知识表示学习的研究

Anonymous

北京航空航天大学计算机学院

摘要 摘要。

关键字： 表示学习，知识图谱，知识表示

1 引言

知识图谱可以结构化的表示真实世界中的实体和实体间关系。

2 知识图谱

知识图谱是用来表示真实世界中的实体和实体间关系的图结构网络。知识图谱由节点和边组成，其中节点代表实体，边代表实体之间的关系，通常用三元组$G = \{E, R, S\}$ 来表示知识图谱，其中E 为实体的集合，R 为关系的集合，S 是实体-关系-实体的三元组集合，即$S \subseteq E \times R \times E$ 。此外，一个实体-关系-实体的三元组也可以表示为(h, r, t) 的格式，其中h 和t 分别代表头实体和尾实体，r 表示头实体和尾实体之间的关系。

3 知识表示学习

知识图谱通常以图的形式存储，这种图结构不利于对数据集的进一步研究，更需要设计特定的算法进行相应的计算。因此通常将图结构转化为向量的形式，便于

通过向量表示实体的方法通常有两种，一种使用独热编码构建向量，该方法对一个有n个数据数据集，需要对每个实体使用一个n维向量来表示，向量中只有1维的值为1，其余值均为0。这种方法在数据集较大时效果较差，又由于向量之间两两正交，不仅无法表示向量的语义信息，更无法表示出向量之间的相似程度。另一种则是通过机器学习算法得到每一个实体或关系

的低维稠密向量。对于一个词向量，单独拿出一维数据可能毫无意义，但组成向量后却可以表示出相应实体的语义信息，从而通过余弦或距离来表示实体在真实世界的相似程度，这就是表示学习。

知识表示学习，即对知识库的表示学习。通过用低维度的稠密向量来表示知识库中的实体或关系，从而进行进一步的计算。

4 Related Work

Among the first to tackle the task of automatically classifying music into emotion-based categories, Li and Ogihara used Support Vector Machines (SVM) with audio-based features (related to timbre, pitch and rhythm) and reported 45% accuracy on a dataset of consisting of 499 music clips and 13 mood categories [**?**].

Starting in 2007, the Audio Music Mood Classification task appeared regularly in the literature to encourage the development of improved music-IR systems. Since then, datasets comprised of hundreds of music tracks were collected and made available to the research community and more than two hundred systems have been evaluated. Despite other supervised methods like Gaussian Mixture Model [**?**], Random Forest and K-Nearest Neighbor, many studies found that SVM combined with spectral features often yield the best results [**?**].

Due to the limiting factors of features based solely on audio [**?**] and because of the semantically rich nature of music lyrics, lyric-based features found their way into emotion-based music classification. Among others, Hu et al. [**?**] investigate the usefulness of low-level text features such as the Bag-of-Words (BoW) representation of lyrics, also parts of speech and function words. They also combine lyric and audio features and report accuracy as high as 72% on a private dataset consisting of 5,585 music tracks and 18 mood categories [**?**]. He et al. [**?**] report that higher-order BoW features such as tf-idf weighted unigram, bigram and trigram, can capture more semantic relations in lyrics for mood classification. Similarly, other lyric features derived from the Affective Norm of English Words also obtain encouraging results [**?**].

There are several ways to combine information from different domains, such as audio and text. The *early fusion* methods simply concatenate audio and lyric features to create feature vectors in a new space [**?**]; in the *late fusion normally* separate classifiers are trained on the features from their own separate domains [**?**]. While Xue et al. [**?**] fused audio and ltext domains through a model fusion scheme. In this work, we follow the idea to use Deep Boltzmann Machines for multimodal learning [**?**] and demonstrate its effectiveness on the largest publicly available music dataset.

5 Bi-Modal Deep Boltzmann Machine Model

Deep Boltzmann Machine (DBM) [**?**] is a deep neural network architecture based on Restricted Boltzmann Machine [**?**]. It contains a set of visible units $\mathbf{v} \in \{0,1\}^D$ and a sequence of layers comprised of hidden units $\mathbf{h}^{(1)} \in \{0,1\}^{F_1}, \mathbf{h}^{(2)} \in \{0,1\}^{F_2}, ..., \mathbf{h}^{(n)} \in \{0,1\}^{F_n}$. The connections are available only between units in adjacent layers, i.e. no connection is allowed between any two units within the same layer or between any two units in non-adjacent layers. The energy of the joint configuration $\{\mathbf{v}, \mathbf{h}\}$ is defined according to $\mathbf{h} = \{\mathbf{h^{(1)}}, \mathbf{h^{(2)}}, ..., \mathbf{h^{(n)}}\}$ and parameters $\theta = \{\mathbf{W^{(1)}}, \mathbf{W^{(2)}}, ..., \mathbf{W^{(n)}}, \mathbf{b}, \mathbf{b^{(1)}}, \mathbf{b^{(2)}}, ..., \mathbf{b^{(n)}}\}$. The DBM assigns probability to a set of visible units according to the Boltzmann distribution:

$$P(\mathbf{v}; \theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} exp(-E(\mathbf{v}, \mathbf{h}^{(1)}, \mathbf{h}^{(2)}; \theta)) \tag{1}$$

where $Z(\theta)$ is the normalising constant.

Multimodal DBM is a generative model for that can create fused representations by combining features from different modalities in a model fusion scheme [**?**]. Fig. 1 illustrates the proposed audio-text aware bi-modal DBM architecture; it consists of two 2-layer DBM networks, with an additional layer of hidden units added on top to join the two DBMs and form a single model.

Let $\mathbf{v}_a \in \mathbb{R}^D$ denote the audio input and $\mathbf{v}_t \in \mathbb{R}^K$ denote the text input, where $K, D \in \mathbb{R}$ is the dimension of audio and text features. Then, the joint

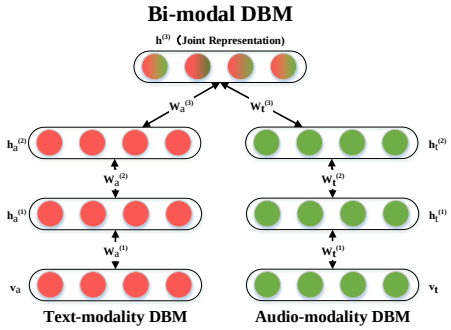

图 1. Bi-modal Deep Boltzmann Machine

distribution of bi-modal input can be then written as:

$$P(\mathbf{v}_a, \mathbf{v}_t; \theta) = \sum_{\mathbf{h_a}^{(2)}, \mathbf{h_t}^{(2)}, \mathbf{h}^{(3)}} P(\mathbf{h_a}^{(2)}, \mathbf{h_t}^{(2)}, \mathbf{h}^{(3)}) \left(\sum_{\mathbf{h_a}^{(1)}} P(\mathbf{v}_a, \mathbf{h_a}^{(1)} \mid \mathbf{h_a}^{(2)}) \right) \left(\sum_{\mathbf{h_t}^{(1)}} P(\mathbf{v}_t, \mathbf{h_t}^{(1)} \mid \mathbf{h_t}^{(2)}) \right) \tag{2}$$

The second term in Eq. 2 denotes the probability distribution of the audio modality, which assigns probability to $\mathbf{v}_a$ in a Gaussian RBM scheme:

$$\begin{aligned} P(\mathbf{v_a}; \theta_a) &= \sum_{\mathbf{h_a}^{(1)}, \mathbf{h_a}^{(2)}} P(\mathbf{v_a}, \mathbf{h_a}^{(2)}, \mathbf{h_a}^{(1)}; \theta_a) \\ &= \frac{1}{Z(\theta_a)} \sum_{\mathbf{h_a}^{(1)}, \mathbf{h_a}^{(2)}} exp\left(- \sum_i \frac{(v_{\mathbf{a}i} - b_{\mathbf{a}i})^2}{2\sigma_i^2} + \sum_{ij} \frac{v_{\mathbf{a}i}}{\sigma_i} W_{\mathbf{a}ij}^{(1)} h_{\mathbf{a}j}^{(1)} + \right. \\ &\quad \left. \sum_{jl} W_{\mathbf{a}jl}^{(1)} h_{\mathbf{a}j}^{(1)} h_{\mathbf{a}l}^{(2)} + \sum_j b_{\mathbf{a}j}^{(1)} h_{\mathbf{a}j}^{(1)} + \sum_l b_{\mathbf{a}l}^{(2)} h_{\mathbf{a}l}^{(2)} \right) \end{aligned} \tag{3}$$

The third term in Eq. 2 denotes the probability distribution of the text modality, where $\mathbf{v} \in \mathbb{N}^{\mathbf{k}}$ denotes a vector of visible units and each v_k is the number of times word k occurs in the lyrics with the dictionary size M. The

model assigns probability to $\mathbf{v}_t$ in a Replicated Softmax RBM scheme:

$$
\begin{aligned}
P(\mathbf{v_t};\theta_t) &= \sum_{\mathbf{h_t^{(1)}},\mathbf{h_t^{(2)}}} P(\mathbf{v_t},\mathbf{h_t^{(2)}},\mathbf{h_t^{(1)}};\theta_t) \\
&= \frac{1}{Z_M(\theta_t)} \sum_{\mathbf{h_t^{(1)}},\mathbf{h_t^{(2)}}} exp\left(\sum_{jk} W^{(1)}_{\mathbf{t}k,j} h^{(1)}_{\mathbf{t}j} v_{\mathbf{t}k} + \sum_{jl} W^{(2)}_{\mathbf{t}jl} h^{(1)}_{\mathbf{t}j} h^{(2)}_{\mathbf{t}l} + \right. \\
&\qquad \left. \sum_k b_{\mathbf{t}k} v_{\mathbf{t}k} + M \sum_j b^{(1)}_{\mathbf{t}j} h^{(1)}_{\mathbf{t}j} + \sum_l b^{(2)}_{\mathbf{t}l} h^{(2)}_{\mathbf{t}l}\right)
\end{aligned} \tag{4}
$$

The parameters of DBM can be initialised randomly. However, here we use a greedy layer-wise pre-training strategy [**?**, **?**].

6 Experimental Study

In our experiments, we use the largest publicly available music dataset, the Million Song Dataset (MSD) [**?**]. It is a conglomeration of several datasets containing different information about the tracks; we use two of its subsets. First, MusiXmatch, contains information about the lyrics, each song is described as a set of words from the recorded top 5,000 frequent words across all lyrics. Second, Last.fm, contains annotations obtained from music listeners in a form of tags, like "happy" and "upbeat"; from it, we select tracks that are described by emotion related tags. Additionally, we obtain already pre-extracted audio-based features from the MSD Benchmarking dataset, which is an extension of MSD and was created for the purposes of comparing different approaches while maintaining invariability in various experimental parameters [**?**]. To capture both modalities, in our experiments, each music track is represented by both lyrics (found in MusiXmatch dataset) and audio-based features (from MSDB dataset), there are 236,486 tracks that satisfy these conditions.

Initially, to test the validity of our approach, we select only the tracks that contain "happy" and "sad" tags. After removing ambiguous tracks that contain both tags, we obtain 7,945 "happy" songs and 5,840 "sad" tracks. To avoid classifier bias due to class imbalance, we perform random subsampling and then conduct a binary emotion classification experiment.

In a multi-class scenario, some songs may cover a variety of emotions, rendering the representation by independent dimensions inadequate. For this reason, we employ Russell's Valence-Arousal model [**?**] and follow Corona's and O'Mahony's scheme of selecting social tags that clearly indicate the song's emotional trend [**?**]. We group the tags according to their quadrants in the Valence-Arousal model and report the final number of tracks tagged by each emotion group in Table 1. We use the tracks that have the emotion-related tags as labelled data for training the classifier, and the remainder as unlabelled data for unsupervised pre-training. Our final dataset contains 41,727 labelled and 194,759 unlabelled tracks.

表 1. Mood quadrants and their corresponding number of songs

Quadrant	**Group**	**Tag**	**Songs**
v^-a^+	G29	aggressive,aggression.	28,168
	G28	anger,angry,choleric,etc.	
v^+a^+	G6	cheerful,jolly,festive,etc.	16,315
	G5	happy,happiness,etc.	
v^-a^-	G15	sad,sadness,unhappy,etc.	10,154
	G16	depressed,blue,dark,gloom,etc.	
	G17	heartbreak,grief,sorrow,etc.	
v^+a^-	G8	brooding,contemplative,etc.	2,629
	G12	calm,comfort,quiet,etc.	

The deep learning architecture is configured as following. The audio pathway is modeled by an RBM with 194 visible units, each taking as input acoustic content descriptors, such as MFCC and SSD features. The visible layer is followed by two layers of hidden units, 100 and 50 each. The text modality is formed by RBM consisting of 5,000-unit visible layer followed by hidden layers of 2,048 and 1,024 units each. A joint layer combines the two modalities and consists of 1,074 hidden units. Its output can be considered as a complex probability estimate of the mood classes. We use the output from our Mulimodal RBM as input to either Softmax or SVM for the final classification decision. Additionally, to test the robustness of our chosen audio features, we expand the audio modality from 194 to 3,456 dimensions

by including additional audio-based features. The hidden layers are also expanded to 2,048 and 1,024 respectively; and the joint layer to 2,048 units.

Because the SVM classifier performs slightly better on average, we omit the Softmax results. In our experiments, we perform k-fold repeated random sub-sampling validation with $k = 5$. In each fold, 60% (6,984) tracks are selected for training and 40% (4,656) for testing. We compute Mean Average Precision (MAP) and Accuracy as metrics to comprehensively evaluate the models. The initial experimental results are shown in Fig. 2, where we also illustrated the baseline SVM performance (no DBM) using early concatenation method to join the two modalities into a single input vector.

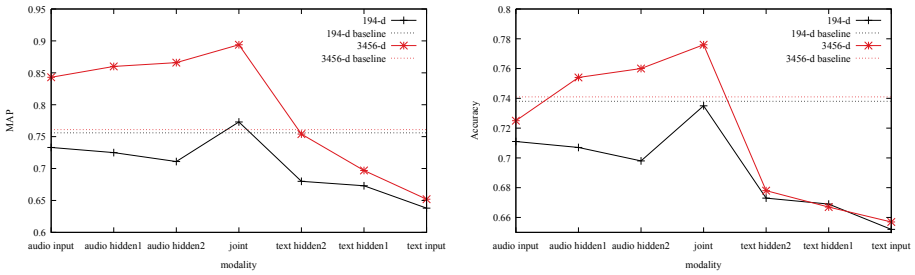

图 2. MAP and Accuracy achieved by the Bi-modal Boltzmann Machine in the "happy"/"sad" binary classification task

As can be seen from Fig. 2, audio-based features indeed outperform the lyric-based features to some extent. We conjecture that this may be because the audio modality is represented by features that were hand-crafted and improved over the years. Meanwhile, the text modality is represented by a shallow BoW statistical measure with large vocabulary, which results in a sparse input vector. This again urges the study on higher level lyric features, which may yield interesting results. We also noticed that the classification performance declined through the audio pathway, which indicates that some valuable information are lost through the extracting process in the audio modality. After expanding the audio modality with additional features, this phenomenon disappears. This indicates the necessity of feature selection. Among all results, the best performance is achieved at the joint layer, which

shows the effectiveness of the fusing ability of the proposed approach. After expanding the audio features from 194 to 3,456, the baseline SVM performance did not improve much.

In addition to using the lyric- and audio-based features with our approach, we also compare the model fusion, early fusion and late fusion methods. In late fusion, we first trained two SVM classifiers to represent the two modalities separately, denoting as p_a and p_t. Then the output mood class is assigned by

$$p = \alpha p_a + (1 - \alpha) p_t \tag{5}$$

where α indicates the relevant importance between audio and lyric features. We set $\alpha = 0.6$, as per Hu et al. [**?**]. As before, in order to avoid classifier bias towards majority class, we attempt to maintain class balance by ensuring that both training and testing instances are equally distributed across mood classes. Results are shown in Table 2.

表 2. Comparison of accuracy achieved by the different fusion models

	audio_only	**text_only**	**early_fusion**	**late_fusion**	**Bi-modal DBM**
v^-a^+	0.645	0.600	0.689	0.666	**0.706**
v^+a^+	0.625	0.607	0.653	0.639	**0.692**
v^-a^-	0.634	0.620	0.661	0.642	**0.704**
v^+a^-	0.730	0.702	0.745	0.729	**0.785**

Our model outperformed other baseline models in every mood category. The moods in v^+a^- quadrant obtain the highest accuracy. This is interesting given that the v^+a^- quadrant has the least number of songs. The reason may be that music pieces in this mood group has many unique lyric terms. Between other mood categories, however, there is no significant differences in the classification accuracy. Moreover, the fusion methods' accuracy all outperformed the accuracy of classification on single modality, affirming the effectiveness of multi-modal mood classification in the same way as many prior studies show.

7 结论

In this work, we used a deep learning architecture, inspired by the work of Srivastava and Salakhutdinov [**?**], to effectively fuse the audio and text modalities for music mood classification. Results show that fusing modalities is indeed advantageous in the music mood classification task. In addition to including information from other domains/modalities, it would be interesting to see how other lyric derived features perform with this and other multimodal approaches in the music-IR literature, we leave this to our future work.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China (No. 61332018), the National Department Public Benefit Research Foundation (No. 201510209), and the Fundamental Research Funds for the Central Universities.

参考文献