ABSTRACT

In this paper, we present a transfer learning approach for music classification and regression tasks. We propose to use a pre-trained convnet feature, a concatenated feature vector using the activations of feature maps of multiple layers in a trained convolutional network. We show how this convnet feature can serve as general-purpose music representation. In the experiments, a convnet is trained for music tagging and then transferred to other music-related classification and regression tasks. The convnet feature outperforms the baseline MFCC feature in all the considered tasks and several previous approaches that are aggregating MFCCs as well as low- and high-level music features.

1. INTRODUCTION

In the field of machine learning, transfer learning is often defined as re-using parameters that are trained on a source task for a target task, aiming to transfer knowledge between the domains. A common motivation for transfer learning is the lack of sufficient training data in the target task. When using a neural network, by transferring pre-trained weights, the number of trainable parameters in the target-task model can be significantly reduced, enabling effective learning with a smaller dataset.

A popular example of transfer learning is semantic image segmentation in computer vision, where the network utilises rich information, such as basic shapes or prototypical templates of objects, that were captured when trained for image classification [37]. Another example is pre-trained word embeddings in natural language processing. Word embedding, a vector representation of a word, can be trained on large datasets such as Wikipedia [35] and adopted to other tasks such as sentiment analysis [27].

There have been several works on transfer learning in Music Information Retrieval (MIR). Hamel et al. proposed to directly learn music features using linear embedding [57] of mel-spectrogram representations and genre/similarity/tag labels [20]. Oord et al. outlines a large-scale transfer learning approach, where a multi-layer perceptron is combined with the spherical K-means algorithm [16] trained on tags and play-count data [54]. After training, the weights are transferred to perform genre classification and auto-tagging with smaller datasets. In music recommendation, Choi et al. used the weights of a convolutional neural network for feature extraction in playlist generation [10], while Liang et al. used a multi-layer perceptron for feature extraction of content-aware collaborative filtering [29].

2. TRANSFER LEARNING FOR MUSIC

In this section, our proposed transfer learning approach is described. A convolutional neural network (convnet) is designed and trained for a source task, and then, the network with trained weights is used as a feature extractor for target tasks. The schematic of the proposed approach is illustrated in Figure 1.

2.1 Convolutional Neural Networks for Music Tagging

We choose music tagging as a source task because i) large training data is available and ii) its rich label set covers various aspects of music, e.g., genre, mood, era, and instrumentations. In the source task, a mel-spectrogram (X), a two-dimensional representation of music signal, is used as the input to the convnet. The mel-spectrogram is selected since it is psychologically relevant and computationally efficient. It provides a mel-scaled frequency representation which is an effective approximation of human auditory perception [36] and typically involves compressing the frequency axis of short-time Fourier transform representation (e.g., 257/513/1025 frequency bins to 64/96/128 Mel-frequency bins). In our study, the number of mel-bins is set to 96 and the magnitude of mel-spectrogram is mapped to decibel scale \((\log_{10} X)\), following [8] since it is also shown to be crucial in [7].

In the proposed system, there are five layers of convolutional and sub-sampling in the convnet as shown in Figure 1. This convnet structure with 2-dimensional \(3 \times 3\) kernels and 2-dimensional convolution, which is often called Vggnet [44], is expected to learn hierarchical time-frequency
patterns. This structure was originally proposed for visual image classification and has been found to be effective and efficient in music classification. 

2.2 Representation Transfer

In this section, we explain how features are extracted from a pre-trained convolutional network. In the remainder of the paper, this feature is referred to as pre-trained convnet feature, or simply convnet feature.

It is already well understood how deep convnets learn hierarchical features in visual image classification. By convolution operations in the forward path, lower-level features are used to construct higher-level features. Sub-sampling layers reduce the size of the feature maps while adding local invariance. In a deeper layer, as a result, the features become more invariant to (scaling/location) distortions and more relevant to the target task.

This type of hierarchy also exists when a convnet is trained for a music-related task. Visualisation and sonification of convnet features for music genre classification has shown the different levels of hierarchy in convolutional layers 1, [9].

Such a hierarchy serves as a motivation for the proposed transfer learning. Relying solely on the last hidden layer may not maximally extract the knowledge from a pre-trained network. For example, low-level information such as tempo, pitch, (local) harmony or envelope can be captured in early layers, but may not be preserved in deeper layers due to the constraints that are introduced by the network structure: aggregating local information by discarding less-relevant information in subsampling. For the same reason, deep scattering networks [6] and a convnet for musical tagging introduced in [28] use multi-layer representations.

Based on this insight, we propose to use not only the activations of the final hidden layer but also the activations of (up to) all intermediate layers to find the most effective representation for each task. The final feature is generated by concatenating these features as demonstrated in Figure 1, where all the five layers are concatenated to serve as an example.

Given five layers, there are \( \sum_{n=1}^{5} C_n = 31 \) strategies of layer-wise combination. In our experiment, we perform a nearly exhaustive search and report all results. We designate each strategy by the indices of layers employed. For example, a strategy named ‘135’ refers to using a \( 32 \times 3 = 96 \)-dimensional feature vector that concatenates the first, third, and fifth layer convnet features.

During the transfer, average-pooling is used for the 1st–4th layers to reduce the size of feature maps to 1×1 as illustrated in Figure 1. Averaging is chosen instead of max-pooling because it is more suitable for summarising the global statistics of large regions, as done in the last layer in [30]. Max-pooling is often more suitable for capturing the existence of certain patterns, usually in small and local regions.

Lastly, there have been works suggesting random-weights (deep) neural networks including deep convnet can work well as a feature extractor [22] [59] (Not identical, but a similar approach is transferring knowledge from an irrelevant domain, e.g., visual image recognition, to music task [19].) We report these results from random convnet features and denote it as random convnet feature. Assessing performances of random convnet feature will help to clarify the contributions of the pre-trained knowledge transfer versus the contributions of the convnet structure and nonlinear high-dimensional transformation.

2.3 Classifiers and Regressors of Target Tasks

Variants of support vector machines (SVMs) [45, 50] are used as a classifier and regressor. SVMs work efficiently in target tasks with small training sets, and outperformed K-nearest neighbours in our work for all the tasks in a preliminary experiment. Since there are many works that use hand-written features and SVMs, using SVMs enables us to focus on comparing the performances of features.

3. PREPARATION

3.1 Source Task: Music Tagging

In the source task, 244,224 preview clips of the Million Song Dataset [5] are used (201,680/12,605/25,940 for training/validation/test sets respectively) with top-50 last.fm tags including genres, eras, instrumentations, and moods. Mel-spectrograms are extracted from music signals in real-time on the GPU using Kapre [12]. Binary cross-entropy is used as the loss function during training.

For more recent information on kernel shapes for music classification, please see [40].

\[ ^2 \text{Since the average is affected by zero-padding which is applied to signals that are shorter than 29 seconds, those signals are repeated to create 29-second signals. This only happens in Task 5 and 6 in the experiment.} \]
The ADAM optimisation algorithm [25] is used for accelerating stochastic gradient descent. The convnet achieves 0.849 AUC-ROC score (Area Under Curve - Receiver Operating Characteristic) on the test set. We use the Keras [14] and Theano [51] frameworks in our implementation.

3.2 Target Tasks

Six datasets are selected to be used in six target tasks. They are summarised in Table 1.

- Task 1: The Extended ballroom dataset consists of specific Ballroom dance sub-genres.
- Task 2: The GTzan genre dataset has been extremely popular, although some flaws have been found [48].
- Task 3: The dataset size is smaller than the others by an order of magnitude.
- Task 4: Emotion prediction on the arousal-valence plane. We evaluate arousal and valence separately. We trim and use the first 29-second from the 45-second signals.
- Task 5: Excerpts are subsegments from tracks with binary labels (‘vocal’ and ‘non-vocal’). Many of them are shorter than 29s. This dataset is provided for benchmarking frame-based vocal detection while we use it as a pre-segmented classification task, which may be easier than the original task.
- Task 6: This is a non-musical task. For example, the classes include air conditioner, car horn, and dog bark. All excerpts are shorter than 4 seconds.

3.3 Baseline Feature and Random Convnet Feature

As a baseline feature, the means and standard deviations of 20 Mel-Frequency Cepstral Coefficients (MFCCs), and their first and second-order derivatives are used. In this paper, this baseline feature is called MFCCs or MFCC vectors. MFCC is chosen since it has been adopted in many music information retrieval tasks and is known to provide a robust representation. Librosa [34] is used for MFCC extraction and audio processing.

The random convnet feature is extracted using the identical convnet structure of the source task and after random weights initialisation with a normal distribution [21] but without a training.

4. EXPERIMENTS

4.1 Configurations

For Tasks 1–4, the experiments are done with 10-fold cross-validation using stratified splits. For Task 5, pre-defined training/validation/test sets are used. The experiment on Task 6 is done with 10-fold cross-validation without replacement to prevent using the sub-segments from the same recordings in training and validation. The SVM parameters are optimised using grid-search based on the validation results. Kernel type/bandwidth of radial basis function and the penalty parameter are selected from the ranges below:

- Kernel type: [linear, radial]
  - Bandwidth $\gamma$ in radial basis function: $[1/2^3, 1/2^5, 1/2^7, 1/2^9, 1/2^{11}, 1/2^{13}, 1/N_f]$
- Penalty parameter $C$: $[0.1, 2.0, 8.0, 32.0]$

A radial basis function is $\exp(-\gamma ||x - x'||^2)$, and $\gamma$ and $N_f$ refer to the radial kernel bandwidth and the dimensionality of feature vector respectively. With larger $C$, the penalty parameter or regularisation parameter, the loss function gives more penalty to misclassified items and vice versa. We use Scikit-learn [38] for these target tasks. The code for the data preparation, experiment, and visualisation are available on GitHub.

4.2 Results and Discussion

Figure 2 shows a summary of the results. The scores of the i) best performing convnet feature, ii) concatenating ‘12345’ convnet feature and MFCCs, iii) MFCC feature, and iv) state-of-the-art algorithms for all the tasks.

In all the six tasks, the majority of convnet features outperforms the baseline feature. Concatenating MFCCs

---

Table 1: The details of the six tasks and datasets used in our transfer learning evaluation.

<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset name</th>
<th>#clips</th>
<th>Metric</th>
<th>#classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1. Ballroom dance genre classification</td>
<td>Extended ballroom [32]</td>
<td>4,180</td>
<td>Accuracy</td>
<td>15</td>
</tr>
<tr>
<td>T2. Genre classification</td>
<td>GTzan genre [53]</td>
<td>1,000</td>
<td>Accuracy</td>
<td>10</td>
</tr>
<tr>
<td>T3. Speech/music classification</td>
<td>GTzan speech/music [52]</td>
<td>128</td>
<td>Accuracy</td>
<td>2</td>
</tr>
<tr>
<td>T4. Emotion prediction</td>
<td>EmoMusic (45-second) [46]</td>
<td>744</td>
<td>Coefficient of determination ($r^2$)</td>
<td>N/A</td>
</tr>
<tr>
<td>T5. Vocal/non-vocal classification</td>
<td>Jamendo [41]</td>
<td>4,086</td>
<td>Accuracy</td>
<td>2</td>
</tr>
<tr>
<td>T6. Audio event classification</td>
<td>urbansound8K [42]</td>
<td>8,732</td>
<td>Accuracy</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 2: Summary of performances of the convnet feature (blue), MFCCs (purple), and state-of-the-art (red) for Task 1-6 (State-of-the-art of Task 5 does not exist).
with ‘12345’ convnet feature usually does not show improvement over a pure convnet feature except in Task 6, audio event classification. Although the reported state-of-the-art is typically better, almost all methods rely on musical knowledge and hand-crafted features, yet our features perform competitively. An in-depth look at each task is therefore useful to provide insight.

In the following subsections, the details of each task are discussed with more results presented from (almost) exhaustive combinations of convnet features as well as random convnet features at all layers. For example, in Figure 3, the scores of 28 different convnet feature combinations are shown with blue bars. The narrow, grey bars next to the blue bars indicate the scores of random convnet features. The other three bars on the right represent the scores of the concatenation of ‘12345’ + MFCC feature, MFCC feature, and the reported state-of-the-art methods respectively. The rankings within the convnet feature combinations are also shown in the bars where top-7 and lower-7 are highlighted.

We only briefly discuss the results of random convnet features here. The best performing random convnet features do not outperform the best-performing convnet features in any task. In most of the combinations, convnet features outperformed the corresponding random convnet features, although there are few exceptions. However, random convnet features also achieved comparable or even better scores than MFCCs, indicating i) a significant part of the strength of convnet features comes from the network structure itself, and ii) random convnet features can be useful especially if there is not a suitable source task.

4.2.1 Task 1. Ballroom Genre Classification

Figure 3 shows the performances of different features for Ballroom dance classification. The highest score is achieved using the convnet feature ‘123’ with 86.7% of accuracy. The convnet feature shows good performances, even outperforming some previous works that explicitly use rhythmic features.

The result clearly shows that low-level features are crucial in this task. All of the top-7 strategies of convnet feature include the second layer, and 6/7 of them include the first layer. On the other hand, the lower-7 are {‘5’, ‘4’, ‘3’, ‘45’, ‘35’, ‘2’, ‘25’}, none of which includes the first layer. Even ‘1’ achieves a reasonable performance (73.8%).

The importance of low-level features is also supported by known properties of this task. The ballroom genre labels are closely related to rhythmic patterns and tempo [32] [49]. However, there is no label directly related to tempo in the source task. Moreover, deep layers in the proposed structure are conjectured to be mostly invariant to tempo. As a result, high-level features from the fourth and fifth layers poorly contribute to the task relative to those from the first, second, and third layers.

The state-of-the-art algorithm which is also the only algorithm that used the same dataset due to its recent release uses 2D scale transform, an alternative representation of music signals for rhythm-related tasks [33], and reports 94.9% of weighted average recall. For additional comparisons, there are several works that use the Ballroom dataset [18]. This has 8 classes and it is smaller in size than the Extended Ballroom dataset (13 classes). Laykartsis and Lerch [31] combines beat histogram and timbre features to achieve 76.7%. Periodicity analysis with SVM classifier in Gkiokas et al. [17] respectively shows 88.9%/85.6 - 90.7%, before and after feature selection.

4.2.2 Task 2. Gtzan Music Genre Classification

Figure 4 shows the performances on Gtzan music genre classification. The convnet feature shows 89.8% while the concatenated feature and MFCCs show only 78.1% and 66.0% of accuracy. Although there are methods that report accuracies higher than 94.5%, we set 94.5% as the state-of-the-art score following the dataset analysis in [48], which shows that the perfect score cannot surpass 94.5% considering the noise in the Gtzan dataset.

Among a significant number of works that use the Gtzan music genre dataset, we describe four methods in more detail. Three of them use an SVM classifier, which enables us to focus on the comparison with our feature. Arabi and Lu [1] is most similar to the proposed convnet features in a way that it combines low-level and high-level features and shows a similar performance. Beniya et al. [4] and Huang et al. [23] report the performances with many low-level features before and after applying feature selection algorithms. Only the latter outperforms the proposed method and only after feature selection.

- Arabi and Lu [1] uses not only low-level features such as {spectral centroid/flatness/roll-off/flux}, but also high-level musical features such as {beat, chord distribution and chord progressions}. The best combination of the features shows 90.79% of accuracy.
- Beniya et al. [4] uses a particularly rich set of statistics such as {mean, standard deviation, skewness, kurtosis,
covariance} of many low-level features including {RMS energy, attack, tempo, spectral features, zero-crossing, MFCC, dMFCC, ddMFCC, chromagram peak and centroid}. The feature vector dimensionality is reduced by MRMR (max-relevance and min-redundancy) \cite{39} to obtain the highest classification accuracy of 87.9%.

- Huang et al. \cite{23} adopts another feature selection algorithm, self-adaptive harmony search \cite{55}. The method uses statistics such as \{mean, standard deviation\} of many features including \{energy, pitch, and timbral features\} and their derivatives. The original 256-dimensional feature achieved 84.3% of accuracy which increases to 92.2% and 97.2% after feature selection.

- Reusing AlexNet \cite{26}, a pre-trained convnet for visual image recognition achieved 78% of accuracy \cite{19}.

In summary, the convnet feature achieves better performance than many approaches which use extensive music feature sets without feature selection as well as some of the approaches with feature selection. For this task, it turns out that combining features from all layers is the best strategy. In the results, ‘12345’ and ‘2345’ are the best configurations, and all of the top-7 scores are from those strategies that use more than three layers. On the contrary, all lower-7 scores are from those with only 1 or 2 layers. This is interesting since the majority (7/10) of the target labels already exists in source task labels, by which it is reasonable to assume that the necessary information can be provided only with the last layer for those labels. Even in such a situation, however, low-level features contribute to improving the genre classification performance.\footnote{On the contrary, in Task 5 – music emotion classification, high-level feature plays a dominant role (see Section 4.2.4).}

Among the classes of target task, classical and disco, reggae do not exist in the source task classes. Based on this, we consider two hypotheses, \(i\) the performances of those three classes may be lower than the others, \(ii\) low-level features may play an important role to classify them since high-level feature from the last layer may be biased to the other 7 classes which exist in the source task. However, both hypotheses are rebutted by comparing the performances for each genres with convnet feature ‘5’ and ‘12345’ as in Figure 5. First, with ‘5’ convnet feature, classical shows the highest accuracy while both disco and reggae show accuracies around the average accuracy reported over the classes. Second, aggregating early-layer features affects all the classes rather than the three omitted classes. This suggests that the convnet features are not strongly biased towards the genres that are included in the source task and can be used generally for target tasks with music different from those genres.

4.2.3 Task 3. Gtzan Speech/music Classification

Figure 6 shows the accuracies of convnet features, baseline feature, and state-of-the-art \cite{47} with low-level features including MFCCs and sparse dictionary learning for Gtzan music/speech classification. A majority of the convnet feature combinations achieve 100% accuracy. MFCC features achieve 99.2%, but the error rate is trivial (0.8% is one sample out of 128 excerpts).

Although the source task is only about music tags, the pre-trained feature in any layer easily solved the task, suggesting that the nature of music and speech signals in the dataset is highly distinctive.

4.2.4 Task 4. Music Emotion Prediction

Figure 7 shows the results for music emotion prediction (Task 4). The best performing convnet features achieve 0.633 and 0.415 \(r^2\) scores on arousal and valence axes respectively.

On the other hand, the state-of-the-art algorithm reports 0.704 and 0.500 \(r^2\) scores using music features with a recurrent neural network as a classifier \cite{56} that uses 4,777 audio features including many functionals (such as quantiles, standard deviation, mean, inter peak distances) of 12 chroma features, loudness, RMS Energy, zero crossing rate, 14 MFCCs, spectral energy, spectral roll-off, etc.

For the prediction of arousal, there is a strong dependency on the last layer feature. All top-7 performances are from the feature vectors that include the fifth layer. The first layer feature also seems important, since all of the top-5 strategies include the first and fifth layer features. For
valence prediction, the third layer feature seems to be the most important one. The third layer is included in all of the top-6 strategies. Moreover, ‘3’ strategy was found to be best performing among strategies with single layer feature.

To summarise the results, the predictions of arousal and valence rely on different layers, for which they should be optimised separately. In order to remove the effect of the choice of a classifier and assess solely the effect of features, we compare our approach to the baseline method of [56] which is based on the same 4,777 features with SVM, not a recurrent neural network. The baseline method achieves .541 and .320 $r^2$ scores respectively on arousal and valence, both of which are lower than those achieved by using the proposed convnet feature. This further confirms the effectiveness of the proposed convnet features.

4.2.5 Task 5. Vocal/non-vocal Classification

Figure 8 presents the performances on vocal/non-vocal classification using the Jamendo dataset [41]. There is no known state-of-the-art result, as the dataset is usually used for frame-based vocal detection/segmentation. Pre-segmented Excerpt classification is the task we formulate in this paper. For this dataset, the fourth layer plays the most important role. All the 14 combinations that include the fourth layer outperformed the other 14 strategies without the fourth layer.

4.2.6 Task 6. Acoustic Event Detection

Figure 9 shows the results on acoustic event classification using Urbansound8K dataset [42]. Since this is not a music-related task, there are no common tags between the source and target tasks, and therefore the final-layer feature is not expected to be useful for the target task.

The strategy of concatenating ‘12345’ convnet features and MFCCs yields the best performance. Among convnet features, ‘2345’, ‘12345’, ‘123’, and ‘234’ achieve good accuracies. In contrast, those with only one or two layers do not perform well. We were not able to observe any particular dependency on a certain layer.

Since the convnet features are trained on music, they do not outperform a dedicated convnet trained for the target task. The state-of-the-art method is based on a deep convolutional neural network with data augmentation [43]. Without augmenting the training data, the accuracy of convnet in the same work is reported to be 74%, which is still higher than our best result (71.4%).

The convnet feature still shows better results than conventional audio features, demonstrating its versatility even for non-musical tasks. The method in [42] with \{minimum, maximum, median, mean, variance, skewness, kurtosis\} of 25 MFCCs and \{mean and variance\} of the first and second MFCC derivatives (225-dimensional feature) achieved only 68% accuracy using the SVM classifier. This is worse than the performance of the best performing convnet feature.

It is notable again that unlike in the other tasks, concatenating convnet feature and MFCCs results in an improvement over either a convnet feature or MFCCs (71.4%). This suggests that they are complementary to each other in this task.

5. CONCLUSIONS

We proposed a transfer learning approach using deep learning and evaluated it on six music information retrieval and audio-related tasks. The pre-trained convnet was first trained to predict music tags and then aggregated features from the layers were transferred to solve genre classification, vocal/non-vocal classification, emotion prediction, speech/music classification, and acoustic event classification problems. Unlike the common approach in transfer learning, we proposed to use the features from every convolutional layers after applying an average-pooling to reduce their feature map sizes.

In the experiments, the pre-trained convnet feature showed good performance overall. It outperformed the baseline MFCC feature for all the six tasks, a feature that is very popular in music information retrieval tasks because it gives reasonable baseline performance in many tasks. It also outperformed the random-weights convnet features for all the six tasks, demonstrating the improvement by pre-training on a source task. Somewhat surprisingly, the performance of the convnet feature is also very competitive with state-of-the-art methods designed specifically for each task. The most important layer turns out to differ from task to task, but concatenating features from all the layers generally worked well. For all the five music tasks, concatenating MFCC feature onto convnet features did not improve the performance, indicating the music information in MFCC feature is already included in the convnet feature. We believe that transfer learning can alleviate the data sparsity problem in MIR and can be used for a large number of different tasks.

\[\text{\textit{Transfer learning targeting audio event classification was recently introduced in [2, 3] and achieved a state-of-the-art performance.}}\]
6. REFERENCES


