of the models easier.

4 Results and evaluation

In order to evaluate the classifier, the manual annotations will be considered as ground truth. Therefore, the prediction made from the classifier will be compared with the ground truth annotation, evaluating *accuracy* and *error rate*. First, the overall accuracy will be presented for mood and miscellaneous task both for *MusicCNN_MSD* and *VGGish_AudioSet* model. Due to some limitations of this evaluation, the single accuracy of each classifier of the two tasks categories will be also presented. After concluding which model perform the best in overall, the confusion matrix of each classifier of the 'winner' model will be showed.

Table 3 and Table 4 showed the overall accuracy for *mood* and *miscellaneous* task.

Model	Norm. Accuracy
MusicCNN_MSD	0.54
VGGish_AudioSet	0.52

Table 3: Mood category task overall accuracy

Model	Norm. Accuracy
MusicCNN_MSD	0.42
VGGish_AudioSet	0.38

Table 4: Miscellaneous category task overall accuracy

As it is possible to observe, *MusicCNN* model seems to performs better for both tasks. Anyway, those results could be considered as biased. In fact, each model is considering more than a single classifier (7 for the *mood* task category and 4 for the *miscellaneous* task category.) The overall accuracy could be biased by a single classifier of the model which could perform very well, increasing the overall accuracy for the model, meanwhile all the others could perform badly. For this reason, it is interesting to consider also the single accuracy for each classifier of the two tasks, showed in Table 5 and Table 6.

From the results, it is possible to see that for the *mood* category, expect for *mood_electronic* and *mood_sad*, the model which perform the best is the *MusicCNN*. On the other hand, for the miscellaneous category, both model performs more or less the same in terms of accuracy, with the *VGGish* models performing better for the classifier *danceability* and *voice_instrumental*, while

Classifier	*MusicCNN_MSD*	*VGGish_AudioSet*
mood_aggressive	**0.89**	0.87
mood_electronic	0.75	**0.76**
mood_happy	**0.70**	0.68
mood_acoustic	**0.68**	0.62
mood_sad	0.38	**0.39**
mood_relaxed	**0.24**	0.18
mood_party	**0.14**	0.14

Table 5: Mood category single classifier accuracy

Classifier	*MusicCNN_MSD*	*VGGish_AudioSet*
voice_instrumental	0.83	**0.86**
danceability	0.67	**0.72**
tonal_atonal	**0.57**	0.40
gender	**0.27**	0.25

Table 6: Miscallaneous category single classifier accuracy

the *MusicCNN* models wins for the classifier *gender* and *tonal_atonal*, even if for those two classifiers the accuracy is not high. A part from this, it is possible to conclude that the difference between the two models is not that big. Also, it is interesting to notice that the same pattern is recognizable for the two models.

In fact, comparing the source tasks for the two models:

- **Mood**: both models are accurate regarding *mood_acoustic*, *mood_aggressive mood_electronic* and *mood_happy* but their performance decreases for *mood_party*, *mood_relaxed* and *mood_sad*. The *MusicCNN_MSD* model has been considered to perform better for the *mood_party* classifier because the results on the tables have been rounded to the second decimal place, but it actually performs slightly better (results can be checked on the Github repository mentioned before).

- **Miscellanous**: both models perform fine for *danceability* and *voice_instrumental*, but the performance decreases for the other two classifiers, with the *gender* being the one that perform the worst for each model. It need to be mentioned that for the classifier *gender*, while on the annotation step three possible values were available (instrumental, male, female), the models were making prediction only on two values (female, male). This could have affect the results of the classifier accuracy.

Anyway, if we count the number of times the *MusicCNN* models performs better than the *VGGish* models, it is possible to conclude that the classification for the source task *mood* using *MusicCNN* models gives the best overall results. For this reason, only the confusion matrices regarding the *MusicCNN* will be reported for the mood source task.

For the *miscellaneous* task instead, since there is not a winning model, both confusion matrices models will be showed.

For each classifier, a part from the confusion matrix, it is also reported the *accuracy* together with the *error rate*, defined as

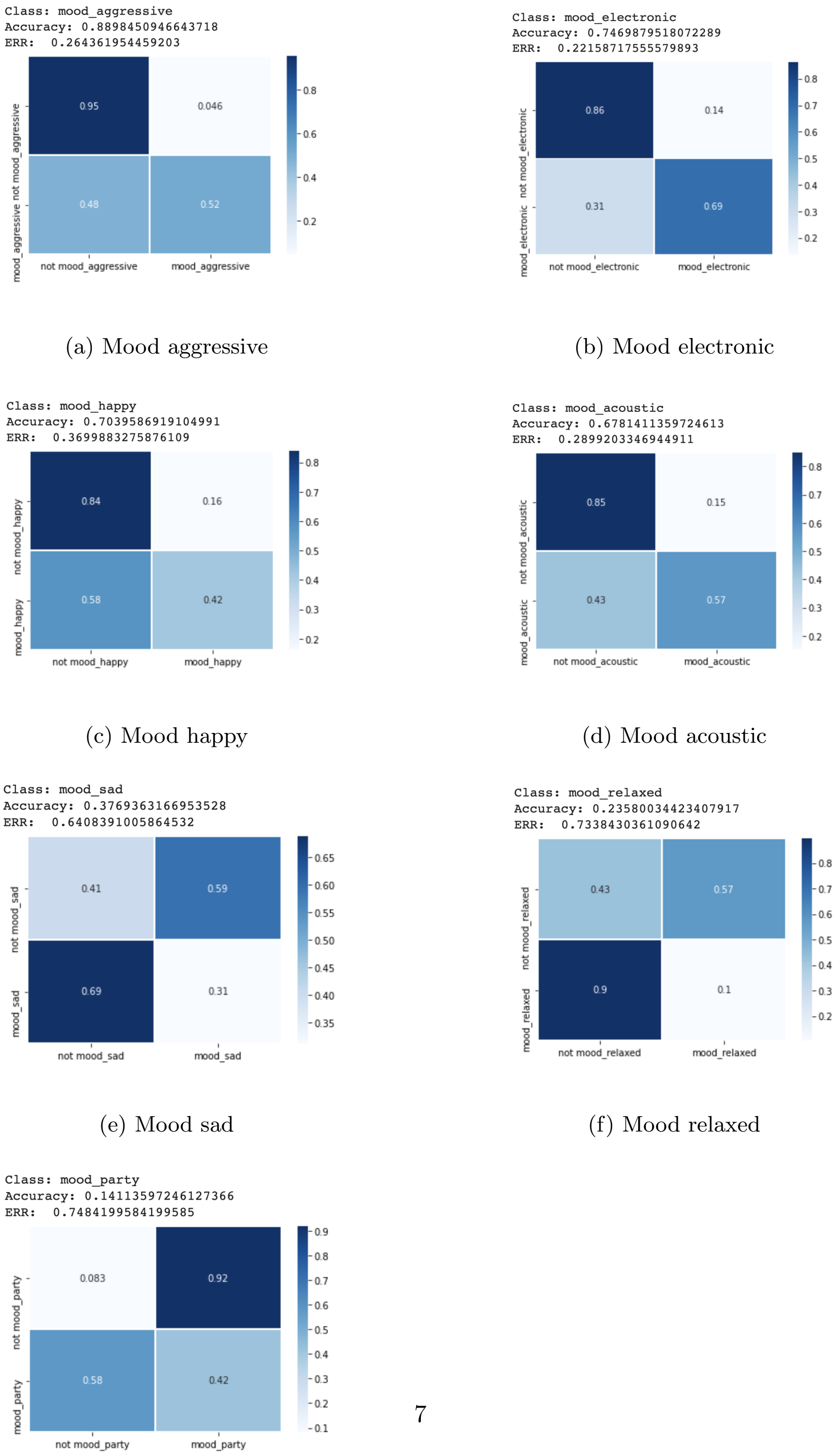
$$\frac{fp + fn}{tn + tp + fp + fn}$$

where:
fp = false positive, *fn* = false negative, *tn* = true negative, *tp* = true positive.

Figure 1 and Figure 2 show the confusion matrices for *MusicCnn* model for *mood* and *miscellaneous* task respectively. Figure 3 presents the confusion matrices for the *miscellaneous* task from the *VGGish* model. From the confusion matrices, it is possible to better understand the accuracy and the reason why for some classes the value is so low, as for example for the class *mood_relaxed* where almost all the songs annotated as relaxed are classified as not relaxed, or *mood_party* where mostly all the songs annotated as not_happy has been classified as happy. It need to be noticed that the confusion matrix for the classifier *gender* is not a 2*2 matrix, but a 3*3 because on the annotation tool three values were available while the prediction was only between male and female. This is also the reason why the third column of the matrix is 0.

5 Conclusion and Discussion

This work aim to analyze and evaluate two different pre-trained Tensorflow model for Essentia, *MusicCNN* and *VGGish*. From the results it is possible to conclude that even if the two models follow the same pattern in terms of accuracy, the model *MusicCNN* performs better for most of the cases. This could be due to the fact that *MusicCNN* is a musically-motivated Convolutional Neural Network [1], which uses vertical and horizontal convolutional filters to capture timbral and temporal patterns, meanwhile *VGGish* is based on an architecture from computer vision. Based on this, *MusicCNN* is expected to learn in a better way music relations. Also, the reason why the accuracy changes so much between a single classifier and another could be a consequence of an annotation based on personal and subjective criteria, which could be totally different from the ones according to which the models were previously trained. Of course, it is also need to be considered limitation due to the task itself, as for example tiredness or distraction while annotating the track during the annotation step. As future work, would be interesting to analyze the task knowing the criteria according to which the models have been previously trained. Also, if would be interesting to see how the gender classifier accuracy would change if either the annotator or the classifier would be correctly changed in order to make either 2 or 3 predictions values and to consider also the timbre classifier.



(a) Mood aggressive

(b) Mood electronic

(c) Mood happy

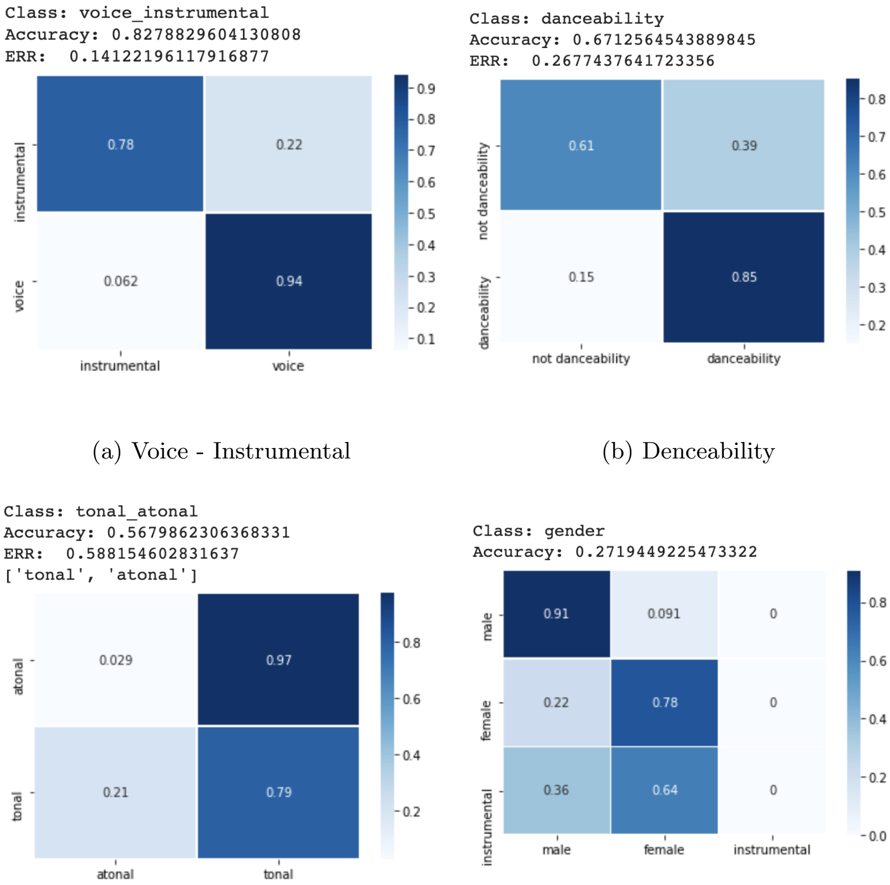
(d) Mood acoustic

(e) Mood sad

(f) Mood relaxed

(g) Mood party

Figure 1: Confusion matrices *MusicCNN* model *mood* task



```
Class: voice_instrumental
Accuracy: 0.8278829604130808
ERR:  0.14122196117916877
```

```
Class: danceability
Accuracy: 0.6712564543889845
ERR:  0.2677437641723356
```

(a) Voice - Instrumental

(b) Denceability

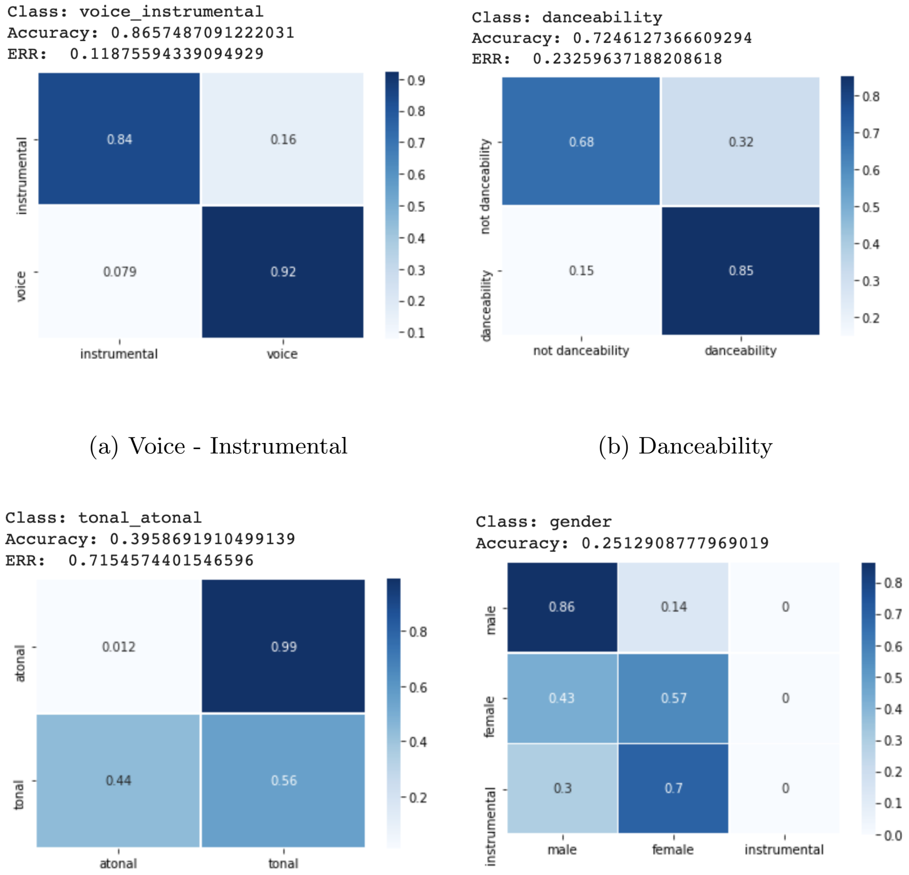
```
Class: tonal_atonal
Accuracy: 0.5679862306368331
ERR:  0.588154602831637
['tonal', 'atonal']
```

```
Class: gender
Accuracy: 0.2719449225473322
```

(c) Tonal - Atonal

(d) Gender

Figure 2: Confusion matrices *MusicCNN* model *miscellaneous* task



```
Class: voice_instrumental
Accuracy: 0.8657487091222031
ERR:  0.11875594339094929
```

```
Class: danceability
Accuracy: 0.7246127366609294
ERR:  0.23259637188208618
```

```
Class: tonal_atonal
Accuracy: 0.3958691910499139
ERR:  0.7154574401546596
```

```
Class: gender
Accuracy: 0.2512908777969019
```

(a) Voice - Instrumental

(b) Danceability

(c) Tonal - Atonal

(d) Gender

Figure 3: Confusion matrices *VGGish* model *miscellaneous* task

References

[1] https://mtg.github.io/essentia-labs/news/2020/01/16/tensorflow-models-released/.

[2] http://millionsongdataset.com/lastfm/.

[3] https://research.google.com/audioset/.

[4] https://github.com/MTG/mtg-jamendo-dataset.

[5] https://github.com/MTG/mtg-jamendo-annotator.

[6] Jordi Pons and Xavier Serra. musicnn: Pre-trained convolutional neural networks for music audio tagging. *arXiv preprint arXiv:1909.06654*, 2019.