The Effects of Noisy Labels on Deep Convolutional Neural Networks for Music Tagging

Keunwoo Choi ⦿, György Fazekas, *Member, IEEE*, Kyunghyun Cho, and Mark Sandler, *Fellow, IEEE*

Abstract—**Deep neural networks (DNNs) have been successfully applied to music classification including music tagging. However, there are several open questions regarding the training, evaluation, and analysis of DNNs. In this paper, we investigate specific aspects of neural networks, the effects of noisy labels, to deepen our understanding of their properties. We analyze and (re-)validate a large music tagging dataset to investigate the reliability of training and evaluation. Using a trained network, we compute label vector similarities, which are compared to groundtruth similarity. The results highlight several important aspects of music tagging and neural networks. We show that networks can be effective despite relatively large error rates in groundtruth datasets, while conjecturing that label noise can be the cause of varying tag-wise performance differences. Finally, the analysis of our trained network provides valuable insight into the relationships between music tags. These results highlight the benefit of using data-driven methods to address automatic music tagging.**

Index Terms—**Music tagging, convolutional neural networks.**

I. INTRODUCTION

MUSIC tags are descriptive keywords that convey various types of high-level information about recordings such as mood ('sad', 'angry', 'happy'), genre ('jazz', 'classical') and instrumentation ('guitar', 'strings', 'vocal', 'instrumental') [1]. Tags may be associated with music in the context of a folksonomy, i.e., user-defined metadata collections commonly used for instance in online streaming services, as well as personal music collection management tools. As opposed to expert annotation, these types of tags are deeply related to listeners' or communities' subjective perception of music. In the aforementioned tools and services, a range of activities including search, navigation, and recommendation may depend on the existence of tags associated with tracks. However, new and rarely accessed tracks often lack the tags necessary to support them, which leads to well-known problems in music information management [2]. For instance, tracks or artists residing in the long tail of popularity distributions associated with large music catalogues may have insufficient tags, therefore they are rarely recommended or accessed and tagged in online communities. This leads to a circular problem. Expert annotation is notoriously expensive and intractable for large catalogues, therefore content-based annotation is highly valuable to bootstrap these systems. Music tag prediction is often called *music auto-tagging* [3]. Content-based music tagging algorithms aim to automate this task by learning the relationship between tags and the audio content.

Music tagging can be seen as a multi-label classification problem because music can be correctly associated with more than one true label, for example, {'rock', 'guitar', 'happy', and '90s'}. This example also highlights the fact that music tagging may be seen as multiple distinct tasks. Because tags may be related to genres, instrumentation, mood and era, the problem may be seen as a combination of genre classification, instrument recognition, mood and era detection, and possibly others. In the following, we highlight three aspects of the task that emphasise its importance in music informatics research (MIR).

First, collaboratively created tags reveal significant information about music consumption habits. Tag counts show how listeners label music in the real-world, which is often very different from the decision of a limited number of experts (see Section III-A) [4]. The first study on automatic music tagging proposed the use of tags to enhance music recommendation [3] for this particular reason. Second, the diversity of tags and the size of tag datasets make them relevant to several MIR problems including genre classification and instrument recognition. In the context of deep learning, tags can particularly be considered a good *source task* for transfer learning [5], [6], a method of reusing a trained neural network in a related task, after adapting the network to a smaller and more specific dataset. Since a music tagger can extract features that are relevant to different aspects of music, tasks with insufficient training data may benefit from this approach. Finally, investigating trained music tagging systems may contribute to our understanding of music perception and music itself. For example, analysing subjective tags such as mood and related adjectives can help building computational models for human perception of music (see Section III-C).

Albeit its importance, there are several issues one faces when analysing music tags. A severe problem, particularly in the context of deep learning, is the fact that sufficiently large training datasets are only available in the form of folksonomies. In these

K. Choi, G. Fazekas, and M. Sandler are with the Centre for Digital Music, Electric Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: keunwoo.choi@qmul.ac.uk; g.fazekas@qmul.ac.uk; marksandler@qmul.ac.uk).

K. Cho is with the Center for Data Science, New York University, New York, NY 10003 USA (e-mail: kyunghyun.cho@nyu.edu).

user-generated metadata collections, tags not only describe the content of the annotated items, for instance, well-defined categories such as instruments that appear on a track or the release year of a record, but also subjective qualities and personal opinions about the items [7]. Tags are often related to organisational aspects, such as self-references and personal tasks [8]. For instance, users of certain music streaming services frequently inject unique tags that have no significance to other users, i.e., they label music with apparently random character sequences which facilitate the creation of virtual personal collections, misappropriating this feature of the service. While tags of this nature are relatively easy to recognise and disregard using heuristics, other problems of folksonomies are not easily solved and constitute a great proportion of noise in these collections. Relevant problems include mislabelling, the use of highly subjective tags, such as those pertaining to genre or mood, as well as heterogeneity in the taxonomical organisation of tags. Researchers have been proposing to solve these problems either by imposing pre-defined classification systems on social tags [7], or providing tag recommendation based on context to reduce tagging noise in the first place [9]. While the benefits of such organisation or explicit knowledge of tag categories have been shown to benefit automatic music tagging systems e.g. in [10], most available large folksonomies still consist of noisy labels.

In this paper, we do not directly address the above issues, but perform data-driven analyses instead, focussing on the effects of noisy labels on deep convolutional neural networks for automatic tagging of popular music. Label noise is unavoidable in most real-world applications, therefore it is crucial to understand its effects. We hypothesise that despite the noise, neural networks are able to learn meaningful representations that help to associate audio content with tags and show that these representations are useful even if they remain imperfect. The insights provided by our analyses may be relevant and valuable across several domains where social tags or folksonomies are used to create automatic tagging systems or in research aiming to understand social tags and tagging behaviour. The primary contributions of this paper are as follows: *i)* An analysis of the largest and most commonly used public dataset for music tagging including an assessment of the distribution of labels within this dataset. *ii)* We validate the groundtruth and discuss the effects of noise, e.g. mislabelling, on both training and evaluation. Finally, *iii)* we provide a novel perspective on using specific network weights to analyse the trained network and obtain valuable insight into how social tags are related to music tracks. This analysis utilises parts of the weights corresponding to the final classifications. We termed these *label vectors*.

The rest of the paper is organised as follows. Section II outlines relevant problems and related works. Section III presents an analysis of a large tag dataset from three different but related perspectives. First, tag co-occurrence patterns are analysed and our findings are presented in Section III-A. Second, we validate the dataset labels and discuss the effects of label noise on neural network training and evaluation in Section III-B. We then assess the capacity of the trained network to represent musical knowledge in terms of similarity between predicted labels and co-occurrences between ground truth labels in Section III-C.

Finally, we draw overall conclusions and discuss cross-domain applications of our methodology in Section IV.

II. BACKGROUND AND RELATED WORK

Music tagging is related to common music classification and regression problems such as genre classification and emotion prediction. The majority of prior research have focussed on extracting relevant music features and applying a conventional classifier or regressor. For example, the first auto-tagging algorithm [3] proposed the use of mid-level audio descriptors such as Mel-Frequency Cepstral Coefficients (MFCCs) and an AdaBoost [11] classifier. Since most audio features are extracted frame-wise, statistical aggregates such as mean, variance and percentiles are also commonly used. This is based on the assumption that the features adhere to a pre-defined or known distribution which may be characterised by these parameters. However, hand-crafted audio features do not necessarily obey known parametric distributions [12], [13]. Consequently, vector quantisation and clustering was proposed e.g. in [14] as an alternative to parametric representations.

A recent trend in music tagging is the use of data-driven methods to *learn* features instead of designing them, together with non-linear mappings to more compact representations relevant to the task. These approaches are often called *representation learning* or *deep learning*, due to the use of multiple layers in neural networks that aim to learn both low-level features and higher-level semantic categories. Convolutional Neural Networks (denoted 'ConvNets' hereafter) have been providing state-of-the-art performance for music tagging in recent works [15], [1], [6]. In the rest of this section, we first review the datasets relevant to the tagging problem and highlight some issues associated with them. We then discuss the use of ConvNets in the context of music tagging.

A. Music Tagging Datasets and Their Properties

Training a music tagger requires examples, i.e., tracks labelled by listeners, constituting a groundtruth dataset. The size of the dataset needed for creating a good tagger depends on the number of parameters in its machine learning model. Using training examples, ConvNets can learn complex, non-linear relationships between patterns observed in the input audio and high-level semantic descriptors such as generic music tags. However, these networks have a very high number of parameters and therefore require large datasets and efficient training strategies. Creating sufficiently large datasets for the general music tagging problem is difficult for several reasons. Compared to genre classification for instance, which can rely mostly on metadata gathered from services such as MusicBrainz[1] or Last.fm[2], tagging often requires listening to the whole track for appropriate labelling, partly because of the diversity of tags [16], i.e., the many different kinds of tags listeners may use or may be interested in while searching.

---

[1] See https://musicbrainz.org: a crowd-sourced music meta-database.
[2] See http://www.last.fm: a personalised online radio.

Tagging is often seen as an inherently ill-defined problem since it is subjective and there is an almost unconstrained number of meaningful ways to describe music. For instance, in the Million Song Dataset (MSD) [17], one of the largest and most commonly used groundtruth sets for music tagging, there are 522,366 tags. This is outnumbering the 505,216 unique tracks present in the dataset. In fact, there is no theoretical limit on the number of labels in a tag dataset, since users often 'invent' specific labels of cultural significance that cannot be found in a dictionary, yet become widely associated with niche artistic movements, styles, artists or genres. Peculiar misspellings also become commonplace and gain specific meaning, for instance, using 'nu' in place of 'new' in certain genre names ('nu jazz', 'nu metal') suggests music with attention to pop-culture references or particular fusion styles, or the use of 'grrrl' refers to bands associated with the underground feminist punk movement of the 90s. Given a degree of familiarity with the music, listeners are routinely able to associate songs with such tags, even if the metadata related to the artist or the broader cultural context surrounding a particular track is not known to them. This leads to a hypothesis underlying most auto tagging research, that is, audio can be sufficient to assign a reasonably broad range of tags to music automatically. We note that our approach, like other generic auto tagging methods, does not aim to cover the kinds of highly personal tags mentioned in Section I.

Tags are also of different kinds and a single tag may often convey only a small part of what constitutes a good description. Tagging, therefore, is a multi-label classification problem. Consequently, the number of unique output vectors in a set increases exponentially with the number of items, while that of single-label classification only increases linearly. Given K binary labels, the size of the output vector set can increase up to 2^K. In practice, this problem is often alleviated by limiting the number of tags, usually to the top-N tags given the number of music tracks a tag is associated with, or the number of users who applied them.

The prevalence of music tags is also worth paying attention to because datasets typically exhibit an unbalanced distribution with a long-tail of rarely used tags. Regarding the diversity of the music and from the training perspective, there is an issue with non-uniform genre distributions too. In the MSD for example, the most popular tag is 'rock' which is associated with 101,071 tracks. However, 'jazz', the 12th most popular tag is used for only 30,152 tracks and 'classical', the 71st popular tag is used 11,913 times only, even though these three genres are on the same hierarchical level.

B. Labelling Strategies

We finally have to consider a number of labelling strategies. Audio items in a dataset may be 'strongly' or 'weakly' labelled, which may refer to several different aspects of tagging. First, there is a question of whether only positive labels are used. A particular form of weak labelling means that only positive associations between tags and tracks are provided. This means, given a finite set of tags, a listener (or annotator) applies a tag in case s/he recognises a relation between the tag and the music.

In this scenario, no negative relations are provided, and as a result, a tag being positive means it is 'true', but a tag being negative, i.e. not applied, means 'unknown'. The most common tags are about positiveness–labels usually explain the existence of features, not the non-existence of them. Exceptions that describe negativeness include 'instrumental', which may indicate the lack of vocals. Typical crowd-sourced datasets are weakly labelled, because it is the only practical solution to create a large dataset. Furthermore, listeners in online platforms cannot be reasonably expected to provide negative labels given the large number of possible tags. Strong labelling in this particular context would mean that disassociation between a tag and a track confirms negation, i.e., a zero element in a tag-track matrix would signify that the tag does not apply. To the best of our knowledge, *CAL500* [18] is the biggest music tag dataset (500 songs) that is strongly labelled. Most recent research has relied on collaboratively created, and therefore weakly-labelled datasets such as *MagnaTagATune* [19] (5,405 songs) and the *MSD* [17] containing 505,216 songs if only tagged items are counted.

The second aspect of labelling relates to whether tags describe the whole track or whether they are only associated with a segment where a tag is considered to be true. Time-varying annotation is particularly difficult and error prone for human listeners, therefore it does not scale. Multiple tags may be applied on a fixed-length segment basis, as is done in smaller datasets such as MagnaTagATune for 30s segments. The MSD uses only track-level annotation, which can be considered a form of weak labelling. From the perspective of training, this strategy is less adverse in case of particular tags than it is for some others. Genre or era tags are certainly more likely to apply to the whole track consistently than instrument tags for instance. This discrepancy may constitute noise in the training data. Additionally, often only preview clips are available to researchers. This forces them to assume that tags are correct within the preview clip too, which constitutes another source of groundtruth noise. In this work, we train ConvNets to learn the association between track-level labels and audio recordings using preview clips associated with the MSD.

Learning from noisy labels is an important problem, therefore several studies address this in the context of conventional classifiers such as support vector machines [20]. In deep learning research, [21] assumes a binary classification problem while [22] deals with multi-class classification. Usually, auxiliary layers are added to learn to fix the incorrect labels, which often requires a noise model and/or an additional clean dataset. Both solutions, together with much other research, are designed for single-class classifications, and there is no existing method that can be applied for music tagging when considered as multi-label classification and when the noise is highly skewed to negative labels. This will be discussed in Section III-B.

C. Convolutional Neural Networks

ConvNets are a special type of neural network introduced in computer vision to simulate the behaviour of the human vision system [23]. ConvNets have convolutional layers, each

of which consists of convolutional kernels. The convolutional kernels sweep over the inputs, resulting in weight sharing that greatly reduces the number of parameters compared to conventional layers that do not sweep and are fully-connected instead. Kernels are trained to find and capture local patterns that are relevant to the task using error backpropagation and gradient descent algorithms. Researchers in music informatics are increasingly taking advantage of deep learning techniques. ConvNets have already been used for chord recognition [24], genre classification [25], onset detection [26], music recommendation [27], instrument recognition [28] and music tagging [1], [15], [6].

In MIR, the majority of works use two dimensional time-frequency representations as inputs, e.g., short-time Fourier transform or mel-spectrograms [29]. Recently, several works proposed learning 2D representations by applying one dimensional convolution to the raw audio signal [15], [30]. It is possible to improve performances by learning more effective representations, although the approach requires increasingly more data, which is not always available. Moreover, these approaches have been shown to learn representations that are similar to conventional time-frequency representations that are cheaper to compute [15], [30].

ConvNets have been applied to various music and audio related tasks, assuming that certain relevant patterns can be detected or recognised by cascaded one- or two dimensional convolutions. They provide state-of-the-art performance in several music information retrieval tasks including music segmentation [31], beat detection [32] and tagging [6], as well as in non-music tasks such as acoustic event detection [33]. There are several possible arguments to justify the use of ConvNets for music tagging. First, music tags are often considered among the topmost high-level features representing song-level information above mid-level or intermediate musical features such as chords, beats, and tonality. This hierarchy fits well with ConvNets as they can learn hierarchical features over multilayer structures. Second, the invariance properties of ConvNets such as translation, distortion and local invariances can be useful for learning musical features when the relevant feature can appear at any time or frequency range with small time and frequency variances.

There are many different architectures for music tagging, but many share a common training scheme. They follow the supervised learning framework with backpropagation and stochastic gradient descent, they regard the problem as a regression problem. Many of them also use cross-entropy or mean square error as a loss function, which is empirically minimised using a training set with the maximum likelihood approach. The analyses presented in this paper aim at understanding the behaviour of ConvNets using supervised learning with noisy labels. This aspect of the research is tangential to the variations of ConvNet structures. Therefore, we omit results related to the different possible ConvNet structures. Particularly with respect to the analysis of the effect of label noise on tagging performance, a major contribution of this paper, different convnet structures have previously shown an almost identical trend in tag-wise performances [34].

D. Evaluation of Tagging Algorithms

There are several methods to evaluate tagging algorithms. Since the target is typically binarised to represent if the ith tag is true or false ($y_i \in \{0, 1\}$), classification evaluation metrics such as 'Precision' and 'Recall' can be used if the prediction is also binarised. Because label noise is mostly associated with negative labels, as we quantify in Section III-B, using recall is appropriate since it ignores incorrect negative labels. We have to note that metrics such as recall cannot be used as a loss function since they are not differentiable. They can be used instead as an auxiliary method of assessment after training. This strategy can work well because it prevents the network from learning trivial solutions for those metrics. For instance, predicting all labels to be 'True' to obtain a perfect recall score. Optimal thresholding for binarised prediction is an additional challenge however and discards information. The network learns a maximum likelihood solution with respect to the training data, which is heavily imbalanced, therefore the threshold should be chosen specifically for each tag. This introduces an entirely new research problem which we do not address here. The area under curve - receiver operating characteristic (AUC-ROC, or simply AUC) works without binarisation of predictions and is often used as an evaluation metric. A ROC curve is created by plotting the true positive rate against the false positive rate. As both rates range between $[0, 1]$, the area under the curve also ranges between $[0, 1]$. However, the effective range of AUC is $[0.5, 1]$ since random classification yields 0.5 when the true positive rate increases at the exact same rate of false positives.

III. EXPERIMENTS AND DISCUSSIONS

In this section, we present the methods and the results of experiments that analyse the Million Song Dataset (MSD) and a network trained on it. We select the MSD as it is the largest public dataset available for training music taggers. It also provides crawlable track identifiers for audio signals, which enables us to access the audio and re-validate the tags manually by listening. The analyses are divided into three parts and discussed separately in the following subsections. Section III-A is concerned with mutual relationships between tags. In Section III-B, we re-validate the groundtruth of the dataset to ascertain the reliability of research that uses it. Section III-C discuss properties of the trained network.

The tags in the MSD are collected using the Last.fm API which provides access to crowd-sourced music tags. We use the top 50 tags sorted by popularity (occurrence counts) in the dataset. The tags include genres ('rock', 'pop', 'jazz', 'funk'), eras ('60s' – '00s') and moods ('sad', 'happy', 'chill'). There are 242,842 clips with at least one of the top 50 tags. The tag counts range from 52,944 ('rock') to 1,257 ('happy') and there are 12,348 unique tag vectors represented as a joint variable of 50 binary values.

Throughout this paper, particularly in Sections III-B2, III-B3, and III-C, we use a ConvNet named 'compact-convnet'. As mentioned earlier, the proposed analysis is structure-agnostic, and we chose this network since it achieves a reasonable performance while being easy to understand and analyse due to its

TABLE I
DETAILS OF THE `compact-convnet` ARCHITECTURE

input (1, 96, 1360)	
Conv2d and Max-Pooling	(32, (3, 3)) and (2, 4)
Batch normalization - ELU activation	
Conv2d and Max-Pooling	(32, (3, 3)) and (4, 4)
Batch normalization - ELU activation	
Conv2d and Max-Pooling	(32, (3, 3)) and (4, 5)
Batch normalization - ELU activation	
Conv2d and Max-Pooling	(32, (3, 3)) and (2, 4)
Batch normalization - ELU activation	
Conv2d and Max-Pooling	(32, (3, 3)) and (4, 4)
Batch normalization - ELU activation	
Fully-connected layer	(50)
output (50)	

2-dimensional convolutional layer is specified by (channel, (kernel lengths in frequency, time)). Pooling layer is specified by (pooling length in frequency and time). Batch normalization layer [35] and exponential linear unit activation [36] are used after all convolutional layers.
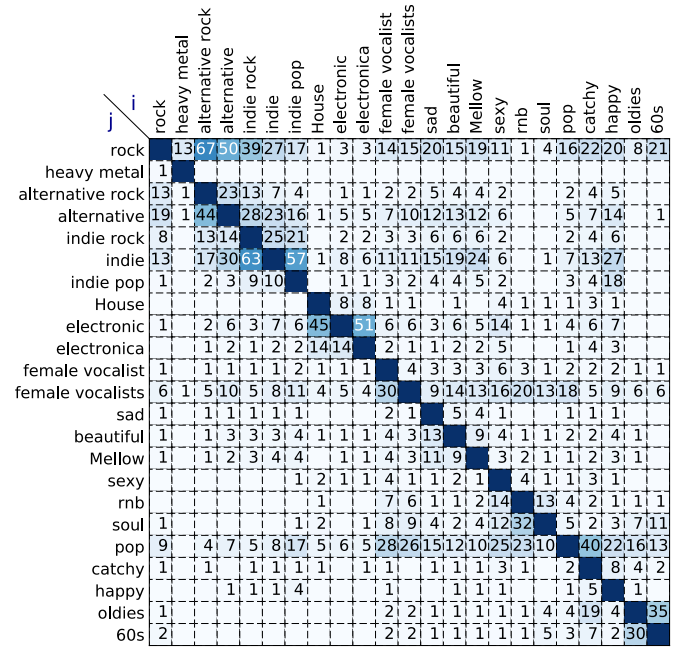

Fig. 1. Normalised tag co-occurrence pattern of the selected 23 tags from the training data. For the sake of visualisation, we selected 23 tags out of 50 that have high co-occurrences and represent different categories; genres, instruments and moods. The values are computed using (1) (and are multiplied by 100, i.e., shown in percentage), where y_i and y_j respectively indicate the labels on the x-axis and y-axis.

simple structure. Table I summarises the hyperparameters which are similar to the network in [1]. The original audio files are encoded in mp3 format with a sampling rate of 22,050 Hz and 64 kbps constant bit-rate. They are decoded, down-mixed to monaural, re-sampled to 12 kHz, and converted into mel-spectrograms with 96 mel-bins through a windowed short-time Fourier transform using 512-point FFT with 50% overlap. The ConvNet consists of homogeneous 5-layer, 3×3 convolutional kernels. On the input side, the mel-spectrogram magnitude is mapped using decibel scaling ($\log_{10} \mathbf{X}$) and adjusted using track-wise zero-mean unit-variance standardisation. We use 201,672/12,633/28,537 tracks as training/validation/test sets respectively, following the set splits provided by the MSD. This network[3] achieves an AUC of 0.845.

A. Tag Co-Occurrences in the MSD

We investigate the distribution and mutual relationships of tags in the dataset. This procedure helps understanding the task. Furthermore, our analysis represents information embedded in the training data. This will be compared to knowledge we can extract from the trained network (see Section III-C).

Here, we investigate the tuple-wise[4] relations between tags and plot the resulting normalised co-occurrence matrix (NCO) denoted $\mathbf{C}$. Let us define $\#y_i := |\{(x, y) \in D | y = y_i\}|$, the total number of the data points with ith label being *True* given the dataset D where (x, y) is an (input, target) pair. In the same manner, $\#(y_i \wedge y_j)$ is defined as the number of data points with both ith and jth labels being *True*, i.e., those two tags *co-occur*. NCO is computed using (1) and illustrated in Fig. 1.

$$C(i, j) = \#(y_i \wedge y_j)/\#y_i. \tag{1}$$

[4]These are not pairwise relations since there is no commutativity due to the normalisation term.

In Fig. 1, the x- and y-axes correspond to i, j respectively. Note that $C(i, j)$ is not symmetric, e.g., ('alternative rock', 'rock') = #(alternative rock $\wedge$ rock)/#alternative rock.

These patterns reveal mutual tag relationships which we categorise into three types: *i)* tags belonging to a genre hierarchy, *ii)* synonyms, i.e., semantically similar words and *iii)* musical similarity. Genre hierarchy tuples include for instance ('alternative rock', 'rock'), ('house', 'electronic'), and ('heavy metal', 'metal'). All first labels are sub-genres of the second. Naturally, we can observe that similar tags such as ('electronica', 'electronic') are highly likely to co-occur. Lastly, we notice tuples with similar meaning from a musical perspective including ('catchy', 'pop'), ('60s', 'oldies'), and ('rnb', 'soul'). Interestingly, $C(i, j)$ values with highly similar tag pairs, including certain subgenre-genre pairs, y_i and y_j are not close to 100% as one might expect. For example the pairs ('female vocalist', 'female vocalists') and ('alternative', 'alternative rock') reach only 30% and 44% co-occurrence values, while the pairs ('rock' and 'alternative rock'), ('rock', 'indie rock') reach only 69% and 39% respectively. This is primarily because *i)* items are weakly labelled and *ii)* there is often a more preferred tag to describe a certain aspect of a track compared to others. For instance, 'female vocalists' appears to be preferred over 'female vocalist' in our data as also noted in [2]. The analysis also reveals that certain types of label noise related to missing tags or taxonomical heterogeneity turn out to be very high in some cases. For instance, only 39% of 'indie rock' tracks are also tagged 'rock'. The effect of such label noise is studied more deeply in Section III-B. Furthermore, the computed NCO under-represents these co-occurring patterns. This effect is discussed in Section III-C.

TABLE II
THE SCORES OF GROUNDTRUTH WITH RESPECT TO OUR STRONGLY-LABELLED MANUAL ANNOTATION (SUBSET100) IN (A)-(D) AND OCCURRENCE COUNTS BY
THE GROUNDTRUTH (E), ESTIMATION (F), AND ON SUBSET400

	Scores of the groundtruth on Subset100				Occurrence counts		
	(a) Error rate, Positive label [%]	(b) Error rate, Negative label [%]	(c) Precision [%]	(d) Recall [%]	(e) In groundtruth (for all items)	(f) Estimate by (2) and Subset100	(g) By our annotation (on Subset400)
instrumental	6.0	12.0	94.0	88.7	8,424 (3.5%)	36,048 (14.9%)	85 (21.3%)
female vocalists	4.0	24.0	96.0	80.0	17,840 (7.3%)	71,127 (29.3%)	94 (23.5%)
male vocalists	2.0	64.0	98.0	60.5	3,026 (1.2%)	156,448 (64.4%)	252 (64.0%)
guitar	2.0	70.0	98.0	58.3	3,311 (1.4%)	170,916 (70.4%)	266 (66.5%)

B. Validation of the MSD as Groundtruth for Auto-Tagging

Next, we analyse the groundtruth noise in the MSD and examine its effect on training and evaluation. There are many sources of noise including incorrect annotation as well as information loss due to the trimming of full tracks to preview clips. Some of these factors may be assumed to be less adverse than others. In large-scale tag datasets, the frequently used weak labelling strategy (see Section II-B) may introduce a significant amount of noise. This is because by the definition of weak labelling, considering numerous tags a large portion of items remain un-labelled, but then these relations are assumed to be negative during training.

Validation of the annotation requires re-annotating the tags after listening to the excerpts, which is not a trivial task for several reasons. First, manual annotation does not scale and requires significant time and effort. Second, there is no single correct answer for many labels–music genre is an ambiguous and idiosyncratic concept, emotion annotation is highly subjective too, so as labels such as 'beautiful' or 'catchy'. Instrumentation labels can be objective to some extent, assuming the annotators have expertise in music. Therefore, we re-annotate items in two subsets using four instrument labels as described below.

- *Labels:* 'instrumental', 'female vocalists', 'male vocalists', 'guitar'.
- *Subsets:*
  Subset100: randomly selected 100 items for each class. All are from the training set and positive/negative labels are balanced as 50/50 respectively.
  Subset400: randomly selected 400 items from the test set.

1) Measuring Label Noise and Tagability: Table II column (a)-(d) summarises the statistics of Subset100. Confidence intervals for precision and recall are computed by bootstrapping [37] and plotted in Fig. 2. The average error rate of negative labels is 42.5%, which is very high, while that of positive labels is 3.5%. As a result, the precision of the groundtruth is high (96.5% on average) while the recall is much lower (71.9% on average). This suggests that the tagging problem should be considered weakly supervised learning to some extent. We expect this problem exists in other weakly-labelled datasets as well, since annotators typically do not utilise all possible labels.

Such a high error rate for negative labels suggests that the tag occurrence counts in the groundtruth are under-represented. This can be related to the *tagability* of labels, a notion which may
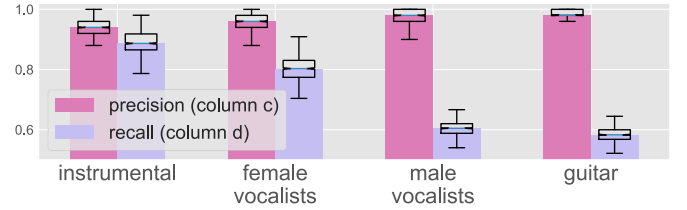

Fig. 2. The precision and recall of groundtruth on Subset100, corresponding to columns (c) and (d) in Table II. They are plotted with 95% confidential interval computed by bootstrapping [37].
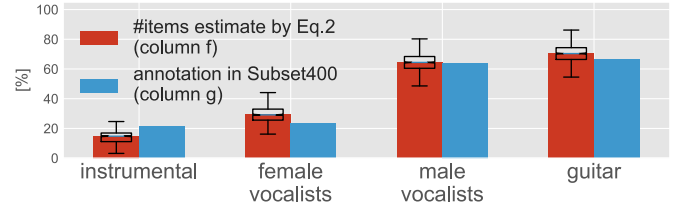

Fig. 3. The estimates of the number of items (red) and the number of items in Subset400 (blue), both in percentage (they correspond to columns (f) and (g) in Table II). The estimates are plotted with 95% confidential interval computed by bootstrapping [37].

be defined as *the likelihood that a track will be tagged positive for a label when it really is positive*. If the likelihood is replaced with the portion of items, tagability is measured by recall as presented in Table II as well as in Fig. 2. For example, bass guitar is one of the most widely used instruments in modern popular music, but it is only the 238th most popular tag in the MSD since tagging music with 'bass guitar' does not provide much information from the perspective of the average listener. Given the scores, we may assume that 'female vocalists' (88.7% of recall) and 'instrumental' (80.0%) are more tagable than 'male vocalists' (60.5%) and 'guitar' (58.3%), which indicates that the latter are presumably considered less unusual.

The correct number of *positive/negative* items can be estimated by applying Bayes' rule with the error rate. The estimated positive label count $\hat{N}^+$ is calculated using (2) as follows:

$$\hat{N}^+ = N^+(1 - p^+) + (T - N^+)p^-, \tag{2}$$

where N^+ is the tag occurrence, T is the number of total items ($T = 242,842$ in our case), and p^+, p^- refers to the error rates of positive and negative labels respectively. Column (f) of Table II and Fig. 3 present the estimated occurrence counts using (2). This estimate is validated using Subset400. Comparing

the percentages in columns (f) and (g) confirms that the estimated counts are more correct than the tag occurrences of the groundtruth. For all four tags, the confidence intervals overlap with the percentage counted in Subset400 as illustrated in Fig. 3. In short, the correct occurrence count is not correlated with the occurrence in the dataset, which shows the bias introduced by tagability. For instance, 'male vocalists' is more likely to occur in music than 'female vocalists', which means it has lower tagability, and therefore it ends up having fewer occurrences in the groundtruth.

2) Effects of Incorrect Groundtruth on the Training: Despite such inaccuracies, it is possible to train networks for tagging with good performances using the MSD, achieving an AUC between 0.85 [1] and 0.90 [6]. This may be because even with such noise, the network is weakly supervised by stochastically correct feedbacks, where the noise is alleviated by a large number of training examples [38]. In other words, given $\mathbf{x}$ is the input and $\mathbf{y}_{true}$, $\mathbf{y}_{noisy}$ are the correct and noisy labels respectively, the network can approximate the relationship $f : \mathbf{x} \rightarrow \mathbf{y}_{true}$ when training using $(\mathbf{x}, \mathbf{y}_{noisy})$.

However, we suggest that the noise affects the training and it is reflected in the performances of different tags. In [34], where different deep neural network structures for music tagging were compared, the authors observed a pattern on the per-tag performances that is common among different structures. This is illustrated in Fig. 5 where the x-axis labels represent tag popularity ranking. The performances were not correlated with the ranking, the reported correlation is 0.077, therefore the question remained unanswered in [34].

We conjecture that tagability, which is related to (negative) label noise can explain tag-wise performance differences. It is obvious that a low tagability implies more false negatives in the groundtruth. Therefore we end up feeding the network with more confusing training examples. For example, assuming there is a pattern related to male vocalists, the positive-labelled tracks provide mostly correct examples. However, many examples of negative-labelled tracks (64% in Subset100) also exhibit the pattern. Consequently, the network is forced to distinguish hypothetical differences between the positive-labelled true patterns and the negative-labelled true patterns, which leads to learning a more fragmented mapping of the input. This is particularly true in music tagging datasets where negative label noise dominates the total noise. This is supported by data both in this paper and [34] as discussed below.

First, tagabilities (or recall) and AUC scores with respect to the groundtruth and our re-annotation are plotted in Fig. 4 using Subset400 items and the `compact-convnet` structure. Both AUC scores are positively correlated to tagability while they are not related to the tag popularity rankings. Although the confidence intervals of 'instrument' vs. 'female vocalists', and 'male vocalists' vs. 'guitar' overlap, there is an obvious correlation. The performances on the whole test set also largely agree with our conjecture.

Second, in Fig. 5, AUC scores for instrument tags are ranked as 'instrumental' > 'female vocalists' > 'guitar' > 'male vocalists' for all three ConvNet structures. This aligns with tagability in Fig. 4 within the confidence intervals.
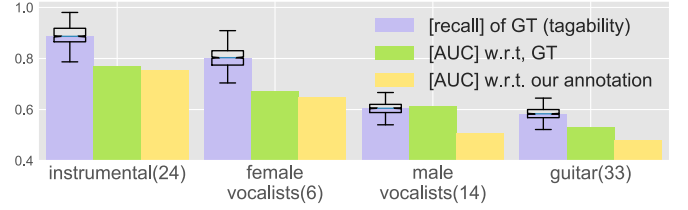

Fig. 4. The recall rates (tagability, pink), AUC scores with respect to the groundtruth (green), and AUC scores with respect to our annotation (yellow), all reported on Subset400. The numbers on the x-axis labels are the corresponding popularity rankings of tags out of 50. The recall rates and their 95% confidential intervals are identical to Fig. 2 but plotted again for comparison with tag-wise AUC scores.

This observation motivates us to expand this approach and assess tags in other categories. Within the Era category in Fig. 5, performance is negatively correlated with the popularity ranking (Spearman correlation coefficient = −0.7). There is a large performance gap between old music groups (60s, 80s, 70s) and the others (90s, 00s). We argue that this may also be due to tagability. In the MSD, older songs (e.g. those released before the 90s) are less frequent compared to modern songs (90s or 00s). According to the year prediction subset of the MSD[5], 84% of tracks are released after 1990. This is also related to the fact that the tag 'oldies' exists while its opposite does not. Hence, old eras seem more tagable, which might explain the performance differences in Era tags. We cannot extend this approach to mood and genre tags because the numbers of tags are much larger and there may be aspects contributing to tag-wise performance differences other than tagability.

3) Validation of the Evaluation: Another problem with using a noisy dataset is evaluation. In the previous section, we assumed that the system can learn a *denoised* relationship between music pieces and tags, $f : \mathbf{x} \rightarrow \mathbf{y}_{true}$. However, the evaluation of a network with respect to $\mathbf{y}_{noisy}$ includes errors due to noisy groundtruth. This raises the question of the reliability of the results. We use our strongly-labelled annotation of Subset400 to assess this reliability.

Let us re-examine Fig. 4. All AUC scores with respect to our annotation are lower than the scores with respect to the groundtruth. Performance for the *guitar* tag is remarkably below 0.5, the baseline score of a random classifier. However, the overall trend of tag-wise performance does not change. Because the results are based only on four classes and a subset of songs, a question arises: How does this result generalise to other tags?

To answer the question, three AUC scores are plotted in Fig. 6: *i)* the scores of the four instrument tags with respect to our annotation (dotted red), *ii)* the scores of the four instrument tags with respect to the given groundtruth (dashed blue), and *iii)* the scores of all tags with respect to the given groundtruth (solid yellow).

The reliability of evaluation is typically assessed with a given groundtruth and can be measured by ρ_1, the Pearson correlation coefficient between AUCs using our annotation and the MSD. The correlation between the four tags and all other tags (shown

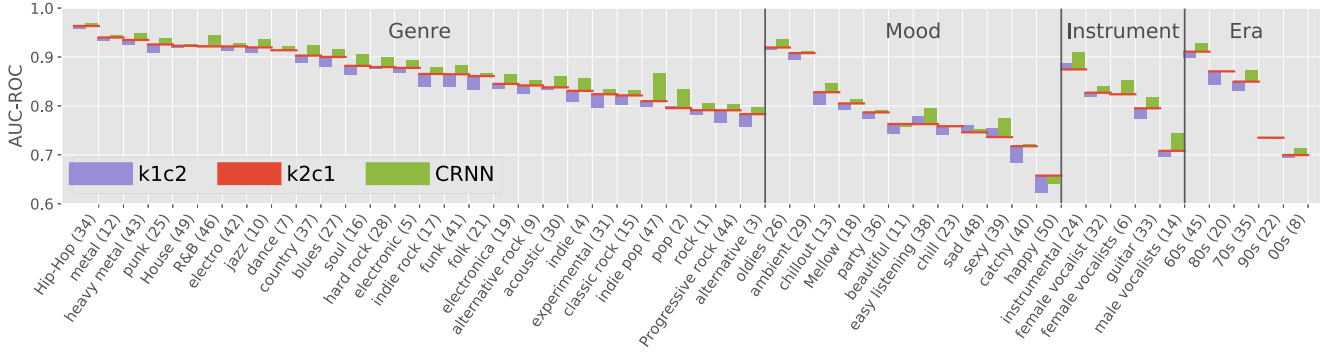[5]https://labrosa.ee.columbia.edu/millionsong/pages/tasks-demos.

Fig. 5. k1c2, k2c1, and CRNN are the names of network structures in [34]. *i)* AUC of each tag is plotted using a bar chart and line. For each tag, the red line indicates the score of k2c1 which is used as a baseline of bar charts for k1c2 (blue) and CRNN (green). In other words, the blue and green bar heights represent the performance gaps, k2c1-k1c2 and CRNN-k2c1, respectively. *ii)* Tags are grouped by categories (genre/mood/instrument/era) and sorted by the score of k2c1. *iii)* The number in parentheses after each tag indicates that tag's popularity ranking in the dataset.
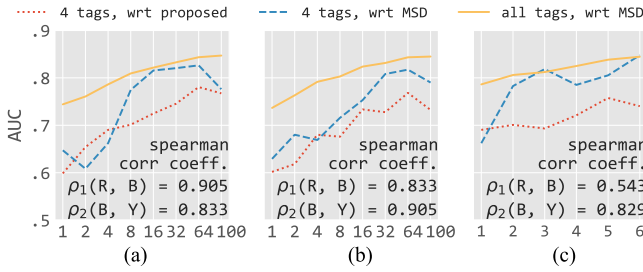

Fig. 6. AUC scores of all tags (yellow, solid) and four instrumentation tags. Instrumentation tags are evaluated using *i)* dataset groundtruth (blue, dashed) and *ii)* our strong-labelling re-annotation (red, dotted). Pearson correlation coefficients between {red vs. blue} and {blue vs. yellow} is annotated on each chart as ρ. In (c), x-axis is experiment index and various audio preprocessing methods were applied for each experiment. (a) Training data used with melgram [%], (b) Training data used with STFT [%], (c) Audio preprocessing methods.
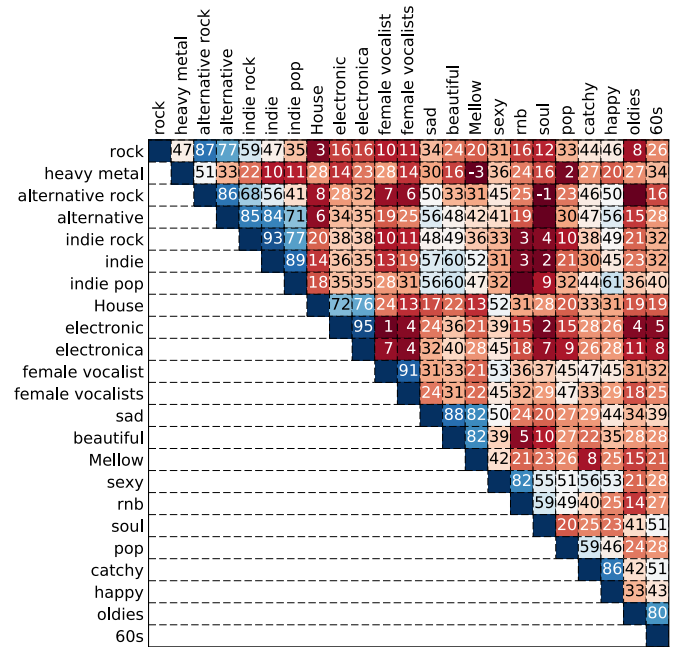

Fig. 7. Label vector similarity matrix by (3) (of manually selected 23 tags, same in Fig. 1, where symmetric components are omitted and numbers are ×100 after dot product for visual clarity.

in blue and yellow), denoted ρ_2, is a measure of how we can generalise our re-annotation result to those concerning all tags.

We selected three sets of tagging results to examine and plotted these in Fig. 6(a)–(c). The first two sets shown in subfigure (a) and (b) are results after training the compact-convnet with varying training data size and different audio representations: (a) melspectrogram and (b) short-time Fourier transform. The third set of curves in Fig. 6(c) compare six results with varying input time-frequency representations, training data size and input preprocessing techniques including normalisation and spectral whitening. The third set is selected to observe the correlations when the performance differences among systems are more subtle than those in (a) and (b).[6] First, the ρ_1 values in (a)–(c) suggest that noisy groundtruth provides reasonably stable evaluation for the four instrument tags. On the first two sets in (a) and (b), the scores of four tags using the MSD groundtruth (in blue) are highly correlated ($\rho_1 = 0.905$ and 0.833) to the scores using our annotation (red). This suggests the evaluation using noisy labels is still reliable. However, in (c), where the scores of all tags with given groundtruth (yellow) are in a smaller range, the correlation between all tags and the four tags (ρ_1) decreases to 0.543. The results imply that the distortion on the evaluation

using the noisy groundtruth may disguise the performance difference between systems when the difference is small. Second, large ρ_2 indicates that our validation is not limited to the four instrument tags but can be generalised to all tags. The correlation coefficients ρ_2 is stable and reasonably high in (a)–(c). It is 0.856 on average.

C. Analysis of Predicted Label Vectors

In the previous sections, groundtruth labels were analysed from various perspectives. It is worth considering how this information is distilled into the network after training, and whether we can leverage the trained network beyond our particular task. To answer these questions we use the trained network weights to assess how the network 'understands' music content by its label. This analysis also provides a way to discover unidentified

---

[6]We omit the details of the preprocessing methods, which are summarised in [39], because the focus is on the correlation of the final scores.

TABLE III
TOP-20 SIMILAR TAG TUPLES BY TWO ANALYSIS APPROACHES

Similar tags by groundtruth labels	(alternative rock, rock)§ (indie rock, indie)# (House, dance)‡‡ (indie pop, indie) (classic rock, rock) (electronica, electronic)* (alternative, rock) (hard rock, rock) (electro, electronic)** (House, electronic)(alternative rock, alternative)¶ (catchy, pop) (indie rock, rock) (60s, oldies)†† (heavy metal, metal)§§ (rnb, soul) (ambient, electronic) (90s, rock) (heavy metal, hard rock)‡ (alternative, indie)‖
Similar tags by label vectors	(electronica, electronic)* (indie rock, indie)# (female vocalist, female vocalists) (heavy metal, hard rock)‡ (indie, indie pop) (sad, beautiful) (alternative rock, rock)§ (alternative rock, alternative)¶ (happy, catchy) (indie rock, alternative) (alternative, indie)‖ (rnb, sexy) (electro, electronic)** (sad, Mellow) (Mellow, beautiful) (60s, oldies)†† (House, dance)‡‡ (heavy metal, metal)§§ (chillout, chill) (electro, electronica)

The first row is by analysing co-occurrence of tags in groundtruth (see Section III-A for details). The second row is by the similarity of trained label vector (see Section III-C for details). Common tuples are annotated with matching symbols.

relationships between labels and the music. The goal of label vector analysis is to better understand network training as well as assess its capacity to represent domain knowledge, i.e., relationships between music tags that are not explicitly shown in the data. In the `compact-convnet` described in Section III, the output layer has a dense connection to the last convolutional layer. The weights are represented as a matrix $\mathbf{W} \in \mathbb{R}^{N \times 50}$, where N is the number of feature maps (N=32 for our case) and 50 is the number of the predicted labels. After training, the columns of $\mathbf{W}$ can be interpreted as N-dimensional latent vectors since they represent how the networks combine information in the last convolutional layer to make the final prediction. We call these *label vector*s.

We compute the pairwise label vector similarity (LVS) using the dot product, i.e., $S(i,j) = w(i) \cdot w(j)$ where $i, j \leq 50$ or equivalently:

$$\mathbf{S} = \mathbf{W}^\top \cdot \mathbf{W}, \qquad (3)$$

which yields a 50×50 symmetric matrix.

LVS is illustrated in Fig. 7. The pattern is similar to the values in NCO (normalised co-occurrence) shown in Fig. 1 (see Section II-B). On average, the similarities in $S(i,j)$ are higher than those in $C(i,j)$. In $\mathbf{S}$, only four pairs show negative values, 'classic rock' – 'female vocalists', and 'mellow' – {'heavy metal', 'hard rock', and 'dance'}. In other words, label vectors are distributed in a limited space corresponding to a 32 dimensional vector space, where the angle θ between $w(i)$ and $w(j)$ is smaller than $\pi/2$ for most of the label vector pairs. This result can be interpreted in two ways: how well the ConvNet reproduce the co-occurrence that was provided by the training set; and if there is additional insight about music tags in the trained network.

First, the Pearson correlation coefficient of the rankings by LVS and NCO is 0.580.[7] The top 20 most similar label pairs are sorted and described in Table III. The second row of the table shows similar pairs according to the label vectors estimated by the network. Eleven out of 20 pairs overlap with the top 20 NCO tuples shown in the top row of the table. Most of these relations can be explained by considering the genre hierarchy. Besides, pairs such as ('female vocalists', 'female vocalists') and ('chillout', 'chill') correspond to semantically similar words. Overall, tag pairs showing high similarity (LVS) reasonably represent musical knowledge and correspond to high NCO values computed from the ground truth. This confirms the effectiveness

of the network to predict subjective and high-level semantic descriptors from audio only.

Second, there are several pairs that are *i)* high in LVS, *ii)* low in NCO, and *iii)* presumably music listeners would reasonably agree with their high similarity. These pairs show the extracted representations of the trained network can be used to measure tag-level musical similarities even if they are not explicitly shown in the groundtruth. For example, pairs such as ('sad', 'beautiful'), ('happy', 'catchy') and ('rnb', 'sexy') are in the top 20 of LVS (6th, 9th, and 12th similarities with 0.88, 0.86, and 0.82 of similarity values respectively). On the contrary, according to the ground truth, they are only 129th, 232nd, 111th co-occurring with 0.13, 0.08, and 0.14 of co-occurrence likelihood respectively.

In summary, the analysis based on LVS indirectly validates that the network learned meaningful representations that correspond to the groundtruth. Moreover, we found several pairs that are considered similar by the network which may help to extend our understanding of the relation between music and tags.

IV. CONCLUSIONS

In this article, we investigated several aspects how noisy labels in folksonomies affect the training and performance of deep convolutional neural networks for music tagging. We analysed the MSD, the largest dataset available for training a music tagger from a novel perspective. We reported on a study aiming to validate the MSD as groundtruth for this task. We found that the dataset is reliable overall, despite several noise sources affecting training and evaluation. Finally, we defined and used *label vectors* to analyse the capacity of the network to explain similarity relations between semantic tags.

Overall, the behaviours of the trained network were shown to be related to the property of the given labels. The analysis showed that tagability, which we measured by recall on the groundtruth, is correlated to the tag-wise performance. This opened a way to explain tag-wise performance differences within other categories of tags such as era. In the analysis of the trained network, we found that the network learns more intricate relationships between tags rather than simply reproducing the co-occurrence patterns in the groundtruth. The trained network is able to infer musically meaningful relationships between tags that are not present in the training data.

Although we focused on music tagging, our results provide general knowledge applicable in several other domains or tasks including other music classification tasks. The analysis method

[7]Because of the asymmetry of $C(i,j)$, rankings of $\max(C(i,j), C(y,y))$ are used.

presented here and the result on the tagging dataset can easily generalise to similar tasks in other domains involving folksonomies with noisy labels or tasks involving weakly labelled datasets, e.g. image tagging, object recognition in video, or environmental sound recognition, where not all sources are necessarily labelled. Future work will explore advanced methods to learn and evaluate using noisy datasets under a structured machine learning framework. Tagability can be understood from the perspective in music cognition research and should be investigated further.

REFERENCES

[1] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proc. 17th Int. Soc. Music Inf. Retrieval Conf.*, New York, NY, USA, 2016, pp. 805–811.

[2] P. Lamere, "Social tagging and music information retrieval," *J. New Music Res.*, vol. 37, no. 2, pp. 101–114, 2008.

[3] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2008, pp. 385–392.

[4] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of musical mood: Comparison between crowd-sourced and curated editorial tags," in *Proc. 2013 IEEE Int. Conf. Multimedia Expo Workshops*, 2013, pp. 1–6.

[5] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proc. 2007 18th Int. Soc. Music Inf. Retrieval Conf.*, Suzhou, China, 2017, pp. 141–149.

[6] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging," *IEEE Signal Process. Lett.*, pp. 1208–1212, 2017.

[7] I. Cantador, I. Konstas, and J. M. Jose, "Categorising social tags to improve folksonomy-based recommendations," *Web Semantics, Sci., Serv. Agents World Wide Web*, vol. 9, no. 1, pp. 1–15, 2011.

[8] A. Chamberlain and A. Crabtree, "Searching for music: Understanding the discovery, acquisition, processing and organization of music in a domestic setting for design," *Pers. Ubiquitous Comput.*, vol. 20, no. 4, pp. 559–571, Aug. 2016.

[9] F. Font, S. Oramas, G. Fazekas, and X. Serra, "Extending tagging ontologies with domain specific knowledge," in *Proc. Int. Semantic Web Conf.*, Riva del Garda, Italy, Oct. 2014, pp. 209–212.

[10] P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot, and M. Sandler, "Genre-adaptive semantic computing and audio-based modelling for music mood annotation," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 122–135, Apr./Jun. 2016.

[11] Y. Freund *et al.*, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, vol. 96, 1996, pp. 148–156.

[12] C. Baume, G. Fazekas, M. Barthet, D. Marston, and M. Sandler, "Selection of audio features for music emotion recognition using production music," in *Proc. 53rd Int. Conf. Audio Eng. Soc. Semantic Audio*, Jan. 2014, pp. 54–62.

[13] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, Apr. 2008.

[14] M. D. Hoffman, D. M. Blei, and P. R. Cook, "Content-based musical similarity computation using the hierarchical dirichlet process," in *Proc. Int. Conf. Music Inf. Retrieval*, 2008, pp. 349–354.

[15] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. 2014 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6964–6968.

[16] J. Lorince, K. Joseph, and P. M. Todd, *Analysis of Music Tagging and Listening Patterns: Do Tags Really Function as Retrieval Aids?* Cham, Switzerland: Springer, 2015, pp. 141–152. [Online]. Available: https://doi.org/10.1007/978-3-319-16268-3_15.

[17] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. 12th Int. Soc. Music Inf. Retrieval Conf.*, Miami, FL, USA, 2011, pp. 591–596.

[18] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the cal500 data set," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 439–446.

[19] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. Proc. 10th Int. Conf. Music Inf. Retrieval*, 2009, pp. 387–392.

[20] B. Frénay, and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 5, pp. 845–869, May 2014.

[21] V. Mnih and G. E. Hinton, "Learning to label aerial images from noisy data," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 567–574.

[22] S. Sukhbaatar and R. Fergus, "Learning from noisy labels with deep neural networks," *Workshop at the 3rd Int. Conf. Learn. Representations*, vol. 2, no. 31, p. 4, 2014.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[24] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, vol. 2, 2012, pp. 357–362.

[25] L. Li, "Audio musical genre classification using convolutional neural networks and pitch and tempo transformations," Master's thesis, Dept. Comput. Sci., , City Univ. Hong Kong, Kowloon Tong, Hong Kong, 2010.

[26] J. Schlüter, and S. Böck, "Musical onset detection with convolutional neural networks," in *Proc. 6th Int. Workshop Mach. Learn. Music*, Prague, Czech Republic, 2013, pp. 79–82.

[27] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.

[28] Y. Han, J. Kim, and K. Lee, "Deep convolutional neural networks for predominant instrument recognition in polyphonic music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 208–221, Jan. 2017.

[29] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," *arXiv preprint arXiv:1709.04396*, 2017.

[30] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," *Proc. 14th Int. Conf. Sound and Music Comput. Conf. (SMC)*, pp. 220–226, 2017.

[31] K. Ullrich, J. Schlüter, and T. Grill, "Boundary detection in music structure analysis using convolutional neural networks," in *Proc. 15th Int. Soc. Music Inf. Retrieval Conf.*, Taipei, Taiwan, 2014, pp. 417–422.

[32] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 255–261.

[33] A. Jansen, D. Ellis, D. Freedman, J. F. Gemmeke, W. Lawrence, and X. Liu, "Large-scale audio event discovery in one million YouTube videos," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 786–790.

[34] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. 2017 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 2392–2396.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[36] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUS)," *The 4th Int. Conf. Learn. Representations*, 2015.

[37] C. Z. Mooney and R. D. Duval, *Bootstrapping: A Nonparametric Approach to Statistical Inference*, nos. 94–95. Newbury Park, CA, USA: Sage, 1993.

[38] L. Torresani, "Weakly supervised learning," in *Computer Vision*. Boston, MA, USA: Springer, 2014, pp. 883–885.

[39] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A comparison on audio signal preprocessing methods for deep neural networks on music tagging," *arXiv preprint arXiv:1709.01922*, 2017.
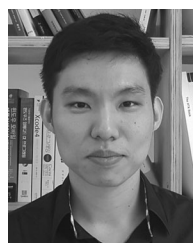
Keunwoo Choi received the B.Sc. and M.Phil. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2009 and 2011, respectively. He has been working toward the Ph.D. degree at the Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K., since 2014. In 2011, he joined the Electrical and Telecommunication Research Institute, Daejeon, South Korea, as a Researcher.

György Fazekas (M'14) received the B.Sc. degree from Kand Klmn College of Electrical Engineering, Faculty of Electrical Engineering, Buda University, Budapest, Hungary, and the M.Sc. and Ph.D. degrees from Queen Mary University of London (QMUL), London, U.K., in 2012. His thesis titled semantic audio analysis-utilities and applications explores novel applications of semantic audio analysis, semantic web technologies, and ontology-based information management. He is a Lecturer with QMUL, working at the Centre for Digital Music, School of Electronic Engineering and Computer Science. He is leading QMUL's research on the EU funded Audio Commons project facilitating the use of open sound content in professional audio production. His research interests include the development of semantic audio technologies and their application to creative music production. He collaborates on several other projects and he is a Member of the AES and ACM.

Mark Sandler (F'14) was born in 1955. He received the B.Sc. and Ph.D. degrees from the University of Essex, Colchester, U.K., in 1978 and 1984 respectively. He is a Professor of signal processing and the Founding Director with the Centre for Digital Music, School of Electronic Engineering and Computer Science, Queen Mary University of London, London, U.K. He has published more than 400 papers in journals and conferences. He is a Fellow of the Royal Academy of Engineering, AES, and IET.
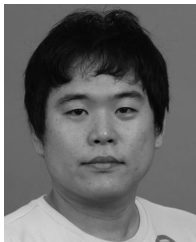
Kyunghyun Cho received the M.Sc. and Ph.D. degrees from Aalto University, Espoo, Finland, in 2014, under the supervision of Prof. J. Karhunen, Dr. T. Raiko, and Dr. A. Ilin. He is an Assistant Professor of computer science and data science with New York University, New York, NY, USA. He was a Postdoctoral Fellow at the University of Montreal, until summer 2015, under the supervision of Prof. Y. Bengio.