# Multi-Level and Multi-Scale Feature Aggregation Using Sample-level Deep Convolutional Neural Networks for Music Classification

**Jongpil Lee** [1] **Juhan Nam** [1]

## Abstract

Music tag words that describe music audio by text have different levels of abstraction. Taking this issue into account, we propose a music classification approach that aggregates multi-level and multi-scale features using pre-trained feature extractors. In particular, the feature extractors are trained in sample-level deep convolutional neural networks using raw waveforms. We show that this approach achieves state-of-the-art results on several music classification datasets.

## 1. Introduction

Learning hierarchical audio representations for music classification in an end-to-end manner is a challenge due to the diversity of music description words. In this study, we combine two previously proposed methods to tackle the problem.

### 1.1. Multi-Level and Multi-Scale Feature Aggregation

Music classification tasks, particularly music auto-tagging among others, have a wide variety of labels in terms of genre, mood, instruments and other song characteristics. In order to address different levels of abstraction that the labels retain, we recently proposed an approach that aggregates audio features extracted in a multi-level and multi-scale manner (Lee & Nam). The method is composed of three steps: extracting features using pre-trained convolutional neural networks (CNNs), feature aggregation and song-level classification. The CNNs are trained in a supervised manner with the tag labels, taking different sizes of input frames. The feature aggregation step extracts multiple-level features using the pre-trained CNNs and summarizes them into a single song-level feature vector. The last step performs final predictions of tags from the aggregated features using a fully-connected neural network. This multi-step architecture has the advantage of capturing local and global characteristics of a song and also has a good accordance with transfer learning. However, our previous approach used mel-frequency spectrograms as input, which are based on the knowledge of pitch perception.

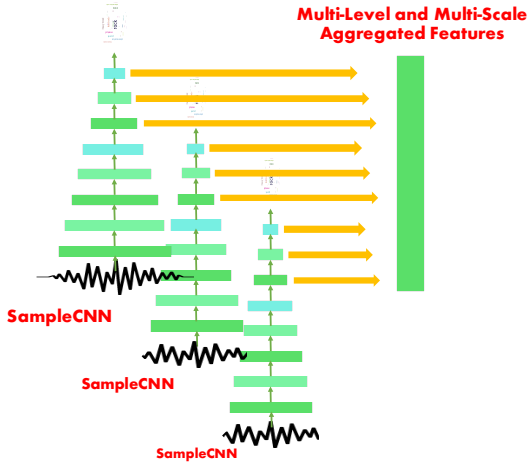


*Figure 1.* Multi-level and multi-scale feature aggregation using sample-level deep convolutional neural networks

### 1.2. Sample-level Deep Convolutional Neural Networks

We recently investigated the possibility of employing raw waveforms as input for deep convolutional neural networks (DCNNs) in music auto-tagging (Lee et al., 2017). They were configured to take a very small grain of waveforms, even 2 or 3 samples, in the bottom-level filters. They show that the "sample-level" representation learning works well and the learned filters in each layer are sensitive to log-scaled frequency along layer such as mel-frequency spectrogram.

### 1.3. The combination

In this study, we combine the two methods to take all the advantages of them. As illustrated in Figure 1, we used the top three hidden layers from the sample-level DCNNs for multi-level feature extraction. The DCNNs take different sizes of input.

[1]Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. Correspondence to: Juhan Nam <juhannam@kaist.ac.kr>.

*Proceedings of the $34^{th}$ International Conference on Machine Learning*, Sydney, Australia, PMLR 70, 2017. 

*Table 1.* Comparison with previous work. Note that we used only multi-level features in the proposed work due to the long training time with MSD (it took about two weeks on the GTX1080Ti GPU to train the DCNN and so we used only the $3^9$ model). Also, "-$n$ LAYER" indicates $n$ layers below from the output.

| | MODEL | GTZAN (ACC.) | MTAT (AUC) | TAGTRAUM (ACC.) | MSD (AUC) |
|---|---|---|---|---|---|
| (LEE & NAM) | MULTI-LEVEL AND MULTI-SCALE FEATURES (PRE-TRAINED WITH MSD) | 0.720 | 0.9021 | 0.766 | **0.8878** |
| (LEE ET AL., 2017) | SAMPLE-LEVEL DCNN ($3^9$ MODEL) | - | 0.9055 | - | 0.8812 |
| PROPOSED WORK | -3 LAYER (PRE-TRAINED WITH MSD) | 0.778 | 0.8988 | 0.760 | 0.8831 |
| (FEATURES FROM | -2 LAYER (PRE-TRAINED WITH MSD) | 0.811 | 0.8998 | 0.768 | 0.8838 |
| SAMPLE-LEVEL | -1 LAYER (PRE-TRAINED WITH MSD) | **0.821** | 0.8976 | 0.768 | 0.8842 |
| DCNN, $3^9$ MODEL) | LAST 3 LAYERS (PRE-TRAINED WITH MSD) | 0.805 | 0.9018 | **0.768** | 0.8842 |

*Table 2.* Comparison of various multi-scale feature combinations. Only MTAT was used.

| FEATURES FROM SAMPLE-LEVEL DCNNS LAST 3 LAYERS (PRE-TRAINED WITH MTAT) | MTAT |
|---|---|
| $3^9$ MODEL | 0.9046 |
| $3^8$ AND $3^9$ MODELS | 0.9061 |
| $2^{13}$, $2^{14}$, $3^8$ AND $3^9$ MODELS | 0.9061 |
| $2^{13}$, $2^{14}$, $3^8$, $3^9$, $4^6$, $4^7$, $5^5$ AND $5^6$ MODELS | **0.9064** |

## 2. Datasets

We validate the effectiveness of the proposed method on different sizes of datasets for genre classification and auto-tagging. The details are as follows[1]:

- GTZAN (fault-filtered version) (Tzanetakis & Cook, 2002; Kereliuk et al., 2015): 930 songs, genre classification (10 genres)
- MagnaTAgaTune (MTAT) (Law et al., 2009): 21105 songs, auto-tagging (50 tags)
- Million Song Dataset with Tagtraum genre annotations (TAGTRAUM, stratified split with 80% training data of CD2C version) (Schreiber, 2015): 189189 songs, genre classification (15 genres)
- Million Song Dataset with Last.FM tag annotations (MSD) (Bertin-Mahieux et al., 2011): 241889 songs, auto-tagging (50 tags)

## 3. Results & Conclusion

We obtained the results from the average of 10 experiments. From Table 1, although the proposed method failed to outperform the best of the previous works on MSD and MTAT, the multi-level and multi-scale aggregation generally improves the performance. The improvement is particularly dominant in GTZAN. From Table 2 where only MTAT is used, the proposed method is superior to the two previous works. Furthermore, we visualize the features at different levels for selected tags in the Figure 2. Songs with genre tag (*Techno*) are more closely clustered in the higher layer (-1 layer). On the other hand, songs with instrument tag (*Piano*) are more closely clustered in the lower layer (-3 layer). This may indicate that the optimal layer of feature representations can be different depending on the type of labels. All of these results show that the proposed feature aggregation method is also effective with the sample-level DCNNs.

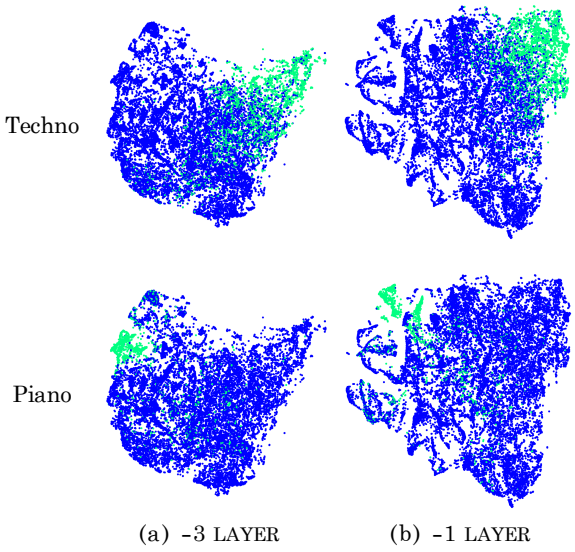


*Figure 2.* This figure shows feature visualization on songs with Piano tag and songs with Techno tag on MTAT using t-SNE. Features are extracted from (a) -3 LAYER and (b) -1 LAYER of the $3^9$ model pre-trained with MSD.

[1]https://github.com/jongpillee/music_dataset_split

## References

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset. In *ISMIR*, volume 2, pp. 10, 2011.

Kereliuk, C., Sturm, B. L., and Larsen, J. Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17(11):2059–2071, 2015.

Law, E., West, K., Mandel, M. I., Bay, M., and Downie, J. S. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pp. 387–392, 2009.

Lee, J. and Nam, J. Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, 24(8):1208–1212.

Lee, J., Park, J., Kim, K. L., and Nam, J. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. *Sound and Music Computing Conference (SMC)*, pp. 220–226, 2017.

Schreiber, H. Improving genre annotations for the million song dataset. In *ISMIR*, pp. 241–247, 2015.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.