AttAE-RL^2: Attention based Autoencoder for Rap Lyrics Representation Learning

Hongru Liang, Qian Li, Haozheng Wang, Hang Li, Jin-Mao Wei, Zhenglu Yang
College of Computer and Control Engineering, Nankai University, Tianjin, China, 300350
{ lianghr, liqian515, hzwang, hangl} @mail.nankai.edu.cn, {weijm, yangzl}@nankai.edu.cn

Abstract

Learning rap lyrics is an important area of music information retrieval because it is the basis of many applications, such as recommendation systems, automatic classification. In this paper, we tackle the issue pertaining to the lack of an effective approach to aggregate various features of lyrics by proposing an attention-based autoencoder for rap lyrics representation learning (AttAE-RL^2). The proposed method appropriately integrates the semantic and prosodic features of rap lyrics. The preliminary experimental results demonstrate that our approach outperforms the state-of-the-art ones.

Keywords

rap lyrics, representation learning, autoencoder

1 Introduction

Among the various music genres, rap is one of the most popular types [3]. Learning rap lyrics is an important task of music information retrieval that has attracted growing interest from researchers. However, rap lyrics are unstructured, and directly using the off-the-shelf natural language processing techniques is infeasible in conducting phonological analysis.

In recent years, many studies have been conducted on rap lyrics analysis [2]. However, the effectiveness of these studies is far from satisfactory for the users, because such studies have either partially utilized the features, e.g., semantic features, or learn ineffective representation of the features, e.g., statistical representations. As far as we know, no study has yet produced general representations of rap lyrics involving both semantic and prosodic information.

To address the aforementioned issues, we introduce an attention-based autoencoder to appropriately aggregate the semantic and prosodic information from the lyrics. Our goal is to learn the overall representation of rap lyrics. The prosodic features are effectively represented by a novel strategy, i.e., rhyme2vec. Moreover, attention mechanism is used to balance the importance among various types of information. All these strategies are integrated into a general framework called the attention-based autoencoder for rap lyrics representation learning (AttAE-RL^2). The performance of our proposed approach is experimentally demonstrated to be superior to the state-of-the-art ones by a large margin. To the best of our knowledge, this is the first study to learn the integrated and distributed representation of rap lyrics.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186902

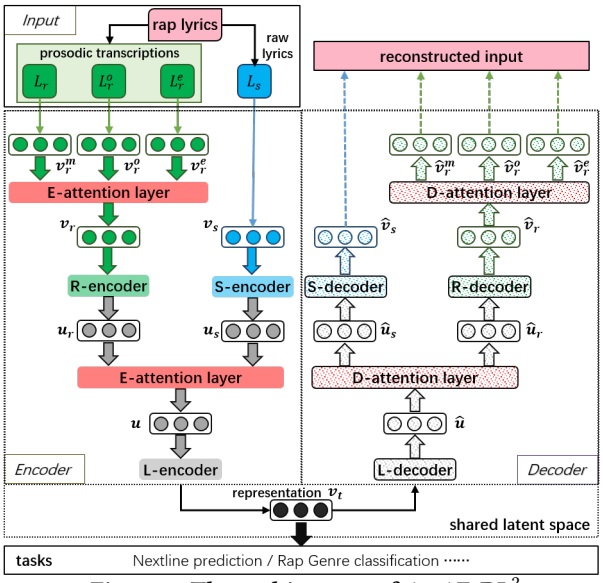

Figure 1: The architecture of AttAE-RL^2

2 Model Description

The model consists of three modules, i.e., Input module, Encoder module, and Decoder module (Fig. 1). The Encoder and the Decoder modules comprise the attention-based autoencoder. Our target is to learn a latent representation for a piece of rap lyric with both semantic and prosodic information.

Input module. As shown at the top of Fig. 1, we utilize rap lyrics as the input. For each piece of rap lyric, let $\boldsymbol{L_s}$ denote its raw lyrics and $\boldsymbol{L_r}$ denote the corresponding phonetic transcriptions, translated by eSpeak[1]. Table 1 shows an example of four consecutive rap lyrics and phonetic transcription from Fort Minor's *Remember the Name*.

Table 1: Rap lyrics with their phonetic transcription

raw lyrics (L_s)	phonetic transcription (L_r)
Put it together himself	p,Ut It t@g,ED3 hIms'Elf
now the picture connects	n'aU D@ p'IktS3 k@n'Ekts
Never asking for someone's help	n'Ev3r- 'aaskIN fO@ s'Vmw0nz h'Elp
to get some respect	t@ gEt s,Vm rI2sp'Ekt

We assume that rap lyrics have both monorhyme and alternate rhyme[2]. For the monorhyme, we treat all consecutive lines as a prosodic block, i.e., $\boldsymbol{L_r}$. For the alternate rhyme, we split the rap lines into two prosodic blocks, namely, $\boldsymbol{L_r^o}$, which includes the odd lines, and $\boldsymbol{L_r^e}$, which includes the even lines.

Attention based autoencoder. Instead of using one-hot codes, we utilize pretrained embeddings as the input. Doc2vec [4] is employed to extract the semantic vector of $\boldsymbol{L_s}$, i.e., $\boldsymbol{v_s}$. Here, $\boldsymbol{v_r^m}$, $\boldsymbol{v_r^o}$,

[1] http://espeak.sourceforge.net/.
[2] Monorhyme is a rhyme scheme in which each line has an identical rhyme. In alternate rhyme, the rhyme is on alternate lines. *Wikipedia*.

and $\boldsymbol{v}_r^e$ are embedded in a similar way of modeling the semantic information, yet at the character level.

In the Encoder module, an E-attention layer, to be explained later, is applied on $\boldsymbol{v}_r^m$, $\boldsymbol{v}_r^o$, and $\boldsymbol{v}_r^e$ to obtain a comprehensive rhyme vector $\boldsymbol{v}_r$. Both $\boldsymbol{v}_r$ and $\boldsymbol{v}_s$ are vectors in different high dimension spaces. To learn the integrated representations of rap lyrics, $\boldsymbol{v}_r$ and $\boldsymbol{v}_s$ are mapped into a shared latent space through fully-connected layers, i.e., R-encoder and S-encoder, respectively. As can be seen, $\boldsymbol{u}_r$ and $\boldsymbol{u}_s$ are the corresponding vectors to $\boldsymbol{v}_r$ and $\boldsymbol{v}_s$ in the shared high dimension space, respectively. Instead of simply adding $\boldsymbol{u}_r$ and $\boldsymbol{u}_s$ up, we employ another E-attention layer to fuse them and obtain a latent vector, $\boldsymbol{u}$. The target representation of rap lyrics, $\boldsymbol{v}_t$, is learned from $\boldsymbol{u}$ through fully-connected layers, L-encoder.

The Decoder module is the inversion of the Encoder module. In this module, the intermediate results and input vectors of the Encoder module should be reconstructed from $\boldsymbol{v}_t$.

Attention mechanism. For an attention layer, noted as E-attention layer in Fig. 1, n m-dimensional vectors serve as inputs. In the E-attention layer, we stack the input vectors as an $n \times m$ matrix, $\boldsymbol{M}$. And an attention vector $\boldsymbol{a}$ is calculated by feeding the concatenation of the input vectors into a fully-connected layer. Every element of $\boldsymbol{a}$ corresponds to one input vector. The output $\boldsymbol{s}$ of an attention layer is calculated with $\boldsymbol{s} = \boldsymbol{a} \cdot \boldsymbol{M}$. A D-attention layer can be regarded as an inverse operation of an E-attention layer. The output of a D-attention layer is calculated as $\hat{\boldsymbol{M}} = \boldsymbol{a}^{+} \cdot \hat{\boldsymbol{s}}$, where $\boldsymbol{a}^{+}$ is the left pseudo inverse of the attention vector $\boldsymbol{a}$ of the corresponding E-attention layer.

Loss function. The loss function contains reconstruction loss and label loss. The reconstruction loss is formulated as $\ell_{ae} = mse(\boldsymbol{v}_s, \hat{\boldsymbol{v}}_s) + mse(\boldsymbol{v}_r^m, \hat{\boldsymbol{v}}_r^m) + mse(\boldsymbol{v}_r^o, \hat{\boldsymbol{v}}_r^o) + mse(\boldsymbol{v}_r^e, \hat{\boldsymbol{v}}_r^e)$, where *mse* stands for mean squared error. To incorporate the label information, we deploy a classifier over $\boldsymbol{v}_t$ as an external tool in fine tuning the model. The label loss, denoted as ℓ_{label}, is a cross entropy function over the classifier. The overall objective function is $\ell = \alpha * \ell_{ae} + (1 - \alpha) * \ell_{label}$, where α is a hyper parameter to balance the importance of the two objectives.

3 Experimental Evaluation

To evaluate the effectiveness of the proposed framework on representing rap lyrics, we conduct two experimental tasks, i.e., NextLine prediction [2] and rap genre classification. The dataset and source code are available at https://github.com/mengshor/attaerl2.

3.1 NextLine Prediction

Given a rap song with a sequence of m lines, the task of NextLine prediction is to predict the (m+1)th line from a set of candidate lines [2]. We obtain a corpus of rap lyrics crawled from the Internet[3], which includes 810567 lines from 16697 songs.

We compare AttAE-RL^2 with other state-of-the-art methods, namely, DopeSemantic, DopeRhyme, and DopeLearning [2], which are collectively known as "Dopes". Doc2vec-rhyme2vec concatenates doc2vec and rhyme2vec representations together.

The performance is evaluated by mean rank and mean reciprocal rank (MRR), as shown in Table 2. Among all methods, AttAE-RL^2 achieves the best performance with a mean rank of 6.6024. This value is much better than that of Dopes (79.9964). Moreover, the MRR of AttAE-RL2 is 0.7278, which is superior to that of Dopes (0.168). The result demonstrates the superiority of AttAE-RL^2 over other methods in that it can represent rap lyrics by fusing more features more effectively .

Table 2: Performance of different models

Algorithm	Mean rank	MRR
DopeSemantic	116.4178	0.0822
DopeRhyme	103.2068	0.1303
DopeLearning	79.9964	0.1680
rhyme2vec	45.8172	0.1906
doc2vec	38.8388	0.2066
doc2vec-rhyme2vec	15.1046	0.4627
AttAE-RL^2	**6.6024**	**0.7278**

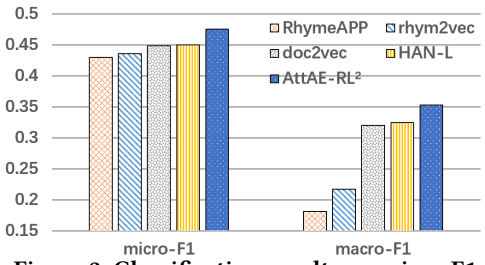

Figure 2: Classification results on micro-F1

3.2 Rap Genre Classification

Given a set of rap songs and a set of genre labels, the rap genre classification task is to predict proper labels for every song based upon the representations learned from raw rap lyrics. We create a dataset consisting of 10167 songs from 9 rap genres, such as alternative rap, grime rap, and so forth.

The baselines evaluated are **RhymeAPP** [1], a lyrical analysis tool that calculates the statistical features of a rap song; **rhyme2vec**, which employs prosodic representations; **doc2vec**, which utilizes merely semantic representations; and **HAN-L** [5], which is a state-of-the-art approach for genre classification of intact lyrics.

The micro-F1 and macro-F1 are employed to evaluate the classification results, as shown in Fig. 2, respectively. As can be seen, AttAE-RL^2 significantly outperforms the other methods, indicating that it has a strong ability to capture prosodic and semantic information in effectively representing rap lyrics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No.U1636116, 11431006, 61772288, and the Research Fund for International Young Scientists under Grant No. 61650110510 and 61750110530.

References

[1] Hussein Hirjee and Daniel G Brown. 2010. Rhyme analyzer: An analysis tool for rap lyrics. In *Proceedings of the 11th International Society for Music Information Retrieval Conference.*

[2] Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. 2016. DopeLearning: A Computational Approach to Rap Lyrics Generation. In *Proceedings of KDD'16.*

[3] Matthias Mauch, Robert M MacCallum, Mark Levy, and Armand M Leroi. 2015. The evolution of popular music: USA 1960–2010. *Royal Society open science* 2, 5 (2015), 150081.

[4] V. Le Quoc and Mikolov Tomas. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of ICML'14.*

[5] Alexandros Tsaptsinos. 2017. Lyrics-Based Music Genre Classification Using a Hierarchical Attention Network. In *Proceedings of ISMIR'17.*

[3] http://ohhla.com/