Recognizing Song Mood and Theme: Clustering-based Ensembles

Maximilian Mayerl[†], Michael Vötter[†], Andreas Peintner[†], Günther Specht, Eva Zangerle
Universität Innsbruck
{firstname.lastname}@uibk.ac.at

ABSTRACT

The *Emotions and Themes in Music* task at MediaEval 2021 has the goal of correctly assigning mood and theme labels to pieces of music. In this paper, we describe our (team UIBK-DBIS) approach solving this task. Last year, we devised an ensemble-based method where we trained multiple neural network models on different partitions of the target labels. This year, we build upon this approach and attempt to automatically generate label partitions based on clustering techniques. This approach achieves a PR-AUC of 0.109 on the test set for the task, which is slightly better than the baseline.

1 INTRODUCTION

The goal of the *Emotions and Themes in Music* as MediaEval 2021 is to detect the moods and themes present in a song based on descriptors of the song's audio properties. In total, there are 56 different mood and theme labels that can be assigned to a song, and each song can have more than one label. The dataset used for this task was created by Bogdanov et al. [2] and is publicly available. Further details about the task itself can be found in the overview paper [5]. Our approach to this year's edition of the task is based on the approach we submitted last year [6]. The basic idea is to train multiple models for distinct subsets of the target labels and then combine the results. Last year, we formed the label subsets by simply splitting the set of labels into equally-sized subsets as well as by manually dividing the labels into *mood* and *theme* labels. This year, we propose to use clustering techniques to generate better label subsets, forming clusters of either similar or dissimilar labels. For clustering similar labels, we use the popular k-means algorithm, and for clustering dissimilar labels, we propose a simple algorithm that can generate such clusters. The code for our implementation is available on GitHub[1].

2 APPROACH

For MediaEval 2020 [6], we proposed an ensemble approach using multiple neural network models trained for handling subsets of the target labels. Our results for this approach showed that using such ensemble models can improve the F_1 score over using a neural network model trained for handling all labels. Building on those results, we make the following changes and additions for this year's edition of the task: (i) Instead of a CRNN architecture, which we used last year, we use a VGG model (taken from the baseline provided with the task dataset [2]) and with a ResNet-18 [3]. (ii) Instead of partitioning the target labels linearly or manually, we employ clustering techniques (see Section 2.2) to find partitions of similar or dissimilar labels.

As every model in the ensemble handles a disjoint subset of target labels, the final prediction results are obtained by concatenation and reordering the label predictions of all models.

2.1 Data Preprocessing

Since the neural network models we use in our approach require mel-spectrograms of equal length for all songs, we extract a spectrogram of size 1366 from the center of each song. This follows the approach by Mayerl et al. [4] for the 2019 edition of the task.

2.2 Clustering

To generate partitions of target labels for our ensemble, we first map labels into a space for clustering, such that each label is represented by one vector. To find the vector for a given label, we take all songs in the training set to which that label is assigned and compute the centroid of the feature vectors for those songs. For this step, we used 22 high-level features extracted with Essentia [1] instead of mel-spectrograms. We then computed label partitions by using two different clustering techniques on the resulting vector space.

To find partitions such that each partition contains similar labels, we used the well-known k-means algorithm. As k-means requires manually setting the number of desired clusters, we used the popular elbow method to determine the best number of clusters, which we found to be four. To find partitions such that each partition contains dissimilar labels, we propose a simple clustering algorithm, which we call dk-means (*dissimilar k-means*). This algorithm is a variation of the k-means algorithm and works as follows:

(1) Randomly chose k points (in our case, corresponding to labels) as seeds. This gives us k clusters, each containing one point.
(2) For each cluster
    (a) Compute the centroid of the points in the cluster.
    (b) Among all the points not yet assigned to a cluster, find the point that has the *highest* euclidean distance to this centroid. Add that point to the cluster.
(3) Repeat (2) until all points are assigned.

A visualization of the clusters produced by these methods is given in Figure 1. For this visualization, the centroids corresponding to each label have been projected to a 2-dimensional space using principal component analysis.

[†] authors contributed equally to this work
[1] https://github.com/dbis-uibk/mediaeval2021

MediaEval'21, December 13-15 2021, Online

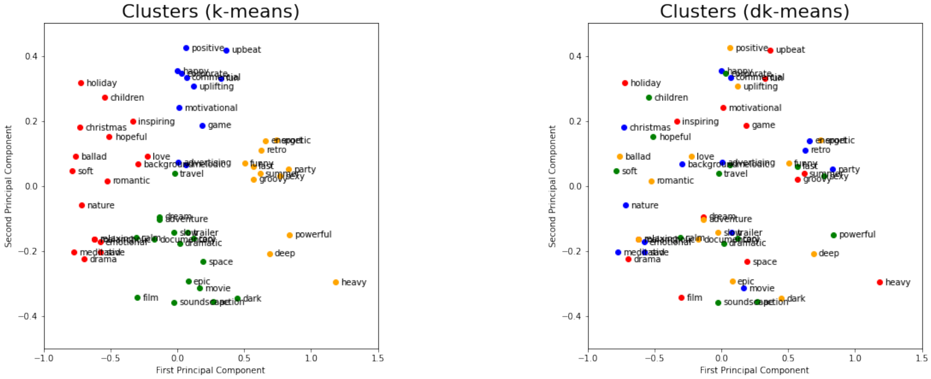

Figure 1: A 2D PCA visualization of the clusters produced by the two clustering methods used. Each color denotes a cluster.

2.3 Submissions

Based on those clustering approaches, we submitted five runs:

- Run #1: k-means clustering and VGG models
- Run #2: k-means clustering and ResNet-18 models
- Run #3: dk-means clustering and VGG models
- Run #4: dk-means clustering and ResNet-18 models
- Run #5: linear label splits and ResNet-18 models

In all the runs, the models were trained for 100 epochs. For ResNet-18 we additionally utilized early stopping.

3 RESULTS AND ANALYSIS

The evaluation results are given in Table 1. The table also includes results for two baseline approaches as well as a run using linear label splits and VGG models, which is included for comparison. The evaluation was done using four evaluation metrics, as defined by the task. The baseline approaches consist of a single model trained on all target labels, i.e. do not use an ensemble. Comparing the results of the submitted approaches to the baselines shows that the submitted approaches generally perform worse than or equal to the baseline. Looking at the results for the approaches using VGG models, we observe a clear performance improvement when using k-means clustering compared to linear splits. Both the ROC-AUC as well as the PR-AUC increase, from 0.684 to 0.705 and from 0.097 to 0.109 respectively, while both F_1 scores stay the same. The same is not true when using dk-means clustering, were the performance remains almost the same compared to linear splits across all four metrics. From this, we conclude that partitioning target labels such that similar labels are handled by the same model in the ensemble is beneficial and results in better performance when using VGG models, at least for the given dataset. The approaches using ResNet-18 show a different behavior. Here, linear splits clearly outperform both splits using k-means and dk-means clustering. This indicates that, with the given dataset, partitioning target labels based on similarity or dissimilarity does not improve performance. Lastly, we can compare approaches using k-means clustering with approaches using dk-means clustering. Here, we can observe a decrease in performance when using dk-means compared to k-means, for both VGG and ResNet-18. This implies that applying models to clusters of similar labels results in better performance than doing the same with dissimilar labels.

Table 1: Evaluation results for our submitted runs and baselines. Best results are in bold.

Approach	Run	ROC-AUC	PR-AUC	F_1 (micro)	F_1 (macro)
resnet18_all	-	**0.715**	0.108	**0.107**	0.110
vgg_all	-	0.707	0.101	**0.107**	**0.112**
resnet_linear	5	0.700	0.092	0.106	0.106
resnet_kmeans	2	0.692	0.091	0.103	0.104
resnet_dkmeans	4	0.681	0.080	0.097	0.098
vgg_linear	-	0.684	0.097	0.104	0.110
vgg_kmeans	1	0.705	**0.109**	0.104	0.110
vgg_dkmeans	3	0.683	0.098	0.103	0.110

4 DISCUSSION AND OUTLOOK

In this paper, we presented our approach for the *Emotions and Themes in Music* task at MediaEval 2021. While our approach only slightly outperformed the baselines, we were still able to show potential benefits in building ensemble models based on partitions of target labels using clustering techniques. For models using VGG-based classifiers, we observed an increase in performance when determining label partitions using k-means clustering.

For future work, one interesting avenue would be to combine the various approaches we have developed for this task over the past few years. In 2019, we introduced a random sampling approach to augment the provided dataset and generate more representative samples for each song. Last year, we further built on this by generating a more balanced dataset by drawing more samples for target labels that are underrepresented. As we did not employ either of these techniques for this year's submissions, it would be interesting to see what results could be accomplished by incorporating them into the new, clustering-based approach. Comparing the performance of our ensemble with the baselines implies that training on disjoint subsets of labels leads to a decrease in performance. Hence, it would be interesting to see if we can increase the performance by using overlapping label sets in our ensemble.

REFERENCES

[1] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Herrera Boyer, Oscar Mayor, Gerard Roma Trepat, Justin Salamon, José Ricardo Zapata González, Xavier Serra, and others. 2013. Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR).*

[2] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. 2019. The MTG-Jamendo Dataset for Automatic Music Tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019).* Long Beach, CA, United States. http://hdl.handle.net/10230/42015

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* p. 770–778.

[4] Maximilian Mayerl, Michael Vötter, Hsiao-Tzu Hung, Bo-Yu Chen, Yi-Hsuan Yang, and Eva Zangerle. Recognizing Song Mood and Theme Using Convolutional Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2019 Workshop, Sophia Antipolis, France, 27-30 October 2019.* CEUR-WS.org. http://ceur-ws.org/Vol-2670

[5] Philip Tovstogan, Dmitry Bogdanov, and Alastair Porter. MediaEval 2021: Emotion and Theme Recognition in Music Using Jamendo. In *Working Notes Proceedings of the MediaEval 2021 Workshop, Online, 13-15 December 2021.* CEUR-WS.org.

[6] Michael Vötter, Maximilian Mayerl, Günther Specht, and Eva Zangerle. Recognizing Song Mood and Theme: Leveraging Ensembles of Tag Groups. In *Working Notes Proceedings of the MediaEval 2020 Workshop, Online, 14-15 December 2020.* CEUR-WS.org. http://ceur-ws.org/Vol-2882