AUDIO-BASED AUTO-TAGGING WITH CONTEXTUAL TAGS FOR MUSIC

Karim M Ibrahim, Jimena Royo-Letelier, Elena V. Epure, Geoffroy Peeters, Gael Richard

AUDIO-BASED AUTO-TAGGING WITH CONTEXTUAL TAGS FOR MUSIC

Karim M. Ibrahim[⋆†] *Jimena Royo-Letelier*[†] *Elena V. Epure*[†]
Geoffroy Peeters[⋆] *Gaël Richard*[⋆]

[⋆] LTCI, Télécom Paris, Institut Polytechnique de Paris
[†]Deezer Research & Development
karim.ibrahim@telecom-paris.fr

ABSTRACT

Music listening context such as location or activity has been shown to greatly influence the users' musical tastes. In this work, we study the relationship between user context and audio content in order to enable context-aware music recommendation agnostic to user data. For that, we propose a semi-automatic procedure to collect track sets which leverages playlist titles as a proxy for context labelling. Using this, we create and release a dataset of ∼50k tracks labelled with 15 different contexts. Then, we present benchmark classification results on the created dataset using an audio auto-tagging model. As the training and evaluation of these models are impacted by missing negative labels due to incomplete annotations, we propose a sample-level weighted cross entropy loss to account for the confidence in missing labels and show improved context prediction results.

Index Terms— music auto-tagging, user context, dataset collection, multi-label classification, missing labels.

1. INTRODUCTION

Recommender systems have gained much attention from the research community [1] and has become ubiquitous in online services handling very large catalogs and serving millions of customers. Generic methods such as Collaborative Filtering or Matrix Factorization have been successfully applied to various types of items (books, movies, etc.) [1]. These methods often build a single model per user, e.g. a low-dimensional projection. Implicitly, this amounts to modeling user tastes as rigid, context-independent concepts. Recent works have challenged this static model [2, 3], and proposed to incorporate factors such as time, location or activity in the model and recommendation process. Indeed, past studies have shown [4, 5] that music listening context heavily influences users' tastes. Intuitively, one adapts the musical choices to the immediate social environment, mood or activity.

Context-aware recommender systems, along with the traditional ones, are common in many services other than music, e.g. in online shopping [6] or movie streaming [7]. However, it is specifically more challenging in the case of music streaming due to the dynamic nature of music listening. Music tracks often have a duration of few minutes while users listen to music for hours in a day. The user context, e.g. activity or location, often changes, leading to a need to anticipate the changes in the user's listening preference too. Since accessing the user context is often not feasible due to privacy issues, allowing users to select a specific context and get recommended related tracks could be an alternative. However, for this, we need to grasp to what extent it is possible to infer user context from audio content only.

Few studies have already addressed the annotation of music datasets with user context tags [8, 9]. However, there has been no standard procedure on how to find context tags relevant to music and how to employ them. Previous studies have focused on a number of contexts that were defined arbitrarily by the authors [4, 9]. Additionally, even scarcer research has investigated the relationship between audio content and user contexts, and the feasibility to automatically predict context from a music track's audio content [9]. Such study is important for automatically generating context-aware playlists [10] or for facilitating music discovery by context tags.

In this paper, we propose the following contributions: 1) a procedure to label music tracks with context tags using playlists titles; 2) a dataset of ∼50k tracks labelled with the 15 most common context tags, which we make available for future research; 3) benchmark results of a trained auto-tagging model to predict context tags using audio content; 4) a strategy to account for the confidence in the sample tags to overcome the problem of missing negative labels, that we observed to hinder the training of the auto-tagging models [11].

2. PREVIOUS WORK

Music can be listened to in various situations and it was shown that music preferences and user context are related [5]. For example, North et al. [4] studied the influence of 17 different listening situations on music preferences and showed that there is a link between them. Gillhofer et al. [12] investigated the affect of user context on the genre and mood of tracks selected by the user. These studies show that there are multiple factors affecting the user's choice of music in a given moment and justify the need for considering user context in music recommendation.

Previous studies on context-aware recommendation systems used a variety of contextual information including location, activity, time, weather, or sensor data collected from the user's phone. Cheng et al. [13] developed a location-aware recommendation system for music. Due to difficulties in accessing the user's location, the approach relied on detecting the location using audio content. However, the trained model had low classification accuracy which would lead to noisy recommendations. Wang et al. developed a recommendation system for different daily activities [9]. The approach was

This project has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

constrained by the limited amount of available data, specifically music tracks that are labelled with certain activities. The study required part of the data to be hand-labelled using human participants, which is time consuming and not scalable. In another study on music for activities [8], Yadati et al. relied on an automated procedure to label the data using Youtube search queries on "music for X", where X is an activity. However, while the procedure is suitable for studying the audio-context relationship, it is quite subjective to the creator of the compiled video and can be noisy. In all past works mentioned so far, there is no common definition of context tags, labelled dataset, or uniform, reliable procedure to annotate new data.

A challenge in collecting labels for context is the so called missing negative labels. Missing labels renders the automatic inference of context tags from music audio content quite challenging. Auto-tagging music tracks with contexts is a multi-label classification problem where a track can be labelled with multiple contexts simultaneously. For training and testing this model, we can select positive examples for a certain context by sampling context-related tracks. However, we do not have access to negative label examples because the absence of a tag could also mean that the annotation was incomplete. The problem of Multi-Label classification with Missing Labels (MLML) is a common challenge, which received attention in past research [14, 15]. Most previous approaches relied on exploiting the correlation between labels to predict the missing negative labels [14, 15, 16]. However, the state-of-the-art approaches in MLML [17, 18] are not simply integratable in cases where a pre-defined model is used. They either rely on jointly learning the correlations between the labels along with the model parameters, require prior extraction of manually engineered features for the task [17], or assume the location of the missing labels is known but the value is missing [18]. However, in cases where a pre-defined model is used, there is no straightforward approach to tackle unknown missing labels.

Consequently, to handle the limitations of the previous works with regard to our problem (the automatic context tagging of music track's audio), we set multiple goals. First, we define a procedure to extract popular contexts in music consumption. Second, we semi-automatically label a dataset with the defined context tags ensuring its label balance. Third, we study the relationship between the audio content and the user context. Specifically, given the audio content of a track, we predict all its suitable contexts using an audio auto-tagging model. Also, we propose an easily-implementable solution for the missing negative labels that can be integrated with predefined models by weighting the loss function. Previous literature [19, 20] has already looked into weighting the loss function. However, to our knowledge, none of these approaches apply confidence-based weight per sample for each of the positive and negative labels independently to overcome the missing labels problem.

3. DATASET CREATION

To infer the contextual listening of different tracks, we rely on playlist titles, as in Pichl et al. [21]. Users often create playlists that are intended for specific contexts and tend to use suggestive titles to reflect these contexts. However, while Pichl et al. tried to automatically extract context clusters from playlists titles, we found that this approach leads to noisy clusters that are not clearly contextual and are often related to genres or other popular, non-context words used in playlists. Instead, we filter the playlists using context keywords.

3.1. Track labelling with context tags using playlist titles

We present the procedure to automatically select the most representative keywords related to context from playlist titles. We started with a set of context-related keywords collected from the literature we described in Section 2 [4, 9]. Then, we added keywords that are semantically similar. Ninety six keywords were categorized in one of four categories, location, activity, time, and mood, similar to the categorization in [22]. To construct the first version of the dataset, we selected the 15 most frequent keywords found in the playlist titles of the Deezer[1] catalogue out of the collected context keywords. The final keywords are *car, chill, club, dance, gym, happy, night, party, relax, running, sad, sleep, summer, work, workout*.

The following step was to filter the playlists to include only the contextual ones. We first collected all the public playlists in the Deezer catalogue that included any of the 15 keywords. Afterwards, we removed all playlists that contained more than 100 tracks, since playlists with many tracks tended to be less focused on a specific context and rather noisy. We also removed all playlists where a single artist or album made up more than 25% of all the tracks in the playlist, to ensure that the playlist was not intended for a specific artist. Finally, we tagged tracks that appeared in more than 3 playlists containing the same context keyword with that context tag. For example, a track that appears in 5 playlists containing the word "dance", 3 playlists containing the word "party", and 1 playlist containing the word "chill", would be labelled as "dance" and "party", but not "chill". This was to increase the confidence that the selected track belonged to the target context.

3.2. Dataset balancing

After applying the previous filtering to the catalogue of Deezer, we retrieved 612k playlists that belonged to one or more of the 15 selected contexts. The playlists contained 198k unique tracks. However, the dataset was highly imbalanced due to the popularity of some contexts compared to others. Hence, we balanced the dataset to keep a nearly equal number of tracks within each context. Since we work in a multi-label setting, i.e. one track can belong to multiple contexts at the same time, it is difficult to have exactly the same number of samples for each label. We applied an iterative approach to add samples to incomplete classes with a limit of 20000 tracks, which was the number of tracks in the least represented context class. The number of tracks dropped to 49929 unique tracks. The balanced dataset contains on average 24k positive samples per label and 7 labels per sample. We distribute the collected dataset to the research community[2], which is composed of the track ID in the Deezer catalogue and the 15 contextual labels. The audio content for each track is available as a 30 seconds snippet through the Deezer API using the track ID.

3.3. Analysis of context co-occurrences

The co-occurrences of context tags enable us to learn about the relationships between contexts. In Figure 1, we show the number of tracks co-labelled with each pair of contexts. We observe some interesting patterns in these co-occurrences. For example, we find that the three contexts "relax", "sad", and "sleep" co-occur more

[1] Deezer is an online music streaming service: https://www.deezer.com/

[2] https://doi.org/10.5281/zenodo.3648287

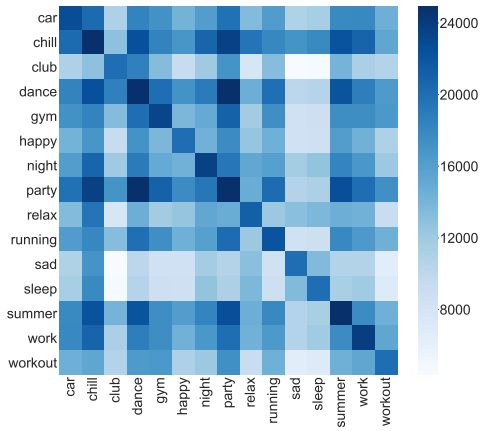

Fig. 1: Tracks co-occurrences between contexts

often together than with other contexts. This matches our expectation about the music style of the tracks related to these contexts to be rather calm and soothing. We also find that contexts such as "club", "dance", and "party" often co-occur together, having most likely associated energetic tracks. We also observe that "chill" often co-occurs with all the other contexts. This indicates that certain contexts are user-specific and would require additional data about users and music listening cases, apart from audio, for being inferred.

4. MULTI-CONTEXT AUDIO AUTO-TAGGING MODEL

4.1. Baseline

Our goals are to predict contexts for a track given its audio content and to assess to what extent this is possible. There has been a number of approaches proposed to auto-tag tracks using audio content. The most recent, best-performing approaches rely on Convolutional Neural Networks (CNNs) applied to the melspectrograms of the input audio [23, 24]. We selected one of the previously proposed and commonly used models by Choi [23], which is such a multi-layer convolutional neural network.

We trained the network with an input size of 646 frames x 96 mel bands, representing 30 seconds from each track cropped after 30 seconds from the start of the track to match the Deezer preview samples. The output corresponds to the 15 context tags. We applied a batch normalization on the input melspectrograms followed by 4 pairs of convolutional and max pooling layers. Each convolutional layer has a fixed filter size (3x3) and (32,64,128,256) filters respectively followed by a ReLu activation function. We used max pooling filter of size (2x2). We pass the flattened output of the last CNN layer to a fully connected layer with 256 hidden nodes with ReLu activation function and apply a dropout with 0.3 ratio for regularization. Finally, we pass the output to the final layer of 15 output nodes and a Sigmoid activation function. Initially we used binary cross entropy as a loss function optimized with Adadelta and a learning rate initialized to 0.1 with an exponential decay every 1000 iterations. We stopped the training after 10 epochs of no improvement on the validation set and retrieved the model with the best validation loss. We selected a split of 65% training, 10% validation, and 25% testing. We applied an iterative sampling scheme to ensure that there is no overlap of artists or albums between the splits, while having same proportional representation of each context tag [25].

Table 1: Results of the CNN model on our context-annotated dataset.

	HL*↓	AUC↑	Recall↑	Precision↑	f1↑	TN rate*↑
Car	0.39	0.65	0.58	0.58	0.58	0.63
Chill	0.27	0.71	0.93	0.75	0.83	0.29
Club	0.24	0.84	0.64	0.74	0.68	0.85
Dance	0.26	0.8	0.83	0.79	0.81	0.58
Gym	0.34	0.72	0.74	0.62	0.67	0.6
Happy	0.34	0.7	0.46	0.61	0.53	0.8
Night	0.44	0.58	0.54	0.54	0.54	0.58
Party	0.26	0.77	0.86	0.76	0.81	0.53
Relax	0.31	0.74	0.66	0.63	0.65	0.7
Running	0.38	0.68	0.64	0.58	0.61	0.61
Sad	0.22	0.85	0.74	0.71	0.73	0.8
Sleep	0.23	0.84	0.73	0.71	0.72	0.8
Summer	0.37	0.67	0.74	0.63	0.68	0.51
Work	0.41	0.62	0.61	0.57	0.59	0.57
Workout	0.3	0.77	0.62	0.63	0.62	0.75
Macro average	0.32	0.73	0.66	0.69	0.67	0.64

* HL : Hamming Loss. TN rate: True negative rate

The initial results showed the model can predict certain contexts fairly well, while others are harder to predict. Table 1 gives the performance of the model on the different contexts with standard multi-label classification evaluation metrics [26]. We find that certain contexts such as "club", "party", "sad", and "sleep" are easier to predict, while contexts such as "car", "work", and "night" are harder to predict. These results confirm the intuition that certain contexts could be more related to the audio characteristics and hence could be inferred from it, such as energetic dance music for "party" and calming soothing music for "sleep". However, for other contexts, the audio does not appear sufficient and the music style which people tend to listen to in a car or at work seems to widely vary. These contexts would potentially need additional information about the user in order to be predicted correctly in a personalized manner.

One drawback of this method is that we do not have explicit negative samples for each label. Hence, it is challenging to fairly evaluate and train the model with missing negative labels. In this work, we mainly focus on the recall because we are confident in the positive labels and would prefer to correctly predict all of them. However, since a classifier that predicts all labels for any given track would give perfect recall, it is important to ensure a balance with the true negative rate and the precision as well. As the missing negative labels are still used in training, they would lead to falsely train the model on false negatives. To counteract this, we propose to modify the loss function as presented in the next section.

4.2. Sample-level weighted cross entropy

We propose to modify the binary cross entropy loss to account for the confidence in the missing labels. This can be done by adding weighting factors to our loss function. We apply confidence-based weight per sample for each of the positive and negative labels independently. We hypothesise that using these weights can improve our model performance in predicting the correct label by giving less weight to samples with low confidence in their label.

Formally, let $\mathbb{X} = \mathbb{R}^d$ denote the d-dimensional space for the instances, $\mathbb{Y} = \{0, 1\}^m$ denote the label space marking the absence or presence of each of the m context classes for each instance. The

task of multi-label classification is to estimate a classifier $f : \mathbb{X} \mapsto \mathbb{Y}$ using the labelled dataset $D = \{(\mathbf{x_i}, \mathbf{y_i}) | 1 \leq i \leq n\}$.

We can describe our classifier as $\mathbf{y_i} = f(\mathbf{x_i}; \theta)$, which tries to estimate the labels $\mathbf{y_i}$ for the given sample $\mathbf{x_i}$, while θ represents the trainable parameters of the model. The model parameters are trained by optimizing a loss function $J(D, \theta)$ that describes how the model is performing over the training examples. In multi-label classification, it is common to use the binary cross entropy loss:

$$CE(\mathbf{x_i}, \mathbf{y_i}) = -\sum_{c=1}^{m} y_{i,c} log(f_c(\mathbf{x_i})) + (1 - y_{i,c}) log(1 - f_c(\mathbf{x_i})) \quad (1)$$

where $y_{i,c}$ is the c^{th} label in $\mathbf{y_i}$ and $f_c(\mathbf{x_i})$ is the output of the f_c classifier corresponding to the c^{th} label.

The cross entropy is made of two terms, one is active when the label is positive while the second is zero, and vice versa. We propose to modify each term to add a weighting factor, one relative to the confidence in the positive label and a second one relative to the expectation of a negative label for each sample.

$$CE_{proposed}(\mathbf{x_i}, \mathbf{y_i}) = -\sum_{c=1}^{m} \omega_{i,c} y_{i,c} log(f_c(\mathbf{x_i})) + \bar{\omega}_{i,c}(1 - y_{i,c}) log(1 - f_c(\mathbf{x_i})) \quad (2)$$

where $\omega_{i,c}$ represents the confidence in the positive label, while $\bar{\omega}_{i,c}$ represents the confidence in the negative label.

4.3. Application to the context prediction problem

For context classification using audio content, since we have no missing positive labels, we only need to add confidence weights to the negative part. However, we also tested using a confidence metric for the positive labels to evaluate its performance. The confidence in the positive label increases as the number of playlists where the track appears increases. To estimate if a negative label is missing, we use the correlations between contexts.

Regarding the negative weights, the confidence is defined as:

$$\bar{\omega}_{i,c} = P(y_{i,c} = 0 | \mathbf{y_i}) \quad (3)$$

which corresponds to the probability of having a negative label for the c^{th} label given the vector of labels $\mathbf{y_i}$ for the point $\mathbf{x_i}$. This probability can be estimated from the ground-truth label matrix based on label co-occurrences. When estimating the weight, it is possible to either ignore the zeros in the labels $\mathbf{y_i}$ since we have lower confidence about them, referred to as `ignore zeros`, or we can condition on the whole label vector including the zeros, `exact match`. We experimented with both of the negative weight schemes.

Regarding the positive weights, we propose using TF-IDF [27]:

$$\omega_{i,c} = \frac{n_{i,c}}{Nc} * \log\left(\frac{m}{\bar{n}_i}\right) \quad (4)$$

where $n_{i,c}$ is the number of times track $\mathbf{x_i}$ appeared in playlists from context class y_c. N_c is the total number of tracks that appeared in playlists of context class y_c. $\bar{n}_i$ is the number of context classes x_i is labelled with. The tf-idf values are naturally very small, hence, we normalize the values with unit-mean unit-variance. We interpret the positive weights as a priority rank to learn predicting important samples first, i.e. the ones with high tf-idf. Since there are no missing labels in the positive samples, we normalize it to have a mean of 1.

Table 2: Classification results for models trained with different weighting schemes computed with macro averaging

	HL*	AUC	Recall	Precision	f1	TN ratio*
No Weights	0.32	0.73	0.69	0.66	0.67	0.64
Negative Weights (exact match)	0.32	0.73	0.77	0.63	0.69	0.55
Negative Weights (ignore zeros)	0.39	0.72	0.94	0.56	0.7	0.27
Both Weights (exact match)	0.32	0.73	0.66	0.67	0.66	0.67
Both Weights (ignore zeros)	0.37	0.73	0.91	0.59	0.71	0.33

* HL : Hamming Loss. TN rate: True negative rate

4.4. Evaluation results

Table 2 shows the results of using different weighting schemes for training the model. We observe that using specific weighting schemes improves the results compared to using the non-weighted loss. It leads to a higher recall with a varying drop in the precision and true negative rate. We find that using the negative weights with ignoring the zeros gives the best results in terms of improving the recall, without pushing the model to outputting all ones. Hence, using the zeros when computing the co-occurrences of labels, even if some of them are missing, leads to better estimation of true and missing labels. We also found that using the tf-idf weighting scheme for the positive samples does not lead to much improvement in the classification results, which is not surprising as there are no missing positive labels in this dataset.

The goal is to have the highest recall with the least drop in precision. However, the missing labels in the ground-truth makes it challenging to objectively evaluate the performance. It is possible that the drop in the precision and true negative rate is due to missing labels in the ground-truth that were regarded as a false prediction while it is a false ground-truth. Our interpretation is that using the weights is useful for correctly predicting more positive samples. However, the balance between the recall and precision is subject to the problem and the use case of the classifier. While the problem of evaluating a model with missing labels is still an open issue, using the sample-level weighting in the loss function seems promising, especially in cases where detecting true positives is prioritized.

5. CONCLUSION

In this paper, we studied the problem of context prediction using the audio content. We proposed a procedure to label and collect a dataset of context-tagged tracks, which we distribute to the community. We evaluated and presented the baseline results of a common audio tagging model on our dataset. We finally proposed a solution for accounting for missing labels using pre-defined models with promising results. We found that certain contexts could easily be inferred from audio, which tend to be either energetic or calm music. Other contexts might require either an improved model or other types of data beside the audio content. Future work will aim at extending the dataset using the proposed procedure to more than 15 classes, investigating the audio features linked to each context, further evaluation of the weighted loss function, and investigating better evaluation metrics for multi-label classification with missing labels.

6. REFERENCES

[1] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.

[2] Norha M Villegas, Cristian Sánchez, Javier Díaz-Cely, and Gabriel Tamura, "Characterizing context-aware recommender systems: A systematic literature review," *Knowledge-Based Systems*, vol. 140, pp. 173–200, 2018.

[3] Khalid Haruna, Maizatul Akmar Ismail, Suhendroyono Suhendroyono, Damiasih Damiasih, Adi Pierewan, Haruna Chiroma, and Tutut Herawan, "Context-aware recommender system: a review of recent developmental process and future research direction," *Applied Sciences*, vol. 7, no. 12, pp. 1211, 2017.

[4] Adrian C North and David J Hargreaves, "Situational influences on reported musical preference.," *Psychomusicology: A Journal of Research in Music Cognition*, vol. 15, no. 1-2, pp. 30, 1996.

[5] Alinka E Greasley and Alexandra Lamont, "Exploring engagement with music in everyday life using experience sampling methodology," *Musicae Scientiae*, vol. 15, no. 1, pp. 45–71, 2011.

[6] George Prassas, Katherine C Pramataris, Olga Papaemmanouil, and Georgios J Doukidis, "A recommender system for online shopping based on past customer behaviour," in *Proceedings of the 14th BLED Electronic Commerce Conference, BLED*, 2001.

[7] V Subramaniyaswamy, R Logesh, M Chandrashekhar, Anirudh Challa, and V Vijayakumar, "A personalised movie recommendation system based on collaborative filtering," *International Journal of High Performance Computing and Networking*, vol. 10, no. 1-2, pp. 54–63, 2017.

[8] Karthik Yadati, Cynthia Liem, Martha Larson, and Alan Hanjalic, "On the automatic identification of music for common activities," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. ACM, 2017.

[9] Xinxi Wang, David Rosenblum, and Ye Wang, "Context-aware mobile music recommendation for daily activities," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012.

[10] Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani, "Recsys challenge 2018: Automatic music playlist continuation," in *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018.

[11] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler, "The effects of noisy labels on deep convolutional neural networks for music tagging," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 139–149, 2018.

[12] Michael Gillhofer and Markus Schedl, "Iron maiden while jogging, debussy for dinner?," in *International Conference on Multimedia Modeling*. Springer, 2015.

[13] Zhiyong Cheng and Jialie Shen, "Just-for-me: an adaptive personalization system for location-aware social music recommendation," in *Proceedings of international conference on multimedia retrieval*. ACM, 2014.

[14] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji, "Multi-label learning with missing labels," in *Proceedings of 22nd International Conference on Pattern Recognition*. IEEE, 2014.

[15] Wei Bi and James T Kwok, "Multilabel classification with label correlations and missing labels," in *Proceedings of 28th AAAI Conference on Artificial Intelligence*, 2014.

[16] Miao Xu, Gang Niu, Bo Han, Ivor W Tsang, Zhi-Hua Zhou, and Masashi Sugiyama, "Matrix co-completion for multi-label classification with missing features and labels," *arXiv preprint arXiv:1805.09156*, 2018.

[17] Zhi-Fen He, Ming Yang, Yang Gao, Hui-Dong Liu, and Yilong Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowledge-Based Systems*, vol. 163, pp. 145–158, 2019.

[18] Jun Huang, Feng Qin, Xiao Zheng, Zekai Cheng, Zhixiang Yuan, Weigang Zhang, and Qingming Huang, "Improving multi-label classification with missing labels by learning label-specific features," *Information Sciences*, vol. 492, pp. 124–146, 2019.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[20] Sankaran Panchapagesan, Ming Sun, Aparna Khare, Spyros Matsoukas, Arindam Mandal, Björn Hoffmeister, and Shiv Vitaladevuni, "Multi-task learning and weighted cross-entropy for dnn-based keyword spotting.," in *Proceddings of Interspeech*, 2016.

[21] Martin Pichl, E. Zangerle, and G. Specht, "Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name?," in *Proceddings of International Conference on Data Mining Workshop (ICDMW)*, 2015.

[22] Marius Kaminskas and Francesco Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2-3, pp. 89–119, 2012.

[23] Keunwoo Choi, George Fazekas, and Mark Sandler, "Automatic tagging using deep convolutional neural networks," *arXiv preprint arXiv:1606.00298*, 2016.

[24] Jordi Pons Puig, Oriol Nieto, Matthew Prockup, Erik M Schmidt, Andreas F Ehmann, and Xavier Serra, "End-to-end learning for music audio tagging at scale," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018.

[25] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas, "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011.

[26] Grigorios Tsoumakas and Ioannis Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[27] Juan Ramos et al., "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, 2003.