Large Scale Music Data Analysis

Yiqun Liu
Audio And Music Processing Lab - Module 1
Universitat Pompeu Fabra

March 24, 2020

Abstract

In the music information retrieval task, when there is a new dataset or classification algorithm, the evaluation of its performance is important. In this assignment, we implement such a process with the MTG-Jamendo dataset and some Essentia auto-tagging models. We manually annotate 563 audio songs from the dataset and use pre-trained MusiCNN and VGGish models from Essentia to predict the annotations. For evaluation, we compute the accuracy of the model prediction, with our manual annotations being the ground truth. The results show some disagreements between the two approaches, which we think comes from the cognitive difference between annotators and the semantic gap between human description and computational representation.

1 Introduction

The development of streaming services of music industry brings new challenges for music information retrieval (MIR). There is much more amount of music data available for the MIR task. Therefore, the analysis of large scale music data is inevitable. In this assignment, we go through a typical workflow of large music data annotation, model prediction, and evaluation. We implement both manually annotation and model prediction on the same dataset that includes 563 audio tracks. Both the manual and model approaches concern 11 characteristics of each audio track. The 11 characteristics can be divided into two groups, mood and miscellaneous. After obtaining the results, we use our manual annotations as the ground truth to evaluate the model performance.

2 Processes

As described above, the assignment has three steps: manual annotation, model prediction, and evaluation.

2.1 Manual annotation

We use the MTG-Jamendo-Annotator to manually annotate 563 songs from chunk No.18 of the MTG-Jamendo dataset. This task is divided into 2 groups including 12 characteristics: mood-acoustic, mood-electronic, mood-aggressive, mood-relaxed, mood-happy, mood-sad,

mood-party, tonal-atonal, danceability, voice-instrumental, gender, and timbre, in which the first 7 characteristics belong to "mood" group and last 5 characteristics belong to "miscellaneous" group (the last one "timbre" is not considered in the following steps). The two groups are annotated separately, so for each song, we obtain 2 JSONs containing the mood annotations and miscellaneous annotations respectively.

2.2 Model prediction

We take advantage of two auto-tagging model architectures from Essentia to process the 563 songs and predict the 11 characteristics mentioned above (without "timbre"). The model architectures are MusiCNN [1] and VGGish [2, 3]. MusiCNN is a set of musically motivated convolutional neural networks (CNN) with vertical and horizontal convolutional filters. VGGish is an altered version of the CNN for computer vision, VGG. They are both pre-trained, MusiCNN on the Million Song Dataset (MSD) and VGGish on the AudioSet. We also dump the prediction into JSON format and obtain 4 JSONs for each song, 2 generated by MusiCNN and 2 generated by VGGish.

2.3 Evaluation

Now we have one group of manual annotations and two groups of model predictions. We use our manual annotations as the ground truth to evaluate the performance of both models. We calculate the accuracies and generate confusion matrices for all 11 tasks. We will describe the results in the next part.

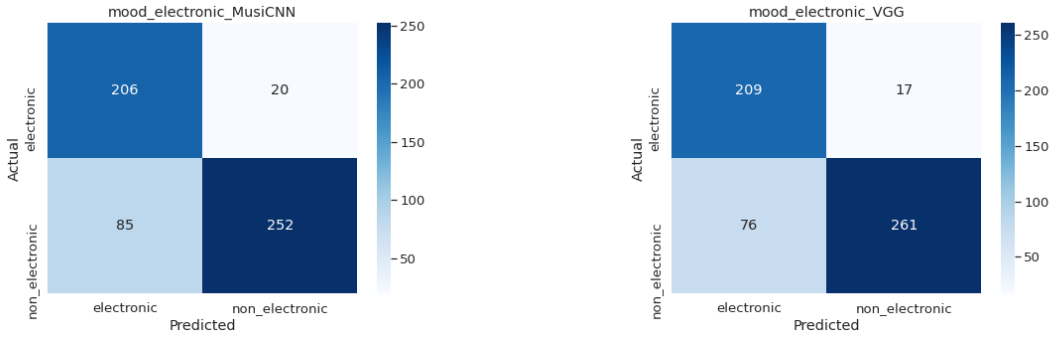
3 Results

By computing the accuracies of the two models for each task, we get the results as shown in Table 1.

Tasks	MusiCNN accuracy	VGGish accuracy
mood_acoustic	0.7211	0.8117
mood_electronic	0.8135	0.8348
mood_aggressive	0.9147	0.8774
mood-relaxed	0.3943	0.2948
mood_happy	0.7513	0.7655
mood_sad	0.3837	0.3659
mood_party	0.1883	0.1829
tonal_atonal	0.2114	0.4085
danceability	0.6448	0.6821
voice_instrumental	0.8455	0.8579
gender	0.9342	0.8421

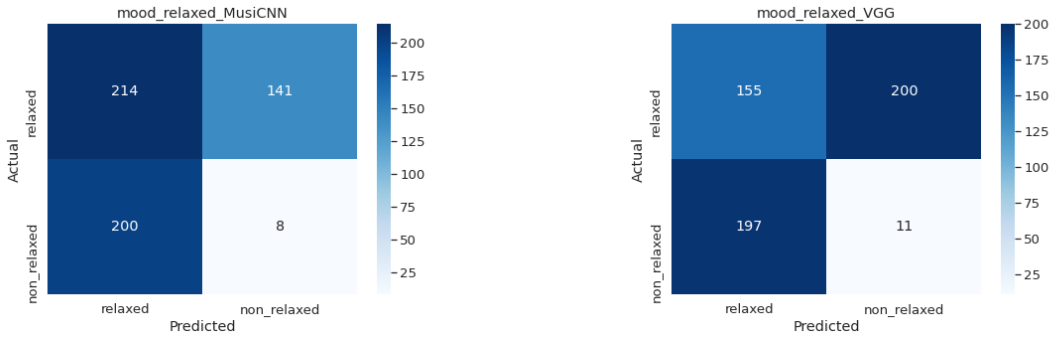
Table 1: Model prediction results

According to Table 1, both MusiCNN and VGGish get relatively higher accuracy on task electronic (Figure 1), aggressive, voice-instrumental, and gender. Besides, they are both biased against our manual annotation in task relaxed (Figure 2), sad, and party.



(a) MusiCNN (b) VGGish

Figure 1: Confusion matrices of task electronic



(a) MusiCNN (b) VGGish

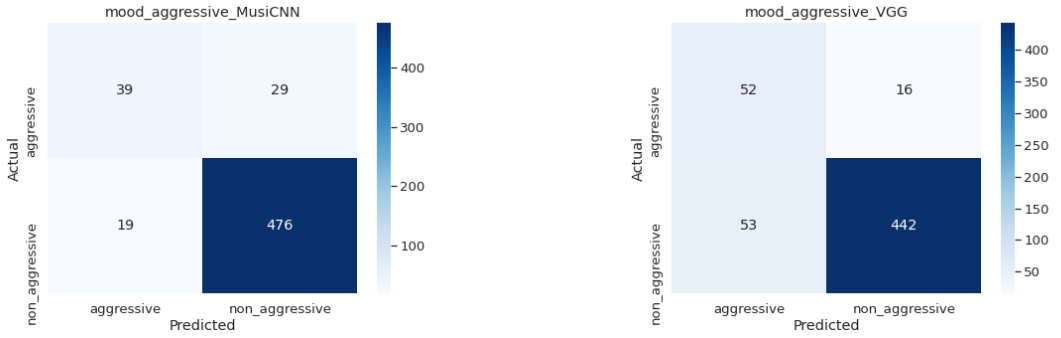
Figure 2: Confusion matrices of task relaxed

Moreover, when comparing the two models, we find they perform roughly similar, especially in task electronic (Figure 1), happy, and voice-instrumental. Although in some tasks they show obvious disagreement, for example, task relaxed (Figure 2), acoustic, and tonal-atonal.

4 Discussion

By observing the tasks, we think that the disagreements between manual annotation and model prediction result from the cognitive difference and semantic gap. On the one hand, it's impossible for every annotator to have the same idea when interpreting a piece of music, especially in terms of emotional level. In other words, the annotators of the dataset that is used for model training may tend to dance with the rhythmic electronic music, while we also consider swing music danceable in our annotation. On the other hand, such disagreements also happen in the computational representation of human description. Therefore, the model

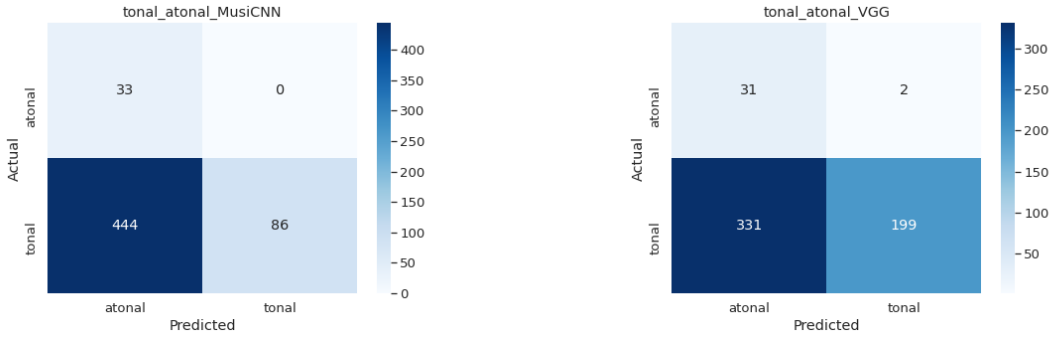
may not be able to fully capture the features of party music (we also find it hard to decide whether a song is suitable for a party or not during the annotation process).
In our results, the two models both give low accuracies in some emotional tasks like relaxed and sad, showing the cognitive difference between annotators. But in those intuitive tasks (e.g. voice-instrument, gender, acoustic, and electronic), people are easier to reach consensuses because of the distinctive timbres. Similarly, the models can also extract timbre information and obtain accurate predictions.
In addition, it is worth noticing that although task aggressive and happy are related to emotions, their accuracies are quite high. In fact, both manual and model approaches classify most songs as "not aggressive" (Figure 3) and "not happy". We think this is because people generally hold the same opinion on aggressiveness and happiness, one usually with extremely fast and heavy percussion, the other with moderately fast tempo and bright timbre. Such music is not common in our dataset.



(a) MusiCNN (b) VGGish

Figure 3: Confusion matrices of task aggressive

However, it is unexpected that the task tonal-atonal has a strong bias between manual annotation and model prediction. According to the confusion matrices (Figure 3), we classified most songs as "tonal". But the models' results are exactly the opposite. We still don't have a proper explanation for such a phenomenon.



(a) MusiCNN (b) VGGish

Figure 4: Confusion matrices of task tonal-atonal

5 Conclusion

In this assignment, we annotated 563 songs through manual approach and model prediction. The results of the two methods are used to evaluate the performance of the models. We conclude that the models are not able to achieve high accuracy in every task due to the cognitive difference and semantic gap. This assignment gives us a general view of large scale music data analysis and helps us get familiar with the Essentia auto-tagging models.

References

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

[2] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017.

[3] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.