Multimodal Learning with Deep Boltzmann Machine for Emotion Prediction in User Generated Videos

Lei Pang, Chong-Wah Ngo
Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
leipang3-c@my.cityu.edu.hk, cscwngo@cityu.edu.hk

ABSTRACT

Detecting emotions from user-generated videos, such as "anger" and "sadness", has attracted widespread interest recently. The problem is challenging as effectively representing video data with multi-view information (e.g., audio, video or text) is not trivial. In contrast to the existing works that extract features from each modality (view) separately followed by early or late fusion, we propose to learn a joint density model over the space of multi-modal inputs (including visual, auditory and textual modalities) with Deep Boltzmann Machine (DBM). The model is trained directly on the user-generated Web videos without any labeling effort. More importantly, the deep architecture enlightens the possibility of discovering the highly non-linear relationships that exist between low-level features across different modalities. The experiment results show that the DBM model learns joint representation complementary to the hand-crafted visual and auditory features, leading to 7.7% performance improvement in classification accuracy on the recently released VideoEmotion dataset.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Performance, Experimentation

Keywords

Emotion analysis; Multimodal learning; Deep Boltzmann Machine

1. INTRODUCTION

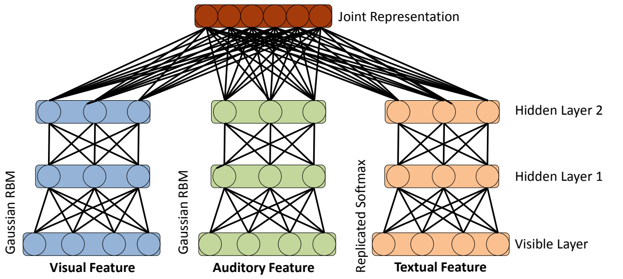

Figure 1: The multimodal DBM that models the joint distribution over visual, auditory and textual features. All layers but the first (bottom) layers use standard binary units. Gaussian RBM model is used to model the distributions over the visual and auditory features. Replicated Softmax topic model is applied for mining the textual features.

While intensive research efforts have been devoted to emotion analysis on documents or images, emotion prediction on user-generated videos is still a relatively new and largely untapped research area. With the popularity of social websites and mobile devices with high quality cameras, there is a large volume of unedited videos generated by users to record all kinds of activities in our life. Automatically understanding the emotions from these videos with diverse contents is in high demand for many applications [5]. For example, when searching information about a resort, the retrieved videos can be reranked based on their emotions to provide implicit comments. In addition, when asking opinion-related questions about hot events, providing emotion tags for retrieved videos helps users more quickly understand the sentiment of public's view.

While there have been significant progresses made on the emotion prediction in images [2, 6, 15], previous efforts on video emotion analysis are mainly conducted in movie domain [1, 11, 12, 14]. In these works, a variety of visual and auditory features have been exploited, including low-level features [1, 5] such as HOG (Histogram of Gradients) from visual features and zero-crossing rate from auditory features, and semantic-level features [2, 5, 6] such as concept features from SentiBank [2]. Although jointly using visual and auditory features improves the performance and shows promising results on emotion classification of Hollywood movies [1, 11, 12, 14], little effort has been devoted to learn a joint representation over visual and auditory modalities. In [12], SVM classifiers are trained on the fused feature vectors that concatenate both visual and auditory features. The same early fusion is also adopted in [14], but with Conditional Random Field (CRF) to model the temporal information for classification. In [11], two Baysian classifiers are trained on visual and auditory features separately and the output scores are linearly fused to generate the final prediction. In [1], late fusion is adopted to linearly combine outputs of SVM classifiers. In short, these works adopt either early or

ICMR'15, June 23–26, 2015, Shanghai, China.
Copyright © 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00.
http://dx.doi.org/10.1145/2671188.2749400.

late fusion to combine the visual and auditory modalities. A more advanced approach is proposed in [5], where multiple kernels defined for individual features are linearly fused to learn a kernel SVM classifier. We argue that these methods may work well for combining different features in the same modality, but poorly for different modalities. The reason is that different modalities have different kinds of representations and correlational structures, making it difficult for a shallow combination to discover relationships across modalities, especially the highly non-linear relationships [10].

In this paper, we focus on how to learn a joint representation of visual and auditory modalities for emotion prediction. To learn the highly non-linear relationships among different modalities with very different statistical properties, we adopt the Multimodal Deep Boltzmann Machine (DBM) [10]. The learning of DBM is unsupervised and thus is suitable for our case as a plenty of unlabeled video data is available on the Web. Among the user-generated videos, the textual information (e.g., title and description) is somewhat invariant to large changes in the visual and auditory modalities [2]. For example, the videos expressing "joy" emotion range from showing a wedding party with romantic music to recording a single person enjoying delicious food, while the textual information always conveys similar words (i.e., happy night). As discussed in [10], this invariance provides a rich learning signal for mapping the variant visual and auditory features to the coherent concept level features. Hence, we expand the multimodal DBM in [10] from two pathways to three pathways, which includes visual, auditory and textual information as shown in Figure 1. In the model, all layers but the first (bottom) layers are binary-valued. All the modalities are first modeled with a two-layer DBM [8] separately. In this stage, the visual and auditory modalities are modeled with Gaussian Restricted Boltzmann Machines (RBMs) [4], which have been widely used for modeling real-valued inputs for speech and vision tasks, whereas the textual modality is modeled with Replicated Softmax model [9], which has been shown to be effective in modeling sparse word count vectors. An additional layer of binary hidden units is added on top of the final layer of each modality to learn the joint representation over the three modalities.

The main contribution of this work is the learning of a joint representation over multiple modalities using unlabeled user-generated videos on the Web. Basically, this work demonstrates a promising way to make use of the large amount of unlabelled user-generated videos for improving the performance of emotion prediction.

2. MULTIMODAL DEEP BOLTZMANN MACHINE

Let $\mathbf{v}^m \in \mathbb{R}^M$ and $\mathbf{v}^a \in \mathbb{R}^A$ denote the real-valued input from visual and auditory inputs. Meanwhile, $\mathbf{v}^t \in \{1, \ldots, K\}$ denotes associated metadata (i.e., title and description) containing N words, with v_k^t denoting the count for the k^{th} word. In addition, the hidden layers of visual pathway are denoted as $\mathbf{h}^{(1m)} \in \{0,1\}^{F_1^m}$, $\mathbf{h}^{(2m)} \in \{0,1\}^{F_2^m}$. Meanwhile, $\mathbf{h}^{(1a)} \in \{0,1\}^{F_1^a}$, $\mathbf{h}^{(2a)} \in \{0,1\}^{F_2^a}$ are for auditory pathway and $\mathbf{h}^{(1t)} \in \{0,1\}^{F_1^t}$, $\mathbf{h}^{(2t)} \in \{0,1\}^{F_2^t}$ are for textual pathway.

As shown in Figure 1, each modality is first modeled with a separate two-layer DBM [8]. The probability that generates a visible vector $\mathbf{v}_m$ by the visual-pathway DBM with Gaussian RBM [4] is given by

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

where $\mathcal{Z}(\theta^m)$ is the partition function and the free energy E is given by

$$\begin{aligned} E(\mathbf{v}^m, \mathbf{h}^{(1m)}, \mathbf{h}^{(2m)}; \theta^m) &= \sum_{i=1}^{M} \frac{(v_{mi} - b_{mi})^2}{2\delta_{mi}^2} \\ -\sum_{i=1}^{M}\sum_{j=1}^{F_1^m} \frac{v_{mi}}{\delta_{mi}} W_{ij}^{(1m)} h_j^{(1m)} &- \sum_{j=1}^{F_1^m}\sum_{l=1}^{F_2^m} h_j^{(1m)} W_{jl}^{(2m)} h_j^{(2m)} \end{aligned} \tag{2}$$

Note that for brevity, the bias terms on the hidden layers are omitted. The probability for an auditory vector $\mathbf{v}_a$ is similar to that of visual-pathway DBM. Here, we only give the formula of free energy E from textual-pathway with Replicated Softmax [9]

$$\begin{aligned} E(\mathbf{v}^t, \mathbf{h}^{(1t)}, \mathbf{h}^{(2t)}; \theta^t) &= -\sum_{k=1}^{K} b_k^t v_k^t - \sum_{k=1}^{K}\sum_{j=1}^{F_1^t} W_{k,j}^{(1t)} h_j^{(1t)} v_k^t \\ -\sum_{j=1}^{F_1^t}\sum_{l=1}^{F_2^t} W_{jl}^{(2t)} h_j^{(1t)} h_l^{(2t)} &- N\sum_{j=1}^{F_1^t} b_j^{(1t)} h_j^{(1t)} - \sum_{l=1}^{F_2^t} b_l^{(2t)} h_l^{(2t)} \end{aligned} \tag{3}$$

The multimodal DBM is formed by adding an additional binary-valued layer $\mathbf{h}^{(3)} \in \{0,1\}^{F_3}$ on top of the second hidden layers from different pathways $(\mathbf{h}^{2m}, \mathbf{h}^{2a}, \mathbf{h}^{2t})$. Then, the joint density distribution over the multimodal inputs is given by

$$\begin{aligned} P(\mathbf{v}^m, \mathbf{v}^a, \mathbf{v}^t; \theta) = &\sum_{h^{(2m)},h^{(2a)},h^{(2t)},h^{(3)}} P(h^{(2m)}, h^{(2a)}, h^{(2t)}, h^{(3)}) \\ &(\sum_{h^{(1m)}} P(v^m, h^{(1m)}, h^{(2m)}))(\sum_{h^{(1a)}} P(v^a, h^{(1a)}, h^{(2a)})) \\ &(\sum_{h^{(1t)}} P(v^t, h^{(1t)}, h^{(2t)})) \end{aligned} \tag{4}$$

2.1 Network training

Multiple layers of hidden units and multiple modalities make the learning difficult [10]. Therefore, the learning is split into two stages. First, each RBM component of multimodal DBM is pretrained by using the greedy layerwise pretraining strategy [8]. Then, the learnt parameters are used to initialize the parameters of all layers in multimodal DBM, and the mutlimodal DBM is trained to finetune different modalities in an unified way. As discussed in [3], 1-step contrastive divergence (CD_1) is adopted for pretraining. Meanwhile, persistent contrastive divergence (PCD) is used for the whole multimodal DBM learning, which learns significantly better models than CD_1 but costs more time than CD_1.

2.2 Feature Extraction and Classification

Based on the joint distribution of multimodal DBM (Equation 4), we can easily infer the conditional distribution over the joint representation layer as the final features extracted from the model by logistic function $\sigma(x)$:

$$\begin{aligned} p(h_p^{(3)} = 1|\mathbf{h}^{(2)}) = \sigma(&\sum_{l=1}^{F_2^m} W_{lp}^{(3m)} h_l^{(2m)} + \sum_{l=1}^{F_2^a} W_{lp}^{(3a)} h_l^{(2a)} \\ &+ \sum_{l=1}^{F_2^t} W_{lp}^{(3t)} h_l^{(2t)} + b_p^{(3)}) \end{aligned}$$

When extracting the features, we adopt the mean-field update to approximate the true posteriors [8]. Since the multimodal DBM is a fully generative model, we can still generate the joint feature representation even when some modalities are missing by unclamping the missing inputs and updating them after each mean-field update. The inputs are updated based on the conditional distribution as follows:

$$v_i^m|\mathbf{h}^{(1m)} \sim \mathcal{N}(\delta_i \sum_{j=1}^{F_1^m} W_{ij}^{(1m)} h_j^{(1m)} + b_i^m, \delta_i^2)$$

$$v_i^a|\mathbf{h}^{(1a)} \sim \mathcal{N}(\delta_i \sum_{j=1}^{F_1^a} W_{ij}^{(1a)} h_j^{(1a)} + b_i^a, \delta_i^2)$$

$$p(v_{ik}^t = 1|\mathbf{h}^{(1t)}) = \frac{exp(\sum_{j=1}^{F_1^t} h_j^{(1t)} W_{jk}^{(1t)} + b_k^t)}{\sum_{q=1}^{K} exp(\sum_{j=1}^{F_1^t} h_j^{(1t)} * W_{jq}^{1t} + b_k^t)}$$

The multimodal DBM can effectively fill in the missing modalities. The generated modalities can serve as a plausible proxy for extracting the final output joint representation [10].

With the joint representation as feature, emotion prediction models can be easily trained with linear or non-linear classifiers. In this paper, we adopt the standard Gaussian RBF kernel SVM, which shows good classification performance on real-valued features [5]. One-against-all strategy is adopted to train a separate classifier for each emotion category, and a test sample is assigned to the category with the highest prediction score.

3. EXPERIMENTS

3.1 Dataset and Model Learning

The multimodal DBM is trained on 156,219 videos crawled from YouTube. To make sure that the dataset consists of emotion-related videos and has a balanced distribution over all kinds of emotion-related high-level concepts, the Adjective Noun Pairs (ANPs) provided by SentiBank [2] are used as searching keywords. These ANPs are all emotion-related concepts, such as "happy kids" or "creepy spider". In this work, all the 3,244 ANPs are used. For each ANP, at most 100 videos are collected and the duration of each video is limited to 2 minutes since a long video usually has multiple different concepts. We adopt the same visual and audio feature sets used in [5]. The visual features include Dense SIFT, HOG, SSIM, GIST, and LBP. The MFCC and Audio-Six are extracted as audio clues. Among these features, Dense SIFT, HOG, SSIM and MFCC are quantized into a bag-of-words representation. The visual features are extracted using the codes from the authors of [13] and the audio features are extracted using the software by [7]. Finally, the visual and auditory information are represented as 20,651-dimensional and 4,000-dimensional features respectively. By adopting the same features as [5], the effectiveness of the joint representation by multimodal DBM can be more objectively evaluated. In addition to visual and auditory information, we also make use of the textual description associated with videos. Each associated metadata is represented using a vocabulary of the 3,413 most frequent words (whose frequencies are larger than 800). The average number of words associated with a video is 7.13 with a standard deviation of 6.21.

The visual-pathway consists of a Gaussian RBM with 20,651 visible units followed by 2 standard binary-valued layers with 2048 and 1024 hidden units. The auditory-pathway has the same settings as visual-pathway except that the number of visible units is 4,000. Meanwhile, the textual-pathway consists of a Replicated Softmax Model with 3,413 visible units followed by 2 layers of 1,024 hidden units. The joint layer contains 3,072 hidden units. As mentioned in Section 2.1, each pathway is pretrained using a stack of modified RBMs by CD$_1$ for initializing the DBM model. During pretraining and DBM learning, each dimension of visual and auditory features is mean-centered and normalized to unit variance to avoid the instability problem [3]. In addition, to avoid running separate Markov chains for each word count to get sufficient statistics for model distribution, all word count vectors are scaled so that they sum to 5 [10]. The model is implemented upon the open source tool "deepnet"[1].

Category	$\mathbf{h}^{(1m)}$	$\mathbf{h}^{(2m)}$	$\mathbf{h}^{(3)}$	$\mathbf{h}^{(2a)}$	$\mathbf{h}^{(1a)}$
Anger	0.297	**0.364**	0.294	0.124	0.112
Anticipation	0.040	**0.080**	0.067	0.024	0.034
Disgust	0.228	**0.277**	0.267	0.311	0.145
Fear	0.415	**0.406**	0.395	0.211	0.233
Joy	0.428	0.397	**0.442**	0.291	0.385
Sadness	0.154	0.176	0.217	0.226	**0.235**
Surprise	**0.676**	0.669	0.666	0.665	0.643
Trust	0.093	0.130	**0.136**	0.117	0.077
Overall	0.350	0.360	**0.371**	0.302	0.290

Table 1: Prediction accuracies for each emotion category of VideoEmotion dataset [5] obtained by applying one-vs-all SVM to representations learned at different layers. The highest accuracy of each category is highlighted.

3.2 Classification Performance

The emotion prediction is conducted on VideoEmotion dataset [5], which consists of 1,101 videos and the videos are classified into 8 categories. To measure the effectiveness of DBM, three different combination settings are adopted. Section 3.2.1 shows the performance achieved on combining visual and auditory features. Section 3.2.2 compares the performance of using visual or auditory features to that of using SentiBank [2]. Section 3.2.3 reports the result of combining DBM with [5] by average late fusion.

3.2.1 Multimodal Inputs

Since the associated metadata of each video is not provided in VideoEmotion [5], the input to the textual-pathway in DBM is initialized to zeros. The model is allowed to update the state of the textual input layer when performing mean-field update. In this experiment, we run the mean-field update for 5 times [10]. Table 1 shows the prediction performance for each emotion category. The representations extracted at different layers are used to train the one-vs-all SVM classifiers based on the training data given in [5]. From the table, we can see that the joint representation from joint hidden layer ($\mathbf{h}^{(3)}$) achieves the best performance on the whole dataset, which improves that of the second visual hidden layer and auditory hidden layer by 3.1% and 22.8% respectively. We have also observed that the classifier trained on the representation from the second hidden layer outperforms that from the first hidden layer in both visual and auditory pathways. Although the joint representation does not always achieve the best performance, the performance on joint representation is the most consistent one, which is

[1] https://github.com/nitishsrivastava/deepnet

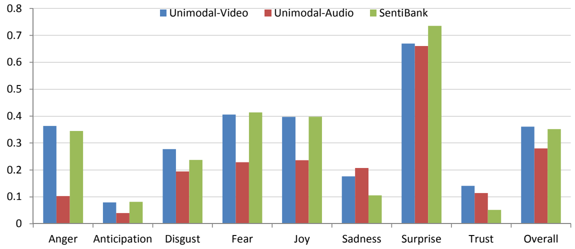

Figure 2: Prediction accuracies. Unimodal-Video and Unimodal-Audio are the classifiers trained on the joint representation generated by using only the visual or auditory information through DBM.

mostly attributed to the well balance over the visual and auditory features in the joint representation.

3.2.2 *Unimodal Input*

In this section, we want to verify the effectiveness of DBM when some modalities are missing. The method proposed in [2] is a good baseline, which also models the emotion prediction task in a hierarchical way by first training ANP detectors and then training SVM-based emotion classifiers on the responses from ANP detectors. In addition, only visual information is used in [2]. For comparison, we generate the joint representation from DBM by using only visual information (Unimodal-Video) or auditory information (Unimodal-Audio). In Unimodal-Video, the auditory and textual information are initialized with zeros and these information are updated when performing mean-field update. Similarly for the Unimodal-Audio, the visual and textual information are initialized with zeros and updated in the same way. As shown in Figure 2, Unimodal-Video outperforms SentiBank by 2.7%. Unimodal-Video exhibits better performance than SentiBank in 4 categories, especially in those categories where Unimodal-Audio achieves better performance. This shows that the DBM model effectively fills in the missing modalities.

3.2.3 *Average Fusion*

To see whether DBM model learns unique correlation between visual and auditory information, the prediction scores from DBM and [5] are averagely fused. Generally speaking, the joint representation from DBM can be treated as latent attributes. Therefore, two fusion settings are adopted. First, the DBM score is combined with the prediction score from [5] using only visual and auditory features (V.+Au.+DBM). Second, the DBM score is fused with the score from [5] using visual, auditory and attribute features (V.+Au.+At.+DBM). From the Table 2, we can see that V.+Au.+DBM achieves better performance than both V.+Au. and V.+Au.+At., which indicates that DBM mines more information than the attribute features used in [5]. In addition, we also observe that V.+Au.+At.+DBM improves the accuracy of V.+Au.+At. by 7.7%, which is so far the best reported performance on this dataset to the best of our knowledge. This result basically indicates that the joint representation is complementary to hand-crafted features, by providing additional latent attributes modeling non-linear relationships among different modalities.

Category	V.+Au.	V.+Au. +DBM	V.+Au. +At.	V.+Au. +At.+DBM
Anger	**0.549**	0.527	0.527	0.509
Anticipation	0.028	**0.067**	**0.067**	0.034
Disgust	0.399	0.381	**0.438**	0.399
Fear	0.396	0.484	0.471	**0.545**
Joy	0.480	0.557	0.484	**0.590**
Sadness	**0.289**	0.274	0.208	0.217
Surprise	0.746	0.802	0.767	**0.828**
Trust	0.311	**0.327**	0.287	0.312
Overall	0.451	0.484	0.463	**0.499**

Table 2: Prediction accuracies by late fusion with the features from [5]. The notations V., Au., and At. represent visual, auditory and attribute features respectively.

4. CONCLUSION

We have presented the learning of joint representation from multiple modalities with DBM. The joint representation has demonstrated the effectiveness of learning unique correlation among different modalities, and being capable of dealing with the problem when some modalities of training or testing examples are missing. Currently, the model still needs hand-crafted features as inputs. We will leverage the strength of RBMs and Convolutional Neural Network to learn features directly from raw inputs in the future.

5. ACKNOWLEDGMENTS

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11210514), and a grant from the National Hi-Tech Research and Development Program (863 Program) of China under Grant 2014AA015102. The Tesla K40 used for this research was donated by the NVIDIA Corporation.

6. REFERENCES

[1] Y. Baveye, J.-N. Bettinelli, E. Dellandreá, L. Chen, and C. Chamaret. A large video database for computational models of induced emotion. In *ACII*, 2013.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[3] G. Hinton. A practical guide to training restricted boltzmann machines. Technical report, 2010.

[4] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504 – 507, 2006.

[5] Y.-G. Jiang, B. Xu, and X. Xue. Predicting emotions in user-generated videos. In *AAAI*, 2014.

[6] B. Jou, S. Bhattacharya, and S.-F. Chang. Predicting viewer perceived emotions in animated gifs. In *ACM MM*, 2014.

[7] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard. Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, 2010.

[8] R. Salakhutdinov and G. Hinton. Deep Boltzmann machines. In *AI Statistics*, 2009.

[9] R. Salakhutdinov and G. Hinton. Replicated softmax: an undirected topic model. In *NIPS*, 2010.

[10] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. *JMLR*, 15:2949–2980, 2014.

[11] R. M. Teixeira, T. Yamasaki, and K. Aizawa. Determination of emotional content of video clips by low-level audiovisual features. *Multimed. tools appl.*, pages 1–29, 2011.

[12] H. L. Wang and L.-F. Cheong. Affective understanding in film. *CSVT*, 16(6):689–704, 2006.

[13] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[14] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui. Hierarchical affective content analysis in arousal and valence dimensions. *Signal Proc.*, 93:2140–2150, 2013.

[15] Q. You, J. Luo, H. Jin, and J. Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *AAAI*, 2015.