The Future of Reality: An Artificial Intelligence based Mixed-Reality

Anynimous

School of Computer Science and Engineering, Beihang University, Beijing, China

Abstract. Bringing Virtual Reality and Augmented Reality to the world is not something new anymore. They can be found in most modern technology related facilities such as medical equipment, film production, games, daily tools, etc. However it is also true that the current Virtual Reality and Augmented Reality are simply technologies that bring more convenient to our daily routine, which means they simply act as tools which has no basic difference with PCs. But followed by the growing trends of both Artificial Intelligence and Human-made Realities, it has also become a new trend to bring them together, creating a true Mixed Reality world that actually intelligently interacts with people. This paper will explore the development of Artificial Intelligence-enhanced Mixed Reality, thoroughly discussing the future of Intelligent Reality in our world.

Keywords: Mixed Reality, Extended Reality, Artificial Intelligence, Virtual Reality, Augmented Reality

1 Introduction

In the current world–and probably in decades of the future–our reality will become something that we call by the Extended Reality (XR). An Extended Reality is a new type of mixed reality that comes from the mix of Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR). It is an unavoidable fact that the reality is not as pure as it was, mixed with the existence of various types of other 'realities' such as Virtual Reality, Augmented Reality, and many others. Among those types of reality, VR and AR play dominant roles, covering most of our current technologies, be it in film production, game development, medical facilities, etc. [11].

Many still have problems in differing between Virtual Reality and Augmented Reality due to their similarities in definitions. However they do have quiet huge differences in action, while Virtual Reality is bringing the "real world" (Hence, the users) to the virtual world, Augmented Reality works the other way, by bringing the virtual world into the real world [13]. As it became evident that using features based on audio alone is not enough, many researchers started combining features from different domains [9]. One such domain, music lyrics, has become a popular source of features for music emotion and mood classification among other music retrieval tasks. Mayer et al. [14] show that, in some emotion categories, when features derived from lyrics are included, the classifier performance improves over using the leading audio features alone. However, Hu et al. [6] reveal that this is not true for all of the mood categories. To further improve the classification performance, some researchers integrate audio features with lyrics together and form hybrid features that could carry information from two different modalities (domains) simultaneously [6]. Accordingly different integration strategies (e.g., early fusion [5], late fusion [10] and model fusion [20]) are proposed in the literature.

In this work, we follow the feature fusion model and use a hybrid model based on Multimodal Deep Boltzmann Machine; in addition to fusing different modalities, it is also able make use of unlabelled data to further improve performance [19]. Additionally, we adopt the commonly used Russell's 2-dimensional Valence-Arousal (V-A) model of affect [15] to capture the emotional content of music lyrics. To show the effectiveness of our approach, we conduct an experimental study on the largest dataset that is publicly available for music retrieval research, the Million Song Dataset [1], from which we are able to use over 230,000 music tracks that contain both lyric and audio features.

2 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 [12].

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 [13], Random Forest and K-Nearest Neighbor, many studies found that SVM combined with spectral features often yield the best results [21].

Due to the limiting factors of features based solely on audio [13] 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. [6] 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 [7]. He et al. [3] 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 [8].

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 [5]; in the *late fusion normally* separate classifiers are trained on the features from their own separate domains [10]. While Xue et al. [20] 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 [19] and demonstrate its effectiveness on the largest publicly available music dataset.

3 Bi-Modal Deep Boltzmann Machine Model

Deep Boltzmann Machine (DBM) [16] is a deep neural network architecture based on Restricted Boltzmann Machine [18]. 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.

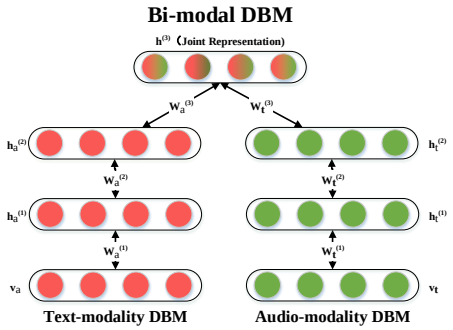

Fig. 1. Bi-modal Deep Boltzmann Machine

Multimodal DBM is a generative model for that can create fused representations by combining features from different modalities in a model fusion scheme [19]. 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 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:

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

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:

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

The parameters of DBM can be initialised randomly. However, here we use a greedy layer-wise pre-training strategy [16, 19].

4 Experimental Study

In our experiments, we use the largest publicly available music dataset, the Million Song Dataset (MSD) [1]. 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 [17]. 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 [15] and follow Corona's and O'Mahony's scheme of selecting social tags that clearly indicate the song's emotional trend [2]. 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.

Table 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.

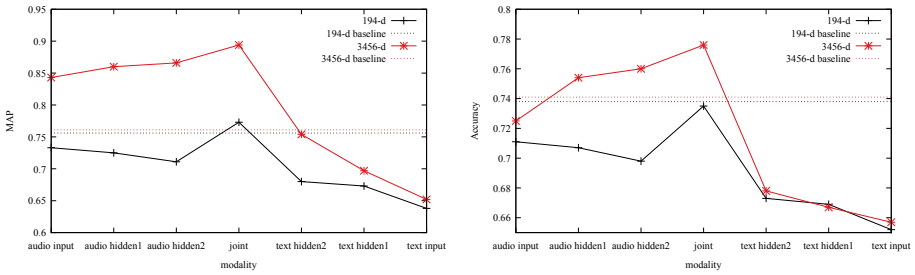

Fig. 2. MAP and Accuracy achieved by the Bi-modal Boltzmann Machine in the "happy"/"sad" binary classification task

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. [4]. 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.

Table 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.

5 Conclusion

In this work, we used a deep learning architecture, inspired by the work of Srivastava and Salakhutdinov [19], 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.

References

1. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: Proceedings of 12th International Society for Music Information Retrieval Conference. pp. 591–596 (2011)
2. Corona, H., O'Mahony, M.P.: An exploration of mood classification in the million songs dataset. In: Proceedings of 12th Sound and Music Computing Conference (2015)
3. He, H., Jin, J., Xiong, Y., Chen, B., Sun, W., Zhao, L.: Language feature mining for music emotion classification via supervised learning from lyrics. In: Proceedings of 3rd International Symposium on Intelligence Computation and Applications. pp. 426–435 (2008)
4. Hu, X., Choi, K., Downie, J.S.: A framework for evaluating multimodal music mood classification. Journal of the Association for Information Science and Technology (2016)
5. Hu, X., Downie, J.S.: When lyrics outperform audio for music mood classification: A feature analysis. In: Proceedings of 11th International Society for Music Information Retrieval Conference. pp. 619–624 (2010)

6. Hu, X., Downie, J.S., Ehmann, A.F.: Lyric text mining in music mood classification. In: Proceedings of 10th International Society for Music Information Retrieval Conference. pp. 411–416 (2009)
7. Hu, X., Downie, J.S., Laurier, C., Bay, M., Ehmann, A.F.: The 2007 MIREX audio mood classification task: Lessons learned. In: Proceedings of 9th International Conference on Music Information Retrieval. pp. 462–467 (2008)
8. Hu, Y., Chen, X., Yang, D.: Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In: Proceedings of 10th International Society for Music Information Retrieval Conference. pp. 123–128 (2009)
9. Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J.J., Speck, J.A., Turnbull, D.: State of the art report: Music emotion recognition: A state of the art review. In: Proceedings of 11th International Society for Music Information Retrieval Conference. pp. 255–266 (2010)
10. Laurier, C., Grivolla, J., Herrera, P.: Multimodal music mood classification using audio and lyrics. In: Proceedings of 7th International Conference on Machine Learning and Applications. pp. 688–693 (2008)
11. Li, T., Mitsunori, O., Tzanetakis, G. (eds.): Music Data Mining. CRC Press (2012)
12. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of 4th International Society for Music Information Retrieval Conference (2003)
13. Lu, L., Liu, D., Zhang, H.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech & Language Processing 14(1), 5–18 (2006)
14. Mayer, R., Neumayer, R., Rauber, A.: Combination of audio and lyrics features for genre classification in digital audio collections. In: Proceedings of 16th International Conference on Multimedia. pp. 159–168 (2008)
15. Russell, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39(6), 1161–1178 (1980)
16. Salakhutdinov, R., Hinton, G.E.: Deep boltzmann machines. In: Proceedings of 12th International Conference on Artificial Intelligence and Statistics. pp. 448–455 (2009)
17. Schindler, A., Mayer, R., Rauber, A.: Facilitating comprehensive benchmarking experiments on the million song dataset. In: Proceedings of 2012 International Society for Music Information Retrieval Conference. pp. 469–474 (2012)
18. Smolensky, P.: Information processing in dynamical systems: Foundations of harmony theory. Tech. rep., DTIC Document (1986)
19. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. Journal of Machine Learning Research 15(1), 2949–2980 (2014)

20. Xue, H., Xue, L., Su, F.: Multimodal music mood classification by fusion of audio and lyrics. In: Proceedings of 21st International Conference on MultiMedia Modeling, Part II. pp. 26–37 (2015)
21. Yang, Y.H., Chen, H.H.: Machine recognition of music emotion: A review. ACM Transactions on Intelligent Systems and Technology 3(3), 338–343 (2012)