11-2015

Deep multimodal learning for affective analysis and retrieval

Lei PANG

Shiai ZHU

Chong-wah NGO
Singapore Management University, cwngo@smu.edu.sg

Citation

1

Deep Multimodal Learning for Affective Analysis and Retrieval

Lei Pang, Shiai Zhu, and Chong-Wah Ngo

***Abstract*—Social media has been a convenient platform for voicing opinions through posting messages, ranging from tweeting a short text to uploading a media file, or any combination of messages. Understanding the perceived emotions inherently underlying these user-generated contents (UGC) could bring light to emerging applications such as advertising and media analytics. Existing research efforts on affective computation are mostly dedicated to single media, either text captions or visual content. Few attempts for combined analysis of multiple media are made, despite that emotion can be viewed as an expression of multimodal experience. In this paper, we explore the learning of highly non-linear relationships that exist among low-level features across different modalities for emotion prediction. Using the deep Bolzmann machine (DBM), a joint density model over the space of multimodal inputs, including visual, auditory, and textual modalities, is developed. The model is trained directly using UGC data without any labeling efforts. While the model learns a joint representation over multimodal inputs, training samples in absence of certain modalities can also be leveraged. More importantly, the joint representation enables emotion-oriented cross-modal retrieval, for example, retrieval of videos using the text query "crazy cat". The model does not restrict the types of input and output, and hence, in principle, emotion prediction and retrieval on any combinations of media are feasible. Extensive experiments on web videos and images show that the learnt joint representation could be very compact and be complementary to hand-crafted features, leading to performance improvement in both emotion classification and cross-modal retrieval.**

***Index Terms*—Cross-modal retrieval, deep Boltzmann machine, emotion analysis, multimodal learning.**

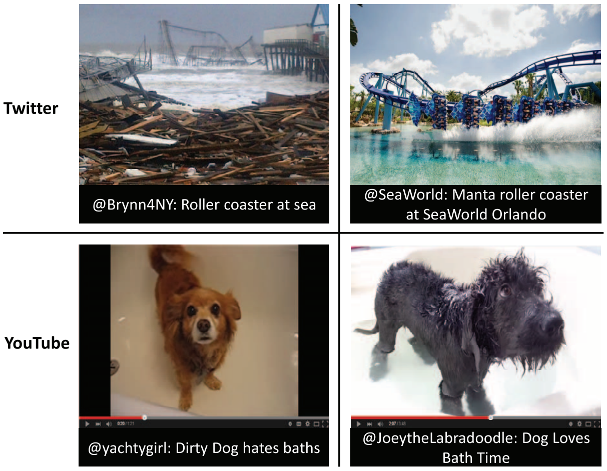

Twitter
YouTube
@Brynn4NY: Roller coaster at sea
@SeaWorld: Manta roller coaster at SeaWorld Orlando
@yachtygirl: Dirty Dog hates baths
@JoeytheLabradoodle: Dog Loves Bath Time

Fig. 1. Examples of emotional images and videos with their associated tags and titles from Twitter and YouTube. The left image in the first row describes the destruction caused by Hurricane Sandy and the right one describes the Manta roller coaster at Sea World Orlando. The textual descriptions for these two images are almost the same and the implicit emotions can only be predicted by the visual information. In contrast, emotion classification in the lower two videos with similar visual appearances is performed using the strong clues in the textual captions and auditory information.

I. Introduction

SOCIAL MEDIA is an opinion-rich knowledge source including plenty of timely user-generated-contents (UGC) with different media types. Automatically understanding the emotional status of users from their uploaded multimedia contents is in high demands for many applications [1]. For example, when searching information about a resort, the retrieved images or videos can be ranked 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. This function can also be used by governments to better understand people's reactions towards their new policies.

The existing works on affective analysis of UGC data are mostly devoted to single media [2]–[10]. For example, linguistic [2], [4] and semantic [3], [5] features are both adopted for text-based analysis. However, inferring perceived emotion signals underlying short messages that are usually sparse in textual description is not easy. Fig. 1 (1st row) shows two images of "sadness" and "happiness" emotions respectively. Nevertheless, no obvious emotion clues are observed by merely reading their text captions, which play the roles of referring visual content (roller coaster) rather than emotion in the images. In other words, the captions convey semantic meanings while the actual emotion signals are buried inside the images. Apparently, visual content such as color contrast and tone provide more vivid clues to reveal the underlying emotions for this example. The second row of Fig. 1 shows a counter example, where the emotions are subject to user perceptions by looking at the snapshots of dogs sampled from two different videos. By turning off the visual signals and only listening to audio effects, the woofing and

Manuscript received April 20, 2015; revised September 11, 2015 and September 22, 2015; accepted September 22, 2015. Date of publication September 25, 2015; date of current version October 20, 2015. This work was supported in part by the Research Grants Council of the Hong Kong Special Administrative Region, China, under Grant CityU 11210514 and Grant CityU 120213, and by the National Hi-Tech Research and Development Program (863 Program) of China under Grant 2014AA015102. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

L. Pang and C.-W. Ngo are with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: leipang3-c@my.cityu.edu.hk; cwngo@cs.cityu.edu.hk).

S. Zhu is with the Multimedia Communications Research Laboratory (MCRLab), School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: zshiai@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2482228

laughing sounds of dogs already convey a strong sense of emotions. In this example, the moods underlying videos are also somewhat captured by the words “dirty”, “hates” and “loves” in the captions.

Due to these limitations, there are several works studying the fusion of multimodal features, including multi-kernel fusion [1], conditional random field [11] and Bayesian rules [12]. These works are based on the standard early or late fusion strategies [11]–[14], despite the employment of different machine learning models. The “shallow” way of combining different features is also questionable in principle, given the diverse representations and correlational structures among different modalities. For example, it remains an open problem on the right way of combining raw pixels and audio waveform, for joint understanding of phonemes and visemes (lip pose and motion), in speech recognition [15]. In brief, the highly non-linear relationships existing between different modalities, particularly, are often overlooked by the existing approaches.

Emotion is also correlated with surrounding context, specifically the objects, sounds and scenes in images or videos. For example, a video with “a man listening to bird chirping in the garden” casts an enjoy mood useful for emotion prediction. SentiBank [6] is one such recent effort that explicitly identifies 3,244 adjective noun pairs (ANPs) for learning large-scale emotion-oriented concept classifiers. Examples of ANPs are “amazing flowers”, “awesome view” and “shy smile”. Nevertheless, due to the open nature of how nouns and adjectives are combined, SentiBank is hard to be generalized to cover the possible ANPs, not even mentioning the daunting efforts required in labeling of training examples for training ANP classifiers free of sample noise.

In this paper, we propose a more generalized framework for unsupervised feature learning based on Deep Boltzmann machine (DBM) [16], aiming to learn features coupling emotion and semantic signals buried in multimodal signals. As we consider eight types of wildly different features in terms of statistical properties, deep-based learning is preferable for inferring non-linear relationships among features within and across modalities. A joint embedded space shared by multimodal signals is expected to capture such relationships with semantic and emotion contexts. As studied in other works [17], a joint space learnt by DBM is capable of preserving both common and modality-specific information. The learning of DBM is unsupervised and thus is suitable for our problem as plenty of weakly labeled training examples are freely available on social media websites. Although these examples are without careful hand-labels, the text captions can still somewhat provide rich learning signals for mapping diverse visual and auditory features to a coherent embedded space shared by different modalities. For example, the text captions in Fig. 1 provide clue connecting two visually dissimilar roller coasters. The word “joy” can possibly link diverse events, such as wedding party and couple hugging, of different audio-visual effects.

Traditionally emotion prediction and semantic classification are treated as two separate tasks. Emotion-oriented queries such as “dog hates bath” always posts challenges for no clear understanding of how classifiers of different natures should be combined. SentiBank [6] could possibly deal with such queries, but is inherently limited by the number of ANP vocabularies. Our model, empowered on the joint space learnt through DBM, can more naturally answer these queries, assuming the availability of a huge number of training examples with wild but rich textual, visual and auditory signals. Due to unsupervised learning, our model has a much better capacity and scalability than SentiBank in capturing emotional experience in multimodal settings. Furthermore, the joint space also enlightens cross-modal retrieval, where for example, either text-to-video or video-to-text search can be performed under our model.

The main contribution of this paper is the proposal of a deep multimodal learning platform that enables a more generalized way of learning features coupled with emotions and semantics. Empirical studies also demonstrate the feasibility of employing such features for multi-modal affective classification and retrieval. The remaining of this paper is organized as follows. Section II presents related works. Section III presents the deep network architecture, followed by Sections IV and V describing the joint space representation and network learning respectively. Section VI and VII presents experimental results on affective analysis and retrieval respectively, and finally Section VIII concludes this paper.

II. Related Work

Affective computation has been extensively studied in the last decades, and many methods are proposed for handling various media types including textual documents [2]–[5], images [6]–[10], music [18]–[20] and movies [11]–[14]. Two widely investigated tasks are emotion detection and sentiment analysis. Both of them are standard classification problems with different state spaces. Usually emotion detection is defined on several discrete emotions, such as anger, sadness, joy etc., while sentiment analysis aims at categorizing data into positive or negative. Since the adopted techniques of these two tasks are quite similar, we will not differentiate them in this section. Previous efforts are summarized mainly based on the modality of the data they are working on.

For textual data, lexicon-based approach using a set of pre-defined emotional words or icons has been proved to be an effective way. In [2], they propose to predict the sentiment of tweets by using the emoticons (e.g., positive emoticon “:)” and negative one “: =(”) and acronyms [e.g., lol (laugh out loudly), gr8 (great) and rotf (rolling on the floor)]. A partial tree kernel is adopted to combine the emoticons, acronyms and Part-of-Speech (POS) tags. In [4], three lexicon emotion dictionaries and POS tags are leveraged to extract linguistic features from the textual documents. In [3], a semantic feature is proposed to address the sparsity of microbloggings. The non-appeared entities are inferred using a pre-defined hierarchical entity structure. For example, “iPad” and “iPhone” indicate the appearance of “Product/Apple”. Furthermore, the latent sentiment topics are extracted and the associated sentiment tweets are used to augment the original feature space. In [5], a set of sentimental aspects, such as opinion strength, emotion and polarity indicators, are combined as meta-level features for boosting the sentiment classification on Twitter messages.

Affective analysis of images adopts a similar framework with general concept detection. In SentiBank [6], a set of visual concept classifiers, which are strongly related to emotions and sentiments, are trained based on unlabeled Web images. Then, a SVM classifier is built upon the output scores of these concept

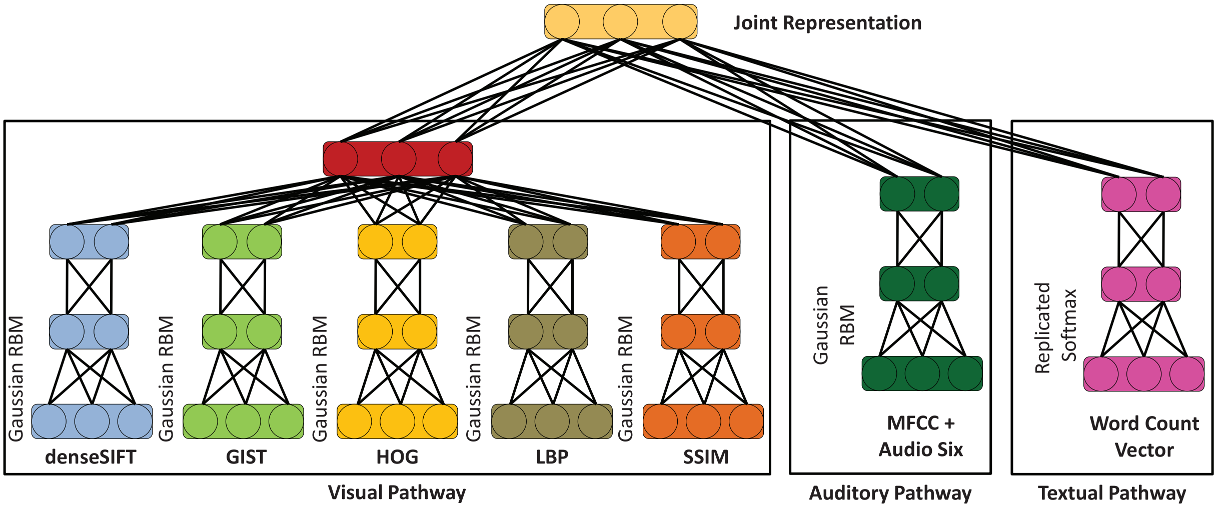


Fig. 2. Multimodal DBM that models the joint distribution over visual, auditory, and textual features. All layers but the first (bottom) layers use standard binary units. The Gaussian RBM model is used to model the distributions over the visual and auditory features. The replicated Softmax topic model is applied on the textual features.

classifiers. The performance of SentiBank is recently improved by using deep convolution neural network (CNN) [7]. Nevertheless, the utility of SentiBank is limited by the number and kind of concepts (or ANPs). Due to the fact that ANPs are visually emotional concepts, selection of right samples for classifier training could be subjective. In addition to the semantic level features, a set of low-level features, such as color-histogram and visual aesthetics, are also adopted in [8]. The combined features are then fed into a multi-task regression model for emotion prediction. In [21], hand-crafted features derived from principles-of-art such as balance and harmony are proposed for recognition of image emotion. In [10], the deep CNN is directly used for training sentiment classifiers rather than using a mid-level consisting of some general concepts. Since Web images are weakly labeled, the system progressively select a subset of the training instances with relatively distinct sentiment labels to reduce the impact of noisy training instances.

For emotional analysis of music, various hand-crafted features corresponding to different aspects (e.g., melody, timbre and rhythm) of music are proposed. In [19], the early fused features are characterized by cosine radial basis function (RBF). In [22], a ListNet layer is added on top of the RBF layer for ranking the music in valence and arousal in Cartesian coordinates. Besides hand-crafted features, the authors in [20] adopt deep belief networks (DBN) on the Discrete Fourier Transforms (DFTs) of music signals. Then, SVM classifiers are trained on the latent features from hidden layers.

In the video domain, most research efforts are dedicated to movies. In [14], a large emotional dataset, which contains about 9,800 movie clips, is constructed. SVM classifiers are trained on different low-level features, such as audio features, complexity and color harmony. Then, late fusion is employed to combine the classifiers. In [13], a set of features are proposed based on psychology and cinematography for affective understanding in movies. Early fusion is adopted to combine the extracted features. Other fusion strategies on auditory and visual modalities are studied in [12]. In [11], a hierarchical architecture is proposed for predicting both emotion intensity and emotion types. CRF is adopted to model the temporal information in the video sequence. In addition to movies, a large-scale Web video dataset for emotion analysis is recently proposed in [1], where a simplified multi-kernel SVM is adopted to combine the features from different modalities.

Different from those works, the approach proposed in this paper is a fully generative model, which defines a joint representation for various features extracted in different modalities. More importantly, the joint representation conveying information from multiple modalities can still be generated when some modalities are missing, which means that our model does not restrict to the media types of user generated contents.

III. Deep Network Design

Fig. 2 shows the proposed network architecture, which is composed of three different pathways respectively for visual, auditory and textual modalities. Each pathway is formed by stacking multiple Restricted Boltzmann Machines (RBM), aiming to learn several layers of increasingly complex representations of individual modality. Similar to [23], we adopt Deep Boltzmann Machine (DBM) [16] in our multimodal learning framework. Different from other deep networks for extracting feature, such as Deep Belief Networks (DBN) [24] and denoising Autoencoders (dA) [25], DBM is a fully generative model which can be utilized for extracting features from data with certain missing modalities. Additionally, besides the bottom-up information propagation in DBN and dA, a top-down feedback is also incorporated in DBM, which makes the DBM more stable on missing or noisy inputs such as weakly labeled data on the Web. The pathways eventually meet and the sophisticated non-linear relationships among three modalities are jointly learned. The final joint representation can be viewed as a shared embedded space, where the features with very different statistical properties from different modalities can be represented in an unified way.

The proposed architecture is more generalized and powerful in terms of scale and learning capacity. In visual pathway, the low-level features amount to 20,651 dimensions, resulting in large number of parameters to be trained if connecting them directly to the hidden layer. Instead, we design a separate pathway for each low-level feature, which requires less parameters and hence more flexible and efficient to train. This advantage makes our system more scalable to handling higher dimensional features, rather than features of 3,875 dimensions used in [23]. We further consider learning the separated pathways in visual modality in parallel. The computational cost can be further reduced. Furthermore, we generate a compact representation which represents the common feature and preserves the unique characteristic of each visual feature. In this way, it will not overwhelm other modalities because of high dimensionality during joint representation learning. Auditory and textual pathway do not suffer from this problem. However, the proposed structure can be easily extended for other modalities.

It is worth noticing that the high-level semantics in visual and auditory modalities can be represented in the final joint representation, by considering the correlations between them and textual inputs during training. Since our model is fully generative, the semantics of input data without textual modality can also be extracted. Other semantic features for visual and auditory data (e.g., SentiBank [6], Classemes [26] and Objectbank [27]) basically adopt the shallow learning models, which learn the local patterns extracted from the data. These methods suffer from information loss [28], [29], and is sensitive to the diverse appearance of input data. In contrast, our model has the capability of mining generative representations from the raw data, which has been proved to be more powerful [17], [30].

IV. Multimodal Joint Representation

Our network is built upon RBMs. A standard RBM has two binary-valued layers, i.e., visible layer (denoted as $\mathbf{v}$) and hidden layer (denoted as $\mathbf{h}$). The probability distribution over the inputs $\mathbf{v}$ is defined as

$$P(\mathbf{v}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{h}} exp(-E(\mathbf{v}, \mathbf{h})). \tag{1}$$

$E(\mathbf{v}, \mathbf{h})$ is the free energy between $\mathbf{v}$ and $\mathbf{h}$, given by

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{ij} v_i W_{ij} h_j - \sum_{i} a_i v_i - \sum_{j} b_j h_j \tag{2}$$

where $\mathbf{W}$ is the weight of link connecting two layers, $\mathbf{a}$ and $\mathbf{b}$ are the bias weights for $\mathbf{v}, \mathbf{h}$ respectively. The feature learning problem is elegantly stated to maximize the probability in (1) or to minimize the free energy in (2). The standard RBM can only handle binary-valued inputs. Other generalized RBMs include Gaussian RBM [31] designed for modeling real-valued inputs and Replicated Softmax [32] for modeling sparse word count vectors. Next, we describe different pathways and their joint representation. Each pathway consists of a stack of RBMs selected according to the property of input data.

A. Visual Pathway

The visual input consists of five complementary low-level features widely used in previous works [1], [6]. As shown in Fig. 2, each feature is modeled with a separate two-layer DBM. Let $\mathcal{K} = \{d, g, o, l, s\}$ denote the set of five features, respectively as DenseSIFT [33], GIST [34], HOG [35], LBP [36] and SSIM [37]. Furthermore, let $\mathbf{v}_{\mathcal{V}} = \{\mathbf{v}^k\}$, $\mathbf{h}^1_{\mathcal{V}} = \{\mathbf{h}^{(1k)}\}$ and $\mathbf{h}^2_{\mathcal{V}} = \{\mathbf{h}^{(2k)}\}$ as the sets of real-valued inputs, first and second hidden layers respectively, where $k \in \mathcal{K}$. For example, $\mathbf{v}^d$ refers to the visible layer for DenseSIFT. In addition, the joint layer in visual pathway (the layer in red in Fig. 2 is denoted as $\mathbf{h}^{(v)}$

The connections between $\mathbf{v}^k$ and $\mathbf{h}^{(1k)}$ are modeled with Gaussian RBM [31] and the connections between $\mathbf{h}^{(1k)}$ and $\mathbf{h}^{(2k)}$ are modeled with standard binary RBM. Hence, the probability distribution over the real-valued input $\mathbf{v}^k$ is given by

$$P(\mathbf{v}^k; \theta^k) = \frac{1}{\mathcal{Z}(\theta^k)} \sum_{\mathbf{h}^{(1k)}, \mathbf{h}^{(2k)}} exp(-E(\mathbf{v}^k, \mathbf{h}^{(1k)}, \mathbf{h}^{(2k)}; \theta^k) \tag{3}$$

where $\mathcal{Z}(\theta^k)$ is the partition function and the free energy E is defined as

$$E\left(\mathbf{v}^k, \mathbf{h}^{(1k)}, \mathbf{h}^{(2k)}; \theta^k\right) = \sum_{i} \frac{\left(v_i^k - b_i^k\right)^2}{2(\delta_i^k)^2} - \sum_{ij} \frac{v_i^k}{\delta_i^k} W_{ij}^{(1k)} h_j^{(1k)} - \sum_{jl} h_j^{(1k)} W_{jl}^{(2k)} h_j^{(2k)} \tag{4}$$

where $\theta^k = \{\mathbf{a}, \mathbf{b}, \mathbf{W}^{(1k)}, \mathbf{W}^{(2k)}\}$ are the model parameters. Note that for brevity, the bias terms $\mathbf{a}$ on the hidden layers are omitted. To generate the joint representation over these five low-level features, we combine the five DBM models by adding an additional layer $\mathbf{h}^{(v)}$ on top of them. Then, the joint density distribution over the five features $\mathbf{v}_{\mathcal{V}}$ is given by

$$P\left(\mathbf{v}_{\mathcal{V}}; \theta^v\right) = \sum_{\mathbf{h}^2_{\mathcal{V}}, \mathbf{h}^{(v)}} P\left(\mathbf{h}^2_{\mathcal{V}}, \mathbf{h}^{(v)}\right) \times \left(\prod_{k \in \mathcal{K}} \left(\sum_{\mathbf{h}^{(1k)}} P\left(\mathbf{v}^k, \mathbf{h}^{(1k)} | \mathbf{h}^{(2k)}\right)\right)\right). \tag{5}$$

The density distribution $P(\mathbf{h}^2_{\mathcal{V}}, \mathbf{h}^{(v)})$ in (5) is given by

$$P(\mathbf{h}^2_{\mathcal{V}}, \mathbf{h}^{(v)}) = \frac{1}{\mathcal{Z}(\theta^v)} \prod_{k \in \mathcal{K}} exp\left(\sum_{pq} W^{(3k)} h_p^{(2k)} h_q^{(v)}\right). \tag{6}$$

The joint distribution $P(\mathbf{v}^k, \mathbf{h}^{(1k)} | \mathbf{h}^{(2k)})$ of $\mathbf{v}^k$ and $\mathbf{h}^{(1k)}$ over $\mathbf{h}^{(2k)}$ in (5) can be easily inferred from (3) as

$$P(\mathbf{v}^k, \mathbf{h}^{(1k)} | \mathbf{h}^{(2k)}) = \frac{1}{\mathcal{Z}(\theta^k)} exp\left(-\sum_{i} \frac{\left(v_i^k - b_i^k\right)^2}{2\delta_i^2} + \sum_{ij} \frac{v_i^k}{\delta_i} W_{ij}^{(1k)} h_j^{(1k)} + \sum_{jp} W_{jp}^{(2k)} h_j^{(1k)} h_p^{(2k)}\right). \tag{7}$$

Until now, all the probability distributions in (5) are provided and the probability distribution over the whole set of input features $\mathbf{v}_{\mathcal{V}}$ in visual pathway can be easily inferred by subscribing these equations.

B. Auditory Pathway

The input features adopted in auditory pathway are MFCC [38] and Audio-Six (i.e., Energy Entropy, Signal Energy, Zero

Crossing Rate, Spectral Rolloff, Spectral Centroid, and Spectral Flux) [1]. The Audio-Six descriptor, which can capture different aspects of an audio signal, is expected to be complementary to the MFCC. Since the dimension of Audio-Six is only six, we directly concatenate the MFCC feature with Audio-Six rather than separating them into two sub-pathways as the design in visual pathway. The correlation between these two features can be learned by the deep architecture of DBM [23]. Let $\mathbf{v}_a$ denote the real-valued auditory features and $\mathbf{h}^{(1a)}$ and $\mathbf{h}^{(2a)}$ represent the first and second hidden layers respectively. Similar to (3), the DBM is constructed by stacking one Gaussian RBM and one standard binary RBM.

C. Textual Pathway

Different from the visual and auditory modalities, the inputs of the textual pathway are discrete values (i.e., count of words). Thus, we use Replicated Softmax [32] to model the distribution over the word count vectors. Let $\mathbf{v}_t$ as a visible unit denoting the associated metadata (i.e., title and description) of a video t, and v_k^t denotes the count of the kth word in a pre-defined dictionary containing K words. The first and second hidden layers are $\mathbf{h}^{(1t)}$ and $\mathbf{h}^{(2t)}$. Then, the probability of generating $\mathbf{v}_t$ by the text-specific two-layer DBM is given by

$$P(\mathbf{v}_t;\theta^t) = \frac{1}{\mathcal{Z}(\theta^t)} \sum_{\mathbf{h}^{(1t)},\mathbf{h}^{(2t)}} exp\left(\sum_{jk} W_{kj}^{(1t)} h_j^{(1t)} v_k^t\right. + \sum_{jl} W_{jl}^{(2t)} h_j^{(1t)} h_l^{(2t)} + N\sum_j b_j^{(1t)} h_j^{(1t)}). \quad (8)$$

Note that the bias term of the first hidden layer $\mathbf{h}^{(1t)}$ is scaled up by the length of the document. This scaling is important for allowing hidden units to behave sensibly when dealing with documents of different lengths. As stated in [23], [32], without the bias scaling, the scale of the weights would be optimized to fit to the average document length. This would induce that the longer documents tend to saturate the units and shorter ones may be ambiguous on activating the hidden units.

D. Joint Representation

To combine the learned representations of DBMs for the three modalities, an additional layer is added on top of the three pathways, which is annotated as "Joint Representation" in Fig. 2. We denote this layer as $\mathbf{h}^{(J)}$. We further use $\mathbf{v} = \{\mathbf{v}_\mathcal{V}, \mathbf{v}_a, \mathbf{v}_t\}$ to represent all the visible inputs. The final joint density distribution over multi-model inputs can be written as

$$\begin{aligned} P(\mathbf{v};\theta) = &\sum_{\mathbf{h}^{(v)},\mathbf{h}^{(2a)},\mathbf{h}^{(2t)},\mathbf{h}^{(J)}} P\left(\mathbf{h}^{(v)},\mathbf{h}^{(2a)},\mathbf{h}^{(2t)},\mathbf{h}^{(J)}\right) \\ &\cdot \left(\sum_{\mathbf{h}_\mathcal{V}^1,\mathbf{h}_\mathcal{V}^2} P\left(\mathbf{v}_\mathcal{V},\mathbf{h}_\mathcal{V}^1,\mathbf{h}_\mathcal{V}^2|\mathbf{h}^{(v)}\right)\right) \\ &\cdot \left(\sum_{\mathbf{h}^{(1a)}} P\left(\mathbf{v}_a,\mathbf{h}^{(1a)}|\mathbf{h}^{(2a)}\right)\right) \\ &\cdot \left(\sum_{\mathbf{h}^{(1t)}} P\left(\mathbf{v}_t,\mathbf{h}^{(1t)}|\mathbf{h}^{(2t)}\right)\right). \end{aligned} \quad (9)$$

By subscribing (5) and (8) into above equation, the probability distribution over the multiple inputs formulated by the proposed network can be easily inferred. We do not show the detail formula here due to the space limitation.

V. Networking Learning and Inferencing

A. Approximate Network Learning

The learning of our proposed model is not trivial due to multiple layers of hidden units and multiple modalities. Inspired by [23], we split the learning process into two stages. First, each RBM component of the proposed multimodal DBM is pretrained by using the greedy layerwise pretraining strategy [16]. In this stage, the time cost for exactly computing the derivatives of the probability distributions with respect to parameters increases exponentially with the number of units in the network. Thus, we adopt 1-step contrastive divergence (CD_1), an approximate learning method, to learn the parameters. In CD_k algorithm, a k-step Markov chain is initialized with the training sample. The stochastic reconstruction of the training sample from Markov chain by Gibbs sampling has a decreased free energy. Hence, this reconstruction can be approximately treated as the distribution generated by the RBM model. The contrast between the training sample and its reconstruction is used to approximate the direction of the change for the parameters. In practice, CD_1 is widely used for RBM training, since good approximation of the changing direction is already obtained when $k = 1$. Note that, CD_1 actually performs poorly in approximating the size of the change in parameters. However, it is accurate enough for learning a RBM to provide hidden features for a high-level RBM training [39]. This is because CD_1 retains most of the information in the inputs.

As discussed in [39], CD_1 is still far from optimal to be used for learning a joint-density model. Therefore, in the joint learning stage, we adopt a more radical departure from CD_1, named as "persistent contrastive divergence" (PCD) [40]. In contrast to initialize each alternating Gibbs Markov chain at a training sample, the states of a number of persistent chains or "fantasy particles" are tracked in PCD. Each persistent chain has its hidden and visible states, which are generated by running mean-field updates with Gibbs sampling for one or a few times after each weight is updated. Then the derivative of the probability distribution is approximated by the difference between the pairwise statistics measured on a mini-batch of data and the persistent chains. Since the weight-updates repel each chain from its current state by raising the energy of that state, the persistent chains mix surprisingly fast [41]. Furthermore, PCD also learns significantly better models than CD_1 or even CD_{10} as reported in [40].

B. Joint Representation Inferring

The representation learnt by the proposed model is a set of distributions over layers conditioned on their adjacent layers. If all the modalities are present, the element $h_p^{(J)}$ in joint representation is inferred by Gibbs sampling as

$$\begin{aligned} &p\left(h_p^{(J)} = 1|\mathbf{h}^{(v)},\mathbf{h}^{(2a)},\mathbf{h}^{(2t)}\right) \\ &\quad = g\left(\sum_q W_{qp}^{(Jv)} h_q^{(v)} + \sum_l W_{lp}^{(Ja)} h_l^{(2a)} + \sum_m W_{mp}^{(Jt)} h_m^{(2t)}\right) \end{aligned} \quad (10)$$

TABLE I

NUMBER OF NEURONS IN EACH LAYER OF OUR ENHANCED MULTIMODAL DBM (E-MDBM). $\mathbf{v}$, $\mathbf{h}^1$, $\mathbf{h}^2$, $\mathbf{h}^{(v)}$ AND $\mathbf{h}^{(J)}$ REPRESENT THE VISIBLE LAYERS, FIRST HIDDEN LAYERS, SECOND HIDDEN LAYERS, JOINT REPRESENTATION LAYER OVER VISUAL FEATURES, AND JOINT REPRESENTATION LAYER OVER VISUAL, AUDITORY, AND TEXTUAL MODALITIES

Features	$\mathbf{v}$	$\mathbf{h}^1$	$\mathbf{h}^2$	$\mathbf{h}^{(v)}$	$\mathbf{h}^{(J)}$
DenseSIFT	6,300	2,048	1,024	2,048	4,096
GIST	512	1,024	1,024		
HOG2x2	6,300	2,048	1,024		
LBP	1,239	2,048	1,024		
SSIM	6,300	2,048	1,024		
MFCC+AudioSix	4,000	2,048	1,024	-	
Word Count Vector	4,314	2,048	1,024	-	

where $g(x) = 1/(1 + exp(-x))$ and bias term is omitted from presentation. For full details of inference at each hidden layer, please refer to Appendix.

There are two ways to generate the joint representation if some modalities are not available. First we can directly generate the joint representation based on the existing modalities only and leave the missing ones out. For example, if text is missing, the joint representation will be computed by $p(h_p^{(J)} = 1|\mathbf{h}^{(v)}, \mathbf{h}^{(2a)})$. The second way is to infer the missing modalities by alternating Gibbs sampling. Meanwhile, the joint representation is updated with the generated data of missing modalities. For example, assuming that the textual modality is missing, the observed visual modality $\mathbf{v}_{\mathcal{V}}$ and auditory modality $\mathbf{v}_a$ are clamped at the inputs and all hidden units are initialized randomly. Alternating Gibbs sampling is used to draw samples from $P(\mathbf{v}_t|\mathbf{v}_{\mathcal{V}}, \mathbf{v}_a)$ by updating each hidden layer given the states of the adjacent layers. As reported in [23], the second method achieves better performance than the first one, which indicates that the mutlimodal DBM can generate meaningful representations of the missing modalities.

C. Discussions

While the proposed architecture follows the principle [23], the main novelty comes from the design of multiple visual pathways. Despite that the architecture may appear more complicated than [23] at first glance, the design indeed simplifies [23] by significantly reducing the number of hyper parameters. With reference to Table I that lists the number of neurons in each layer, the network contains 99,690,496 hyper parameters. While this number is terribly high, it requires only 29% of that parameters in [23], for converting the input features from 20,651 to 2,048 dimensions. Furthermore, the design accelerates learning by allowing parallel training of parameters on each pathway. The locally connected hidden units in the pathways also speed up the PCD learning at the second stage. In our experiments, by using Tesla-K20 GPU, network learning completes in about one week with around 1 million training examples. Given the same amount of time, [23] is only able to train a network with input features of 3,875 dimensions. In short, our proposed network is more scalable in learning and effective in testing (see Section VI) than [23].

Overfitting becomes an issue with large number of hyper parameters to be learnt in the network. As stated in [39], assuming that each image contains 1,000 pixels, using 10,000 training examples to learn weights of a million parameters in one RBM is quite reasonable. In the proposed network, the largest RBM has $6,300 \times 2,048$ or around 1.3 millions of parameters. Using 20,000 training samples, for example, is practically feasible for learning this RBM. Since RBMs in the network are learnt in parallel, the chance of overfitting shall not be high even with only around 20,000 training samples. In our case, there are close to a million of training images and videos (see Section VI-A), and we did not observe tendency of overfitting when learning the parameters.

VI. EXPERIMENT: AFFECTIVE ANALYSIS

This section starts by introducing model training with unlabeled images and videos sampled from social media websites (Section VI-A). Two sets of experiments are conducted for affective analysis (Section VI-B and Section VI-C), respectively emotion prediction on YouTube videos and sentiment classification on Twitter images.

A. Model Learning

We constructed two datasets, E-Flickr and E-YouTube, for DBM learning. The images in E-Flickr are crawled from Flickr by using the 3244 ANPs used in SentiBank [6] as keywards. On average, there are 250 images being retrieved for each ANP. The number of images per ANP is kept to about the same so as not to bias any ANP during DBM learning. All these images, along with their metadata (title, descriptions and tags), are included in E-Flickr. Similarly for E-YouTube, ANPs keywords are issued to YouTube for crawling videos. For each ANP, only top-100 ranked videos are considered, considering that videos further down the ranked list are likely to be irrelevant. Among these videos, lengthy videos with duration more than two minutes are excluded from E-YouTube. Generally lengthy videos are more likely to contain segments with no emotional content. Including these videos into training will practically hurt the learning effectiveness. Since tags are not available for download, each video is crawled along with title and description only. On average, 50 videos are crawled per ANP and this number does not differ by more than 10 across ANPs. To this end, E-Flickr and E-YouTube include 830,580 images and 156,219 videos respectively.

The set of features extracted from the datasets are summarized in Table I. For each video, keyframes are sampled at the rate of one frame per second. Five different visual features, as listed in Table I, are respectively extracted from the keyframes and then averagely pooled to form feature vectors. Audio features are extracted over every 32 ms time-window of audio frames, with 50% overlap between two adjacent windows. Similar as visual features, these features are averagely pooled across time-windows. Among the set of audio-visual features, DenseSIFT, HOG, SSIM and MFCC are further quantized into bag-of-words representation. We followed the same settings as [1], and the dimensions of different features are listed in the second column of Table I. As for textual features, a total of 1,447,612 distinct words are extracted from E-Flickr and E-YouTube after stopword removal and lemmatization using CoreNLP [42]. In the experiments, only words with document frequency larger than 800 are kept. Eventually, textual feature is in 4,314 dimensions, with an average of 13 words per image and 8 words per video.

TABLE II
PREDICTION ACCURACIES FOR EACH EMOTION CATEGORY OF VIDEOEMOTION OBTAINED BY APPLYING LOGISTIC REGRESSION TO REPRESENTATIONS LEARNED AT DIFFERENT HIDDEN LAYERS. THE HIGHEST ACCURACY OF EACH CATEGORY IS HIGHLIGHTED

Category	$\mathbf{h}^{(1d)}$	$\mathbf{h}^{(2d)}$	$\mathbf{h}^{(1g)}$	$\mathbf{h}^{(2g)}$	$\mathbf{h}^{(1o)}$	$\mathbf{h}^{(2o)}$	$\mathbf{h}^{(1l)}$	$\mathbf{h}^{(2l)}$	$\mathbf{h}^{(1s)}$	$\mathbf{h}^{(2s)}$	$\mathbf{h}^{(v)}$	$\mathbf{h}^{(1a)}$	$\mathbf{h}^{(2a)}$	$\mathbf{h}^{(J)}$
Anger	0.152	0.242	0	0.061	0.152	0.273	0.182	0.273	**0.333**	0.242	0.212	0.030	0.030	0.303
Anticipation	0.031	0.094	0	0	0.031	0	0	**0.125**	0.062	0.062	0.031	0.031	0	0
Disgust	0.128	0.205	0	0.103	0.256	0.205	0.179	0.256	0.154	0.154	0.077	**0.308**	0.077	0.282
Fear	0.436	0.418	0.418	0.455	0.364	0.455	0.345	0.309	0.418	0.436	**0.509**	0.145	0.345	**0.509**
Joy	0.373	0.441	0.356	0.339	0.441	0.390	0.458	0.373	0.356	0.373	0.407	0.339	0.339	**0.508**
Sadness	0.176	0.235	0	0	0	0.029	0	0.118	0.059	0.059	0.059	0.206	0.206	**0.265**
Surprise	0.675	0.650	**0.863**	0.775	0.750	0.713	0.762	0.700	0.750	0.762	0.850	0.675	0.725	0.675
Trust	0.062	0.156	0.031	0.031	0.031	0.062	0.031	0.125	0.094	**0.188**	0.094	0.031	0.031	0.156
Overall	0.327	0.365	0.313	0.313	0.338	0.343	0.332	0.346	0.352	0.360	0.374	0.286	0.299	**0.404**

Note that video-level metadata only describes a fraction of video keyframes, and furthermore, feature pooling could possible introduce noise. In contrast, image-level metadata provides relatively more specific description of image content and emotion. For this consideration, the learning of DBM is started from using image samples followed by video sample. Audio-pathway is left out during pre-training using E-Flickr images, but turned on when E-YouTube videos are involved for fine-tuning. During training, each dimension of visual and auditory features is mean-centered and normalized to unit variance to avoid the instability problem [39]. In addition, to avoid running separate Markov chains for each word count to get sufficient statistics for modeling distribution, all word count vectors are scaled so that they sum to 5 [23].

In this section, we evaluate the performance of the joint representation learnt using our multimodel DBM on affective analysis. We name our model as E-MDBM since the architecture has been enhanced with more features and modalities.

B. Video Emotion Detection

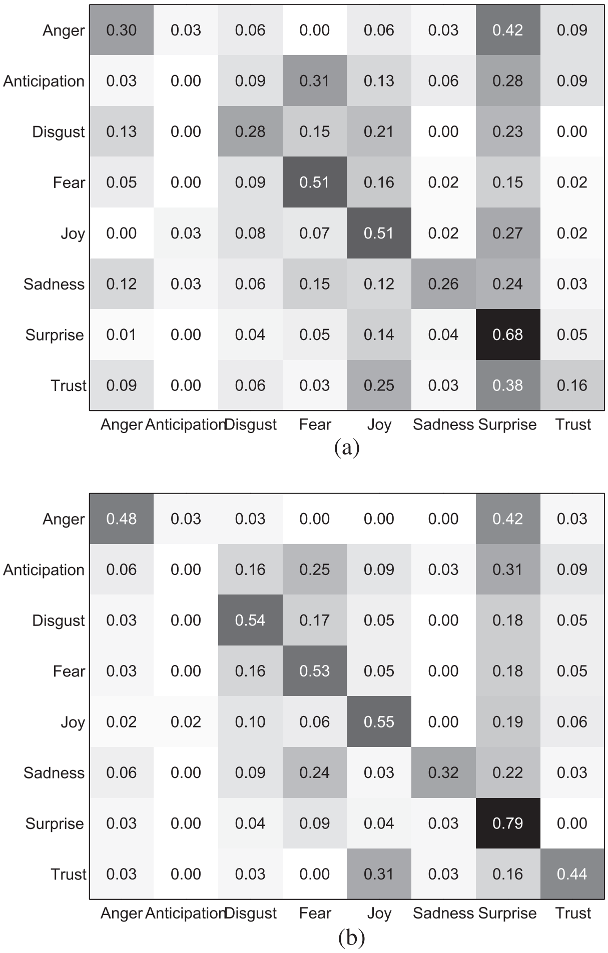
The "VideoEmotion" dataset consists of 1,101 videos which are manually labeled with eight emotional categories. Following [1], the results are evaluated by accuracy. As the textual information of the videos is not provided, we only have visual and auditory modalities in this dataset.

Effect of exploring multimodal relations. We first evaluate the capability of proposed model in learning non-linear relations among different modalities. The input to the textual pathway is missing and initialized to zeros. As described in Section V-B, the model is allowed to update the state of the textual input layer when performing mean-field update by alternating Gibbs sampling. In this experiment, we run the mean-field update for 5 times [16], [23]. The final joint representations (up layer) are drawn from $P(\mathbf{h}^{(J)}|\mathbf{v}_{\mathcal{V}}, \mathbf{v}_a)$ and used for learning a logistic regression model. For comparison, classifiers using the same training data are learned with the representations extracted from different hidden layers in Fig. 2.

We adopt the same settings for train-test splits in [1]. Ten train-test splits are generated, each using 2/3 of the data for training and 1/3 for testing. Table II shows the prediction results. We can see that the joint representation $\mathbf{h}^{(J)}$ achieves the best overall performance. It improves the accuracy over the joint visual representation $\mathbf{h}^{(v)}$ and the representation from second auditory hidden layer $\mathbf{h}^{(2a)}$ by 8.02% and 41.26% respectively. Although single modality may perform slightly better than the joint representation for some emotions, the performance is not consistent. For example, auditory feature is better to recognize "Disgust", but it performs poorly for emotion "Fear". This is because that "Fear" is not apparently conveyed in the auditory signal. However, visual feature achieves the best result on "Fear". Another interesting observation is that the performance using second hidden layer of each pathway is generally better than that of the first hidden layer. As mentioned in Section III, the E-MDBM model is a fully generative model. The neurons of hidden layers will receive messages from both lower layers and higher layers. By using this top-down feedback, the higher hidden layers can deal with the impact from ambiguous inputs, and thus are more robust. In addition, the joint representation ($\mathbf{h}^{(v)}$) on visual pathway leads to 2.5% improvement over the best performance achieved in single visual feature. This indicates that the proposed structure can preserve the capability of learning correlation between different visual features, meanwhile reduce the complexity of the model learning comparing to [23].

Fig. 3(a) further shows the confusion matrix based on the joint representation $\mathbf{h}^{(J)}$. Most categories are confused with the category "Surprise", where similar observation is also noted in [1]. Second, the category "anticipation" is confused particularly by "Fear" and "Surprise". As shown in Table II, almost all features perform poorly on this category. We attribute this unsatisfactory performance to the fact that neither audio nor visual can concretely describe the emotion of "anticipation", for example, in a sport event. Facial expression seems to be the dominant cue in conveying "anticipation". This is probably the reason that LBP, often being applied for face recognition, performs comparatively good for this category.

Impact of missing modality. There is no textual modality in this dataset. We further evaluate the impact of missing certain modality by using either only visual feature or auditory feature as input, which are named as E-MDBM-V and E-MDBM-A respectively. The input of missing modality is initialized with zero and updated in the same way with the missing textual modality. The results are showed in Fig. 4. We also show the result of $\mathbf{h}^{(J)}$ in Table II, which is named as E-MDBM-VA for consistency. Additionally, the performance of SentiBank from [1] and art features (AF) from [21] are also shown here for comparison. In SentiBank, logistic regression model is trained on the scores of the 1,200 ANP classifiers. For AF, 10 features are proposed in [21] for representing 6 artistic principles (i.e., balance, emphasis, harmony, variety, gradation, and movement) and logistic regression model is learnt on these features. Although our model is able to fill in the missing modalities and integrate the information into the final joint representation, E-MDBM-VA using both visual and auditory inputs exhibits the best performance.



	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Anger	0.30	0.03	0.06	0.00	0.06	0.03	0.42	0.09
Anticipation	0.03	0.00	0.09	0.31	0.13	0.06	0.28	0.09
Disgust	0.13	0.00	0.28	0.15	0.21	0.00	0.23	0.00
Fear	0.05	0.00	0.09	0.51	0.16	0.02	0.15	0.02
Joy	0.00	0.03	0.08	0.07	0.51	0.02	0.27	0.02
Sadness	0.12	0.03	0.06	0.15	0.12	0.26	0.24	0.03
Surprise	0.01	0.00	0.04	0.05	0.14	0.04	0.68	0.05
Trust	0.09	0.00	0.06	0.03	0.25	0.03	0.38	0.16

(a)

	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Anger	0.48	0.03	0.03	0.00	0.00	0.00	0.42	0.03
Anticipation	0.06	0.00	0.16	0.25	0.09	0.03	0.31	0.09
Disgust	0.03	0.00	0.54	0.17	0.05	0.00	0.18	0.05
Fear	0.03	0.00	0.16	0.53	0.05	0.00	0.18	0.05
Joy	0.02	0.02	0.10	0.06	0.55	0.00	0.19	0.06
Sadness	0.06	0.00	0.09	0.24	0.03	0.32	0.22	0.03
Surprise	0.03	0.00	0.04	0.09	0.04	0.03	0.79	0.00
Trust	0.03	0.00	0.03	0.00	0.31	0.03	0.16	0.44

(b)

Fig. 3. Confusion matrix based on (a) joint representation and (b) fusion results on the VideoEmotion dataset.

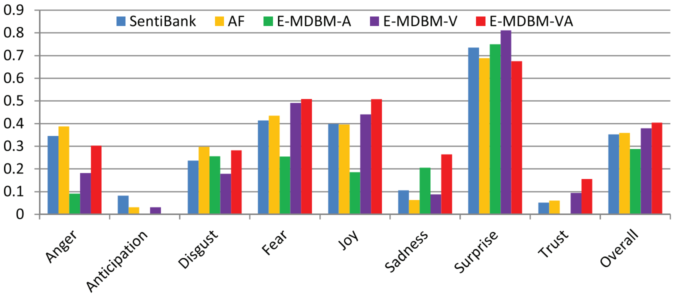


Fig. 4. Prediction accuracies. SentiBank is the attribute feature proposed in [6]. E-MDBM-V and E-MDBM-A represent the joint representation generated by our proposed E-MDBM using only the visual or auditory signals as inputs. Similarly, E-MDBM-VA indicates the joint representation using both visual and auditory modalities.

This is not surprising since the filling of missing modalities generated some noises comparing to the real data. However, the accuracy of E-MDBM-V is approaching that of E-MDBM-VA. For category "surprise", E-MDBM-V is even better. This indicates that the missing data is somehow recovered using this model. In addition, only visual features are used in SentiBank, AF and E-MDBM-V. Our model outperforms SentiBank and AF by 7.7% and 5.6% in accuracy respectively. This is because that our model embeds the information from all the three modalities during unsupervised model training. Thus the correlations between visual modality and other two modalities can be jointly represented, especially the textual modality which helps to explore the semantics in the videos. Comparing to SentiBank, where the extracted semantics are limited to 1,200 predefined ANPs, our model trained on wild Web data is expected to capture more complex semantics. AF, which comprises hand-tuned features dedicated for art photos, still performs reasonably well on the web video domain. This basically gives clue to the correlation between art-based features and emotions. Interestingly, AF even performs slightly better than SentiBank that predefines the set of ANPs for emotion description. E-MDBM-V, which learns features directly from examples while considering multi-modality correlation, has better capacity in dealing with diversities in user-generated videos compared to SentiBank and AF.

We can also observe that the auditory information seems less effective comparing to visual information. However, it works better in "Disgust" and "Sadness" categories, where the visual information cannot provide enough clues for emotion detection. For instance, there is a video showing a cat, which is annotated as "Disgust". The visual appearance actually conveys no emotional information. On the other hand, the background music, which is very sharp and noisy, actually makes people feel uncomfortable and disgust. The same situation exists in the "Sadness" videos. There are only several common objects shown in the video, such as faces and people hugs, whereas, woeful music is used as background. In short, missing of modality will degrade the performance even using our proposed model. However, the joint representation can somehow capture the correlations between different modalities, and is a good compensation when certain modality is not available.

Comparison with state-of-the-arts. We compare our model with the simplified version in [23]. For fair comparison, we extend the model in [23] to handle three modalities by adding a new pathway for textual input. This model is named as MDBM. Same training data and settings described in Section VI-A are employed for learning MDBM. We also show the results reported in [1] here. These results can be considered as the state-of-the-arts as they are produced through fine tuning on the dataset. In [1], various auditory (Au.) features, visual (V.) features, attribute (At.) features, and their combinations are evaluated. Table III shows the best one in each kind of the feature. Furthermore, the combinations of our joint representation and other features are also included. In addition, the performance based on the art features (AF) [21] is also reported.

We can see that E-MDBM consistently outperforms MDBM either evaluated individually or combined with other features. Specifically, E-MDBM leads to 9% performance improvement over MDBM. This indicates that our design can better preserve the unique property of each visual feature during the learning in visual pathway by splitting the architecture into several sub-pathways, each of which corresponds to one feature. In contrast, all the features are concatenated into one feature vector in MDBM, where some visual features may be overwhelmed during the learning. However, V. + Au. performs better than E-MDBM. This is probably because the pre-training of our model is performed on Flickr images, while V. + Au. is tuned on the Web videos in VideoEmotion dataset. The domain gap may influence the learned joint representation. As stated in [1],

TABLE III
PREDICTION ACCURACIES OF THE STATE-OF-THE-ARTS ON VIDEOEMOTION [1]. THE NOTATIONS V., Au., AND At. REPRESENT VISUAL, AUDITORY AND ATTRIBUTE FEATURES RESPECTIVELY

Category	SentiBank	AF	MDBM	E-MDBM	V.+Au.	V.+Au. +MDBM	V.+Au. +E-MDBM	V.+Au.+At.	V.+Au.+At. +MDBM	V.+Au.+At. +E-MDBM
Anger	0.345	0.388	0.294	0.303	**0.549**	0.527	0.515	0.527	0.509	0.485
Anticipation	**0.082**	0.031	0.067	0	0.028	0.067	0.030	0.067	0.034	0
Disgust	0.237	0.298	0.267	0.282	0.399	0.381	0.308	0.438	0.399	**0.538**
Fear	0.414	0.435	0.395	0.509	0.396	0.484	**0.648**	0.471	0.545	0.527
Joy	0.398	0.396	0.442	0.508	0.480	0.557	0.567	0.484	**0.590**	0.542
Sadness	0.106	0.063	0.217	0.265	0.289	0.274	0.229	0.208	0.217	**0.324**
Surprise	0.735	0.688	0.666	0.675	0.746	0.802	**0.861**	0.767	0.828	0.787
Trust	0.052	0.061	0.136	0.156	0.311	0.327	0.250	0.287	0.312	**0.438**
Overall	0.352	0.359	0.371	0.404	0.451	0.484	0.501	0.463	0.499	**0.511**

this may also cause the performance degradation of attribute features (e.g., SentiBank) which are extracted using concept classifiers learned on Web images. Despite the domain gap, E-MDBM consistently achieves better performance than SentiBank, either when being utilized individually or fused with other features. While not performing better than hand-crafted features, E-MDBM can complement these features well. When average lately fused with features in [1], an improvement of 11.09% is attained. Further fusion with attribute features (i.e., SentiBank, Classemes and ObjectBank), an accuracy of 0.511 is attained. The degree of improvement introduced by E-MDBM is greater than that can be offered by MDBM. Fig. 3(b) shows the confusion matrix based on fusion results of E-MDBM. Compared with Fig. 3(a), four categories, especially “Trust”, become less confused by “Surprise” after fusion. Nevertheless, the performance for “Anticipation” remains poor.

C. Sentiment Analysis on Twitter Messages

To avoid the impact of domain shift, we further conduct experiments on “ImageTweets” dataset [6], which includes 596 text-image Twitter messages. In this dataset, only textual and visual modality are available. These messages are manually assigned to either positive or negative sentiment based on affection expressed in the text-image pairs.

We compare our model with several state-of-the-art methods used in [6]. Besides the early fused low-level visual features (“Visual”) and attribute feature (“SentiBank”) extracted from images, a lexicon-based approach (“Lexicon”) used on textual analysis is also selected as baseline. The text information is represented using the sentiment scores of words obtained from SentiStrength [43]. The art features (AF) [21] is also adopted as one baseline. The classifiers for textual feature and visual features are Naive Bayes classifier and logistic regression model respectively. MDBM [23] is utilized on this dataset with visual +textual input. For our proposed E-MDBM, we consider three features E-MDBM-V, E-MDBM-T amd E-MDBM-VT corresponding to visual input, textual input and visual+textual input respectively. For fair comparison, logistic regression model is used upon these joint representations.

In [6], the dataset is equally split into five subsets. In our experiment, each classifier is trained on four subsets and tested on the other. This process is repeated five times. Fig. 5 shows the average results. We first consider the case of single modality input. We can see that Lexicon performs much worse than other methods. This is not surprising since Twitter messages are usually sparse and lack of emotion signals. In comparison, with the same input, our joint representation E-MDBM-T achieves 43.49% improvement over Lexicon. Again, this demonstrates that the common space embedded in the E-MDBM preserves the correlations from multiple modalities with different emotional signals. Thus the sparsity problem can be addressed using the information from other modalities by mapping from textual input to the joint space. For the visual modality, different from the results in Table III, SentiBank and E-MDBM-V

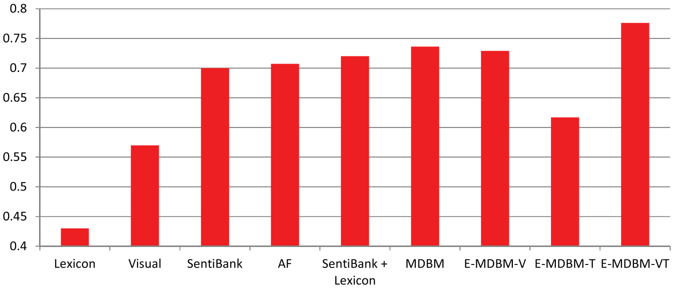


Fig. 5. Prediction accuracies on the ImageTweets. E-MDBM-V and E-MDBM-T are the classifiers trained on the joint representation generated by using only the visual or textual information through E-MDBM. E-MDBM-VT is trained on the joint representations over both visual and textual modalities. MDBM represents the joint representations over both visual and textual modalities but based on the architecture proposed in [23]. Meanwhile, SentiBank represents the classifiers trained on the scores of the concepts classifiers in [6]. Lexicon represents the Naive Bayes classifiers trained based on SentiStrength [43]. AF represents the classifiers trained on the art features proposed in [21].

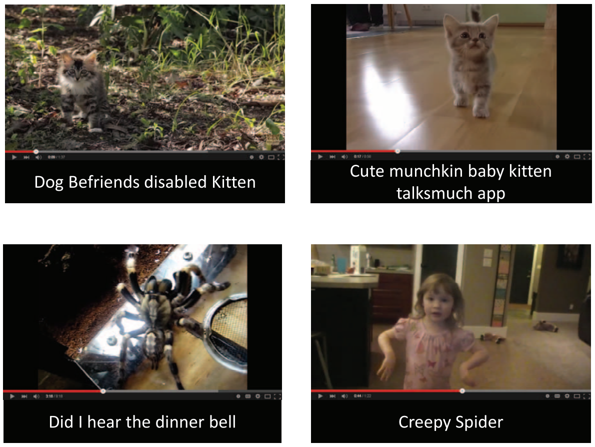

Dog Befriends disabled Kitten
Cute munchkin baby kitten talksmuch app
Did I hear the dinner bell
Creepy Spider

Fig. 6. Examples of video queries. The caption for the first example (top left) describes a disabled kitten, but the video expresses a “joy” emotion with an audio-visual effect. The two videos in the second row show counter examples. The video on the left has a spider but is not mentioned by the caption, as opposed to the video on the right where the spider is mentioned in caption but does not appear in video. The query (top right) expresses consistent emotion and semantics across visual, auditory, and textual modalities.

TABLE IV
MEAN AVERAGE PRECISION@20 OF TEXT-BASED, VIDEO-BASED, AND MULTIMODAL QUERY FOR RETRIEVING EMOTIONAL VIDEOS

Category	Text-based Query		Video-based Query			Multimodal Query		
	Word Count	E-MDBM-T	V.+Au.	SentiBank+Au.	E-MDBM-VA	V.+Au.+Te.	SentiBank+Au.+Te.	E-MDBM-VAT
Anger	0.222	0.254	0.345	0.340	**0.489**	0.340	0.404	0.398
Anticipation	0.202	0.234	0.437	0.384	0.393	0.384	0.393	**0.634**
Disgust	0.389	0.411	0.446	0.492	0.481	**0.496**	0.492	0.465
Fear	0.417	0.356	0.423	0.436	**0.487**	0.436	0.507	0.464
Joy	0.401	0.521	0.490	0.528	0.584	0.530	0.584	**0.673**
Sadness	0.337	0.337	**0.492**	0.422	0.410	0.422	0.410	0.400
Surprise	0.508	0.579	0.554	0.640	0.585	**0.641**	0.588	0.572
Trust	0.333	0.333	0.400	0.443	0.400	0.443	0.400	**0.467**
Overall	0.353	0.402	0.450	0.470	0.489	0.472	0.492	**0.533**

perform much better than low level visual features (Visual) on image dataset. This is because that both our model and attribute feature represent the semantics in the images, which can narrow down the gap between low-level features and high level human perceptions. In addition, AF achieves a similar performance to SentiBank, indicating that the Tweet images also partially follow the artistic principles. Again, E-MDBM-V improves over SentiBank and AF by 4.14% and 3.11% in accuracy respectively. We then compare the performances of different methods using multiple modalities. For comparison, we also show the result of lately fused Lexicon and SentiBank (SentiBank+Lexicon) features. We can see that E-MDBM-VT exhibits the best results, which leads to 8% and 5% improvements over SentiBank+Lexicon and MDBM respectively. Finally, similar to the conclusion made on VideoEmotion dataset, the performances of E-MDBM-V and E-MDBM-T are worse than that of E-MDBM-VT, which also indicates that there are some noises in the generated missing modalities.

VII. EXPERIMENT: RETRIEVAL

This section experiments the feasibility of the proposed model for video retrieval (Section VII-A) and cross-media retrieval (Section VII-B). A new dataset including 1,139 reference videos was constructed using 23 ANPs in [6] as queries. Both videos and their surrounding metadata are downloaded from YouTube. With this dataset, we can perform Web video retrieval using different types of queries. In specific, we randomly select 50 videos covering all the eight emotions in [1] as queries. In some of the queries, the content and emotion in different modalities are not always aligned. Fig. 6 shows some examples of queries. The selected 23 ANPs include 18 emotion words[1] and 16 general concept words.[2] We first annotate all the test and query videos for the 34 words. A video is considered to be relevant to the query if they share at least one emotion category or concept. In other words, two videos can be semantically or emotionally relevant. In this way, we can generate ground-truth for each query. Note that as emotion words are extracted from ANPs, their categories are inferred from which the ANPs belong to.

TABLE V
CROSS-MODAL RETRIEVAL: MEAN AVERAGE PRECISION@20 ON FOUR DIFFERENT TYPES OF QUERIES AGAINST FOUR DIFFERENT VERSIONS OF DATASETS

Modality	DB-T	DB-V	DB-A	DB-VA
E-MDBM-T	-	0.366	0.371	0.368
E-MDBM-V	0.437	-	0.358	-
E-MDBM-A	0.351	0.365	-	-
E-MDBM-VA	0.430	-	-	-

A. Video Retrieval

Three sets of experiments are conducted by using the text, video and multimodal (text+video) queries. For each query, a joint representation is extracted using the proposed E-MDBM. Note that there are missing modalities for text or video queries. For the reference videos, we assume that all the three modalities are available when extracting the joint representations. The baselines used in the experiments depend on the modality of queries. For text query, we compute Jaccard coefficient between the word count vector of a query and the surrounding text of a reference video. For video query, we select two baseline methods that utilize visual and auditory information by fusing low level visual feature and SentiBank respectively with auditory feature. For multimodel query, the two baselines are further augmented using textual feature. In this way, for a given type of query, the compared methods leverage same set of input features.

Table IV shows the results of video retrieval on eight different categories of emotions in terms of mean average precision (MAP). E-MDBM achieves consistently better performances than the baselines in all three types of queries. As baseline methods consider only matching the modalities of same type during similarity measure, reference videos with the searched content or emotion exists in a modality different from the type of query cannot be retrieved. For example, although the top-left video in Fig. 6 shows a "joy" emotion with a dog befriends a cat, text-only query is misled by the word "disable". For this particular example, the performance is poor even for multimodal query with late fusion strategy. Similarly for the two queries shown in the second row of Fig. 6, where there are mismatches between the concepts in captions and video content. Late fusion of multiple modalities helps little for these example queries. In contrast, by non-linearly projecting all the reference videos into a joint space, E-MDBM has generated features that inherently capture the complex relationship among different modalities of videos. Hence, cross-modal matching between queries and

[1] List of emotion words: cold, broken, fantastic, curious, classic, fat, crazy, lazy, creepy, haunted, clear, bright, natural, heavy, dirty, adorable, shiny, tasty

[2] List of concept words: morning, chair, fence, architecture, bird, spider, cat, tree, castle, moon, spring, rain, dog, star, food, body

TABLE VI
CROSS-MODAL RETRIEVAL: MEAN AVERAGE PRECISION@20 FOR EIGHT EMOTION CATEGORIES UNDER DIFFERENT SCENARIOS. THE BEST PERFORMANCE FOR EACH CATEGORY IS HIGHLIGHTED

Category	T→V+A	T→V	T→A	A→T	A→V	V→A	V→T	V+A→T
Anger	0.222	0.224	0.280	0.233	0.223	0.253	**0.376**	**0.376**
Anticipation	0.205	0.200	0.218	0.200	0.200	0.202	**0.246**	0.200
Disgust	0.383	0.384	0.394	0.390	0.383	0.385	**0.475**	0.473
Fear	0.374	0.356	0.473	0.401	0.356	0.345	**0.667**	0.656
Joy	**0.475**	0.472	0.411	0.386	0.471	0.394	0.456	0.450
Sadness	0.333	0.333	0.333	0.333	0.333	0.337	**0.380**	0.372
Surprise	0.500	0.500	0.504	0.519	0.500	0.504	0.560	**0.566**
Trust	0.333	0.333	0.333	0.333	0.333	**0.444**	0.333	0.333

reference videos are implicitly performed during retrieval. For the examples in Fig. 6, E-MDBM shows improvement in large margin, for example, the AP for the top-left query achieved by V. + Au. + Te. and SentiBank + Au. + Te. are 0.193 and 0.247 resepctively. In contrast, our joint representation exhibits much better AP with 0.394. Finally, it is worth mentioning that using E-MDBM, each video is represented as a feature in 4,096 dimensions, in contrast to 28,965 dimensions of hand-crafted features. The compact feature representation will greatly save time and space in video retrieval.

B. Cross-Modal Retrieval

We further evaluate our method on cross-modal retrieval. Different from the experiment in VII-A where the feature extracted for database videos adopts all the three modalities, cross-modal retrieval assume that query and the candidate data have different form of media types. Many applications can benefit from the cross-modal retrieval. For example, given a video without text description, the retrieved textual documents are helpful for automatically understanding the videos. Comparing to traditional semantic concept detection, which aims at recognizing some general concepts from visual appearance or auditory signals, the generated description using cross-modal retrieval would be more comprehensive.

Using the same dataset as Section VII-A, we generate four different versions of datasets, where each with one or two modalities of reference videos being purposely omitted. The four datasets are named as DB-X, with X-only modality being generated by the joint representation, where X includes T (text), V (visual), A (audio) and VA (visual-audio) modalities. During experiments, E-MDBM-X is experimented to search on a dataset without modality X. For example E-MDBM-T (text-only query) is searched against the dataset DB-V where the text and audio signals of all videos are ignored. Table V summarizes the results of cross-modal retrieval for 50 queries. As can be observed, the performances are very encouraging. For example, E-MDBM-T can achieve MAPs of 0.366 and 0.371 on DB-V and DB-A respectively when metadata of reference videos are ignored. These results are better than text-based query using word count (MAP $= 0.353$) when the metadata are involved in similarity comparison, as shown in Table IV. Using E-MDBM-V (visual-only query) against DB-T can attain MAP of 0.437, which is fairly impressive given that only the metadata of reference videos are kept for similarity measure. The result is close to SentiBank (AP $= 0.446$), which compares similarity with reference videos directly based on visual modality.

We further detail the results for each emotion category in Table VI, where every column corresponds to a different cross-modal retrieval scenario. For example, $\mathrm{T} \rightarrow \mathrm{V}+\mathrm{A}$ means the use of text-based query for searching against videos with only visual and audio modalities. Interestingly, query-by-visual often exhibits better performance even though the visual modality of reference videos is not considered. In addition, visual seems to be correlated better with text than audio modality, resulting in better performance in 7 out of eight categories for $\mathrm{V} \rightarrow \mathrm{T}$. The only exception is the category "Trust". This is because many videos expressing the trust emotion in this dataset are about *trust test* with laughing and cheering sounds. Comparing Table VI with Table IV, which can be considered as the up-bound of cross model retrieval, there is still a performance gap. However, considering that we are matching the data from different modalities with different statistic properties, basically the results in Table V can demonstrate the ability of our method for modeling the joint representations.

VIII. CONCLUSION AND FUTURE WORK

We have presented a deep model for learning multimodal signals coupled with emotions and semantics. Particularly, we propose a multi-pathway DBM architecture dealing with low-level features of various types and more than twenty-thousand dimensions, which is not previously attempted to the best of our knowledge. The major advantage of this model is on capturing the non-linear and complex correlations among different modalities in a joint space. The model enjoys peculiarities such as learning is unsupervised and can cope with samples of missing modalities.

Compared with hand-crafted features, our model generates much more compact features and allows natural cross-modal matching beyond late or early fusion. As demonstrated on ImageTweets datasets, the features generated by mapping single-modality samples (text or visual) into the joint space consistently outperform hand-crafted features in sentiment classification. In addition, we show the complementary between deep and hand-crafted features for emotion prediction on VideoEmotion dataset. Among the eight categories of emotion, nevertheless, the categories "anticipation" and "surprise" remain difficult either with learnt or hand-tuned features. For video retrieval, our model shows favorable performances, convincingly outperforms hand-crafted features over different types of queries. Encouraging results are also obtained when applying the deep features for cross-modal retrieval, which is not possible for hand-crafted features. Compared to SentiBank, our model has the edge of not limiting to a predefined set of vocabularies.

Hence, the learning is fully generative and the model is more expressive, as we show in the experiments that our model is able to perform better than SentiBank in both classification and retrieval tasks. Finally, our model also consistently outperforms the early version of MDBM [23] in all the experiments conducted in this paper.

Appendix

The conditional distributions over different layers, using "DenseSIFT" and "Textual" pathway as example, is inferred as follows:

$$
\begin{aligned}
v_i^d|\mathbf{h}^{(1d)} &\sim \mathcal{N}\left(\delta_i \sum_j W_{ij}^{(1d)} h_j^{(1d)} + b_i^d, \delta_i^2\right) \\
p\left(h_j^{(1d)} = 1|\mathbf{v}^d,\ \mathbf{h}^{(2d)}\right) & \\
&= g\left(\sum_i W_{ij}^{(1d)} \frac{v_i^d}{\delta_i} + \sum_l W_{jl}^{(2d)} h_l^{(2d)}\right) \\
p\left(h_l^{(2d)} = 1\ |\mathbf{h}^{(1d)}, \mathbf{h}^{(v)}\right) & \\
&= g\left(\sum_j W_{jl}^{(2d)} h_j^{(1d)} + \sum_q W_{lq}^{(vd)} h_q^{(v)}\right) \\
p\left(h_q^{(v)} = 1|\mathbf{h}_{\mathcal{V}}^2,\ \mathbf{h}^{(J)}\right) & \\
&= g\left(\sum_l W_{lq}^{(vd)} h_l^{(2d)} + \sum_l W_{lq}^{(vg)} h_l^{(2g)}\right) \\
&\quad + \sum_l W_{lq}^{(vo)} h_l^{(2o)} + \sum_l W_{lq}^{(vl)} h_l^{(2l)} \\
&\quad + \sum_l W_{lq}^{(vs)} h_l^{(2s)} \\
p\left(v_{ik}^t = 1|\mathbf{h}^{(1t)}\right) &= \frac{exp\left(\sum_j h_j^{(1t)} W_{jk}^{(1t)} + b_k^t\right)}{\sum_{q=1}^K exp\left(\sum_j h_j^{(1t)} W_{jq}^{(1t)} + b_k^t\right)} \\
p\left(h_j^{(1t)} = 1|\mathbf{v}^t,\ \mathbf{h}^{(2t)}\right) & \\
&= g\left(\sum_{k=1}^K W_{kj}^{(1t)} v_k^t + \sum_l W_{jl}^{(2t)} \mathbf{h}_l^{(2t)} + N b_j^{(1t)}\right) \\
p\left(h_l^{(2t)} = 1|\mathbf{h}^{(1t)}, h^{(J)}\right) & \\
&= g\left(\sum_j W_{jl}^{(2t)} h_j^{(1t)} + \sum_p W_{lp}^{(Jt)} h_p^{(J)}\right)
\end{aligned} \tag{11}
$$

where $g(x) = 1/(1 + exp(-x))$ is the logistic function. When inferring the distributions, the observed modalities are clamped at the inputs and Gibbs sampling is performed for updating the states of each layer. As mentioned in Section V-B, mean-field update is adopted for state updating. Since each hidden layer is influenced by its higher and lower layers, alternating sampling is conducted for update all the necessary states to approximate the distribution.

Acknowledgment

The Tesla K40 used for this research was donated by the NVIDIA Corporation.

References

[1] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proc. AAAI*, 2014, pp. 73–79.

[2] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proc. Workshop Languages Social Media*, 2011, pp. 30–38.

[3] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for Twitter sentiment analysis," in *Proc. Workshop Making Sense Microposts Co-Located WWW* , 2012, pp. 2–9.

[4] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good, the bad and the omg!," in *Proc. ICWSM*, 2011, pp. 538–541.

[5] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting twitter sentiment analysis," in *Proc. 2nd Int. Workshop Issues Sentiment Discovery Opinion Mining*, 2013, pp. 2:1–2:9.

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM MM*, 2013, pp. 223–232.

[7] T. Chen, D. Borth, T. Darrell, and S. Chang, "DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1410.8586

[8] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, Sep. 2011.

[9] B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated gifs," in *Proc. ACM MM*, 2014, pp. 213–216.

[10] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI*, 2015, pp. 381–388.

[11] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Process.*, vol. 93, pp. 2140–2150, 2013.

[12] R. M. Teixeira, T. Yamasaki, and K. Aizawa, "Determination of emotional content of video clips by low-level audiovisual features," *Multimedia Tools Appl.*, vol. 61, pp. 1–29, 2011.

[13] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[14] Y. Baveye, J.-N. Bettinelli, E. Dellandreá, L. Chen, and C. Chamaret, "A large video database for computational models of induced emotion," in *Proc. ACII*, 2013, pp. 13–18.

[15] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[16] R. Salakhutdinov and G. Hinton, "Deep Boltzmann machines," in *Proc. AI Statist.*, 2009, pp. 448–455.

[17] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[18] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, "Aggregate features and ADABOOST for music classification," *Mach. Learn.*, vol. 65, no. 2–3, pp. 473–484, 2006.

[19] Y.-H. Yang and H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 762–774, May 2011.

[20] P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf.*, 2010, pp. 339–344.

[21] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 47–56.

[22] Y.-H. Yang, W. H. Hsu, and H. H. Chen, "Online reranking via ordinal informative concepts for context fusion in concept detection and video search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 12, pp. 1880–1890, Dec. 2009.

[23] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, pp. 2949–2980, 2014.

[24] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[25] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[26] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 776–789.

[27] L. jia Li, H. Su, L. Fei-fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, vol. 23, pp. 1378–1386.

[28] I. Goodfellow, Q. Le, A. Saxe, H. Lee, and A. Y. Ng, "Measuring invariances in deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, vol. 22, pp. 646–654.

[29] Y. Bengio, "Learning deep architectures for ai," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.

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

[32] R. Salakhutdinov and G. Hinton, "Replicated softmax: An undirected topic model," in *Proc. NIPS*, 2010, pp. 1607–1614.

[33] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[34] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, pp. 145–175, 2001.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[36] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[37] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[38] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proc. ISMIR*, 2010, pp. 441–446.

[39] G. Hinton, "A practical guide to training restricted Boltzmann machines," in *Neural Networks: Tricks of the Trade* , 2nd ed. Berlin, Germany: Springer, 2012, vol. 7700, pp. 599–619.

[40] T. Tieleman, "Training restricted Boltzmann machines using approximations to the likelihood gradient," in *Proc. ICML*, 2008, pp. 1064–1071.

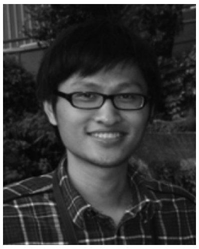
[41] T. Tieleman and G. Hinton, "Using fast weights to improve persistent contrastive divergence," in *Proc. 26th Int. Conf. Mach. Learn.*, 2009, pp. 1033–1040.

[42] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2014, pp. 55–60.

[43] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010.

Lei Pang received the B. Eng degree from Nankai University, Tianjin, China, in 2010, and is currently working toward the Ph.D. degree in computer science at the City University of Hong Kong, Hong Kong.

He is currently with the VIREO Group, City University of Hong Kong. His research interest lies in multimedia content analysis, including Web video face naming, multimedia question answering, and emotion prediction on Web videos.



Shiai Zhu received the B.S. degree from the Civil Aviation University of China, Tianjin, China, in 2005, the M.S. degree from the University of Science and Technology of China, Hefei, China, in 2008, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2013.

He is currently a Post-Doctoral Researcher with the Multimedia Communications Research Laboratory (MCRLab), University of Ottawa, Ottawa, ON, Canada. Before joining the University of Ottawa, he was a Staff Researcher with the Image and Visual Computing Lab (IVCL), Lenovo Group, Hong Kong. His research interests include social media, multimedia analysis, and machine learning, particularly the research of image and video content understanding and its applications.

Chong-Wah Ngo received the B.Sc. and M.Sc. degrees in computer engineering from Nanyang Technological University, Singapore, and the Ph.D. in computer science from the Hong Kong University of Science and Technology (HKUST), Hong Kong.

He is currently a Professor with the Department of Computer Science, City University of Hong Kong, Hong Kong. Before joining the City University of Hong Kong, he was a Postdoctoral Scholar with the Beckman Institute, University of Illinois at Urbana–Champaign (UIUC), Urbana, IL, USA. He was also a Visiting Researcher with Microsoft Research Asia, Beijing, China. His research interests include large-scale multimedia information retrieval, video computing, multimedia mining, and visualization.

Prof. Ngo was the Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (2011–2014). He was the Conference Co-Chair of the ACM International Conference on Multimedia Retrieval 2015 and the Pacific Rim Conference on Multimedia 2014. He also served as Program Co-Chair of ACM Multimedia Modeling 2012 and ICMR 2012. He was the Chairman of ACM (Hong Kong Chapter) from 2008 to 2009.