Music Emotion Recognition via End-to-End Multimodal Neural Networks

Byungsoo Jeon*
Carnegie Mellon University
Pittsburgh, PA, USA
jbsimdicd@gmail.com

Chanju Kim
Clova, NAVER Corp.
Seongnam, Korea
chanju.kim@navercorp.com

Adrian Kim
Clova, NAVER Corp.
Seongnam, Korea
adrian.kim@navercorp.com

Dongwon Kim
Clova, NAVER Corp.
Seongnam, Korea
dongwon.kim@navercorp.com

Jangyeon Park
Clova, NAVER Corp.
Seongnam, Korea
jangyeon.park@navercorp.com

Jung-Woo Ha†
Clova, NAVER Corp.
Seongnam, Korea
jungwoo.ha@navercorp.com

ABSTRACT

Music emotion recognition (MER) is a key issue in user context-aware recommendation. Many existing methods require hand-crafted features on audio and lyrics. Here we propose a new end-to-end method for recognizing emotions of tracks from their acoustic signals and lyrics via multimodal deep neural networks. We evaluate our method on about 7,000 K-pop tracks labeled as positive or negative emotion. The proposed method is compared to end-to-end unimodal models using audio signals or lyrics only. The experimental results show that our multimodal model achieves the best accuracy as 80%, and we discuss the reasons of these results.

KEYWORDS

Music Emotion Recognition, Music Recommendation, Multimodal Neural Network

ACM Reference format:
Byungsoo Jeon, Chanju Kim, Adrian Kim, Dongwon Kim, Jangyeon Park, and Jung-Woo Ha. 2017. Music Emotion Recognition via End-to-End Multimodal Neural Networks. *RecSys '17 Poster Proceedings, Como, Italy, August 27–31, 2017,* 2 pages.

1 INTRODUCTION

Music emotion recognition (MER) is a core technology of context-aware music recommendation. Users usually want music to amplify their emotions while partying or driving, for examples. Music recommendation using content-based MER allows music's emotion to be aligned with that of users in these scenarios. However, this is challenging because it is still unclear how music is causing emotions. It is known that numerous factors such as tone, pace, and lyrics are related to determine music emotion.

Existing studies tackle MER in various ways. They mainly formulate MER as either a classification or a regression problem. Laurier et al. use four emotion categories in [5] while Hu et al. use 18 categories in [3]. Both of them require additional feature engineering, such as rhythmic and tonal feature extractions and psychological feature extractions from words, while our model doesn't. [1] propose a convolutional recurrent neural network for music tagging, inspiring us to extend it to multimodal neural network for MER. [2] and [4] also suggest other neural network models for MER while formulating regression based on unsupervised learning. [6] and [7] tackle sentence-level MER problem, but we formulate song-level MER problem because it is more reasonable to recommend to users.

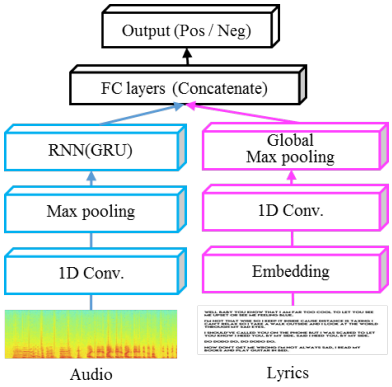

Figure 1: Model structure for multimodal music emotion classification from acoustic signals and lyrics of tracks

Here we simplify MER as a polarity emotion (positive / negative) classification of tracks to reduce the uncertainty from many emotion categories, considering an application to simple music recommendation scenario. We propose an end-to-end multimodal neural network models without an additional feature engineering process. We also create new dataset including tracks served on a Korean music streaming service to guarantee the high-quality data.

2 DATA DESCRIPTION

We describe our dataset from a famous Korean music streaming service, NAVER Music [0]. This consists of 3,742 positive and 3,742 negative tracks with their lyrics, and they are represented into mel-spectrograms. How could we seperate a positive and negative track? There are tracks tagged by editors in NAVER Music. We use a predefined emotion word dictionary to seperate positive and negative tags. For instance, positive emotion words are 'happy', and 'cheerful', while negative emotion words are 'sad', and 'lonely'. Then, we filter out the tracks whose tags include both positive and

*This work was performed in NAVER Corp.
†Corresponding author

[0] http://music.naver.com

negative words. We reject the tracks whose length is less than a minute or lyrics include less than 30 words. We use the first one minute of each mel-spectrograms, and only use noun, verb, adjective, and adverb in words of lyrics. Finally, each mel-spectrogram is represented as a 128 by 1024 matrix including 128 mels and 1024 time slots corresponding to one minute length of acoustic signals. Also, we have (27,496, 400) word vectors where vocabulary size $|V| = 27,496$ and the maximum length of word sequences is 400.

3 MUTIMODAL DEEP NETWORKS FOR MUSIC EMOTIONAL RECOGNITION

Figure 1 illustrates our end-to-end multimodal neural network model that directly predicts track's emotion from audio and lyrics. Our model has audio and lyrics branches which take a (128, 1024) mel-spectrogram (audio) and a (27496, 400) padded word vector (lyrics) as an input, respectively. At the bottom of the audio branch, there are five 1D convolution and max pooling layers to understand mel-spectrogram as a sequence of 1D vector $\mathbf{x_a}$ whose length is 128.

$$\mathbf{u_a} = [Maxpooling(Conv(\mathbf{x_a}))]^5 \quad (1)$$

Five 1D convolution layers whose filter sizes are all 3 have 128, 128, 128, 64, and 64 output of filters, respectively. Filter sizes of five max pooling layers are 3, 3, 3, 2, and 2. We use the exponential linear unit (elu) function as a non-linear function of convolution layers.

On top of that, we put two RNN layers (GRU) whose output dimensionality is 64 and one fully connected layer whose weight matrix is ${}^{FCa}W$ is to build an audio embedding vector $\mathbf{v_a}$ with length 64 before merging with lyrics branch.

$$\mathbf{v_a} = {}^{FCa}W\{GRU^2(\mathbf{u_a})\} \quad (2)$$

At the bottom of the lyrics branch, there is an embedding layer (weight matrix: ${}^{FCe}W$) whose output dimensionality is 200 followed by an 1D convolution layer whose filter size and number of output are 3 and 250 where an input word vector is $\mathbf{x_l}$. On top of that, we put global 1D max pooling layer because it is more robust to noise words than non-global one. As in the audio branch, there is one fully connected layer whose weight matrix is ${}^{FCl}W$ on it to build an lyrics embedding vector with length 64.

$$\mathbf{v_l} = {}^{FCl}W\{GlobalMaxpooling(Conv({}^{FCe}W\mathbf{x_l}))\} \quad (3)$$

We concatenate two branches because it shows the best performance. Lastly, we produce final output by taking this concatenated vector as an input of one fully connected layer (weight matrix: ${}^{FCm}W$) whose output dimensionality is 64 followed by a softmax layer to compute the binary cross-entorpy loss.

$$\mathbf{o} = Softmax(ReLU({}^{FCm}W\{Concatenate(\mathbf{v_a}, \mathbf{v_l})\})) \quad (4)$$

As a result, output vector $\mathbf{o}$ includes two values, the probabilities of a positive and negative emotional track.

4 EVALUATION

We compare the classification accuracies of models using audio, lyrics, and both. We test five unimodal models and one multimodal model which consists of the best unimodal models for audio and lyrics as in Figure 1. To test these models, we implemented them with Keras on Tensorflow while using Tesla M40 GPU. Table 1 shows the classification accuracies for each model. We randomly split the dataset into 90% for training and the rest for validation, and obtain the results shown in Table 1 after 5 runs of the test for each model. 1D CNN + RNN model is the best among models using audio. 1D CNN model is the best among models using lyrics. The reason why 1D CNN is better than RNN to predict from lyrics may be that the word sequences are too long. It is also notable that the model for lyrics works better than that for audio. Overall, the multimodal model using audio and lyrics shows the best accuracy, 0.8046. Figure 2 presents validation accuracy and loss of the best model for each modality (audio, lyrics, both). The model for lyrics and both shows little more stable convergence than that for audio.

Data	Model	Accuracy
Audio	CNN	0.6479
	RNN	0.6303
	CNN+RNN	**0.6619**
Lyrics	**CNN**	**0.7815**
	RNN	0.7716
Both	**CNN+RNN, CNN**	**0.8046**

Table 1: Classification accuracies using audio, lyrics, and both

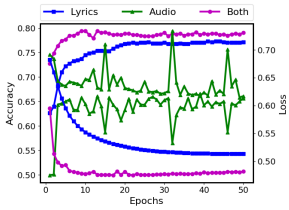

Figure 2: Validation accuracies and losses of the best model for each modality

5 CONCLUSION

We define MER as a polarity emotion classification and propose a multimodal neural network model trained in an end-to-end manner without additional feature engineering. We present lyrics are better features than audio on our problem, and our multimodal models proves the best accurcacy, 80% compared to unimodal models. We will further investigate end-to-end deep learning strategies with more tracks and emotional categories. Furthermore, we will apply our method to the context-aware music recommendation service of Clova, Cloud-based AI-assistant platform developed as a collaboration project by NAVER-LINE[1].

REFERENCES

[1] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. 2016. Convolutional Recurrent Neural Networks for Music Classification. *arXiv preprint arXiv:1609.04243* (2016).
[2] Eduardo Coutinho, George Trigeorgis, Stefanos Zafeiriou, and Björn W. Schuller. 2015. Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*
[3] Xiao Hu and J. Stephen Downie. 2010. When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *11th ISMIR 2010.* 619–624.
[4] Moyuan Huang, Wenge Rong, Tom Arjannikov, Nan Jiang, and Zhang Xiong. 2016. Bi-Modal Deep Boltzmann Machine Based Musical Emotion Classification. In *ICANN 2016.* 199–207.
[5] Cyril Laurier, Jens Grivolla, and Perfecto Herrera. 2008. Multimodal Music Mood Classification Using Audio and Lyrics. In *7th ICMLA 2008.* 688–693.
[6] Bin Wu, Erheng Zhong, Andrew Horner, and Qiang Yang. 2014. Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning. In *ACM MM 2014.* 117–126.
[7] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. 2008. A Regression Approach to Music Emotion Recognition. *IEEE Trans. Audio, Speech & Language Processing* 16, 2 (2008), 448–457.

[1] https://clova.ai