MediaEval 2018 AcousticBrainz Genre Task: A baseline combining deep feature embeddings across datasets

Sergio Oramas[1], Dmitry Bogdanov[2], Alastair Porter[2]
[1]Pandora Media Inc., US
[2]Universitat Pompeu Fabra, Spain
soramas@pandora.com,dmitry.bogdanov@upf.edu,alastair.porter@upf.edu

ABSTRACT

In this paper we present a baseline approach for the MediaEval 2018 AcousticBrainz Genre Task that takes advantage of stacking multiple feature embeddings learned on individual genre datasets by simple deep learning architectures. Although we employ basic neural networks, the combination of their deep feature embeddings provides a significant gain in performance compared to each individual network.

1 INTRODUCTION

This paper describes our baseline submission to the MediaEval 2018 AcousticBrainz Genre Task [1]. The goal of the task is to automatically classify music tracks by genres based on pre-computed audio content features provided by the organizers. Four different genre datasets coming from different annotation sources with different genre taxonomies are used in the challenge. For each dataset, training, validation, and testing splits are provided. This allows to build and evaluate classifier models for each genre dataset independently (Subtask 1) as well as explore combinations of genre sources in order to boost performance of the models (Subtask 2).

For this baseline, we decided to focus on demonstration of possibilities of merging different genre ground truth sources using a simple deep learning architecture. To this end, we explore how stacking deep feature embeddings obtained on different datasets can benefit genre recognition systems.

2 RELATED WORK

Submission to the previous edition of the task have explored late fusion of predictions made by classifier models trained for each genre source individually. In order to predict genres following a taxonomy of a target source, the proposed solutions applied genre mapping between taxonomies, either by computing genre co-occurrences on the intersection of all four training genre datasets [4], or by textual string matching [6].

In our baseline, we propose an alternative early fusion approach, similar to the one proposed in [7] for multimodal genre classification. The approach incorporates knowledge across datasets by stacking deep feature embeddings learned on each dataset individually and using those as an input to predict genres for each test dataset.

MediaEval'18, 29-31 October 2018, Sophia Antipolis, France

3 APPROACH

3.1 Input features

We use all available features extracted from music audio recordings using *Essentia* [2] and provided for the challenge. As a preprocessing step, we apply one-hot encoding for a few categorical features related to tonality (`key_key`, `key_scale`, `chords_key`, and `chords_scale`) and standardize all features (zero mean, unit variance). In total, this amounts to 2669 input features.

3.2 Neural network architecture

A simple feedforward network is used to predict the probabilities of each genre given a track. The network consists of an input layer of 2669 units (the size of the feature vector for an input recording), followed by a hidden dense layer of 256 units with ReLu activation, and the output layer where the number of units coincide with the number of genres to be predicted in each dataset. Dropout of 0.5 is applied after the input and the hidden layer. As the targeted genre classification task is multi-label, the output layer uses sigmoid activations and is evaluated with a binary cross-entropy loss.

Mini-batches of 32 items are randomly sampled from the training data to compute the gradient, and the Adam [3] optimizer is used to train the models, with the default suggested learning parameters. The networks are trained for a maximum of 100 epochs with early stopping. Once trained, we extract the 256-dimensional vectors from the hidden layer for the training, validation, and test sets.

The model architecture is used to train a multi-label genre classifier on each of the four datasets. The models are trained on 80% of the training set and validated after each epoch using the other 20% using the split script with release-group filtering provided by the organizers. Predictions are computed for the validation and test sets.

3.3 Embedding fusion approach

Following the described methodology, one model per dataset is trained and these models serve for predictions in Subtask 1. Then, the given models are used as feature extractors. All four models share the same input format, so input feature vectors from one dataset can be used as input to a model trained on other dataset. Thus, for each model we feed all tracks from the training, validation and test sets of each dataset, and obtain the activations of the hidden layer as a 256-dimensional feature embedding. Therefore, for each track in each dataset we obtain four different feature embeddings, coming from each of the four previously trained models.

Given the four feature embeddings of each track, we apply $l2$-norm to each of them and then stack them together into a single 1024-dimensional feature vector. Following this process, we obtain

Table 1: ROC AUC on validation datasets

	AllMusic	Discogs	Lastfm	Tagtraum
Subtask 1	0.6476	0.7592	0.8276	0.8017
Subtask 2	0.8122	0.8863	0.9064	0.8874

new feature vectors for every track in the training, validation and test sets of each dataset. Then, we use these feature vectors as input of a simple network where the input layer is directly connected to the output layer. Dropout of 0.5 is applied after the input layer. The output layer is exactly the same as in the network described in above, where sigmoid activation and binary cross-entropy loss are applied. The new network is trained following the same methodology described before, with Adam as the optimizer and mini-batches of 32 items randomly sampled. The network is trained on 80% of the training data and validated on the other 20%. Following this approach, we train a network per dataset, and obtain the genre probability predictions of the validation and test sets for Subtask 2.

3.4 Predictions thresholding

The predictions made by each model contain continuous values, while the task requires binary prediction of genre labels. We therefore apply a plug-in rule approach thresholding the prediction values in order to maximize the evaluation metrics. We decided to maximize the macro F-score, and applied thresholds individual for each genre label that we estimated on the validation data [5].

4 RESULTS AND ANALYSIS

We evaluated a single run for both Subtask 1 and 2. Table 1 presents the ROC AUC metric on the validation sets. Table 2 presents the final results after applying thresholding on the test datasets. As the general pattern, we can clearly see the benefit of models based on embedding fusion approach compared to the models trained individually on each dataset. While the individual models (Subtask 1) are hardly usable compared to the random and popularity baselines, the combined models got a significant improvement in performance, being competitive with last years' second ranked submission [6].

In our experiments, we focused on optimizing macro F-score, however choosing this metric for threshold optimization can have a negative effect on micro-averaged metrics. In the case of infrequent subgenre labels and an uninformative classifier, an optimal, but undesirable strategy may involve predicting those labels always [5]. Indeed, this was the case for the individual models, but the hybrid models did not have this issue.

5 DISCUSSION AND OUTLOOK

In our baseline approach we focused on Subtask 2 and demonstrated the advantage of fusing feature embeddings learned on individual genre datasets on the example of a simple feedforward network architecture. We may expect further improvements in performance by means of a more sophisticated network architecture (for example [4]). The code of the baseline is available online.[1]

[1]https://github.com/MTG/acousticbrainz-mediaeval-baselines

Table 2: Precision, recall and F-scores on test datasets

Average per		Dataset			
		AllMusic	Discogs	Lastfm	Tagtraum
Subtask 1 (individual models)					
Recording (all labels)	P	0.0147	0.0591	0.0976	0.0992
	R	0.5753	0.5263	0.4512	0.5017
	F	0.0280	0.1035	0.1506	0.1623
Recording (genres)	P	0.2786	0.6305	0.3974	0.2991
	R	0.6960	0.7289	0.5966	0.6630
	F	0.3399	0.6382	0.4455	0.4040
Recording (subgenres)	P	0.0114	0.0256	0.0518	0.0636
	R	0.4861	0.3497	0.3295	0.4164
	F	0.0219	0.0467	0.0856	0.1083
Label (all labels)	P	0.0225	0.0744	0.0732	0.0951
	R	0.4943	0.2588	0.2330	0.2412
	F	0.0324	0.0935	0.0947	0.1141
Label (genres)	P	0.1658	0.3733	0.2321	0.2551
	R	0.4729	0.4229	0.3484	0.3573
	F	0.1938	0.3801	0.2546	0.2676
Label (subgenres)	P	0.0184	0.0595	0.0572	0.0764
	R	0.4949	0.2506	0.2213	0.2276
	F	0.0278	0.0792	0.0786	0.0962
Subtask 2 (fusion models)					
Recording (all labels)	P	0.1340	0.2775	0.2718	0.2972
	R	0.4809	0.5432	0.4762	0.5127
	F	0.1880	0.3320	0.3066	0.3451
Recording (genres)	P	0.5689	0.6877	0.5407	0.6061
	R	0.6905	0.7473	0.6335	0.6885
	F	0.5880	0.6845	0.5602	0.6243
Recording (subgenres)	P	0.0946	0.1472	0.1570	0.2022
	R	0.3251	0.3703	0.3368	0.4148
	F	0.1343	0.1892	0.1911	0.2465
Label (all labels)	P	0.0614	0.1087	0.1108	0.1235
	R	0.1640	0.2226	0.2168	0.2324
	F	0.0736	0.1247	0.1314	0.1400
Label (genres)	P	0.2907	0.4404	0.3077	0.2878
	R	0.3713	0.4713	0.3735	0.3565
	F	0.3080	0.4393	0.3246	0.3053
Label (subgenres)	P	0.0550	0.0922	0.0909	0.1043
	R	0.1582	0.2102	0.2009	0.2179
	F	0.0670	0.1089	0.1119	0.1206

ACKNOWLEDGMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No 688382 (AudioCommons) and 770376-2 (TROMPA), as well as the Ministry of Economy and Competitiveness of the Spanish Government (Reference: TIN2015-69935-P).

REFERENCES

[1] Dmitry Bogdanov, Alastair Porter, Julián Urbano, and Hendrik Schreiber. 2018. The MediaEval 2018 AcousticBrainz Genre Task: Content-based Music Genre Recognition from Multiple Sources. In *MediaEval 2018 Workshop*. Sophia Antipolis, France.

[2] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J.R. Zapata, and X. Serra. 2013. Essentia: An Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval (ISMIR'13) Conference*. Curitiba, Brazil, 493–498.

[3] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[4] Khaled Koutini, Alina Imenina, Matthias Dorfer, Alexander Rudolf Gruber, and Markus Schedl. 2017. MediaEval 2017 AcousticBrainz Genre Task: Multilayer Perceptron Approach. In *MediaEval 2017 Workshop*. Dublin, Ireland.

[5] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal thresholding of classifiers to maximize F1 measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 225–239.

[6] Benjamin Murauer, Maximilian Mayerl, Michael Tschuggnall, Eva Zangerle, Martin Pichl, and GÄijnther Specht. 2017. Hierarchical Multilabel Classification and Voting for Genre Classification. In *MediaEval 2017 Workshop*. Dublin, Ireland.

[7] Sergio Oramas, Francesco Barbieri, Oriol Nieto, and Xavier Serra. 2018. Multimodal Deep Learning for Music Genre Classification. *Transactions of the International Society for Music Information Retrieval* 1, 1 (2018).