Automatic Music Mood Detection Using Transfer Learning and Multilayer Perceptron

Bhuwan Bhattarai and Joonwhoan Lee

Chonbuk National University, Jeonju, Korea

Ijfis

Abstract

This paper proposes an automatic mood detection of music with a composition of transfer learning and multilayer. The five layered convolutional neural network pre-trained on Million Song dataset is used to extract the features from EmoMusic dataset. We obtain a set of features from the different five layers, which is fed into multilayer perceptron (MLP)-based regression. Through the regression network we estimate the mood of music on Thayer's two-dimensional emotion space, which consists of the axes corresponding to arousal and valence. Because the EmoMusic dataset does not provide enough number of data for training, we augment the data by time stretching to make it tripled. We perform the experiment with the augmented data as well as the original EmoMusic dataset. Box and whisker plot along with the mean of 10-fold cross-validation has been used for evaluating the proposed mood detection. In terms of the percentage of R^2 score for measure of accuracy, the proposed MLP shows state-of-the-art estimates for the augmented EmoMusic dataset.

Keywords: Multilayer perceptron (MLP), Mood detection, Music, Transfer learning, Global average pooling

1. Introduction

People enjoy music in their leisure time. Now a days there is more and more music on personal computers, on mobile phone and on the Internet. The huge number of music can be specified by mood, which can help us to easily retrieve a proper set of music. There are two major aspects for automatically evaluating mood of music; one is mood classification [1, 2] and the other is mood regression. In the mood classification the category named by an adjective term is automatically given to a song. In the mood regression the mood is evaluated by values on the proper scales, each of which represents the degree of a specific emotion state.

Whatever types of the detection we take, it is necessary to specify an appropriate emotion space to evaluate the mood. Even though there is no standardized space, however, Thayer's two-dimensional emotion space model as in Figure 1 is well recognized as such space, in which the fundamental aspects of music mood are represented in a two-dimensional space of valence and arousal [3]. In the model, valence axis describes the continuous scale from pleasantness (happy, pleased, hopeful, etc.) to unpleasantness (unhappy, annoyed, hopeful, etc.). The arousal axis represents the degree of calming or exciting. Each axis spans the bipolar scale values from -1 to 1 . The emotion is represented by a point or region [4] based on the values of coordinates, each of which corresponds to the positive or negative degrees of the feelings. Sometimes the mood categories can be designated in the two-dimensional emotion

Received: Mar. 1, 2019
 Revised : Jun. 7, 2019
 Accepted: Jun. 21, 2019

Correspondence to: Joonwhoan Lee
 (chlee@chonbuk.ac.kr)
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Figure 1. Thayer's two-dimensional emotion space named arousal-valence space [3].

space as in Figure 1 [5].

The authors [2, 6] concludes that number of musical features like tempo, pitch, rhythm and harmony makes listener perceive and correctly identify a specific intended emotional expressions of mood like sad, fear, humorous, happy, and exiting. The study of lyric-based music mood recognition [7] has been performed to observe the relationship between content of lyrics and mood using OPM songs. KeyGraph keyword generation algorithm and word level features like TF-IDF have been examined by using various parameters and threshold to extract the keywords from lyrics. Automatic music mood recognition using support vector regression [8, 9] is studied by mapping the various music features into Thayer's two-dimentional emotion space model which can later predict the values for valence and arousal. Similarly a 5-layer convolutional neural network (CNN) trained on Million Song dataset [10] for music tagging is used to extract the features to detect the music mood with respect to valence and arousal [11]. Similar audio network [?] pre-trained on Million Song dataset is used to classify the GTZAN, 1517-Artists, Unique and Magnatagatune datasets.

The deep neural network needs large amount of labeled dataset, but music information retrieval community (MIR) still have scarcity of such data especially for the music mood. Transfer learning also called inductive transfer is one of the way that can resolve the problem of limited number of labelled data set. Transfer learning reuses knowledge of the source domain to solve a new task in the target domain. For example, knowledge

gained while recognize cars could apply when trying to recognize trucks [13]. Transfer learning has been applied successfully in many domains such as visual object recognition [14], web document classification [15, 16], and human action recognition [17]. There were several studies [18–21] in the field of machine learning using transfer learning approach. A natural RGB images to malignancy classification is used in [22] for contrast-enhanced MR images. The POS tagging on Penn Treebank data is used to transfer the knowledge of source task to target task POS tagging on microblogs with limited available annotation for sequence tagging data [23]. Similarly, the patterns of human epithelial type 2 (HEp-2) cells are studied by applying transfer learning from a pre-trained deep CNN to extract the generic features [24].

In this work, we extract the features of original and augmented EmoMusic dataset by reusing the layer-wise weights from the pre-trained network [11]. The pre-trained network is trained on Million Song dataset for music tagging. After extracting features, multilayer perceptron (MLP) with three or four hidden layers trained with 50% dropout is used to detect the scores in valence and arousal. In the experiment the number of data is increased by augmenting the original EmoMusic dataset using time stretching. The ground truth label for augmented data in valence and arousal dimension is set to be identical as original data. The experimental result of 10-fold cross-validation exhibits that MLP with the original EmoMusic dataset is similar to state-of-the-art accuracy in SOA(1) [11] and SOA(2) [25]. The highest R^2 -score of 79.64% for valence and 85.61% for arousal is achieved by the augmented EmoMusic dataset which is better than the state of art in EmoMusic dataset.

The organization of this paper is as follows. In Section 2, the methods of audio, feature extraction, mood detection using the MLP for regression and data augmentation method are described. Section 3 presents about the experimental results. Finally, we draw the conclusion in Section 4.

2. Methods

2.1 Feature Extraction

The audio file from the EmoMusic dataset is preprocessed using Librosa library [26] to generate the Mel-spectrogram. The single channel audio files amounts to the length of 29 seconds, with sampling rate of 12,000 Hz. The size of FFT is 512 with the hop length of 256, which produces 1,360 frames for a song. The number of Mel-bins is 96 so that the size of the generated

Mel-spectrogram is 96×1360 , which is fed into the pre-trained network for feature extraction.

2.1.1 Pre-trained network

The pre-trained CNN has the goal to automatically generate multiple tagging. The network has been trained with Million Song data set, which gives the tags among 50. In each layer this pre-trained convolutional network consists of consecutive 5 blocks, each of which has a set of 3×3 convolution filters, max-pooling operation, and batch normalization, and then followed by a fully connected layer for the multiple tagging [11]. Every unit in each layer has exponential linear unit (ELU) as activation function. The 50 units of the last fully connected layer have the sigmoidal activation functions to generate class-belongings corresponding to the tags.

Figure 2 shows the CNN after removing the fully connected layer for the tag generation which is nothing to do with the feature extraction. The structure of convolutional network for feature extraction is summarized in Table 1.

2.1.2 Feature extraction of EmoMusic from pre-trained network

After preprocessing the 96×1360 Mel-spectrogram of EmoMusic dataset is fed into the pre-trained network to generate feature maps. Each feature maps has its own size and each units

Figure 2. The structure of CNN for feature extraction.

in a map has its own receptive field in Mel-frequency bin and time dimensions.

We assume that the whole music has consistently continued the same mood. Note that one can divide the whole music into several segments, try to detect the mood for each time segments, and summarize every mood of the segment to obtain the mood of whole music.

Because there are 32 channels with different set of filters in each layer, we can take 32 features by global average pooling. The 32 features from each of 5 layers can be separately exploited or concatenated to obtain a higher dimensional features. Therefore, the maximum dimension of the output feature is 160 when we concatenate all five layers.

Note that one can divide a feature map into several groups according to the Mel-frequency bin. In this case, we could take more features for a map. In the paper we do not consider the method, because there might be too huge number of features and it is necessary to consider the feature reduction technique.

2.2 Mood Detection by Multilayer Perceptron and Its Training

After obtaining the features from the pre-trained network, the mood regressor is constructed by a MLP. By the empirical experiment we set three or four hidden layers, and one output layer as shown in Figure 3. The numbers of units in the hidden layers are also empirically set according to the data size in the experiment. The last layer has two nodes, each of which evaluates the mood in the Thayer's two-dimensional scales of valence and arousal. Because the value of each bipolar dimension spans $[-1, 1]$, the tangent hyperbolic is adapted as activation function [27]. We also construct same size MLP for both original and augmented data in our experiment. However, we didn't mention the result because the amount of original data is not sufficient for huge size MLP so that resulting accuracy is decrease and network seems to be overfitted.

To avoid over fitting in the training, 50% of dropout [28] is applied to the units in each hidden layer. The network is trained

Table 1. Structure of CNN for feature extraction

Layer	Max-pooling	Filter-size	Channels	Activation
1	2×4	3×3	32	ELU
2	4×4	3×3	32	ELU
3	4×5	3×3	32	ELU
4	2×4	3×3	32	ELU
5	4×4	3×3	32	ELU

Figure 3. Multilayer perceptron for mood detector (See Tables 3 and 4).

with 10-fold cross-validation with stratified splits. Adam optimizer [29] is used for updating weight and bias to minimize the mean squared error (MSE) loss function. During training batch size of 100 and epochs of 300 are used.

As mentioned earlier there are 32 features from each layer of pre-trained network by taking the global average pooling. In the experiment, we compare the performance for the hierarchical features from each different layer of pre-trained CNN. In addition to the 5 sets of 32-dimensional features, a set of all the concatenated features from 5 layers is considered to compare the performance.

For the comparison, we use each of these six sets of features for training and obtain the percentage of R^2 scores, separately. The R^2 score in Eq. (2) is the normalized version of the mean squared error in Eq. (1). In the equations, y_i , $\hat{y}_i$, and $\bar{y}$ represent the i -th true value, i -th predicted value, and the average of truth values over n data, respectively.

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1)$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (2)$$

2.3 Data Augmentation

In the data-driven approach, the large number of data is essential to avoid overfitting. Therefore, the augmentation is required

to artificially increase the number of data, especially when the number of data is not sufficient. Because the expensive psychological experiment is necessary to obtain the set of labelled data for mood detection, the size of labeled set is usually not so big for the data-driven approach.

The EmoMusic dataset has only 744 labelled data for mood regression. Therefore, we try to increase the number of EmoMusic dataset by time stretching with the modest scaling rates. Note that the too much shortened or stretched audio data can change pitch and tempo, which again makes the mood change differently from the original data. We use a randomly selected scaling factor in the small interval around 1.0 as $[0.95, 1.05]$. Note that the larger(less) than 1 scaling factor produces time stretched(shortened) audio data. After augmentation, we can have 3×744 new EmoMusic data including two times of time stretched and the original data.

We use Librosa library for increasing the number of data by the time stretching. For time-stretching the EmoMusic dataset, at first, short time Fourier transform (STFT) of audio data is computed. Then we draw two randomly chosen stretching or shrinking rates from the uniform distribution in the interval.

In order to perform the time stretching, we take STFT of audio data. Then the STFT matrix of time-frequency representation is stretched or shorten based on the phase vocoder implemented in [30]. After the time stretching, the resulting time-frequency representation can be transformed back into the audio data in time dimension with the inverse STFT (ISTFT).

Each music in EmoMusic dataset used in our experiment has its own emotionally evaluated scores in two scales of valence and arousal in Thayer's emotion space. There are eight genres of music included in EmoMusic data set including Blues, Electronic, Rock, Classical, Folk, Jazz, Country and Pop, which is annotated via crowdsourcing named Amazon's Mechanical Turk.

The two types of dataset are used in our experiment as in Table 2. One is original EmoMusic dataset [31] and another is our augmented EmoMusic data. The original dataset consists of 744 audio data and another data set additionally includes the 2×744 time stretched data. The emotional score for augmented data is the same as original data, which is a pair of continuous values from 1 to 9 in arousal and valence scales. In order to adapt to tangent hyperbolic activation function, each of the score is normalized from -1 to 1 .

3. Experimental Results with Discussion

Table 2. EmoMusic and augmented data sets for experiment

Experiments	Included data	No. of data
Experiment 1	Original EmoMusic	744
Experiment 2	Original EmoMusic +Augmented data	2232

Table 3. MLP structure for experiment 1 of original data

Layer	No. of nodes	Activation	Dropout (%)
Input	32 or 60	-	-
Hidden 1	200	ReLu	50
Hidden 2	100	ReLu	50
Hidden 3	20	ReLu	50
Output	2	Tanh	-

3.1 Experimental Results

The R^2 scores of ten-fold cross-validation is visualized using box and whisker plot [32], which is convenient for quickly summarizing the results. A box and whisker plot for a symmetrically distributed data has the mean close to the median line [33].

In each experiments in Table 2, *layer_1*, *layer_2*, *layer_3*, *layer_4* and *layer_5* represent the separate 32-dimensional features from layer one to layer five, and *layer_all* denotes the concatenated 160-dimensional features of all layers as in Figure 2. The mean R^2 score over 10-folds is labelled by a green triangle located inside the box in both valence and arousal plots.

3.1.1 Results from experiment 1

The original data of EmoMusic contains only 744 excerpts of music which is limited with MLP for large number of hidden nodes. So, we take a small sized network in number of hidden layers and the number of nodes in each hidden layer. The structure of MLP from the original data is summarized in Table 3.

The results of experiment 1 are shown in Figure 4. The percentage of R^2 score for the mean over 10-folds is calculated by Eq. (2). As denoted in Figure 4, *layer_all* for both valence and arousal achieves the highest score of 51.88% and 67.07%, respectively. All the layers in arousal provides a similar result, that is, *layer_1*, *layer_2*, *layer_3*, *layer_4* and *layer_5* has only 7%, 4%, 2%, 1% and 3% less than *layer_all*, respectively. Similarly, *layer_1* and *layer_5*, in valence produce relatively low scores which are 12% and 8% less than *layer_all*, respectively. The result of valence in 10-folds has a large variation compared with arousal. *layer_3* and *layer_4* of valence has the highest

Figure 4. Results of MLP for experiment 1 of the original data.

variation whereas, *layer_1* *layer_4* and *layer_5* has lowest variation in arousal. We can conclude that the assorted features from all layers provide the best results, and there is no distinctive difference in the features from the layers except *layer_1* for mood estimation in the MLP regressor.

3.1.2 Results from experiment 2

Table 4 summarizes the structure of MLP for experiment 2. We add an additional hidden layer and increase the number of nodes in the hidden layers, because the number of data is increased in the experiment.

The results of the experiment are shown in Figure 5. The highest scores of the mean over 10-folds are achieved by *layer_all*, which are 79.64% in valence and 85.61% in arousal. In addition, the lowest score is obtained by *layer_1* which is 37.75% in valence and 55.70% in arousal. Similarly, the intermediate layers of *layer_2*, *layer_3*, and *layer_4* produce 50.32%, 62.39% and 72.04%, respectively. For arousal those intermediate layers achieve 63.45%, 70.28% and 78.79%, respectively. A close result is obtained from *layer_5* compared with *layer_all*, which is around 3% less for both valence and arousal. The box plot indicates that the distribution of 10-fold results in arousal is more symmetric and lumped than that in valence.

3.2 Discussion

From the results of our experiments we can conclude that higher layer features as well as assorted features of all layers are better than lower layer features. The box and whisker plot in Figures 4 and 5 defines the distribution of R^2 score from 10-fold data. In

Table 4. MLP structure for experiment 2 of augmented data

Layer	No. of nodes	Activation	Dropout (%)
Input	32 or 160	-	-
Hidden 1	1000	ReLu	50
Hidden 2	500	ReLu	50
Hidden 3	100	ReLu	50
Hidden 4	20	ReLu	50
Output	2	Tanh	-

Figure 5. Results of MLP for experiment 2 of augmented data.

most of the cases, for augmented data, the mean approximately coincides with the median line, which means the distribution of the 10-fold results is symmetric. However, the valence in experiment 1, the interquartile range (IQR) and length of two whisker is so large, which implies the results obtained from 10-folds have large variations. In contrast, there is less variation in the results of experiment 2.

The results from experiment 2 are smaller IQR and shorter length of two whiskers than those from experiment 1, which implies the augmented data by time stretching is valid to obtain stable mood detection in the MLP-based regressor. This results coincide well with the conclusion that deep learning and other modern nonlinear machine learning techniques can obtain better result with data augmentation [34, 35].

Considering the stable results from experiment 2, we can carefully conclude that the features from the higher layer provides better results in mood estimation of music. Usually, the CNN features in higher layer have wider receptive fields. This implies the features extracted from longer time durations are more effective to detect the mood of music.

Table 5. Comparison with SOA(1) and SOA(2)

	Valence	Arousal
MLP (original)	51.88	67.07
MLP (augmented)	79.64	85.61
SOA(1)	45.72	65.31
SOA(2)	50.0	70.0

3.3 Result of Comparison

We compare the mean over 10-fold results from *layer_all* in Table 4 with SOA(1) [11] and SOA(2) [25]. The SOA(1) adopted transfer learning technique to solve the mood detection of EmoMusic dataset and achieved the highest R^2 score of 65.6% and 46.2% in valence and arousal scales, respectively. On the other hand, SOA(2) reports R^2 score of 70.4% and 50.0% using music features with a recurrent neural network as a classifier. SOA(2) exploits 4777 audio features including quantiles, mean, standard deviation, zero crossing rate, MFCC, spectral energy, etc.

In Table 5, MLP with the augmented data performs the best scores of 79.64% and 85.61% in valence and arousal scales, respectively. Similarly, MLP with original data in experiment 1 also performs better result in valence, which is around 6% higher than SOA(1), and 2% higher than SOA(2). But the score in arousal is 3% less than SOA(2) and 2% higher than SOA(1).

4. Conclusion

This paper proposes an automatic mood detection of music with a composition of transfer learning and multilayer perceptron. The 5-layered CNN pre-trained on Million Song dataset is used to extract the features from EmoMusic dataset. We obtain a set of features from the different five layers, which is fed into MLP-based regression. Through the regression network we estimate the mood of music on Thayer's two-dimensional emotion space, which consists of the axes corresponding to arousal and valence. Because the EmoMusic dataset does not provide enough number of data for training, we augment the data by time stretching to make it tripled. We perform the experiment with the augmented data as well as the original EmoMusic dataset. Box and whisker plot along with the mean of ten-fold cross-validation has been used for evaluating the proposed mood detection. In terms of the percentage of R^2 score for measure of accuracy, the proposed MLP shows 79.64% and 85.61% in valence and arousal scales, respectively, which is the state of art for EmoMusic dataset.

Based upon our results we conclude that data augmentation

technique plays a significant role for stable and efficient detection of music mood. In addition, the features taken from the convolution filters with larger receptive field is more effective for the mood detection in MLP-based regressor.

We believe that the performance can be further improved by the ensemble strategy of time segmented emotional evaluations, and the deliberate feature selection method.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The research leading to these result, authors would like to thank Korean Ministry of Education for funding. The National Research Foundation of Korea (NRF) support us under Basic Science Research Program (NRF-2015R1D1A1A01058062).

References

- [1] C. Laurier and P. Herrera, "Audio music mood classification using support vector machine," Music Technology Group of the Universitat Pompeu Fabra, Barcelona, Spain, 2007.
- [2] B. K. Baniya and J. Lee, "Rough set-based approach for automatic emotion classification of music," *Journal of Information Processing Systems*, vol. 13, no. 2, pp. 400-41, 2017. <http://doi.org/10.3745/JIPS.04.0032>
- [3] R. E. Thayer, *The Biopsychology of Mood and Arousal*. New York, NY: Oxford University Press, 1989.
- [4] L. Zhang, D. Tjondronegoro, and V. Chandran, "Representation of facial expression categories in continuous arousal-valence space: feature and correlation," *Image and Vision Computing*, vol. 32, no. 12, pp. 1067-1079, 2014. <https://doi.org/10.1016/j.imavis.2014.09.005>
- [5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980. <http://doi.org/10.1037/h0077714>
- [6] D. Vastfjall, "Emotion induction through music: a review of the musical mood induction procedure," *Musicae Scientiae*, vol. 5(1_suppl), pp. 173-211, 2001. <http://doi.org/10.1177/10298649020050S107>
- [7] I. V. Ascalon and R. Cabredo, "Lyric-based music mood recognition," in *Proceedings of the DLSU Research Congress*, Manila, Philippines, 2015.
- [8] M. Sarode and D. G. Bhalke, "Automatic music mood recognition using support vector regression," *International Journal of Computer Applications*, vol. 163, no. 5, pp. 32-35, 2017.
- [9] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448-457, 2008. <http://doi.org/10.1109/TASL.2007.911513>
- [10] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, Miami, FL, 2011.
- [11] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," in *Proceedings of the 18th International Society of Music Information Retrieval Conference*, Suzhou, China, 2017.
- [12] A. Van Den Oord, S. Dieleman, and B. Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," in *Proceedings of the 15th International Society for Music Information Retrieval Conference*, Taipei, Taiwan, 2014.
- [13] J. West, D. Ventura, and S. Warnick, "Spring research presentation: a theoretical foundation for inductive transfer," College of Physical and Mathematical Sciences, Brigham Young University, Provo, UT, 2007.
- [14] T. Tommasi, N. Quadrianto, B. Caputo, and C. H. Lampert, "Beyond dataset bias: multi-task unaligned shared knowledge transfer," in *Computer Vision-ACCV 2012*. Heidelberg: Springer, 2012, pp. 1-15. https://doi.org/10.1007/978-3-642-37331-2_1
- [15] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007, pp. 759-766. <http://doi.org/10.1145/1273496.1273592>
- [16] H. Al-Mubaid and S. A. Umair, "A new text categorization technique using distributional clustering and learning logic," *IEEE Transactions on Knowledge & Data*

- Engineering*, vol. 18, no. 9, pp. 1156-1165, 2006. <http://doi.org/10.1109/TKDE.2006.135>
- [17] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *Proceedings of 2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 463-469. <http://doi.org/10.1109/IJCNN.2017.7965890>
- [18] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2009. <http://doi.org/10.1109/TKDE.2009.191>
- [19] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang, "Transfer learning using computational intelligence: a survey," *Knowledge-Based Systems*, vol. 80, pp. 14-23, 2015. <https://doi.org/10.1016/j.knosys.2015.01.010>
- [20] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 2208-2217.
- [21] P. Hamel, M. E. Davies, K. Yoshii, and M. Goto, "Transfer learning in MIR: sharing learned latent representations for music audio classification and similarity," in *Proceedings of the 14th International Conference on Music Information Retrieval*, Curitiba, Brazil, 2013, pp. 9-14.
- [22] C. Haarbuerger, P. Langenberg, D. Truhn, H. Schneider, J. Thuring, S. Schrading, C. K. Kuhl, and D. Merhof, "Transfer learning for breast cancer malignancy classification based on dynamic contrast-enhanced MR images," in *Bildverarbeitung für die Medizin 2018*. Heidelberg: Springer Vieweg, 2018, pp. 216-221. https://doi.org/10.1007/978-3-662-56537-7_61
- [23] Z. Yang, R. Salakhutdinov, and W. W. Cohen, "Transfer learning for sequence tagging with hierarchical recurrent networks," in *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France, 2017.
- [24] H. T. H. Phan, A. Kumar, J. Kim, and D. Feng, "Transfer learning of a convolutional neural network for HEp-2 cell image classification," in *Proceedings of 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, Prague, Czech Republic, 2016, pp. 1208-1211. <https://doi.org/10.1109/ISBI.2016.7493483>
- [25] F. Weninger, F. Eyben, and B. Schuller, "On-line continuous-time music mood regression with deep recurrent neural networks," in *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 5412-5416. <https://doi.org/10.1109/ICASSP.2014.6854637>
- [26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, Austin, TX, 2015, pp. 18-25.
- [27] P. M. M. R. Pushpa and K. Manimala, "Implementation of hyperbolic tangent activation function in VLSI," *International Journal of Advanced Research in Computer Science & Technology*, vol. 2, pp. 225-228, 2014.
- [28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929-1958, 2014.
- [29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, 2015.
- [30] D. Ellis, "A phase vocoder in MATLAB," 2003, Available <https://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc/>
- [31] M. Soleymani, M. N. Caro, E. M. Schmidt, C. Y. Sha, and Y. H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, Barcelona, Spain, 2013, pp. 1-6. <https://doi.org/10.1145/2506364.2506365>
- [32] R. L. Nuzzo, "The box plots alternative for visualizing quantitative data," *PM&R*, vol. 8, no. 3, pp. 268-272, 2016. <https://doi.org/10.1016/j.pmrj.2016.02.001>
- [33] D. C. Li, W. T. Huang, C. C. Chen, and C. J. Chang, "Employing box plots to build high-dimensional manufacturing models for new products in TFT-LCD plants," *Neurocomputing*, vol. 142, pp. 73-85, 2014. <https://doi.org/10.1016/j.neucom.2014.03.043>
- [34] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," 2017, Available <https://arxiv.org/abs/1712.04621>

- [35] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification,” *Neurocomputing*, vol. 321, pp. 321-331, 2018. <https://doi.org/10.1016/j.neucom.2018.09.013>

Bhuwan Bhattarai receives his B.S. degree from Patan Multiple Campus (affiliation of Tribhuvan University) of Nepal, in computer science and information technology in 2015. He receives his M.S. degree in computer science and engineering from Chonbuk National University, Korea, in 2019. He is currently a doctoral student in Artificial Intelligence Laboratory in Chonbuk National University. His research includes music information retrieval (MIR), image processing, object detection in images and music source separation.
E-mail: bhubon240@gmail.com

Joonwhoan Lee received the B.E. degree in electronic engineering from Hanyang University, Korea, in 1980, the M.S. degree in electric and electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), in 1982 and Ph.D. degree in electrical and computer engineering from University of Missouri, Columbia, MO, in 1990, respectively. In 1985, he joined the Department of Electronic Engineering, Chonbuk National University, Korea, and has been working as a professor in the Department of Computer Science and Engineering of the same university from 1998. Also, he has stayed as a visiting scholar in the School of Computing Science of Simon Fraser University, Canada during his sabbatical leave from August 2013 to August 2014. Now, he is the director of AI research Institute of Chonbuk National University. His research interests include machine learning applications of signal processing, and fuzzy and rough sets.
E-mail: chlee@chonbuk.ac.kr